

Bioinformatics for Glycobiology and Glycomics

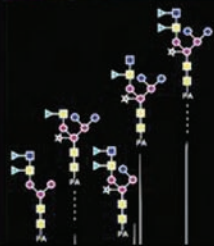
An Introduction

Editors

Claus-Wilhelm von der Lieth, Thomas Lütteke and Martin Frank



 WILEY



Bioinformatics for Glycobiology and Glycomics

Bioinformatics for Glycobiology and Glycomics: an Introduction

Edited by

Claus-Wilhelm von der Lieth

Formerly at the Molecular Structure Analysis Core Facility
Deutsches Krebsforschungszentrum (German Cancer Research Center)
Heidelberg, Germany

Thomas Lütteke

Faculty of Veterinary Medicine
Institute of Biochemistry and Endocrinology
Justus-Liebig University Gießen
Gießen, Germany

and

Martin Frank

Molecular Structure Analysis Core Facility
Deutsches Krebsforschungszentrum (German Cancer Research Center)
Heidelberg, Germany

 **WILEY-BLACKWELL**

A John Wiley & Sons, Ltd, Publication

This edition first published 2009, © 2009 by John Wiley & Sons Ltd.

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Other Editorial Offices:

9600 Garsington Road, Oxford, OX4 2DQ, UK

111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Bioinformatics for glycobiology and glycomics : an introduction / edited by Claus-Wilhelm von der Lieth, Thomas Lütteke, and Martin Frank.

p. ; cm.

Includes bibliographical references and index.

ISBN 978-0-470-01667-1

1. Glycomics. 2. Glycoconjugates—Research—Data processing. 3. Bioinformatics.

I. Lieth, Claus-Wilhelm von der. II. Lütteke, Thomas. III. Frank, Martin, 1963—

[DNLM: 1. Glycomics—methods. 2. Carbohydrates—chemistry. 3. Computational Biology—methods.

4. Glycoproteins—chemistry. QU 75 B6155 2009]

QP702.G577B56 2009

572'.56—dc22

2009023580

A catalogue record for this book is available from the British Library.

Set in 10/12pt Times by Aptara Inc., New Delhi, India.

Printed in Singapore by Markono Print Media Pte Ltd.

First Impression 2009

Contents

List of Contributors	ix
Preface	xv
<i>Claus-Wilhelm von der Lieth</i>	
Section 1: Introduction	
1. Glycobiology, Glycomics and (Bio)Informatics	3
<i>Claus-Wilhelm von der Lieth</i>	
Section 2: Carbohydrate Structures	
2. Introduction to Carbohydrate Structure and Diversity	23
<i>Stephan Herget, René Ranzinger, Robin Thomson, Martin Frank and Claus-Wilhelm von der Lieth</i>	
3. Digital Representations of Oligo- and Polysaccharides	49
<i>Stephan Herget and Claus-Wilhelm von der Lieth</i>	
4. Evolutionary Considerations in Studying the Sialome: Sialic Acids and the Host–Pathogen Interface	69
<i>Amanda L. Lewis and Ajit Varki</i>	
Section 3: Carbohydrate-active Enzymes and Glycosylation	
5. Carbohydrate-active Enzymes Database: Principles and Classification of Glycosyltransferases	91
<i>Pedro M. Coutinho, Corinne Rancurel, Mark Stam, Thomas Bernard, Francisco M. Couto, Etienne G. J. Danchin and Bernard Henrissat</i>	
6. Other Databases Providing Glycoenzyme Data	119
<i>Thomas Lütteke and Claus-Wilhelm von der Lieth</i>	
7. Bioinformatics Analysis of Glycan Structures from a Genomic Perspective	125
<i>Kiyoko F. Aoki-Kinoshita and Minoru Kanehisa</i>	
8. Glycosylation of Proteins	143
<i>Claus-Wilhelm von der Lieth and Thomas Lütteke</i>	

9. Prediction of Glycosylation Sites in Proteins 163
*Karin Julenius, Morten B. Johansen, Yu Zhang,
Søren Brunak and Ramneek Gupta*

Section 4: Experimental Methods – Bioinformatic Requirements

10. Experimental Methods for the Analysis of Glycans and Their Bioinformatics Requirements 195
Claus-Wilhelm von der Lieth
11. Analysis of *N*- and *O*-Glycans of Glycoproteins by HPLC Technology 203
Anthony H. Merry and Sviatlana A. Astrautsova
12. Glycomic Mass Spectrometric Analysis and Data Interpretation Tools 223
Niclas G. Karlsson and Nicolle H. Packer
13. Software Tools for Semi-automatic Interpretation of Mass Spectra of Glycans 257
Kai Maass and Alessio Ceroni
14. Informatics Concepts to Decode Structure-Function Relationships of Glycosaminoglycans 269
Rahul Raman, S. Raguram and Ram Sasisekharan
15. NMR Databases and Tools for Automatic Interpretation of Spectra of Carbohydrates 295
Claus-Wilhelm von der Lieth
16. Automatic Spectrum Interpretation Based on Increment Rules: CASPER 311
Roland Stenutz
17. Interpretation of ¹³C NMR Spectra by Artificial Neural Network Techniques (NeuroCarb) 321
Andreas Stoeckli, Matthias Studer, Brian Cutting and Beat Ernst

Section 5: 3D Structures of Complex Carbohydrates

18. Conformational Analysis of Carbohydrates – A Historical Overview 337
Martin Frank
19. Predicting Carbohydrate 3D Structures Using Theoretical Methods 359
Martin Frank
20. Synergy of Computational and Experimental Methods in Carbohydrate 3D Structure Determination and Validation 389
Thomas Lütteke and Martin Frank

Section 6: Protein–Carbohydrate Interaction	
21. Structural Features of Lectins and Their Binding Sites <i>Remy Loris</i>	415
22. Statistical Analysis of Protein–Carbohydrate Complexes Contained in the PDB <i>Thomas Lütteke and Claus-Wilhelm von der Lieth</i>	433
Section 7: Appendices	
Appendix 1: List of Available Websites	449
Appendix 2: Glossary	453
Index	461

List of Contributors

Kiyoko F. Aoki-Kinoshita

Department of Bioinformatics, Faculty of Engineering, Soka University, Tokyo 192-8577, Japan

Sviatlana A. Astrautsova

Department of Microbiology, Medical University of Grodno, Grodno, Belarus

Thomas Bernard

Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités d'Aix-Marseille I & II, 13402 Marseille cedex 20, France

Søren Brunak

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, 2800 Lyngby, Denmark

Alessio Ceroni

Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, UK

Pedro M. Coutinho

Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités d'Aix-Marseille I & II, 13402 Marseille cedex 20, France

Francisco M. Couto

Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal

Brian Cutting

Institute of Molecular Pharmacy, Pharmacenter of the University of Basel, 4056 Basel, Switzerland

Etienne G. J. Danchin

Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités d'Aix-Marseille I & II, 13402 Marseille cedex 20, France

Beat Ernst

Institute of Molecular Pharmacy, Pharmacenter of the University of Basel, 4056 Basel, Switzerland

Martin Frank

Deutsches Krebsforschungszentrum (German Cancer Research Center), Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany

Ramneek Gupta

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, 2800 Lyngby, Denmark

Bernard Henrissat

Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités d'Aix-Marseille I & II, 13402 Marseille cedex 20, France

Stephan Herget

Deutsches Krebsforschungszentrum (German Cancer Research Center), Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany

Morten B. Johansen

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, 2800 Lyngby, Denmark

Karin Julenius

Division of Matrix Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 17177 Stockholm, *and* Stockholm Bioinformatics Center, SCFAB, Stockholm University, 10691 Stockholm, Sweden

Minoru Kanehisa

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan

Niclas G. Karlsson

Centre for BioAnalytical Sciences, Chemistry Department, NUI Galway, Galway, Ireland

Amanda L. Lewis

Glycobiology Research and Training Center, Departments of Medicine, Biological Sciences and Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA 92093, USA

Remy Loris

Structural Biology Brussels, Vrije Universiteit Brussel and Department of Molecular and Cellular Interactions, VIB, Pleinlaan 2, B-1050 Brussels, Belgium

Thomas Lütteke

Faculty of Veterinary Medicine, Institute of Biochemistry and Endocrinology, Justus-Liebig University Gießen, 35392 Gießen, Germany

Kai Maass

Institute of Biochemistry, University of Gießen, 35392 Gießen, Germany

Anthony H. Merry

Glycosciences Consultancy, Charlbury OX7 3HB, UK

Nicolle H. Packer

Biomolecular Frontiers Research Centre, Department of Chemistry and Biomolecular Sciences, Macquarie University, North Ryde, Sydney, NSW 2109, Australia

S. Raguram

Biological Engineering Division, Harvard-MIT Division of Health Sciences and Technology, Center for Biomedical Engineering, Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Rahul Raman

Biological Engineering Division, Harvard-MIT Division of Health Sciences and Technology, Center for Biomedical Engineering, Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corinne Rancurel

Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités d'Aix-Marseille I & II, 13402 Marseille cedex 20, France

René Ranzinger

Deutsches Krebsforschungszentrum (German Cancer Research Center), Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany

Ram Sasisekharan

Biological Engineering Division, Harvard-MIT Division of Health Sciences and Technology, Center for Biomedical Engineering, Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Mark Stam

Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités d'Aix-Marseille I & II, 13402 Marseille cedex 20, France

Roland Stenutz

Department of Organic Chemistry, Stockholm University, 106 91 Stockholm, Sweden

Andreas Stoeckli

Institute of Molecular Pharmacy, Pharmacenter of the University of Basel, 4056 Basel, Switzerland

Matthias Studer

Institute of Molecular Pharmacy, Pharmacenter of the University of Basel, 4056 Basel, Switzerland

Robin Thomson

Institute for Glycomics, Griffith University - Gold Coast Campus, Queensland 4222, Australia

Ajit Varki

Glycobiology Research and Training Center, Departments of Medicine, Biological Sciences and Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA 92093, USA

Claus-Wilhelm von der Lieth

Formerly at Deutsches Krebsforschungszentrum (German Cancer Research Center), Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany

Yu Zhang

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, 2800 Lyngby, Denmark

Dr Claus-Wilhelm “Willi” von der Lieth 1949–2007

On November 16, 2007, Dr Claus–Wilhelm (Willi) von der Lieth – the ‘father’ of this book – passed away unexpectedly at the age of 58. He is greatly missed as a friend and colleague.

Willi was born in 1949 in Bremervörde, Germany. He studied chemistry at the Technical University of Hannover where he received his Diploma in 1977. He continued his studies in theoretical chemistry at the University of Heidelberg and received his doctoral degree in 1980. In the same year he joined the German Cancer Research Center (DKFZ Heidelberg) where he became the head of the Molecular Modeling group within the Central Spectroscopy Department in 1987. At the DKFZ, he developed a computer-based spectroscopic information system and provided a wide variety of modeling services for many years. In the late 1980s Willi entered the carbohydrate field through the application of modeling methods to support conformational analysis of complex oligosaccharides by NMR. Being a visionary, Willi realized that the Internet offered an opportunity to provide, without barriers, scientific information and tools to a large community. This led in the 1990s to the first web-based molecular builder for carbohydrate 3D structures (SWEET), which is still used by scientists today. In the late 1990s Willi initiated the SWEET-DB project which aimed to use modern web techniques to make existing carbohydrate-related data collections available over the Internet, and to create an interface that allowed glycoscientists to find important data for compounds of interest in a compact and well-structured representation. The data sources of SWEET-DB were structures and literature references from the – at that time discontinued – Complex Carbohydrate Structure Database (CarbBank), NMR data taken from SugaBase, and 3D co-ordinates generated with SWEET-II. An automated link to the NCBI PubMed service was established which allowed scientists to find literature for a particular carbohydrate structure very easily. Over the years more services and tools were developed which ultimately led to the GLYCOSCIENCES.de web portal, which is currently one of the largest scientific resources for carbohydrate structure related data.

Willi was internationally recognized as a pioneer, and a global leader, in the field of glycoinformatics. In recent years Willi was very active in international efforts to integrate and crosslink existing carbohydrate databases based on the philosophy of open access. He was Coordinator of the EUROCarbDB project (an EU funded design study to create the foundations of databases and bioinformatics tools for glycobiology and glycomics), a co-director of HGPI/HUPO (the Human Disease Glycomics/Proteome Initiative), and a member of the US Consortium for Functional Glycomics.

Willi served as Treasurer of the German Chapter of the Molecular Graphics Society and as Referee for several journals, was a member of the Editorial Board of Carbohydrate Research, and contributed over 100 journal articles and book chapters in the fields of molecular modeling, computer-based information systems, and glycoscience.

Willi’s sudden death was a great professional and personal loss for all those who had the privilege to work with him. Willi was an exceptional personality; his boundless enthusiasm, creativity, and active interest in so many research areas was contagious. His uncomplicated interactions with colleagues and students revealed his openness, modesty and generosity.

This book is dedicated to his memory.



Preface

Why Is It Timely to Publish a Book on (Bio)informatics for Glycobiology and Glycomics?

The four essential molecular building blocks of cells are nucleic acids, proteins, lipids, and carbohydrates, often also referred to as glycans. Nucleotide and protein sequences are the heart of nearly all bioinformatics applications and research, whereas glycan and lipid structures have been widely neglected. Glycans are the most abundant and diverse of Nature's biopolymers. Complex carbohydrates are often covalently attached to proteins and lipids and thus constitute a significant amount of the mass and structural variation in biological systems. The field of glycobiology is focused upon understanding the structure, chemistry, biosynthesis, and biological function of glycans and their derivatives.

It has long been known that carbohydrates encode biological information. For example, it was shown already in 1952 that variation of blood group determinants is a consequence of glycosylation, that is, the addition of complex carbohydrates to proteins and lipids. Thus, the chemistry, biochemistry, and biology of carbohydrates were prominent areas of research during the beginning and the middle of the last century. Nevertheless, determining the biological consequences of glycosylation has been extremely difficult. In an editorial article in the March 2001 edition of *Science* devoted to glycobiology, it was described as a “Cinderella field” of research. This means that it is “an area (of research) that involves much work but, alas, does not get to show off at the ball with her cousins, the genomes and proteins.”

With the awareness that the human genome encodes for a significantly smaller number of genes than estimated from genomes of lower organisms such as yeast [1], it became obvious that each gene can be used in different ways depending on how it is regulated. Consequently, the study of post-translational protein modifications, which can alter the functions of proteins, has entered an era of renaissance and come increasingly into the scientific focus. Glycobiology research has attracted increasing attention because glycosylation is the most complex and most frequently occurring post-translational modification. Similar to the developments in genomics and proteomics, high-throughput glycomics projects to decipher the role of carbohydrates in health and disease are emerging [2–9]. With the increasing amount of experimental data, the need to develop appropriate glycan related databases and bioinformatics tools is obvious. However, until recently, informatics have been only poorly represented in glycobiology [10–12].

Unfortunately, most of the tools and applications developed in bioinformatics for the description and analysis of DNA (RNA) and protein sequences cannot be directly applied to carbohydrates. This is mainly due to the fact that oligosaccharides can exhibit various ways to link together their building blocks, the monosaccharides. In nature, monosaccharides are

found which are connected to up to four others. This has the consequence that branched structures can be formed. Such structures can no longer be described as linear sequences, but rather need to be described as topologies of specifically connected building blocks. In this respect, the encoding of complex carbohydrates based on building blocks is more similar to the digital description of organic molecules, which are described in cheminformatics through the topology of atoms.

As with proteomics, mass spectrometry (MS) has become a key technology for the identification of glycans. The experimental procedures in MS-based glycomics analysis are similar to those applied in proteomics. However, until recently, no efficient software tools were available to interpret the spectra. Therefore, the annotation of spectra and assignment of glycan structures to the mass peaks had to be done manually by an expert.

The lack of efficient automatic assignment procedures is still the major bottleneck for an automatic high-throughput analysis of glycans in glycomics projects. Consequently, several computational attempts to overcome this unfavorable situation have been published in recent years. The developed algorithms were mainly implemented by experimentalists focused on solving the specific needs of their experimental setup and scientific questions. However, it is obvious that glycomics calls for more general solutions, highly sophisticated algorithmic approaches, and standardization.

Since the beginning of this century, a small but rather active community of researchers emerged with the aim of working out the foundations of the informatics for glycobiology and glycomics. The development and use of informatics tools and databases for glycobiology and glycomics research have thus increased considerably in recent years. However, this field must still be considered as being in its infancy compared with genomics and proteomics.

It is the aim of this book to give an introduction to this emerging field of science for the experimentalist working in glycobiology and glycomics, and also for the computer scientists looking for new challenges in the development of highly sophisticated algorithmic approaches.

Glycomics: an Exotic and Somewhat Forgotten Area of Bioinformatics

The European Bioinformatics Institute (EBI) (www.ebi.ac.uk/) [13, 14] is one of the largest centers world-wide for research and services in bioinformatics, managing a broad range of freely available databases of biological sequences, information, and knowledge. However, in 2007, the EBI did not provide access to any collection containing glycan structures. The US National Center for Biotechnology Information (NCBI) [15] (www.ncbi.nlm.nih.gov/) provides the PubChem service (pubchem.ncbi.nlm.nih.gov/), a fairly new service containing information on the biological activities of small molecules, which also includes carbohydrate structures. However, the main focus of PubChem is to provide access to glycans that can be used as chemical probes or ligands to study the functions of genes, cells, and biochemical pathways. The Japanese Kyoto Encyclopedia of Genes and Genomes (KEGG) (www.genome.ad.jp/) [16] is a suite of databases and associated software that aims to integrate the current knowledge on molecular interaction networks in biological processes with the information about the universe of genes and proteins, in addition to information about the universe of chemical compounds and reactions. KEGG contains a GLYCAN database with about 11 000 structures [17]. Most of the data were taken from CarbBank [18, 19], the largest attempt to build up a comprehensive collection of all known carbohydrate structures

during the 1980s and 1990s (see also the paragraph on the history of glyco-related databases in Chapter 1). KEGG is currently the most developed project which links glycan structures with available proteomics and glycomics data through biosynthetic pathways.

The Bioinformatics Links Directory (www.bioinformatics.ca/links_directory/) [20, 21], an online community resource that contains a directory of freely available tools, databases, and resources for bioinformatics and molecular biology research – which is compiled by collecting applications which have been published in the annual *Nucleic Acids Research Web* issues – lists only six tools dealing with the bioinformatics of carbohydrate structures. This is in contrast to the 376 useful resources for DNA sequence analyses reported, and the 651 links to useful resources for protein sequence and structure analyses, which include also tools for phylogenetic analyses, prediction of protein structures and functions, and analyses of protein–protein interactions. On the other hand, collections of links compiled with special emphasis on glyco-related web applications (see, e.g., www.eurocarbdb.org) already show more than 60 dedicated websites. This discrepancy reflects the current points of view from both sides: while the bioinformatics community widely ignores the existence of macromolecules other than DNAs, RNAs, and proteins, scientists developing software applications for glycobiology do not regard themselves as part of the bioinformatics community.

The Role of (Bio)informatics in Glycobiology Research

The involvement of (bio)informatics in glycobiology research can be divided into:

1. The application of classical bioinformatics tools to analyze the DNA and protein sequences, which have a relation to carbohydrates. Such sequences may be the proteins to which glycans are attached, the enzymes which build or modify oligosaccharides, or the lectins which recognize a certain sugar epitope.
2. Applications and databases where an explicit description of the glycan structure is required. All analytical tools to determine glycan structures and structure–function relations depend heavily on appropriate encoding of glycan structures.
3. Atomic descriptions of carbohydrate–protein recognition processes in which the spatial structures of complex carbohydrates, their conformational preferences, and their energetics are analyzed.

In this book, we try to provide a comprehensive overview of all three areas of active research. The chapters are written by active researchers who have made essential contributions to the development of the field of glycobioinformatics in recent years.

Use of Classical Bioinformatics Databases and Tools

Chapter 1, *Glycobiology, Glycomics, and (Bio)Informatics*, briefly describes the biological role of carbohydrates, their biosynthesis, and the enzymes which are responsible for the stepwise synthesis of the branched oligosaccharides – the glycosyltransferases. Special emphasis is placed on the role of bioinformatics in accelerating the identification of human carbohydrate active enzymes with the help of bioinformatics databases and appropriate alignment tools.

Chapter 5, *Carbohydrate-active Enzymes Database: Principles and Classification of Glycosyltransferases*, gives a comprehensive overview of the enzymes responsible for the biosynthesis of the glycosidic bonds in living organisms. The authors develop and maintain the CAZy database, the world's most complete resource describing the families of structurally related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. The subsequent chapter gives a short overview of the focus of other existing data resources which provide access to glyco-related enzymes.

Chapter 7, *Bioinformatics Analysis of Glycan Structures from a Genomic Perspective*, describes how informatics can help to elucidate the biological function of glycans, which are intertwined with the rest of the biological system such as interacting proteins and chemical compounds. In this chapter, an approach is presented based on the data of the Kyoto Encyclopedia of Genes and Genomes (KEGG) including a comprehensive glycan data resource called KEGG GLYCAN. It encompasses all aspects of the biological system, incorporating genomic information integrated with pathways and reactions and also chemical compounds.

Chapter 8, *Glycosylation of Proteins*, gives a short introduction to this phenomenon, and Section 8.2 therein, *GlySeq: an Analysis of Experimentally Determined Occupied Glycosylation Sites*, describes services which provide access to the respective glyco-related experimental data contained in the Protein Data Bank (PDB) [22] and SWISS-Prot. The *GlySeq* service – although focusing on analyzing the data from the viewpoint of carbohydrates – provides access to both moieties: the carbohydrates and the proteins. Glycosylation is known to affect protein folding, localization, and trafficking, protein solubility, antigenicity, biological activity and half-life. Chapter 9, *Prediction of Glycosylation Sites in Proteins*, summarizes the ideas of pattern recognition for protein glycosylation site prediction from peptide sequence alone. It provides a general introduction to data-driven prediction methods to solving this problem, including a discussion on artificial neural networks.

Informatics Applications Where a Special Encoding of Glycan Structures is Required

Due to the structural complexity of complex carbohydrates which can form highly branched structures, most of the tools and applications developed in classical bioinformatics for the description and analysis of DNA (RNA) and protein sequences cannot be directly applied to carbohydrates.

Chapter 2, *Introduction to Carbohydrate Structure and Diversity*, provides a short description of the major types of carbohydrate structural motifs found in nature. The structural diversity of carbohydrates that is currently available in databases has been analyzed and is also presented. As excellent reference books are available summarizing the cellular location and biological function of the various types of glycan, the Editors decided not to repeat this information here and recommend that readers consult the existing compendia for further reading.

Special encoding schemes are required which are able to describe all structural features of complex carbohydrates found in nature. Unfortunately, no standard description existed until recently that was capable of coping with all structural features of carbohydrates as needed, for example, for the emerging glycomics projects. Chapter 3, *Digital Representations of*

Oligo- and Polysaccharides, gives a comprehensive overview of all structural features which have to be encoded to cope with all carbohydrate-specific structural elements. It discusses various often used digital formats, which have so far been suggested and are used for specific applications or to encode carbohydrate structures in databases. The chapter gives an outline of future directions of developments.

The introductory part of the book is rounded off by a chapter on evolutionary considerations in studying the structural diversity of the most important class of terminal monosaccharides – the sialic acids. The contribution suggests that the structural diversity of sialic acids reflects the often conflicting pressures of evading pathogens, while simultaneously maintaining endogenous functions. Since most pathogens replicate much faster than their hosts, they can rapidly evolve different ways to target or mimic structures that are critical for host processes, which may be especially relevant to glycans.

An in-depth understanding of the biological functions of complex carbohydrates requires a detailed knowledge of all structural features of their primary sequence and also the conformational space that they can access. Analysis of carbohydrates has proved to be difficult in the past. Fortunately, modern analytical methods have the ability to elucidate most structural details at the concentration levels required for glycomics projects. However, at present, informatics tools give only limited support to enable an automatic, reliable interpretation of the vast amount of data recorded by the analytical methods. This deficiency currently represents a severe bottleneck for the practical implementation of high-throughput glycomics projects. Therefore, it is not surprising that the development of algorithms and tools to interpret analytical data constitutes the most active field of software design in the glycomics field.

The chapters included in Section IV, *Experimental Methods – Bioinformatics Requirements*, summarize the status of analytical procedures used in glycomics and the algorithms, software tools, and services which are available to support the interpretation of data. Similarly to proteomics, MS, HPLC and NMR are the main experimental techniques in glycomics research. However, the concrete experimental procedures applied by various researchers vary considerably, so that it is necessary to outline the analytical procedures since otherwise the informatics requirements would be difficult to explain.

Spatial Structures of Carbohydrates and Atomic Descriptions of Carbohydrate–Protein Recognition Processes

The elucidation of conformational preferences of complex carbohydrate structures has a long history, which dates back to the early 1980s. Chapter 18, *Conformational Analysis of Carbohydrates – A Historical Overview*, reviews these developments, and is followed by Chapter 19, *Predicting Carbohydrate 3D Structures Using Theoretical Methods*, where the various commonly used theoretical approaches and simulation methods are introduced and their applications to predict 3D structures of carbohydrates are discussed. Section 19.4, *Generation of 3D Structures of Glycoproteins*, describes a service available through the GLYCOSCIENCES.de portal [23], which generates 3D structures of glycoproteins. Chapter 20, *Synergy of Computational and Experimental Methods in Carbohydrate 3D Structure Determination and Validation*, describes methods to find and analyze experimental 3D structures of carbohydrates in the Protein Data Bank.

Lectins are carbohydrate-binding proteins or glycoproteins which often recognize a specific sugar epitope and are thus important for a broad variety of specific recognition processes and signaling events. Crystal structures of members of the different animal and plant lectin families have revealed a wide variety of lectin folds and carbohydrate binding site architectures. Despite this large variability, a number of interesting cases of both convergent and divergent evolution among plant, animal, and bacterial lectins are noted. These similarities are reviewed in Chapter 21, *Structural Features of Lectins and Their Binding Sites*.

Chapter 22, *Statistical Analysis of Protein–Carbohydrate Complexes Contained in the PDB*, provides a detailed overview of the interactions, which specific carbohydrates or classes of glycans exhibit with proteins in the available experimentally determined protein–carbohydrate complexes.

Current Status of Informatics for Glycosciences

Recent years have seen a variety of new databases and software tools emerging in the field of glycomics. However, the current situation in glycobioinformatics is characterized by the existence of multiple disconnected and incompatible islands of experimental data, data resources, and specific applications, managed by various consortia, institutions, or local groups. For example, no comprehensive carbohydrate data collections similar to those currently available for genomic and proteomic data have been compiled so far. There is currently no location where information about all carbohydrates reported in peer-reviewed scientific papers is systematically stored. Procedures (similar to those for protein sequences) have not yet been established for scientists to report the observation of specific glycan structures in specific environments and to store these observations in a generally accepted database.

These are the reasons why we have chosen not to describe in detail the currently available glyco-related databases. It can be anticipated that the development of databases will be subject to rather rapid changes within the next few years, so that the descriptions provided could quickly become obsolete. As compensation, we provide a comprehensive list of web links to glyco-related databases, services, consortia, and communities. As many of the listed URLs are maintained by larger consortia and longer lasting bioinformatics projects, it is the hope that these will provide a more sustainable solution for this book.

Acknowledgments

The authors thank Dr Robin Thomson, Institute for Glycomics, Griffith University, for carefully reading the manuscript and many useful suggestions to improve its readability.

Heidelberg, June 2007

Claus-Wilhelm von der Lieth
Martin Frank
Thomas Lütteke

References

1. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004, **431**:931–945.

2. Haslam SM, North SJ, Dell A: Mass spectrometric analysis of N- and O-glycosylation of tissues and cells. *Curr. Opin. Struct. Biol.* 2006, **16**:584–591.
3. Mechref Y, Novotny MV: Miniaturized separation techniques in glycomic investigations. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* 2006, **841**:65–78.
4. Alvarez RA, Blixt O: Identification of ligand specificities for glycan-binding proteins using glycan arrays. *Methods Enzymol.* 2006, **415**:292–310.
5. Blixt O, Head S, Mondala T, Scanlan C, Huftejt ME, Alvarez R, Bryan MC, Fazio F, Calarese D, Stevens J, *et al.*: Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc. Natl. Acad. Sci. USA* 2004, **101**:17033–17038.
6. Feizi T, Chai W: Oligosaccharide microarrays to decipher the glyco code. *Nat. Rev. Mol. Cell Biol.* 2004, **5**:582–588.
7. Stevens J, Blixt O, Paulson JC, Wilson IA: Glycan microarray technologies: tools to survey host specificity of influenza viruses. *Nat. Rev. Microbiol.* 2006, **4**:857–864.
8. Prescher JA, Bertozzi CR: Chemical technologies for probing glycans. *Cell* 2006, **126**:1–4.
9. Pratt MR, Bertozzi CR: Synthetic glycopeptides and glycoproteins as tools for biology. *Chem. Soc. Rev.* 2005, **34**:58–68.
10. von der Lieth CW, Bohne-Lang A, Lohmann K, Frank M: Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform.* 2004, **5**:164–178.
11. von der Lieth CW, Lütteke T, Frank M: The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. *Biochim. Biophys. Acta* 2006, **1760**:568–577.
12. von der Lieth CW: An endorsement to create open databases for analytical data of complex carbohydrates. *J. Carbohydr. Chem.* 2004, **23**:277–297.
13. Brooksbank C, Camon E, Harris MA, Magrane M, Martin MJ, Mulder N, O'Donovan C, Parkinson H, Tuli MA, Apweiler R, *et al.*: The European Bioinformatics Institute's data resources.. *Nucleic Acids Res.* 2003, **31**:43–50.
14. Brooksbank C, Cameron G, Thornton J: The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* 2005, **33**:D46–D53.
15. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvermin V, Church DM, Dicuccio M, Edgar R, Federhen S, *et al.*: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2007, **35**:D3–D4.
16. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 2004, **32**:D277–280.
17. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M: KEGG as a glycome informatics resource. *Glycobiology* 2006, **16**:63R–70R.
18. Doubet S, Bock K, Smith D, Darvill A, Albersheim P: The Complex Carbohydrate Structure Database. *Trends Biochem. Sci.* 1989, **14**:475–477.
19. Doubet S, Albersheim P: CarbBank. *Glycobiology* 1992, **2**:505.
20. Fox JA, Butland SL, McMillan S, Campbell G, Ouellette BF: The Bioinformatics Links Directory: a compilation of molecular biology web servers. *Nucleic Acids Res.* 2005, **33**:W3–W24.
21. Fox JA, McMillan S, Ouellette BF: A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res.* 2006, **34**:W3–W5.
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res.* 2000, **28**:235–242.
23. Lütteke T, Bohne-Lang A, Loss A, Götz T, Frank M, von der Lieth CW: GLYCOCICIENCES.de: an internet portal to support glycomics and glycobiology research. *Glycobiology* 2006, **16**:71R–81R.

Section 1: Introduction

1 Glycobiology, Glycomics and (Bio)Informatics

Claus-Wilhelm von der Lieth

Formerly at the Central Spectroscopic Unit, Deutsches Krebsforschungszentrum (German Cancer Research Center), 69120 Heidelberg, Germany

1.1 The Role of Carbohydrates in Life Sciences Research

Despite their nearly complete neglect in databases and ‘traditional’ bioinformatics projects, carbohydrates are the most abundant and structurally diverse biopolymers formed in nature. Historically, the chemistry, biochemistry, and biology of carbohydrates were very prominent areas of research over a long period of time during the beginning and the middle of the last century. However, during the initial phase of the development of molecular biology, focusing on DNA, RNA, and proteins, studies of carbohydrates lagged far behind. Among the main reasons for this were the inherent structural complexity of carbohydrates, the difficulty in easily determining their structure, the fact that their biosynthesis cannot be directly predicted from the DNA template, and that no methods are available to amplify complex carbohydrate sequences. The more recent development of a variety of new and highly sensitive analytical tools for exploring the structures of oligosaccharides and for producing larger amounts of pure complex carbohydrates has opened up a new frontier in molecular biology. The term glycobiology, which was introduced in the late 1980s [1], reflects the coming together of the traditional disciplines of carbohydrate chemistry and biochemistry, with modern understanding of the cellular and molecular biology of complex carbohydrates, which are often named glycans in this context. The more recently introduced term “glycomics” [2] describes an integrated systems approach to study structure–function relationships of complex carbohydrates – the glycome – produced by an organism such as human or mouse. The glycome can be described as the glycan complement of the cell or tissue as expressed by a genome at a certain time and location. It includes all types of glycoconjugates: glycoproteins, proteoglycans, glycolipids, peptidoglycans, lipopolysaccharides, and so on. The aim of glycomics projects is to create a cell-by-cell catalog of glycosyltransferase (GT) expression and detected glycan structures using high-throughput techniques such as DNA glycogene chips, glycan microarray screening and mass spectrometric (MS) glycan profiling, combined with efficient bioinformatics tools.

Until recently, the role of complex carbohydrates to function as carriers and/or mediators of biological information was a widely neglected and unexplored area in science. However,

with the awareness that the human genome encodes for a significantly smaller number of genes than was estimated from genomes of lower organisms such as yeast [3], it became obvious that each gene can be used in a variety of different ways depending on how it is regulated. Consequently, the study of post-translational protein modifications, which can alter the functions of proteins, came increasingly into scientific focus. Since then, with glycosylation being the most complex and most frequently occurring co- and post-translational modification, glycobiology research has attracted increasing attention.

About 70% of all sequences deposited in the SWISS-PROT [4] protein sequence database include the potential *N*-glycosylation consensus sequence Asn-X-Ser/Thr (where X can be any amino acid except proline) and thus may be glycoproteins. However, it is well known that not all potential sites are actually glycosylated. Based on an analysis of well-annotated and characterized glycoproteins in SWISS-PROT, it was concluded that more than half of all proteins are glycosylated [5, 6]. However, this number should be regarded as a very crude estimation since this study was hampered by the paucity of reliable, experimentally determined, and carefully assigned glycosylation sites.

The glycans are exposed on the surface of biomolecules and cells. They form flexible, branched structures that can extend 30 Å or further into the solvent. With a molecular weight of up to 3 kDa each, the oligosaccharide groups of mammalian glycoproteins frequently make up a sizable proportion of the mass of a glycoprotein and can cover a large fraction of its surface. The carbohydrate moiety of “proteins” may amount to a few percent of the molecular weight, but can be as much as 90% in some cases. *O*-Linked mucin-type glycoproteins are usually large (more than 200 kDa) with attached *O*-glycan chains at a high density. As many as one in three amino acids may be glycosylated and 50–80% of the total mass is due to carbohydrates [7]. An analysis of the available three-dimensional structures of glycoproteins contained in the PDB [8] revealed that the glycan and the protein parts of glycoproteins behave like semi-independent moieties. This behavior has several important biological consequences:

- *N*-Glycans can be modified without appreciable effects on the protein. Every *N*-linked glycan is subject to extensive modifications. This allows cells to fine-tune the biophysical and biological properties of glycoproteins and to generate the microheterogeneity [9] that is so characteristic of glycoproteins.
- The semi-independent nature of glycans also allows cell types and cells in different stages of differentiation and transformation to imprint on their glycoprotein pool their own specific biochemical characteristics, and thus give their exposed surface a “corporate identity.”
- This “corporate identity” [10] exposed on their surface makes cells recognizable to other cells in a multicellular environment. It allows self-recognition and provides a central theme in development, differentiation, physiology, and disease.

1.2 Glycogenes, Glycoenzymes and Glycan Biosynthesis

The biosynthesis of carbohydrates attached to proteins or to lipids – called glycoconjugates – is fundamentally different to the expression of proteins. Whereas the enzymes required for the translation of the genetic information into a polypeptide chain in the ribosome are always the same for all proteins and amino acids, the subsequent glycosylation is a

non-template-driven process where dozens of different enzymes are involved in the synthesis of the sugar chains attached to proteins or lipids. Depending on which of these enzymes are expressed in the cell that synthesizes a glycoprotein, various different glycan chains can be attached to the protein or lipid. Glycoproteins generally exist as populations of glycosylated variants – called glycoforms – of a single polypeptide [11, 12]. Although the same glycosylation machinery is available to all proteins in a given cell, most glycoproteins emerge with a characteristic glycosylation pattern and heterogeneous populations of glycans at each glycosylation site.

Glucose and fructose are the major carbon and energy sources for organisms as diverse as yeast and human beings (see, e.g., [7]: Monosaccharide Metabolism chapter). Organisms can derive the other monosaccharides needed for glycoconjugate synthesis from these major suppliers. It is important to appreciate that not all of the biosynthetic pathways are equally active in all types of cells.

The biosynthesis of oligosaccharides is primarily determined by sequentially acting enzymes, the glycosyltransferases (GTs), which assemble monosaccharides into linear and branched sugar chains. For this purpose, the monosaccharides must be either imported into the cell or derived from other sugars within the cell. However, a common factor is that all glycoconjugate syntheses require activated sugar nucleotide donors. It has long been known that a nucleotide triphosphate such as uridine triphosphate (UTP) reacts with a glycosyl-1-P to form a high-energy donor sugar nucleotide that can participate in glycoconjugate synthesis. Once the sugar nucleotides have been synthesized in the cytosol (or, in the case of CMP-Neu5Ac, in the cell nucleus), they are topologically translocated, since most glycosylation occurs in the endoplasmic reticulum (ER) and Golgi apparatus. As the negative charge of the sugar nucleotides prevents them from simply diffusing across membranes into these compartments, eukaryotic cells have devised no-energy-requiring sugar nucleotide transporters that deliver sugar nucleotides into the lumen of these organelles [7].

1.2.1 Biosynthetic Pathways

In eukaryotes, more than 10 biosynthetic pathways that link glycans to proteins and lipids [13, 14] are known. The KEGG PATHWAY resource [15, 16] – a collection of pathway maps representing current biochemical knowledge of the molecular interaction and reaction networks – has encoded 18 pathways for the biosynthesis of complex carbohydrates and their metabolism (see Figure 1.1), and 20 pathways for metabolism where carbohydrates are involved. More than 200 enzymes are involved in the biosynthesis of carbohydrate structures found on proteins and lipids. More than 30 different enzymes may participate directly in the synthesis of a single glycan. One of the best-characterized pathways is the biosynthesis of complex oligosaccharides that are subsequently attached to a protein through the side-chain nitrogen atom of the amino acid asparagine (Asn) to give glycoproteins [10, 17, 18] (described in Section 8.1 in Chapter 8). Glycosylation of proteins occurs in all eukaryotes and in many archaea but only exceptionally in bacteria.

O-Linked glycosylation, where carbohydrates are attached to serine (Ser) and threonine (Thr), takes place post-translationally in the Golgi apparatus. The monosaccharides are added one by one in a stepwise series of reactions (Figure 1.2). This is in contrast to the *N*-linked glycosylation pathway where a preformed oligosaccharide is transferred *en bloc* to Asn. A second important difference is that there are no known consensus sequence

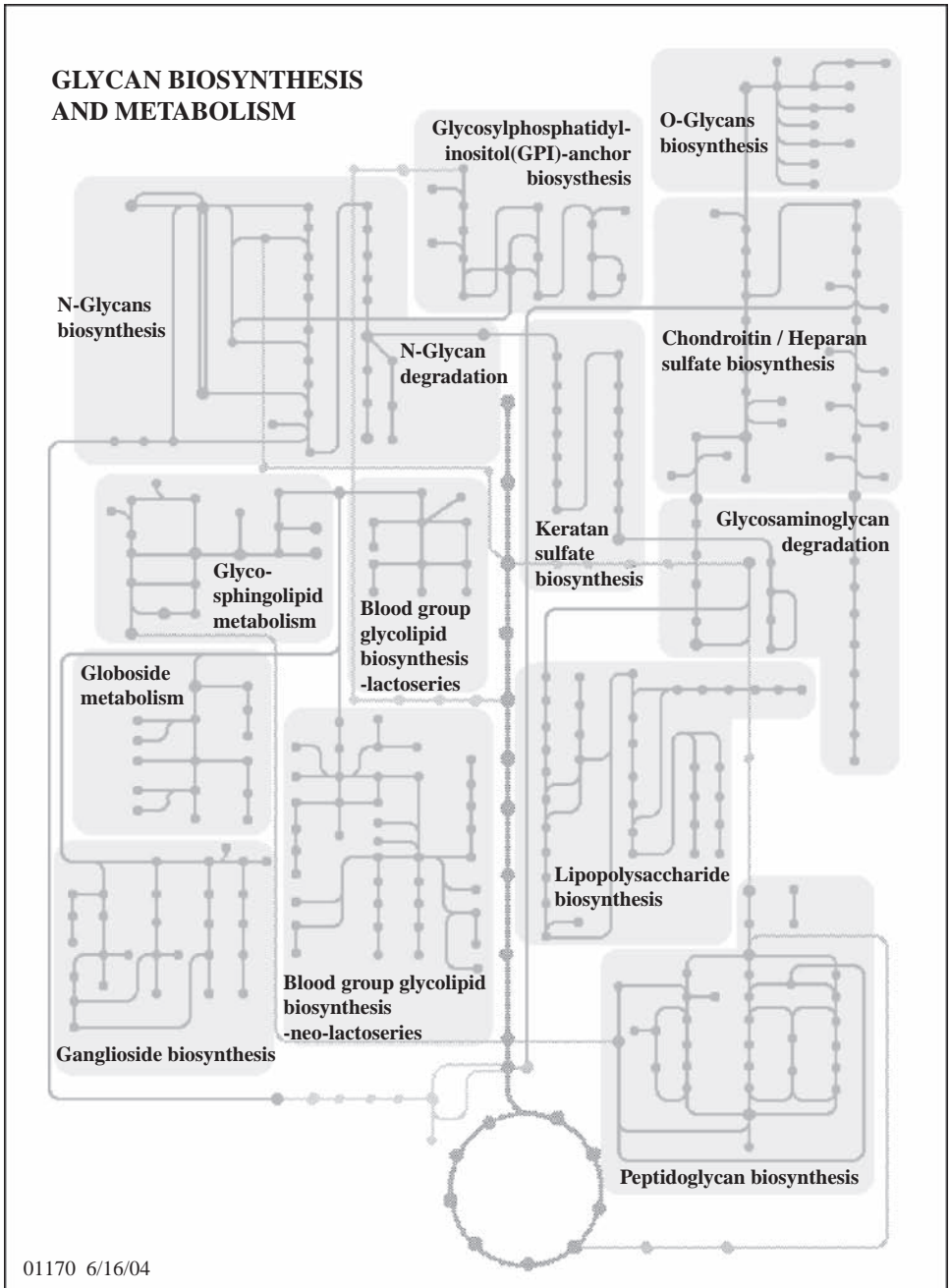


Figure 1.1 Illustration of the pathways for the biosynthesis of complex carbohydrates and their metabolism encoded in KEGG PATHWAY [15, 16] available at: www.genome.jp/kegg/pathway/map/map01170.html.

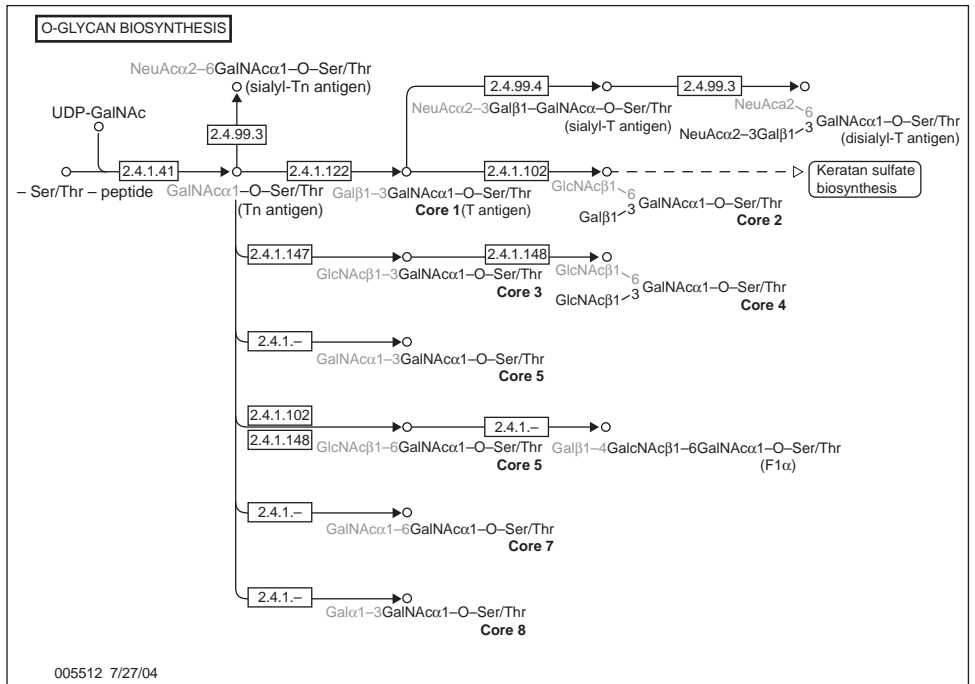


Figure 1.2 Known biosynthesis pathways for carbohydrates attached to the oxygen atom of the side chain of the amino acids serine or threonine as encoded in the KEGG PATHWAY resource [15, 16] (www.genome.jp/kegg/pathway/map/map01170.html). An IUPAC like nomenclature (see Chapter 3) is used to characterize the monosaccharides and linkages. The enzymes are given in the square boxes by their corresponding Enzyme Commission (EC) numbers, which are based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB).

motifs that define an *O*-linked glycosylation site analogous to the Asn-X-Ser/Thr motif for *N*-linked glycosylation.

1.2.2 The Role of Bioinformatics in Identifying Glyco-related Genes

The enzymes required for the biosynthesis of complex carbohydrates can be classified into those needed for the conversion of monosaccharide building blocks to activated sugar nucleotides and their transport within the cell, and those which are used to build (glycosyltransferases) and remodel (glycosidases) glycoconjugates [19]. Many, but not all, of the latter enzymes are found within the ER-Golgi pathway for export of newly synthesized glycoconjugates.

The first mammalian GT gene was reported in 1986 [20]. The progress in identifying new GT genes at that time was slow because they had to be cloned by identifying the partial amino acid sequence of the purified enzyme, which was the limiting step. Thereafter, from the beginning of the 1990s when methods of expression cloning and PCR cloning with degenerated primers were employed, several novel GT genes were detected each year. It became obvious that GTs can be classified into several subfamilies which contain

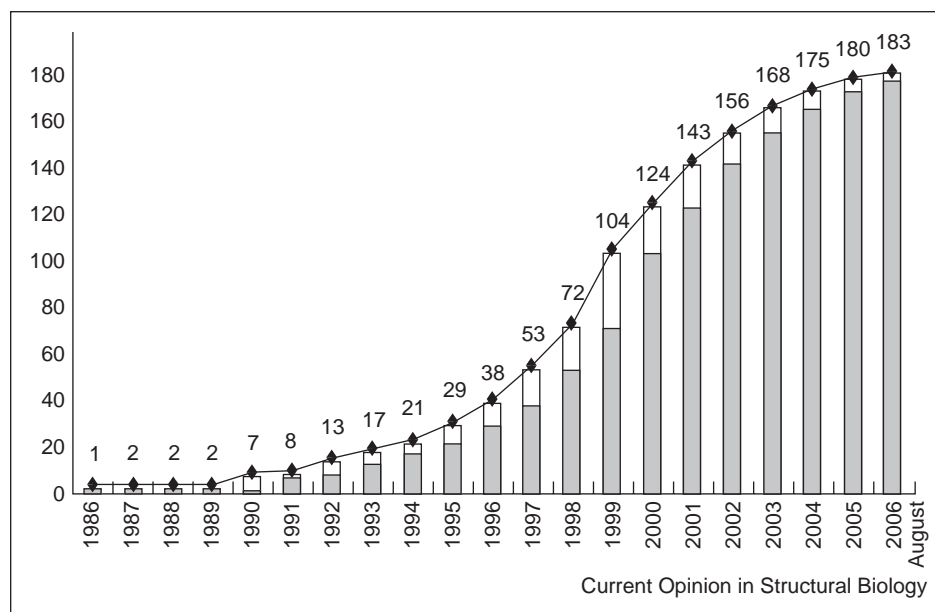


Figure 1.3 Progress in the cloning of glyco-related enzymes (including GTs, sulfotransferases, and sugar–nucleotide transporters). Filled columns indicate the cumulative number of glyco-related enzymes reported during the past two decades. Open columns indicate the number of novel enzymes reported in each year. Reprinted from [24] with permission from Elsevier.

well-conserved sequence motifs. Based on this knowledge and the increasing availability of gene sequences and the development of appropriate bioinformatics searching algorithms, the *in silico* identification of GT genes could be successfully applied [21, 22]. During the middle of the 1990s, the number of newly reported GT genes began to increase significantly, reaching a peak in 1999 (Figure 1.3). This was due to the substantial increase in sequenced genes and the ease of finding new GT genes by homology searching using well-known BLASTN searches. The number of newly identified GT genes began to decrease gradually after 1999 to only five by August 2006, suggesting that mammalian GT gene cloning seems to be approaching its completion. During the past two decades, more than 180 human glyco-related enzymes have been cloned and their substrate specificities analyzed using biochemical approaches [23, 24]. The current status of knowledge compiled for these human GT genes and their links with orthologous genes in other species is summarized in the GlycoGene database [22].

As demonstrated for the identification of GT genes, the application of classical bioinformatics tools and also the use of genomic databases had and will continue to have a significant impact on the rapid development of glycobiology research [25]. The same is true when searching for all lectins with similar binding affinity for a specific carbohydrate, which was also significantly accelerated through systematic analysis of gene sequences for the corresponding sequence motifs [26–29].

However, the use of (bio)informatics in glycobiology research has to be divided between those applications where an explicit description of the glycan structure is required, and those where the proteins to which carbohydrates are attached, the enzymes which build and

modify carbohydrates, or the lectins which recognize a certain sugar epitope, are analyzed. The latter type of applications can be performed using well-known bioinformatics tools such as sequence alignment techniques and attempts to understand the evolutionary relationships through phylogenetic analysis. Where an encoding of the carbohydrate structure is required, however, for example when looking at carbohydrate specificity of a lectin or classification of the glycome of an organism, classical bioinformatics approaches cannot be directly applied.

1.3 Intrinsic Problems of Glycobiology Research

Glycobiologists have to deal with several intrinsic problems, making their research difficult and time consuming, as well as ambitious.

1.3.1 *Carbohydrates Have to Be Analyzed at Physiological Concentrations*

The first major challenge is to develop highly sensitive analytical methods. Since the biosynthesis of complex carbohydrates requires a variety of enzymes, which have to act in a defined and consecutive way, there are currently no methods available to amplify glycans readily in the sense that DNA is amplified using polymerase chain reaction (PCR) techniques. Consequently, highly sensitive analytical methods have to be applied, which are able to detect the small amounts of material found in cells. The chapters on experimental methods will discuss the central analytical methods – mass spectrometry, HPLC and NMR – which are used in different areas of glycobiology to identify glycan structures.

1.3.2 *Complexity of Glycan Structures*

The second major challenge lies in the complexity of glycan structures: each pair of monosaccharide residues can be linked in several ways, and one residue can be connected to three or four others (giving branched structures). The information content which can be potentially encoded by glycans in a given sequence is therefore high. The four nucleotides in DNA can be combined to give 256 four-unit structures, and the 20 amino acids in proteins yield 160 000 four-unit configurations. However, the number of naturally occurring residues is much larger for glycans which have the potential to assemble into more than 15 million four-unit arrangements.

Although oligosaccharides potentially carry this high capacity to store biological information, only a small part thereof is actually used in nature. A recent analysis of the KEGG glycan database [15] containing 4107 unique glycan entries [30], which consist of nine frequently occurring monosaccharides (glucose, galactose, mannose, *N*-acetylglucosamine, *N*-acetylgalactosamine, fucose, xylose, glucuronic acid, and sialic acid) showed, that only 302 (54%) of the 558 (nine monosaccharides, two anomers, 31 substitution possibilities) theoretically possible disaccharides appear in the database. Furthermore, while an enormous number of reaction pattern combinations are theoretically possible, only 2178 of these combinations actually appear in the database. These numbers suggest that the structural diversity of glycans is indeed large, but that the combination of reaction patterns which actually exist in a given cellular environment is limited by the availability of the glyco-related enzymes which build and modify the glycan structures.

1.3.3 Structural Heterogeneity

The third major challenge is the structural heterogeneity and “fuzziness” of glycans. Glycoproteins normally exhibit various glycoforms when isolated from cells and tissues [12, 31, 32]. Often several tens of different glycoforms for a given glycosylation site have been identified. Analytical techniques and also databases and bioinformatics applications have to cope with this phenomenon. Non-stoichiometric modifications to position and amount of chemical substitutions are another unique feature of complex carbohydrates, which requires the development of new concepts to analyze, encode, and handle, for example, the statistical occurrence of sulfate groups at specific positions in glycosaminoglycans such as heparin and heparan sulfate [33, 34].

1.3.4 Multivalent Interactions with Proteins

Glycan-binding proteins mediate diverse aspects of cell biology, including pathogen recognition of host cells, cell trafficking, endocytosis, and modulation of cell signaling [35]. However, the assignment of biological function to carbohydrates in recognition events is complex because individual glycan structures exhibit only very weak interactions with a protein surface. For example, the binding affinity of monovalent oligosaccharide ligands to their respective viral receptors is rather weak, with dissociation constants (K_d) of around 10^{-3} – 10^{-4} M. This low affinity is in strong contrast to the high K_d values of 10^{-8} – 10^{-12} M determined for the binding of complete virions to cell surfaces [36]. It is widely assumed that this high affinity is contributed to by multivalent binding of the repetitive virion surface carbohydrate-recognizing proteins/receptors to repetitive oligosaccharide structures on the cell surface. Unlike protein–protein interactions, which can be generally viewed as “digital” in regulating function, glycan–protein interactions impinge on biological functions in a more “analog” fashion that can in turn “fine-tune” a biological response. This fine-tuning by glycans is achieved through the graded affinity, avidity, and multivalency of their interactions.

1.3.5 New Insights Through Highly Sensitive Analytical Techniques

Much of the increase in a better understanding of the versatile regulatory role of glycans in life can be credited to improvements in existing, and the development of new, highly sensitive analytical techniques. The details of the current status of the analytical techniques will be discussed in detail in the chapters on experimental methods. Here, an especially impressive example will be briefly described, where the combination of modern biomolecular and analytical techniques was used to provide detailed insights into the molecular basis of the receptor specificity of the 1918 so-called Spanish flu.

It is well known that infection with viruses and bacteria often starts with specific interactions with glycans on the surfaces of host cells. For example, the host specificity of influenza A virus infection is mediated by the viral surface glycoprotein hemagglutinin (HA), which binds to host-cell receptors containing glycans with terminal sialic acids.

The impact of influenza infection is felt globally each year, as this disease develops in approximately 20% of the world’s population. The 1918 “Spanish” influenza pandemic represents the largest recorded outbreak of any infectious disease, causing about 20 million deaths. At the end of the 1990s, an American research team was able to detect fragments of the viral genome in lung samples taken from the body of an Inuit woman victim of

the pandemic buried in the Alaskan tundra and a number of preserved samples taken from American soldiers of the First World War. Using modern biomolecular amplification techniques, the entire coding sequence of 1701 nucleotides for the viral surface HA was amplified in 22 overlapping fragments such that the sequences for matching primers could be confirmed [37].

The HA of influenza virus mediates receptor binding and membrane fusion, the first stages of virus infection. The sequences found for the 1918 HA did not reveal any characteristics that were obviously responsible for the extreme virulence of the 1918 pandemic. Independently, two research groups succeeded in growing crystals of the 1918 HA and analyzed its binding properties [38, 39]. The carbohydrate recognition of influenza virus HA is highly specific: whereas human viruses infect epithelial cells in the lungs and upper respiratory tract which have α 2,6-linked sialic acids on their surfaces, avian viruses preferentially bind to α 2,3-linked sialic acids [40]. This slight structural difference in the recognized sugar epitope obviously prevents a spread of the influenza virus infection across species. Analysis of the binding specificity of the highly virulent 1918 influenza virus HA using the glycan array of the US Consortium for Functional Glycomics (CFG) revealed a clear preference for α 2,6-linked sialylgalactose motifs [41]. Glycan microarrays are a relatively new and highly specific technology that allows rapid determination of glycan-binding protein interactions and specificities.

Subsequently it was shown [42] that a single amino acid substitution in the 1918 human influenza virus HA – Asp225 to Gly – changes receptor binding specificity from an HA which preferentially binds to the human α 2,6-sialylgalactose motif to one which binds both the human α 2,6- and the α 2,3-sialylgalactose motif of the avian cellular receptors. Mutation of a further single amino acid back to the avian consensus – Asp190 to Glu – resulted in a preference for the avian receptor. Thus, the species barrier, as defined by the receptor specificity preferences, of 1918 human viruses compared with likely avian virus progenitors, can be circumvented by changes at only two positions in the HA receptor binding site.

A combination of highly sophisticated new techniques revealed that the HA of the 1918 influenza virus might be more like that found in avian influenza than was previously thought. Usually, avian influenza strains do not affect humans directly because bird-adapted HA proteins are not able to bind well to human receptors. Until very recently, it was thought that to make the leap to humans successfully a bird strain must pass through an intermediate animal that contains both bird and human receptors, such as a pig. The new findings suggest that minimal changes in the receptor binding domain of an avian HA may have been enough to broaden its binding targets to include the major sialic acid receptor expressed on human respiratory epithelium.

Modern molecular biology techniques and highly sensitive analytical tools have helped, 80 years after its outbreak, to give some new insights into why the 1918 influenza virus was so devastating. Additionally, glycan microarray technology has been proven to have the ability to detect rapidly strains which have the potential capability to cross species barriers, a major goal for worldwide influenza surveillance [43].

1.4 Carbohydrates as a New Frontier in Pharmaceutical Research

Except for sulfated glycan heparin [44], which belongs to the class of glycosaminoglycans (GAGs), synthetic carbohydrates have not been widely used as therapeutics. One obvious

reason is that complex carbohydrates are difficult to synthesize. The recent development of a (semi-)automated oligosaccharide synthesizer greatly accelerates the assembly of complex, naturally occurring carbohydrates and also chemically modified oligosaccharide structures (mimetics) and promises to have major impact on the field of glycobiology [45, 46]. Synthetic carbohydrates and glycoconjugates will be more readily available for broad use, and will advance the study of their roles in biologically important processes such as inflammation, cell–cell recognition, immunological response, metastasis, and fertilization. Tools such as microarrays, surface plasmon resonance spectroscopy, and fluorescent carbohydrate conjugates to map interactions of carbohydrates in biological systems are available [47–49] and can be used to evaluate systematically the binding specificity and strength of naturally occurring carbohydrates and also mimetics thereof.

1.4.1 Carbohydrates in Drug and Vaccine Development

Bacteria, viruses, and parasites are the major agents leading to disease. All cells in nature are covered with a dense and complex coat of glycans. A wide variety of pathogens initiate infection by binding to the surface glycans of host cells. This is not surprising as cell-surface glycans are the first molecules encountered by pathogens when they contact potential host cells or their secretions. Outer, terminal glycan sequences such as those carrying sialic acid residues are even more likely to be preferred targets, as they are the first residues that pathogens encounter. Examples of disease in which cell-surface glycan recognition is involved include influenza virus infection of the lung and upper respiratory tract, erythrocyte invasion by the malaria parasite *Plasmodium falciparum*, *Helicobacter pylori* infection of the stomach, and intestinal diarrhea caused by the toxin of *Vibrio cholerae*.

In the case of influenza virus, as described above, infection is mediated by the viral surface glycoprotein hemagglutinin which binds to host-cell receptors containing glycans with terminal sialic acids. Surface binding is followed by penetration of the cellular membrane. Complex glycans are involved in cellular adhesion, internalization, and the release of newly formed virus particles, all of which are of high interest for preventive medicine and drug design. Highly potent inhibitors of the viral enzyme neuraminidase, which facilitates release of progeny influenza virus from infected host cells, have been designed with the help of computational chemistry methods using 3D structures of the enzyme. The neuraminidase inhibitors mimic the form of sialic acid seen in the transition state of the enzyme reaction, the cleavage of terminal sialic acid residues from glycans. Neuraminidase inhibitors have been shown to be effective against all neuraminidase subtypes and, therefore, against all strains of influenza, a key point in epidemic and pandemic preparedness. These new drugs have great potential for diminishing the effects of influenza infection [50, 51].

Glycoconjugate vaccines provide effective prophylaxis against bacterial infections. However, only a few vaccines have been developed by chemical synthesis of the key carbohydrate antigens. In Cuba, it was demonstrated that a conjugate vaccine composed of a synthetic capsular polysaccharide antigen of *Haemophilus influenzae* type b (Hib) elicited long-term protective antibody titers [52, 53]. This demonstrates that access to synthetic complex carbohydrate-based vaccines is feasible and provides a basis for further development of similar approaches for other human pathogens [54]. Hib was the leading cause of bacterial meningitis in many parts of the world before the introduction of conjugate vaccines. The use of vaccines against Hib in developing countries is expected to be an important tool for the reduction of vaccine-preventable morbidity and mortality among children less than 5 years old.

About 40% of the world's population live with the risk of contracting malaria. Although only about 1% of all malaria cases are lethal, malaria continues to claim the lives of over two million people annually. No viable vaccine candidate has been developed for malaria. Glycosylphosphatidylinositol (GPI) anchors are a class of naturally occurring glycolipids that link proteins and glycoproteins via their C-terminus to cell membranes. The malarial parasite *Plasmodium falciparum* expresses GPI in protein anchored and free form on the cell surface: the GPI constitutes a toxin which is implicated in the pathogenesis and fatalities of malaria in humans [55]. Recently, it could be demonstrated that mice vaccinated with a synthetic GPI glycan conjugated to a carrier protein produced anti-GPI antibodies and had a greatly improved chance of survival upon infection with *P. falciparum*. Between 60 and 75% of vaccinated mice survived, compared with 0–9% of sham-immunized mice. The parasite levels observed in the blood of the vaccine and control groups did not differ significantly, thus indicating that the synthetic GPI glycan conjugate serves as an anti-toxin vaccine [56].

1.4.2 Carbohydrates Play a Key Role in Many Diseases

Many diseases are caused by disruption of regulatory and control mechanisms within a particular organism. For example, a DNA point mutation may result into a single amino acid replacement in a protein, which may completely change or obliterate the function of the protein. Such mutations may occur in somatic cells of adult individuals, or they may be inherited, resulting in inborn defects, such as congenital disorders of glycosylation (CDGs) – defects in glycan biosynthesis, lysosomal storage diseases – defective glycan catabolism, and von Willebrand disease. Cancer and some autoimmune diseases, such as rheumatism, are other examples of diseases caused by failure of the organism's regulation and control system. Cancer is associated with changes in glycosylation of proteins exposed on the outer cell surface. Therefore, monitoring of temporal changes in glycosylation has potential as a diagnostic tool and as a prognostic indicator. Furthermore, cancer cell-specific complex glycans may also serve as targets for tissue- or cell-selective delivery of agents that can kill tumor cells.

In addition to the effects of altered glycan biosynthesis/catabolism in disease, complex carbohydrate epitopes play key roles in allergy and immune reactions against parasites. They are also of great significance in xeno-transplantation, where species-specific carbohydrate structures can be recognized as non-self and promote tissue rejection. On the other hand, synthetic manipulation of glycosylation patterns is being used to advantage in the biotechnological production of recombinant therapeutic glycoproteins; for example, an increase in the number of sialylated glycans on erythropoietin (EPO) increases its serum half-life [57].

An emerging area of research is so-called metabolic oligosaccharide engineering, the goal of which is a biosynthetically altered cell-surface repertoire through the introduction of unnatural sugar residues into cellular glycans. Such engineered cell surfaces are extremely useful systems for studying biochemistry and cell biology in a broad range of contexts, such as cell–cell interactions.

1.5 A Short History of Databases and Informatics for Glycobiology

It can be expected that the rapid evolution of glycomics, including glycan array technologies, will result in very large data collections that will have to be organized, analyzed, and compared, requiring standards for structural representation. The development and use

of informatics tools and databases for glycobiology and glycomics research has increased considerably in recent years; however, it can still be considered as being in its infancy when compared with the genomics and proteomics areas. The intrinsic factors which make the development of informatics for glycobiology and glycomics a challenging task have been described above. However, there is a general consensus within the community of glycoscientists that the availability of comprehensive and up-to-date carbohydrate databases, and also efficient software to retrieve and handle the data, will be a prerequisite for successfully conducting large-scale glycomics projects aimed at deciphering new, so far unknown, biological functions of glycans. Here, a short overview of the history of databases and informatics for glycobiology will be given.

1.5.1 *The Early Days: CarbBank*

Before information technologies were available, it was a rather time-consuming task to cope with all structures of complex carbohydrates detected in nature, which were published in various journals using different ways to describe structural details. Normally, only a few specialists in the field could successfully access the available knowledge. When digital documentation systems and search engines were introduced into science during the 1980s, it was recognized that this new technology could also be very useful for encoding and retrieving all published glycan structures using a language which was well understood by glycoscientists. In light of this, the Complex Carbohydrate Structure Database (CCSD) [58, 59] – often referred to as *CarbBank* according to the retrieval software used to access the data – was established in the mid-1980s, the main purpose of which was to allow the user to find easily all publications in which specific carbohydrate structures were reported. The CCSD was developed and maintained by the Complex Carbohydrate Research Center of the University of Georgia (USA) and funded by the National Institutes of Health (NIH). The need to develop CarbBank as an international effort was clearly recognized and resulted in worldwide curation teams responsible for specific classes of glycans. During the 1990s, a Dutch group assigned NMR spectra to CCSD entries (*SugaBase*) [60, 61]. This was the first attempt to create a carbohydrate NMR database that complemented CCSD entries with proton and carbon chemical shift values.

For a variety of reasons, including disagreement on the best ways to integrate the CCSD into the growing bioinformatics environment and the need to provide more user-friendly tools compatible with new, Internet-based approaches, the funding for the CCSD stopped during the second half of the 1990s. In a letter sent to the provider of CarbBank in 1998, I wrote concerning the infrequent use of the database: “*One rather obvious reason for this situation is that carbohydrate data collections only rarely exhibit cross-referencing to other available data on the net. CarbBank uses efficient algorithms to provide rapid access to references following the input of a query expressed in terms of the carbohydrate nomenclature. Unfortunately, only pure bibliographic information such as authors, journal, and title are displayed, but not abstracts. Using modern Web techniques, it should be rather straightforward to send a request to the public WEB-Medline (PubMed) and provide elegant access to abstracts. One of the big disadvantages of essentially all carbohydrate Web applications is that there are no annotated and/or cross-referenced implementations, which allow glycoscientists to find important data for the compound of interest in a compact and well-structured representation. Most carbohydrate applications on the Web are designed to answer just one special question.*”

Unfortunately, CarbBank was not developed further and, beyond 1996, the CCSD was no longer updated. An attempt to transfer responsibility for updating the CCSD to a volunteer team of glycoscientists around the world obviously failed. Nevertheless, with 49 897 entries, which correspond to 23 118 distinct glycan structures, the CCSD is still the largest publicly available repository of glycan-related data. All subsequent open access projects initiated at the beginning of the new century made use of the CCSD data.

1.5.2 *Beyond CarbBank*

The collapse of CarbBank was extremely frustrating, especially for those who were involved in this international venture. There was very little support for renewal of the project, as the bioinformatics field – concentrating at that time on the sequencing of the human genome – completely ignored the potential of carbohydrates as a repository of biological information.

A small informatics oriented group of scientists at the DKFZ (German Cancer Research Center) in Heidelberg, initially interested in elucidating the conformational space of complex carbohydrates, first put forward the imperative to develop informatics for glycobiology as an independent sub-branch of bioinformatics. This group also realized the need to make the CCSD entries publicly available using modern Internet-based tools and to cross-reference the glyco-related data with proteomics and glycomics information. These ideas led to the development of the *GLYCOSCIENCES.de* [62, 63] portal and the *EUROCarbDB* (www.eurocarbdb.org) project.

At the beginning of the new century, when the gap between encoded and published glycan structures became obvious, several companies started to provide commercial access to glyco-related data, which they extracted from the literature. However, due to limited commercial success, most of these services stopped. The Australian GlycoSuite [64], the only one of these services that survives today, is willing to provide academic users with free access to the data they have extracted from literature.

1.5.3 *Glycomics Initiatives – a New Stimulus for Glycoinformatics Development*

An important stimulation for glycoinformatics development was the establishment of the Consortium for Functional Glycomics (www.functionalglycomics.org) in 2001. This was the first large-scale project that clearly emphasized the need for informatics to manage and annotate automatically the vast amount of experimental data generated by glycomics research. The development of algorithms for the automatic interpretation of mass spectra – a severe bottleneck that hampers the rapid and reliable interpretation of MS data in high-throughput glycomics projects – is critical for all glycomics projects [65]. This is still the most active area of software development, where various primarily experimentally oriented groups have been developing software solutions and algorithms to solve their specific scientific questions.

Another important step was the integration of glyco-related biological pathways into the schemata of the first ‘classical’ bioinformatics initiative – the Kyoto Encyclopedia of Genes and Genomes (KEGG). Subsequent development of associated databases for glycan structures led to the KEGG GLYCAN [16, 66] approach, which elegantly established the connection between glycan structures and the knowledge of enzymatic reactions to

build the glycan structures. Additionally, the KEGG group made significant progress in applying bioinformatics algorithms to the tree-like structures of glycans for comparison and alignment, to develop similarity scores, and to establish a global view of all glycans belonging to related pathways (see also Chapter 7).

As a consequence of the increasing interest in glycomics research, various new databases were started in recent years (for examples, see the link list at www.eurocarbodb.org/links/). Among these, the EUROCarbDB project (a distributed bottom to top initiative for primary experimental data), the Russian Bacterial Carbohydrate Structure Database (aiming to cover all known structures produced in bacteria), and the *Bioinformatics for Glycan Expression* initiative (development of glyco-related ontologies) of the Complex Carbohydrate Research Center are the largest ones. In general, the development of glyco-related tools and databases can be described as a small but fairly active field of research.

1.5.4 *The current situation*

The current situation in glycoinformatics is characterized by the existence of multiple disconnected and incompatible islands of experimental data, data resources, and specific applications, managed by various consortia, institutions, or local groups. These resources rarely provide communication mechanisms that would permit the widest advantage to be taken of these data by allowing their combination and comparison. However, approaches to link the distributed data have been conceptually worked out and examples are already being implemented. The collaborative spirit recently exhibited by all of the major glycomics initiatives will significantly help to overcome the current unfavorable situation. This positive spirit has recently led to an important milestone, the agreement of an XML standard format for the exchange of glycan structures (GLYDE-II) [67].

None of the existing initiatives has the capacity to fulfill completely the goal of CarbBank at the beginning of the 1990s, that is, to provide comprehensive access to all published carbohydrate structures. In particular, the existing initiatives do not have the worldwide resources to fill the gap of published glycan structures that were not included in CarbBank after its termination in the mid-1990s.

It is likely that the tendency to set up local databases designed to support specific areas of research in glycobiology will continue in the near future. The existence of a centralized glycan structure database would substantially increase the ability to annotate and cross-reference local data with other bioinformatics resources. Offering clear guidelines describing the minimal requirements of data exchange formats, which are required for databases to communicate with each other, will hopefully lead to strong interconnections and compatibility among glycobiology and glycomics databases.

1.5.5 *The Future*

It is clear that there is an urgent need to develop databases and informatics for glycobiology and glycomics. The developments in glycomics will produce an enormous amount of data and there is a need to cut across multiple datasets to understand fully the structure–function relationships of complex carbohydrates. A critical component that will facilitate this process is a bioinformatics platform to store, integrate, and process the recorded data, to condense them to information and knowledge. Several statements of international scientific institutions underpin this direction:

- The European Science Foundation has published a statement *Structural Medicine: The Importance of Glycomics for Health and Disease* (see www.eurocarbdb.org), which emphasizes the need to develop glyco-related databases further.
- In September 2006, the NIH organized a workshop, *Frontiers in Glycomics*. The workshop was the largest meeting focused on the development of databases and informatics for glycomics and glycobiology. A white paper was compiled which set priorities for the most important steps to develop the field [67].
- The outcome of this meeting was, on the one hand, an agreement to accept a standard exchange format for glycan structures called GLYDE-II, and on the other, a list of the most urgent needs – top priority is a centralized, comprehensive, and highly curated carbohydrate structure database.
- The European Strategy Forum for Research Infrastructures (<http://cordis.europa.eu/esfri/>) published a roadmap emphasizing that “*modern science is inconceivable without recourse to well structured, continuously upgraded (...) and freely accessible databases (...). The bioinformatics infrastructure (...) will continue to expand, requiring successive investments for major upgrades, and will remain the depository of biological information for as long as we now can foresee.*”

References

1. Rademacher TW, Parekh RB, Dwek RA: Glycobiology. *Annu Rev Biochem* 1988, **57**:785–838.
2. Hirabayashi J, Arata Y, Kasai K: Glycome project: concept, strategy and preliminary application to *Caenorhabditis elegans*. *Proteomics* 2001, **1**:295–303.
3. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004, **431**:931–945.
4. Apweiler R, Bairoch A, Wu CH: Protein sequence databases. *Curr Opin Chem Biol* 2004, **8**:76–80.
5. Apweiler R, Hermjakob H, Sharon N: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999, **1473**:4–8.
6. Ben-Dor S, Esterman N, Rubin E, Sharon N: Biases and complex patterns in the residues flanking protein N-glycosylation sites. *Glycobiology* 2004, **14**:95–101.
7. Varki A, Cummings R, Esko J, Freeze H, Hart G, Marth J: *Essentials of Glycobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1999.
8. Petrescu A-J, Milac A-L, Petrescu SM, Dwek RA, Wormald MR: Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure and folding. *Glycobiology* 2004, **14**:103–114.
9. Rudd PM, Wormald MR, Stanfield RL, Huang M, Mattsson N, Speir JA, DiGennaro JA, Fetrow JS, Dwek RA, Wilson IA: Roles for glycosylation of cell surface receptors involved in cellular immune recognition. *J Mol Biol* 1999, **293**:351–366.
10. Helenius A, Aebi M: Roles of N-linked glycans in the endoplasmic reticulum. *Annu Rev Biochem* 2004, **73**:1019–1049.
11. Haslam SM, North SJ, Dell A: Mass spectrometric analysis of N- and O-glycosylation of tissues and cells. *Curr Opin Struct Biol* 2006, **16**:584–591.
12. Rudd PM, Dwek RA: Glycosylation: heterogeneity and the 3D structure of proteins. *Crit Rev Biochem Mol Biol* 1997, **32**:1–100.
13. Spiro RG: Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* 2002, **12**:43R–56R.
14. Freeze HH: Genetic defects in the human glycome. *Nat Rev Genet* 2006, **7**:537–551.

15. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004, **32**:D277–D280.
16. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006, **34**:D354–D357.
17. Kornfeld R, Kornfeld S: Assembly of asparagine-linked oligosaccharides. *Annu Rev Biochem* 1985, **54**:631–664.
18. Helenius A, Aebi M: Intracellular functions of N-linked glycans. *Science* 2001, **291**:2364–2369.
19. Taniguchi N, Miyoshi E, Jianguo G, Honke K, Matsumoto A: Decoding sugar functions by identifying target glycoproteins. *Curr Opin Struct Biol* 2006, **16**:561–566.
20. Narimatsu H, Sinha S, Brew K, Okayama H, Qasba P: Cloning and sequencing of cDNA of bovine N-acetylglucosamine (beta 1–4)galactosyltransferase. *Proc Natl Acad Sci USA* 1986, **83**:4720–4724.
21. Kikuchi N, Kwon YD, Gotoh M, Narimatsu H: Comparison of glycosyltransferase families using the profile hidden Markov model. *Biochem Biophys Res Commun* 2003, **310**:574–579.
22. Kikuchi N, Narimatsu H: Bioinformatics for comprehensive finding and analysis of glycosyltransferases. *Biochim Biophys Acta* 2006, **1760**:578–583.
23. Narimatsu H: Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj J* 2004, **21**:17–24.
24. Narimatsu H: Human glycogene cloning: focus on beta 3-glycosyltransferase and beta 4-glycosyltransferase families. *Curr Opin Struct Biol* 2006, **16**:567–575.
25. Schachter H: Protein glycosylation lessons from *Caenorhabditis elegans*. *Curr Opin Struct Biol* 2004, **14**:607–616.
26. Drickamer K, Taylor ME: Identification of lectins from genomic sequence data. *Methods Enzymol* 2003, **362**:560–567.
27. Drickamer K, Fadden AJ: Genomic analysis of C-type lectins. *Biochem Soc Symp* 2002, **69**:59–72.
28. Houzelstein D, Goncalves IR, Fadden AJ, Sidhu SS, Cooper DN, Drickamer K, Leffler H, Poirier F: Phylogenetic analysis of the vertebrate galectin family. *Mol Biol Evol* 2004, **21**:1177–1187.
29. Amado M, Almeida R, Schwientek T, Clausen H: Identification and characterization of large galactosyltransferase gene families: galactosyltransferases for all functions. *Biochim Biophys Acta* 1999, **1473**:35–53.
30. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M: Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* 2005, **21**:3976–3982.
31. Chalabi S, Panico M, Sutton-Smith M, Haslam SM, Patankar MS, Lattanzio FA, Morris HR, Clark GF, Dell A: Differential O-glycosylation of a conserved domain expressed in murine and human ZP3. *Biochemistry* 2006, **45**:637–647.
32. Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA: The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu Rev Immunol* 2007, **25**:21–50.
33. Coombe DR, Kett WC: Heparan sulfate–protein interactions: therapeutic potential through structure–function insights. *Cell Mol Life Sci* 2005, **62**:410–424.
34. Capila I, Linhardt RJ: Heparin–protein interactions. *Angew Chem Int Ed* 2002, **41**:390–412.
35. Collins BE, Paulson JC: Cell surface biology mediated by low affinity multivalent protein–glycan interactions. *Curr Opin Chem Biol* 2004, **8**:617–625.
36. Herrmann M, von der Lieth CW, Stehling P, Reutter W, Pawlita M: Consequences of a subtle sialic acid modification on the murine polyomavirus receptor. *J Virol* 1997, **71**:5922–5931.
37. Reid AH, Fanning TG, Hultin JV, Taubenberger JK: Origin and evolution of the 1918 “Spanish” influenza virus hemagglutinin gene. *Proc Natl Acad Sci USA* 1999, **96**:1651–1656.
38. Stevens J, Corper AL, Basler CF, Taubenberger JK, Palese P, Wilson IA: Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science* 2004, **303**:1866–1870.

39. Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, Ha Y, Vasisht N, Steinhauer DA, Daniels RS, Elliot A, Wiley DC, Skehel JJ: The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science* 2004, **303**:1838–1842.
40. Connor RJ, Kawaoka Y, Webster RG, Paulson JC: Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. *Virology* 1994, **205**:17–23.
41. Stevens J, Blixt O, Glaser L, Taubenberger JK, Palese P, Paulson JC, Wilson IA: Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. *J Mol Biol* 2006, **335**:1143–1155.
42. Glaser L, Stevens J, Zamarin D, Wilson IA, García-Sastre A, Tumpey TM, Basler CF, Taubenberger JK, Palese P: A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *J Virol* 2005, **79**:11533–11536.
43. Stevens J, Blixt O, Paulson JC, Wilson IA: Glycan microarray technologies: tools to survey host specificity of influenza viruses. *Nat Rev Microbiol* 2006, **4**:857–864.
44. Volpi N: Therapeutic applications of glycosaminoglycans. *Curr Med Chem* 2006, **13**:1799–1810.
45. Sears P, Wong CH: Toward automated synthesis of oligosaccharides and glycoproteins. *Science* 2001, **291**:2344–2350.
46. Seeberger PH, Werz DB: Automated synthesis of oligosaccharides as a basis for drug discovery. *Nat Rev Drug Discov* 2005, **4**:751–763.
47. Wang D: Carbohydrate microarrays. *Proteomics* 2003, **3**:2167–2175.
48. de Paz JL, Horlacher T, Seeberger PH: Oligosaccharide microarrays to map interactions of carbohydrates in biological systems. *Methods Enzymol* 2006, **415**:269–292.
49. Feizi T, Chai W: Oligosaccharide microarrays to decipher the glyco code. *Nat Rev Mol Cell Biol* 2004, **5**:582–588.
50. Wilson JC, von Itzstein M: Recent strategies in the search for new anti-influenza therapies. *Curr Drug Targets* 2003, **4**:389–408.
51. Stiver G: The treatment of influenza with antiviral drugs. *CMAJ* 2003, **168**:49–56.
52. Torano G, Toledo ME, Baly A, Fernandez-Santana V, Rodriguez F, Alvarez Y, Serrano T, Musachio A, Hernandez I, Hardy E, Rodriguez A, Hernandez H, Aguila rA, Sanchez R, Diaz M, Muzio V, Dfana J, Rodriguez MC, Heynngnezz L, Verez-Bencomo V: Phase I clinical evaluation of a synthetic oligosaccharide-protein conjugate vaccine against *Haemophilus influenzae* type b in human adult volunteers. *Clin Vaccine Immunol* 2006, **13**:1052–1056.
53. Fernandez Santana V, Pena Icart L, Beurret M, Costa L, Verez Bencomo V: Glycoconjugate vaccines against *Haemophilus influenzae* type b. *Methods Enzymol* 2006, **415**:153–163.
54. Werz DB, Seeberger PH: Carbohydrates as the next frontier in pharmaceutical research. *Chem Eur J* 2005, **11**:3194–3206.
55. Kwon Y-U, Soucy RL, Snyder DA, Seeberger PH: Assembly of a series of malarial glycosylphosphatidylinositol anchor oligosaccharides. *Chem Eur J* 2005, **11**:2493–2504.
56. Schofield L, Hewitt MC, Evans K, Simos M-A, Seeberger PH: Synthetic GPI as a candidate anti-toxic vaccine in a model of malaria. *Nature* 2002, **418**:785–789.
57. Vansteenkiste J, Rossi G, Foote M: Darbepoetin alfa: a new approach to the treatment of chemotherapy-induced anaemia. *Expert Opin Biol Ther* 2003, **3**:501–508.
58. Doubet S, Bock K, Smith D, Darvill A, Albersheim P: The Complex Carbohydrate Structure Database. *Trends Biochem Sci* 1989, **14**:475–477.
59. Doubet S, Albersheim P: CarbBank. *Glycobiology* 1992, **2**:505.
60. van Kuik JA, Vliegenthart JF: Databases of complex carbohydrates. *Trends Biotechnol* 1992, **10**:182–185.
61. van Kuik JA, Hård K, Vliegenthart JFG: A ¹H NMR database computer program for the analysis of the primary structure of complex carbohydrates. *Carbohydr Res* 1992, **235**:53–68.
62. Loss A, Bunsmann P, Bohne A, Loss A, Schwarzer E, Lang E, von der Lieth CW: SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res* 2002, **30**:405–408.

63. Lütteke T, Bohne-Lang A, Loss A, Götz T, Frank M, von der Lieth CW: GLYCOCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* 2006, **16**:71R–81R.
64. Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH: GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res* 2003, **31**:511–513.
65. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R: Advancing glycomics: implementation strategies at the Consortium for Functional Glycomics. *Glycobiology* 2006, **16**:82R–90R.
66. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M: KEGG as a glycome informatics resource. *Glycobiology* 2006, **16**:63R–70R.
67. Packer NH, von der Lieth CW, Aoki-Kinoshita KF, Lebrilla CB, Paulson JC, Raman R, Rudd P, Sasisekharan R, Taniguchi N, York WS: Frontiers in glycomics: Bioinformatics and biomarkers in disease. An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11–13, 2006). *Proteomics* 2008, **8**:8–20.

Section 2: Carbohydrate Structures

2 Introduction to Carbohydrate Structure and Diversity

**Stephan Herget¹, René Ranzinger¹, Robin Thomson²,
Martin Frank¹ and Claus-Wilhelm von der Lieth¹**

¹Deutsches Krebsforschungszentrum (German Cancer Research Centre), Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany

²Institute for Glycomics, Griffith University - Gold Coast Campus, Queensland 4222, Australia

Introduction

Carbohydrates – often also called saccharides or glycans – are the most abundant biological molecules, and are ubiquitously present in the living world [1, 2]. They fulfill numerous functions in Nature. For many organisms, such as insects and plants, carbohydrate chains, for example chitin and cellulose, form the principal structural components. Carbohydrates linked to proteins and lipids – so-called glycoconjugates – play important structural roles and are also often involved in cell communication and signaling events. Protein glycosylation, the covalent attachment of oligosaccharides to proteins during biosynthesis, is the most frequent co- and post-translational modification (PTM) (see also Chapters 8 and 9). A number of excellent reviews and books exist, in which the structural diversity of carbohydrates is discussed with respect to their properties and also biological occurrence and physiological functions (see, for example, [3–7]).

From a bioinformatics point of view, the logical way to approach the diversity of carbohydrate structures starts with a database analysis. Consequently, in this chapter, we present an overview of carbohydrate structures stored in publicly available databases, with a focus on those occurring in mammals, although we also touch on structural aspects of bacterial and plant saccharides. It has to be kept in mind that – at the time of writing – the content of current carbohydrate databases does not represent the existing knowledge manifested in scientific publications since not all discovered carbohydrate structures are stored in a database. Nevertheless, the digitally available carbohydrate structures are appropriate to describe the frequently occurring structural features of classes of glycans found in various species.

Classical bioinformatics is essentially focused on storing and analyzing sequences of the two well-characterized classes of macromolecules, DNA/RNA and proteins, which are composed of a limited number of building blocks (residues) – four nucleotides each for DNA and RNA and 20 amino acids for proteins. A complete constitutional description of these

Table 2.1 Major structural differences between nucleotide, protein, and carbohydrate sequences.

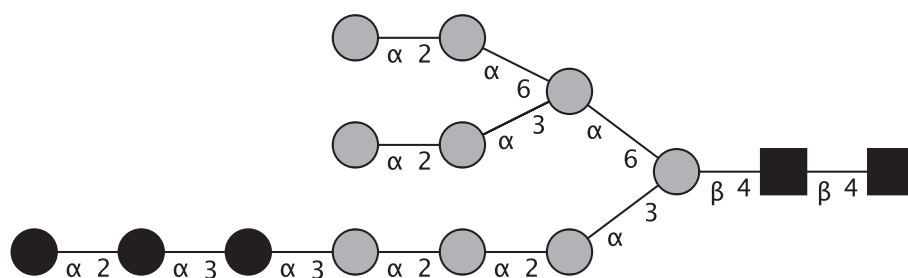
Property	DNA/RNA/protein	Carbohydrates
Topology	Linear	Linear, branched, cyclic
Linkage	Phosphodiester/peptide bond	Glycosidic linkage. Potentially multiple per residue, so position information is crucial for carbohydrate sequences
Alphabet	4/4/20 distinct building blocks	Alphabet of >500 different building blocks; for certain structural and taxonomic subsets considerably smaller
Stereochemistry	L-Configuration for amino acids	Stereoconfigurations can vary; however, dominant configuration typically exists
Size	Variable	Variable, oligomers 1–40 residues, polymers built from repetitive small elements up to 10^3 – 10^4 residues
Structure/conformers	Secondary, tertiary, and quaternary structures	Definite secondary structures are rare, molecules are flexible, ensemble of structures; exception: structural polysaccharides

macromolecules can be accomplished in sequences of characters by abbreviating the building blocks as letters. Although the concept of defining building blocks (monosaccharides) which receive unique names is also widely applied in glycobiology, the encoding of carbohydrate structures differs in several fundamental aspects from linear DNA/RNA and protein sequences (Table 2.1). Carbohydrates, with the exception of polysaccharides, are smaller than typical proteins, but a greater variety of building blocks is observed. The fact that carbohydrates can form branched structures renders many of the classical bioinformatics approaches inapplicable to carbohydrates and has led to the establishment of a new field, originally called glyco-bioinformatics, now termed *glycoinformatics*.

One of the challenges in glycobiology is the accurate and thorough description and/or representation of the complex structures of carbohydrates, and the sequences of polysaccharides. Both mono- and oligosaccharides can be described using a number of textual (e.g. IUPAC) or symbolic formats (e.g. Figure 2.1). From a glycoinformatics point of view, different encoding schemes for monosaccharides are in use, many of which have some limitations (see Chapter 3). Recently, a novel standard sequence format for carbohydrates has been developed (GlycoCT [8], see Chapter 3) which is used in GlycomeDB [9] (www.glycome-db.org) to harmonize carbohydrate sequences in globally distributed carbohydrate structure databases. As a result, GlycomeDB can collate information from the seven major open-access databases. The following sections provide an overview of the structural diversity of carbohydrate sequences¹ derived from this data source.

This chapter presents an overview of carbohydrate nomenclature and structure definitions, and provides an insight into the theoretical and actual diversity of carbohydrate structures. The first section deals with the conventions for naming monosaccharides, and provides an overview of monosaccharide diversity. The second section deals with the constitution of carbohydrate sequences, and is followed by a section describing the major classifications of carbohydrates.

¹ In the context of storing or analyzing carbohydrate structures in databases, the term “carbohydrate sequence” is frequently used instead of the term “carbohydrate structure”.



Symbol	Abbrev.	Monosaccharide	Symbol	Abbrev.	Monosaccharide
○	Gal	Galactose	●	Man	Mannose
□	GalNAc	<i>N</i> -Acetylgalactosamine	■	ManNAc	<i>N</i> -Acetylmannosamine
◻	GalN	Galactosamine	◻	ManN	Mannosamine
◊	GalA	Galacturonic acid	◊	ManA	Mannuronic acid
●	Glc	Glucose	◆	Kdn	2-Keto-3-deoxynononic acid
■	GlcNAc	<i>N</i> -Acetylglucosamine	◆	Neu5Ac	<i>N</i> -Acetylneuraminic acid
◻	GlcN	Glucosamine	☆	Xyl	Xylose
◊	GlcA	Glucuronic acid	◊	IdoA	Iduronic acid
▲	Fuc	Fucose	◇	Neu5Gc	<i>N</i> -Glycolylneuraminic acid

Figure 2.1 Symbolic representation of a carbohydrate structure. The circles and squares represent the monosaccharides. The glycosidic linkages are indicated by the connecting lines and the linkage type annotation ($\alpha 2$, $\alpha 3$, $\beta 4$, ...). Throughout this chapter, the definitions of the Consortium for Functional Glycomics (CFG) will be used. The figure shows the symbols for the monosaccharides in black-and-white, next to the abbreviation and the full monosaccharide name. Fucose and iduronic acid are in the L-configuration, all other basic monosaccharides are in D-configuration. Color versions of the symbols can be found on the CFG homepage (www.functionalglycomics.org).

2.1 Monosaccharide Nomenclature and Diversity

This section will start with a brief introduction to the concepts and nomenclature of monosaccharides, the basic building blocks of carbohydrate sequences. We will then examine from a more theoretical perspective the maximum diversity on the level of monosaccharides. The section is concluded with results from a carbohydrate sequence database analysis identifying and contrasting the prevalent monosaccharides in mammals and bacteria.

2.1.1 Systematic Nomenclature for Monosaccharides

Historically, the term carbohydrate was coined for chemical structures with a net formula $C_n(H_2O)_n$, where formally each carbon atom is associated with a water molecule. IUPAC

recommendations [10] provide standard descriptions to name both simple (i.e. monosaccharide) and complex carbohydrates systematically. They use a more structurally oriented definition of a monosaccharide: a chain of three or more carbon atoms, containing at least one carbonyl and one hydroxyl group, is a carbohydrate. The number of carbon atoms in the main chain defines classes of related carbohydrates. These classes receive a name (e.g. a six-carbon chain = hexose) and a three-letter abbreviation (e.g. Hex; see Table 2.3, Section 2.1.1.2: columns class and abbreviated class names). These class names are frequently used directly if structure elucidation has left ambiguities regarding the exact monosaccharide identity.

2.1.1.1 Definition of Terms and Properties of Monosaccharides. Monosaccharides are chiral compounds, as they contain asymmetric tetrahedral carbon atoms [3, 5, 6], which gives rise to stereoisomers. The pioneering work of Fischer systematically explored this chirality and his definitions are still in use today. According to Fischer, each monosaccharide has a basic configuration, either D or L. The relative spatial orientation of the hydroxyl groups defines the “stem” type (stereochemical identifier) of the monosaccharide (Section 2.1.1.2).

Monosaccharides tend to form an intramolecular ring closure, resulting in hemiacetal forms. The common five- and six-membered ring forms are called furanose and pyranose, respectively. Generated by this cyclization are a new asymmetric C-atom, called the anomeric C-atom, and a new hydroxyl function. This hydroxyl function (or other substituents at this position) can be oriented in two different ways, which are designated α or β (often converted to “a” and “b” in digital representations of carbohydrate sequences) (Figure 2.2a).

Connection between two monosaccharides (formation of a glycosidic bond) occurs by condensation between the activated anomeric C-atom of the hemiacetal form and a hydroxyl function of another monosaccharide. As there is more than one acceptor hydroxyl function in normal monosaccharides, multiple condensations on to the one monosaccharide are possible, leading to a variety of different (branched) oligo- and polysaccharides. A non-carbohydrate connected to a monosaccharide via a glycosidic bond is termed an aglycone (Figure 2.2b). The aglycones may be proteins, lipids or small molecules. Abstractly speaking, the glycosidic bond can be seen as an analog to the commonly known peptide bond of proteins or phosphodiester bonds in nucleic acids.

The basic monosaccharides may lose chirality at certain C-atoms through deoxygenation (OH group replaced by H), a phenomenon frequently observed in natural monosaccharides. Other typical modifications are introduction of acidic functions, double bond formation, shifts of the carbonyl function or the attachment of small chemical groups (substituents). Often the resulting monosaccharides receive trivial names (Figure 2.2c, d).

Summarizing the general properties of monosaccharides, five different attributes are sufficient to identify systematically the basic monosaccharides that are encountered in nature (Table 2.2).

2.1.1.2 Carbohydrate Stem Types. The IUPAC recommendations contain definitions of stem types (or stem names) for monosaccharides with three to six carbon atoms (equivalent to one to four stereogenic centers) for all possible stereoisomers [10]. IUPAC suggests three-letter codes for stem types starting at the pentose level. Smaller stem types have commonly accepted abbreviations (Table 2.3). Each of these stem types belongs to either the D- or the L-series.

Table 2.2 Five attributes are sufficient to define the properties of all basic unsubstituted monosaccharides.

Attribute	Explanation	Example values
Configuration	Fischer systematic series	D, L
Stem type	Stereochemistry of hydroxyl groups	Gal, Glc, Man, Ara, Xyl, . . . (see Section 2.1.1.2)
Ring size	Cyclic hemiacetal	Pyranose (<i>p</i>), furanose (<i>f</i>) (or open chain)
Anomer	Orientation of anomeric hydroxyl group	α , β
Modifiers	Altering the stereochemistry	Deoxygenation

A carbohydrate containing more than four chiral centers is named by composite “stem names” (derived from the IUPAC name in Table 2.3). “Stem names” are assigned in order of the chiral centers in groups of four, beginning with the group proximal to C1 (Figure 2.3). In the composite stem name, the portion relating to the group of carbon atoms furthest from C1 (which may contain less than four atoms) is given first.

2.1.1.3 Frequent Substituents. Frequently, the hydroxyl group of a monosaccharide is functionalized (e.g. sulfated) or replaced by another substituent (Figure 2.4). One of the

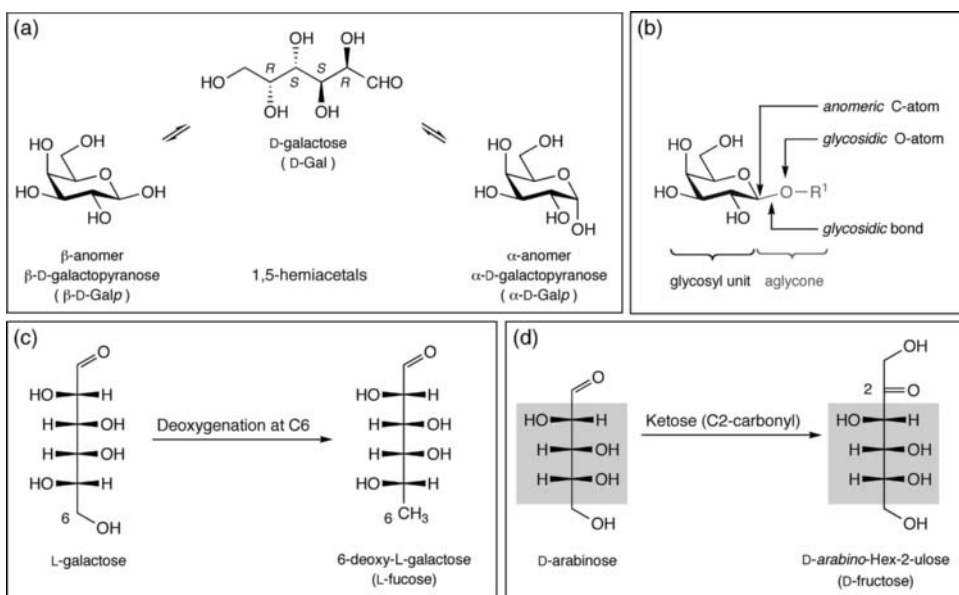


Figure 2.2 (a) The pyranose ring forms of D-galactose. The two anomeric configurations (hemiacetals) and the open form are depicted. Isolated monosaccharides typically show an equilibrium mixture of different cyclic forms in aqueous solution (furanose rings may also be formed through 1,4-cyclization). (b) A monosaccharide bound to a non-monosaccharide molecule (aglycone). The anomeric configuration is fixed to either α or β by this connection. (c) The deoxygenation at C6 (a non-chiral C-atom) of L-galactose yields L-fucose, a frequently occurring natural monosaccharide. (d) Ketoses, monosaccharides with non-terminal carbonyl functions, are systematically named with the suffix -ulose. The chiral C-atoms (gray boxes) define the stereochemical identifier, here the stem type arabinose. The resulting ketose is fructose, a component of the commonly known table sugar.

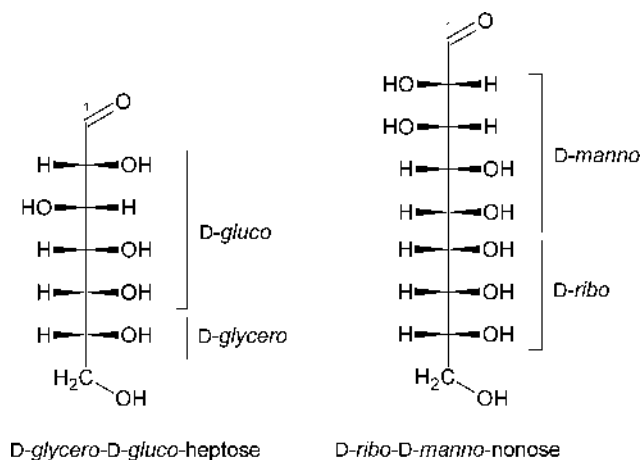
Table 2.3 Stem types for monosaccharides with up to four stereogenic centers. The D- or L- series is not indicated in this table.

IUPAC 3-letter code	IUPAC name	Class	Class name (abbreviated)	Number of C-atoms
Gro	Glyceraldehyde	Triose	TRI	3
Ery	Erythrose	Tetrose	TET	4
Tre	Threose	Tetrose	TET	4
Ara	Arabinose	Pentose	PEN	5
Rib	Ribose	Pentose	PEN	5
Lyx	Lyxose	Pentose	PEN	5
Xyl	Xylose	Pentose	PEN	5
All	Allose	Hexose	HEX	6
Alt	Altrose	Hexose	HEX	6
Gal	Galactose	Hexose	HEX	6
Glc	Glucose	Hexose	HEX	6
Gul	Gulose	Hexose	HEX	6
Ido	Idose	Hexose	HEX	6
Man	Mannose	Hexose	HEX	6
Tal	Talose	Hexose	HEX	6

most commonly found substitutions is the acetylamino group NHC(O)CH_3 , abbreviated NAc, typically at C2 in hexoses. By convention, the C2 atom is the assumed position for this substitution in a hexose when a positional description is not given (e.g. D-GlcpNAc and *not* D-Glcp-2NAc). The number of substituents found in carbohydrate sequence databases is large, with individual glycobiological communities with different taxonomic foci contributing different sets of substituents.

2.1.2 Naturally Occurring Monosaccharides Based on Database Analysis

The total number of all potential monosaccharides is very large. For hexoses, eight stereochemical stem types exist, each in two configurations (D/L) and two anomeric configurations

**Figure 2.3** IUPAC defines a consistent naming scheme for carbohydrates with more than four stereogenic centers, resulting in composite names.

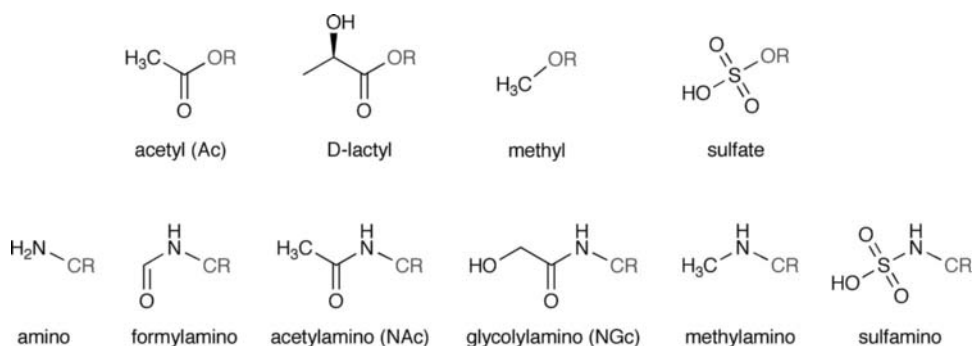


Figure 2.4 Nine frequent substituents found in carbohydrate structure databases. The attachment site to the carbohydrate stem is indicated with OR (i.e. functionalization of the oxygen atom) or CR (replacement of the hydroxyl group).

(α/β), which results in a total of 32 possible hexoses. With each new stereogenic C-atom, this number is increased twofold, and a total of 506 basic stereochemical entities for monosaccharides with less than 10 C-atoms exist. An additional level of complexity is introduced by substitutions and modifications, leading to an immense number of theoretically possible monosaccharides. However, the large combinatorial potential of monosaccharides is not populated by natural organisms. Each living organism uses a distinctly smaller set determined ultimately by its genetic repertoire.

2.1.2.1 Mammalian Monosaccharides. Considering the main saccharide classes – *N*-glycans, *O*-glycans, and components of the extracellular matrix – as deposited in carbohydrate sequence databases, the taxonomic class of *Mammalia* contains 10 monosaccharides based on seven stem types (Figure 2.5) [3, 11, 12]. The 25 most prevalent mammalian

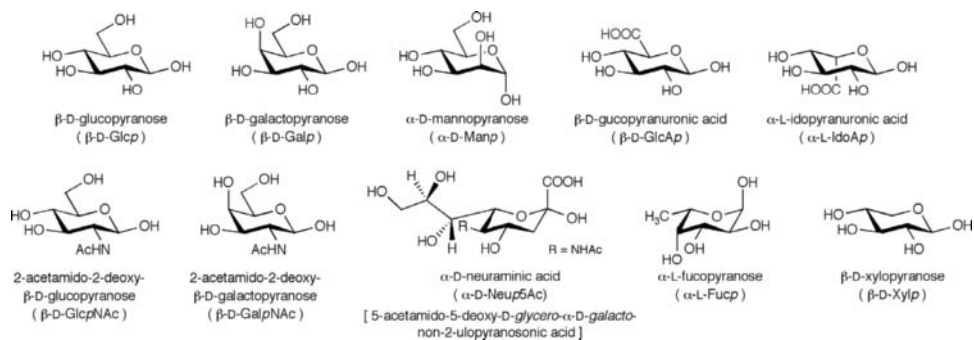


Figure 2.5 The mammalian sequences as reported in carbohydrate databases can be constructed with 10 monosaccharides. The monosaccharides shown here are depicted in their dominant anomeric configuration, although the opposite configuration also might exist. Neuraminic acid is a nonose and exists with multiple substitution patterns. With these 10 monosaccharides, the majority of the digitally available carbohydrate sequences in mammals can be constructed. Taking into account the common substitutions of *O*- and *N*-sulfate and *N*-glycolyl, an almost complete coverage can be achieved (Table 2.4).

Table 2.4 The 25 most prevalent mammalian monosaccharides. The data are based on 4936 sequences with 43 780 monosaccharides in GlycomeDB [9]. The total number of different monosaccharides is 170. This high number can be attributed to incompletely defined monosaccharides and obvious database errors.

No.	IUPAC name	Stem type	Proportion (%)
1	β -D-GlcpNAc	D-Glc	26.36
2	β -D-Galp	D-Gal	21.32
3	α -D-Manp	D-Man	14.42
4	α -Neup5Ac	SIA	7.68
5	α -L-Fucp	L-Gal	6.92
6	β -D-Manp	D-Man	6.03
7	β -D-GalpNAc	D-Gal	1.22
8	β -D-Glcp	D-Glc	1.00
9	α -D-Galp	D-Gal	0.87
10	α -D-GalpNAc	D-Gal	0.71
11	α -Neup5Gc	SIA	0.50
12	α -D-Glcp	D-Glc	0.48
13	β -D-GlcpNAc6Sulfate	D-Glc	0.44
14	β -D-GlcpA	D-Glc	0.31
15	α -D-GlcpNAc	D-Glc	0.27
16	β -D-Galp6Sulfate	D-Gal	0.22
17	β -D-GalpNAc4Sulfate	D-Gal	0.21
18	α -D-GlcpN-Sulfate6Sulfate	D-Glc	0.16
19	α -L-IdopA-Sulfate	L-Ido	0.16
20	α -D-GlcpN-Sulfate	D-Glc	0.10
21	β -D-Galp3Sulfate	D-Gal	0.09
22	β -D-Xylp	D-Xyl	0.08
23	α -L-IdoAp	L-Ido	0.08
24	α -D-GlcpNAc6Sulfate	D-Glc	0.05
25	α -Neup4Ac5Ac	SIA	0.04

monosaccharides., based on an analysis of carbohydrate sequences in GlycomeDB [9], are shown in Table 2.4.

2.1.2.2 Bacterial Monosaccharides. Different organisms have different capabilities of synthesizing and using monosaccharides in oligo- and polysaccharides. The prokaryotic repertoire of monosaccharides is different from that of mammalian systems. The bacterial saccharide sequences have a greater diversity of monosaccharides, with certain monosaccharides being specific to certain groups of bacteria. An overall comparison of the monosaccharide distribution in digitally available bacterial sequences with the mammalian situation reveals a higher content of monosaccharides with more than six C-atoms, a higher degree of phosphorylation and unusual deoxygenations (Table 2.5) [13].

2.2 The Oligosaccharide Assembly Level

In this section, we will examine the constitution of carbohydrate structures. Data derived from available glycan sequences in GlycomeDB will provide insights into the complexity of natural carbohydrate sequences.

Table 2.5 The 25 most prevalent monosaccharides in bacteria. The data are based on 4720 structures with 24 953 monosaccharides in GlycomeDB [9]. The total number of different monosaccharides is 1025.

No.	IUPAC name	Stem type	Proportion (%)
1	β -D-Glcp	D-Glc	8.94
2	β -D-Galp	D-Gal	7.43
3	α -L-Gro-D-Man-Hepp	L-Gro-D-Man-Hep	7.09
4	α -D-Glcp	D-Glc	6.80
5	α -L-Rhap	L-Man	6.60
6	α -D-Galp	D-Gal	4.68
7	β -D-GlcpNAc	D-Glc	4.54
8	α -Kdop	KDO	4.36
9	α -D-Manp	D-Man	3.11
10	α -D-GlcpNAc	D-Glc	2.04
11	α -D-GlcpN	D-Glc	1.53
12	β -D-GlcpA	D-Glc	1.50
13	α -D-Gro-D-Man-Hepp	D-Gro-D-Man-Hep	1.41
14	β -D-GalpNAc	D-Gal	1.37
15	α -Neup5Ac	SIA	1.16
16	α -D-GalpA	D-Gal	1.15
17	α -D-GalpNAc	D-Gal	1.14
18	α -L-Gro-D-Man-Hepp6phosphoethanolamine	L-Gro-D-Man-Hep	0.93
19	β -D-Manp	D-Man	0.92
20	α -L-Fucp	L-Gal	0.92
21	α -D-Rhap	D-Man	0.85
22	β -D-Glcf	D-Glc	0.72
23	β -D-GlcpN4phosphate	D-Glc	0.67
24	β -D-GlcpN	D-Glc	0.67
25	β -L-Rhap	L-Man	0.66

2.2.1 Constitution of Carbohydrate Structures and Sequences

Almost every carbohydrate sequence has a direction caused by the asymmetric character of the glycosidic linkage. Following a chain along the direction $C_{\text{anomeric}} \rightarrow O_{\text{glycosidic}}$ will lead to the monosaccharide whose anomeric center (often connected to an aglycone) is termed the “reducing end” of the sequence (Figure 2.6a) (this relates to the ability of the free anomeric position, which is an aldehyde or ketone in the open-chain form, to reduce Cu^{2+} to Cu^+). Biological chain elongation, however, normally proceeds from the reducing end to the non-reducing end of a chain. A typical hexose has four hydroxyl groups (e.g. at carbons 2, 3, 4, and 6) that can serve as monosaccharide acceptors during the biological elongation process of the nascent carbohydrate chain. Consequently, branch points can be introduced into the chain if a monosaccharide is involved in three or more glycosidic linkages, a feature that is normally not present in other well-known biopolymers. The branching leads to an enormous increase in structural isomers even for small oligomers (Table 2.6).

A number of descriptors can be used to define a carbohydrate structure (Figure 2.6a). The most basic description is that of “composition” – a simple list of the monosaccharides in the structure. The availability of several attachment points on a monosaccharide requires that the linkage description includes position information. Each glycosidic linkage in a natural carbohydrate sequence is the product of a distinct enzyme reaction, and the linked

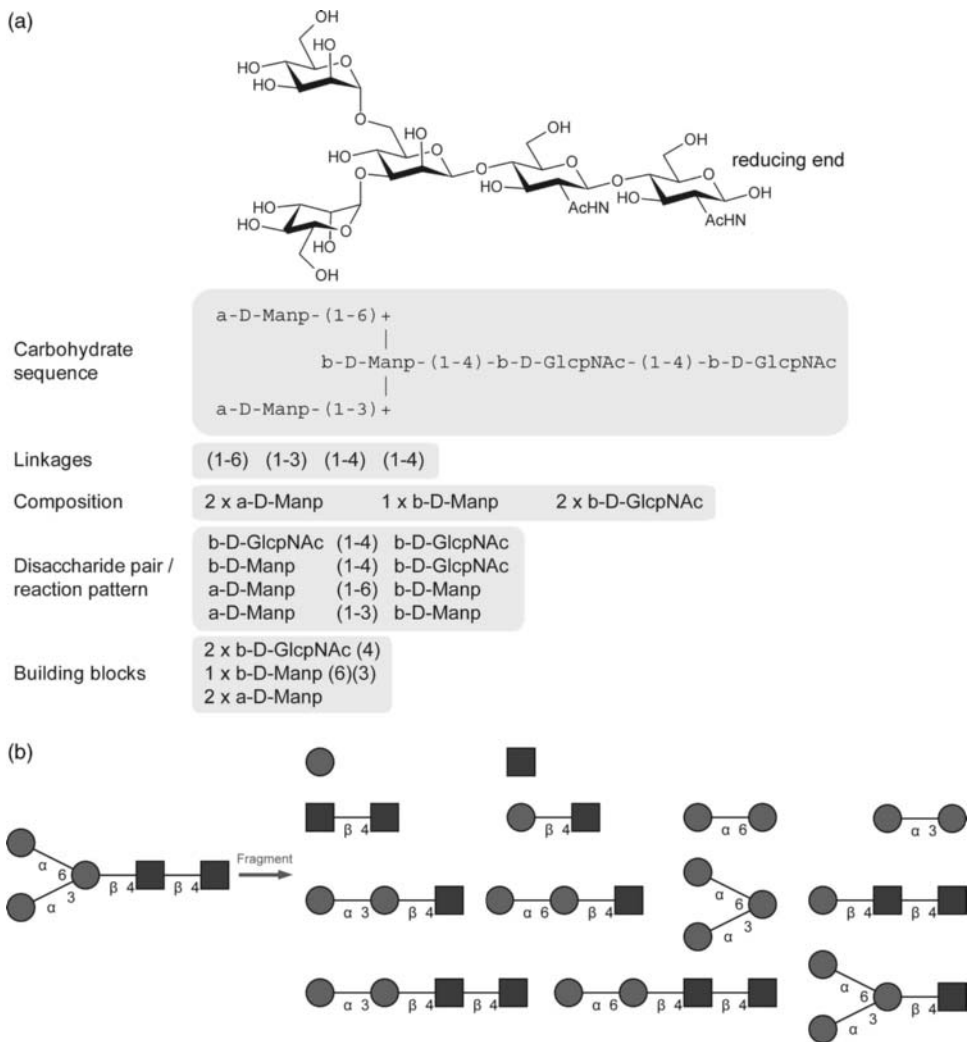


Figure 2.6 (a) The chemical structure and IUPAC 2D sequence of a frequently occurring motif, the *N*-glycan core. Different views of this sequence are used for different purposes (see text for explanation). (b) Pictorial representation of the structure shown in (a) and its fragmentation into all possible connected pieces.

entities are frequently termed reaction patterns or disaccharide pairs [14] (Figure 2.6a). As carbohydrate sequences are secondary gene products, the distribution of reaction patterns provides a direct link to the enzymes needed to construct a given glycan. In the area of chemical synthesis of complex carbohydrates, the building blocks required to construct a carbohydrate are of particular interest. The requirement for particular attachment points drives the use of protecting group chemistry, the transient masking of hydroxyl groups during a multi-step synthesis. Deconstructing the glycan into all possible fragments is a frequently applied procedure (Figure 2.6b). It can be used for diverse applications such as motif and pattern recognition, substructure searches, and mass spectrometric analysis.

Table 2.6 Possible isomers for oligonucleotides, peptides, and oligosaccharides for different oligomer sizes [39]. The numbers given for carbohydrates are based on the 10 mammalian monosaccharides with two different anomeric configurations.

Oligomer size	No. of different oligomers		
	Nucleotides	Peptides	Carbohydrates
1	4	20	20
2	16	400	1360
3	64	8000	126 080
4	256	160 000	13 495 040
5	1024	3 200 000	1 569 745 920
6	4096	64 000 000	192 780 943 360

2.2.2 Not All Sequences Are Completely Defined

So far, it has been assumed that all structural features of the monosaccharides and the topology of a given sequence are completely known and defined. However, a considerable number of structures have been reported in the literature with structurally incomplete descriptions. This “fuzziness” can be attributed either to the biosynthesis or to limitations of the applied experimental method used for structure elucidation.

The biosynthesis of complex carbohydrates is dependent on the coordinated action of modifying enzymes. The catalyzed reactions can be incomplete, leading to an intrinsic microheterogeneity of the complex carbohydrate. Prominent examples of this phenomenon are the sulfated polysaccharides; they are typical representatives of non-stoichiometrically substituted compounds (see Section 2.3.3.1 on proteoglycans). Other examples are found in bacteria and plants, as they synthesize macromolecular carbohydrates with very high masses. These polysaccharides are typically built from smaller precursor units which can be identified as repeating units. In most cases these polysaccharides are an ensemble of polymers with varying chain length.

As there is no direct sequencing method for carbohydrates, the results of different experimental methods are combined to elucidate the structure. Frequently, mass spectrometry is used to gain structural information. This technology cannot distinguish between stereoisomers such as galactose or glucose, which differ only in the stereoconfiguration at C4. Often, an empirical formula (composition) of a glycan is the result of a structure elucidation attempt. Table 2.7 presents an overview of properties which may be missing or undetermined in a carbohydrate sequence.

2.2.3 Statistics of Carbohydrate Sequences Based on Database Analysis

The following data are based on statistical analysis of the GlycomeDB [9] and provide insights into different statistical parameters of typical carbohydrate sequences. First aggregate parameters such as size, length, and branching are examined, and then the dominant linkages of the carbohydrate sequences are shown.

The size of a sequence is defined as the number of monosaccharides contained, and its maximum chain length is the longest path from the reducing end to a distal residue. A branching point is defined as a residue with more than one child residue, resulting in a furcation of the sequence. The size distribution as deposited in databases is broad,

Table 2.7 Incomplete descriptions for complex carbohydrates found in the literature.

Missing property	Example
Identity of monosaccharide	Hex, HexNAc, ?-Galp, D-Glcp or D-Galp
Ring size	α -D-Galp?
Anomeric configuration	?-D-Galp
Linkage	α -D-Galp-(1-?)- β -D-GlcpNAc
Topology	Composition: 5 \times Hex, 2 \times HexNAc
Substitution quantifier	15% of IdoA is sulfated at position C2
Terminal residue location	Two of three <i>N</i> -glycan antennas carry a terminal sialic acid residue
Repeat unit size	Size of polymer varies from 150 to 300 repeating units

ranging from carbohydrate sequences with a single monosaccharide to the largest sequence with more than 100 monosaccharides (Figure 2.7a). The average size of the carbohydrate sequences in GlycomeDB is seven monosaccharides. The maximum chain length shows a broad distribution, with the highest frequency seen for 2–10 monosaccharides as the longest path in the sequences (Figure 2.7b). The longest chain of a sequence, excluding repeating units, is 16 monosaccharides. The branching point calculation reveals the topological complexity of the sequences; 38% of the oligosaccharide structures in the GlycomeDB are linear sequences, with no branching point at all. Only a few carbohydrates contain five or more branches (Figure 2.7c).

2.2.3.1 Topology: Linkages. All linkages between monosaccharides must contain a positional identifier for the attachment sites. The most dominant linkages are found to be 1–4, 1–3, 1–6, and 1–2 connections in this order. The anomeric configuration can be seen as part of the linkage, as the configuration of the linkage is generated by strictly stereospecific glycosyltransferases during glycosidic bond formation. The anomeric configuration of the glycosidic bond is constrained by geometric factors, and their distribution pattern varies depending on the taxonomic class (Table 2.8).

2.2.3.2 Topology: Disaccharide Pairs. Carbohydrate sequences are secondary gene products constructed by an ensemble of enzymes, namely the coordinated action of elongating glycosyltransferases, truncating glycosidases, and other modifying enzymes. The enzymes act typically in a spatially and temporally controlled fashion in specialized compartments of the cell [11, 12]. For each elongation step, at least one enzyme is needed, but frequently isoenzymes exist, so the one-enzyme–one-linkage paradigm is not strictly followed for carbohydrate sequences. Nonetheless, the disaccharide pair distribution for a given organism does provide a direct link to a minimum set of expressed glycosyltransferases and can be correlated with expression data [14, 15]. A summary of existing bacterial and mammalian disaccharide pairs can be found in [13].

2.3 Major Empirical Structural Classifications

This section describes commonly used classifications for oligosaccharides. The major classifications rely on the substructures connected to the reducing end of the carbohydrates, thus defining groups of biosynthetically related sequences by conjugation type. Further breakdown of these classes is often achieved by identifying motifs or topological patterns

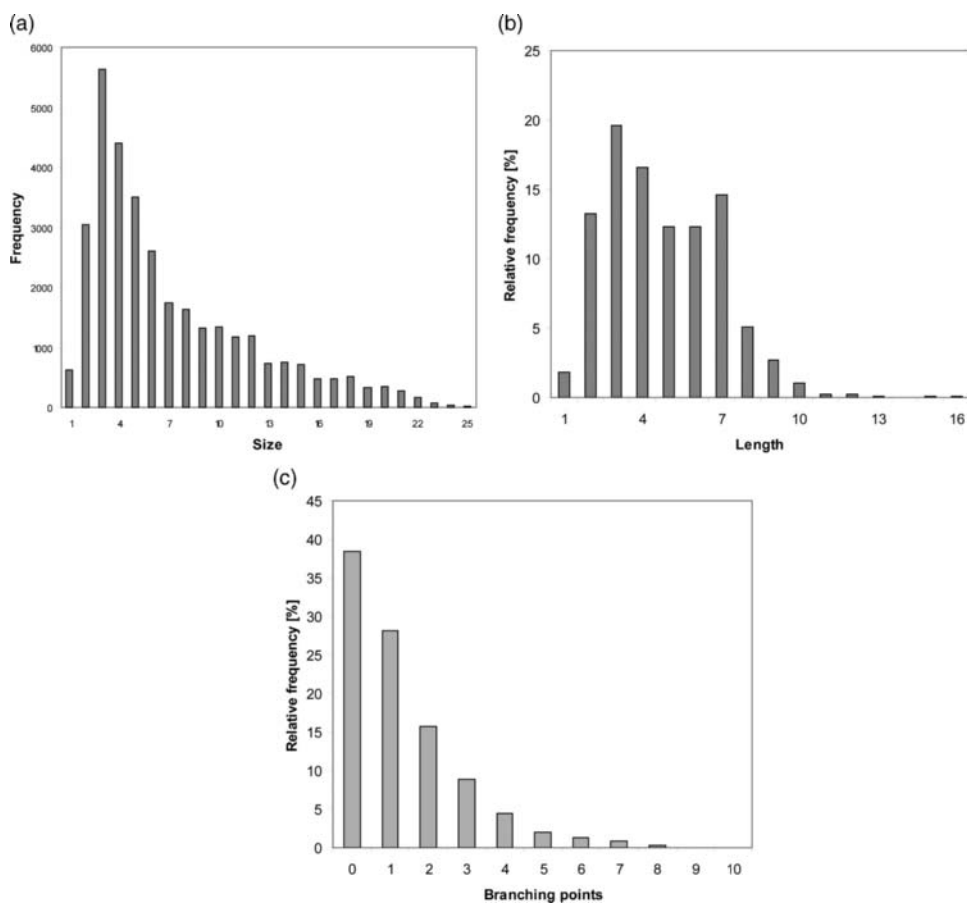


Figure 2.7 Statistical analysis of carbohydrate sequences deposited in GlycomeDB. (a) Frequency distribution of sequence size (number of monosaccharides, statistics up to 25 residues are shown). (b) Frequency distribution of the maximum chain length of the sequences. (c) Frequency of branching points of the sequences. A value of zero stands for linear sequences.

within these groups. This section will introduce important classes of saccharides as defined by glycobiologists and their specific structural features.

2.3.1 Glycoproteins

2.3.1.1 N-Glycans. The most prominent vertebrate post-translational modification is the covalent attachment of an oligosaccharide to specific sites in proteins which is due to processing which begins in the endoplasmic reticulum (ER) [16, 17]. About 50% of the proteins processed in the ER and the Golgi apparatus are thought to be *N*-glycosylated [18]. Protein sequences containing Asn–Xaa–Thr/Ser sequons [19] (where Xaa denotes any amino acid except Pro) are frequently, but not always, glycosylated at the asparagine residue by a cotranslational mechanism controlled by a multi-enzyme complex, the oligosaccharyl transferase (OST) [20]. The exact initiation mechanism of this multi-enzyme complex

Table 2.8 Dominant linkages in GlycomeDB [9] with anomeric configuration. Three taxonomic subsets have been analyzed regarding their linkage distribution (Mammalia with 4936 structures containing 38 957 linkages, bacteria with 4720 structures and 21 643 linkages, and 713 plant structures with 3304 linkages). Unknown anomeric configurations and other ambiguities have been excluded from this table, resulting in the low percentage portion not shown.

Linkage	Mammalia (%)	Bacteria (%)	Plant (%)
$\beta(1-4)$	37.43	16.79	33.44
$\alpha(1-3)$	10.76	21.16	14.32
$\alpha(1-6)$	10.22	3.54	9.38
$\beta(1-2)$	10.58	3.99	13.83
$\beta(1-3)$	9.62	11.22	4.15
$\alpha(1-2)$	3.70	12.06	
$\alpha(1-4)$	1.75	8.82	1.36
$\beta(1-6)$	4.07	4.14	0.15
$\alpha(2-3)$	4.74	1.43	13.80
$\alpha(2-6)$	3.72	1.64	5.87
$\alpha(1-5)$		5.08	
$\alpha(2-8)$	0.22	0.24	
$\alpha(2-4)$	0.06	2.20	0.42
$\alpha(1-7)$		1.88	
$\beta(1-7)$		1.01	
$\beta(1-5)$		0.28	
$\beta(2-6)$		0.19	1.48
$\beta(1-8)$		0.32	
$\beta(2-3)$		0.10	
Sum	96.87	96.09	98.20

is as yet unknown. The OST catalyzes transfer of an oligosaccharide of the composition $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$ (Figure 2.1) *en bloc* from an activated dolichol pyrophosphate precursor to the nascent polypeptide chain. This intermediate oligosaccharide is subsequently trimmed in a process which is correlated with correct folding of the polypeptide chain [16, 17] (see also Chapter 8).

All *N*-glycans share as a minimum structural feature the so-called *N*-glycan core ($\text{Man}_3\text{GlcNAc}_2$, Figure 2.8a). The *N*-glycans are generally classified further into three topological classes, which share common motifs. The high-mannose type contains exclusively mannose residues attached to the *N*-glycan core (Figure 2.8b). Complex types do not contain additional mannose residues, but rather GlcNAc directly attached to the two branches of the *N*-glycan core. Additionally, sialic acids, galactose, and fucose residues can be attached to the complex *N*-glycans (Figure 2.8c, e). Hybrid types have both a high-mannose and a complex branch attached to the core structure (Figure 2.8d). Other empirical distinctive features of *N*-glycans in use are the existence of core-fucosylation, the bisecting criterion, and the number of antennas in a given glycan (Figure 2.8e, f).

2.3.1.2 Ser/Thr *O*-Glycans with Initiating α -D-GalNAc. The general definition for *O*-glycosylation implies the addition of a mono- or oligosaccharide to amino acid residues with functional hydroxyl groups of a protein. By historical consensus, the classical understanding of *O*-glycosylation is confined to the addition of a GalNAc in the α -configuration to serine or threonine residues by the action of a member of the family of GalNAc

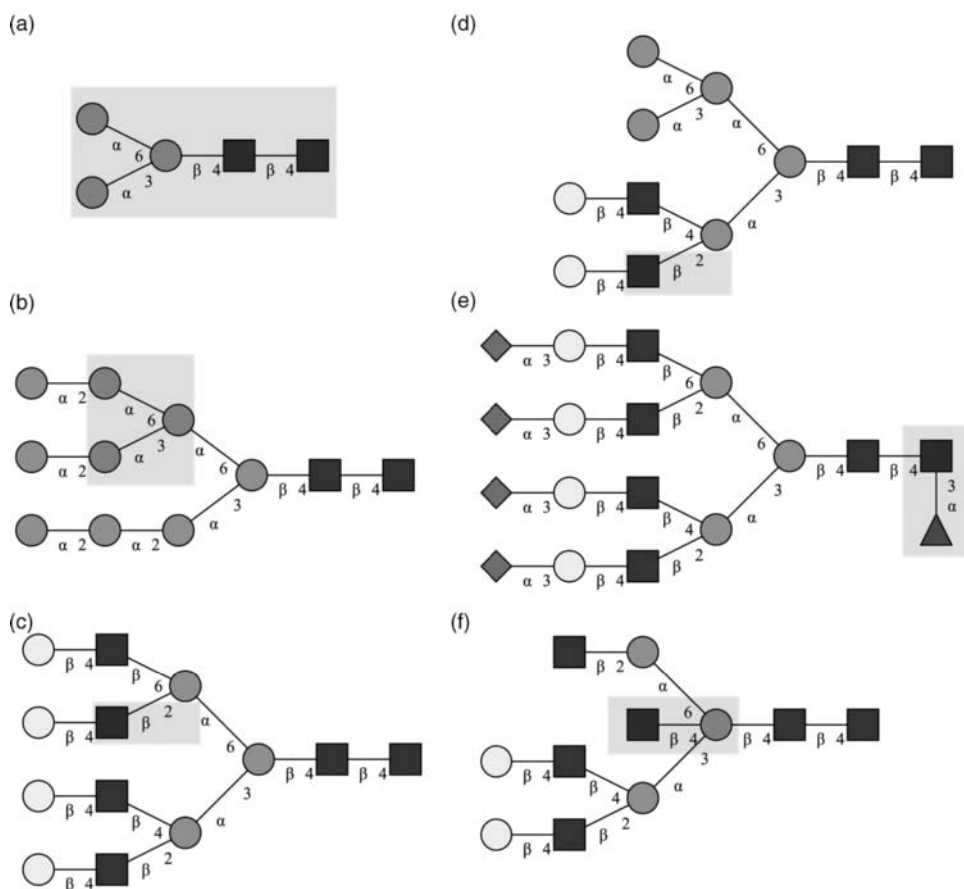


Figure 2.8 Common classes of *N*-glycans. The gray boxes indicate the distinctive features that lead to the respective classifications. (a) The *N*-glycan core structure is the distinctive feature of *N*-glycans. (b) High-mannose types are remains, and further extensions, of the initial $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$ structure (Figure 2.1) transferred *en bloc* to the Asn of the protein. (c) The *N*-glycan core is extended by attachment of GlcNAc and Gal residues in the complex class. (d) The hybrid class contains features of both high-mannose and complex types. (e) A tetra-antennary *N*-glycan with an $\alpha(1-3)$ L-fucose connected to the core region (gray box). (f) A tri-antennary bisected *N*-glycan. Bisecting is defined as the attachment of a $\beta(1-4)$ D-GlcNAc to the central core mannose.

transferases in the secretory pathway during protein maturation in the ER and Golgi apparatus. Adopting this view, the *O*-glycans defined can be subdivided into eight subclasses by identifying common structural motifs at the reducing end (Figure 2.9).

2.3.1.3 Other Types. Historically, protein glycosylation was believed to be restricted to *N*- and *O*-glycosylation carried out in the ER or Golgi apparatus. Recent findings indicate ubiquitous cellular protein glycosylation events [21] and cellular localizations of proteoglycans [22]. The most prominent non-standard glycosylation reaction is the nucleocytoplasmic *O*-GlcNAc modification, which presumably has regulative function [23].

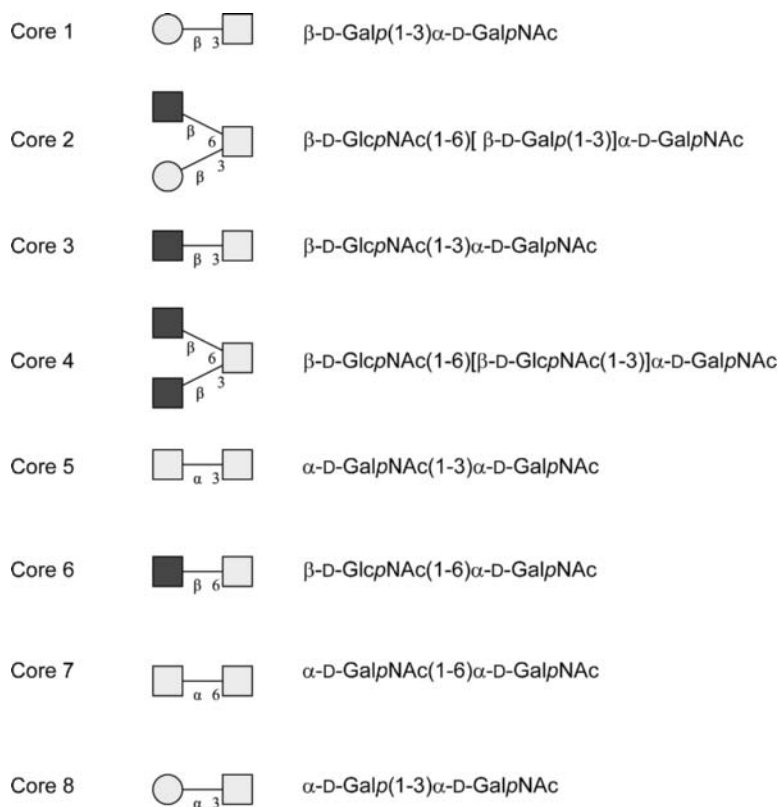


Figure 2.9 The eight core structures for the D-GalNAc *O*-glycosylation are identified by common structural elements at the reducing end.

Other less well characterized glycosylation reactions on proteins have been summarized in the literature [21, 24].

2.3.2 Glycolipids

2.3.2.1 Glycosphingolipids. The substance class of glycosphingolipids is defined by the conjugation of a mono- or oligosaccharide to a ceramide (Figure 2.10) [25]. Glycosphingolipids occur in the outer leaflet of cell membranes and show tissue- and organism-specific variations on both the lipid and the carbohydrate portions. The ceramide unit serves as a membrane anchor and is suggested to play a crucial role in the formation of lipid rafts [26]. Depending on the structural motif connected to the ceramide, basic classes have been defined (Figure 2.11). Also, physicochemical parameters (acidic, neutral) are applied to subdivide the glycosphingolipids. Glycosphingolipids are sometimes simply called glycolipids, as they were the first carbohydrate–lipid conjugates found in Nature. Despite an existing IUPAC recommendation [27], a specific nomenclature is still in widespread use. Originally defined by Svennerholm [28], it results in compact names (e.g. GM1).

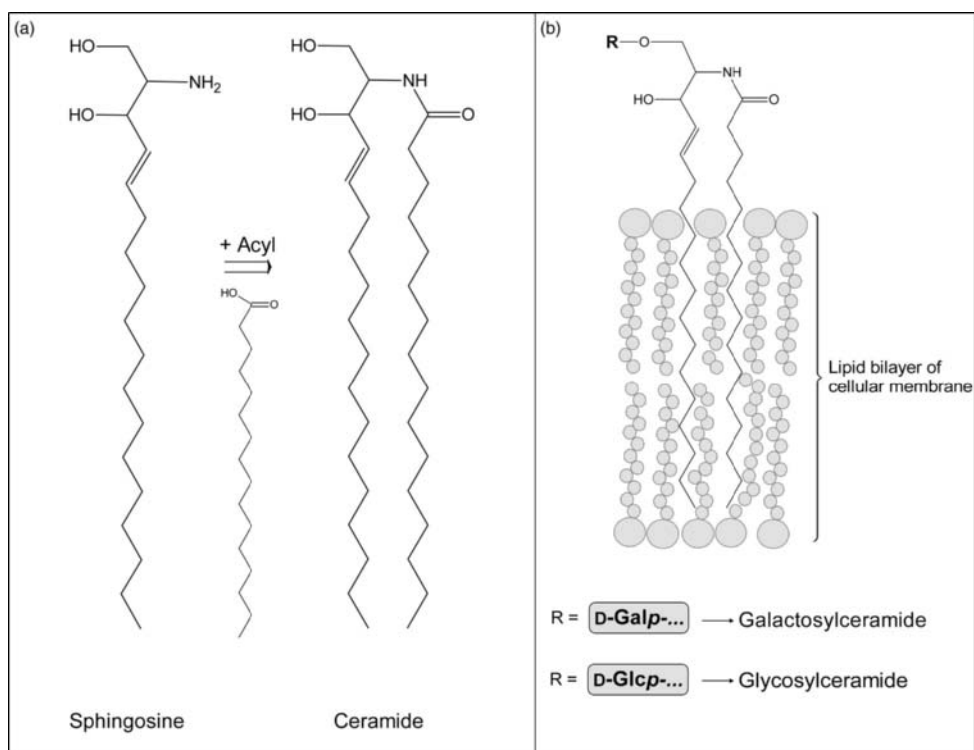


Figure 2.10 (a) Sphingosine is the biosynthetic precursor of ceramide, which serves as the lipid anchor of glycosphingolipids. (b) Schematic drawing of a glycosphingolipid with its membrane insertion. The attachment of the saccharide residues is indicated with R.

2.3.2.2 Glycosylphosphatidylinositols (GPIs). The glycosylphosphatidylinositol (GPI) anchor is capable of anchoring proteins in the cellular membrane [29, 30]. The biosynthesis of GPI-anchored proteins involves a transfer reaction within the endoplasmic reticulum to proteins carrying a carboxy-terminal signal sequence. Structurally, the GPI anchor is phosphatidylinositol glycosidically linked to an unsubstituted glucosamine (GlcN), which is further linked to a linear chain of three mannose residues. The terminal mannose is connected to phosphoethanolamine, which links in turn to the C-terminus of a protein.

2.3.2.3 Glycoglycerolipids. In both plant and animal organisms, glycoconjugates of mono- or oligosaccharides with mono- and diacylglycerols have been detected [31]. The reducing end monosaccharides seem to be restricted to galactose and glucose (Figure 2.12).

2.3.3 Polymeric Structures

Polymeric saccharides are frequently divided into homo- and heteropolysaccharides. Homopolysaccharides are built up from only one kind of monosaccharides, whereas heteropolysaccharides contain at least two different monosaccharides, mostly combined in repetitive units with a size from two to six monosaccharides. Commonly known homopolysaccharides serve heterogeneous functions such as cellulose as a structural cell

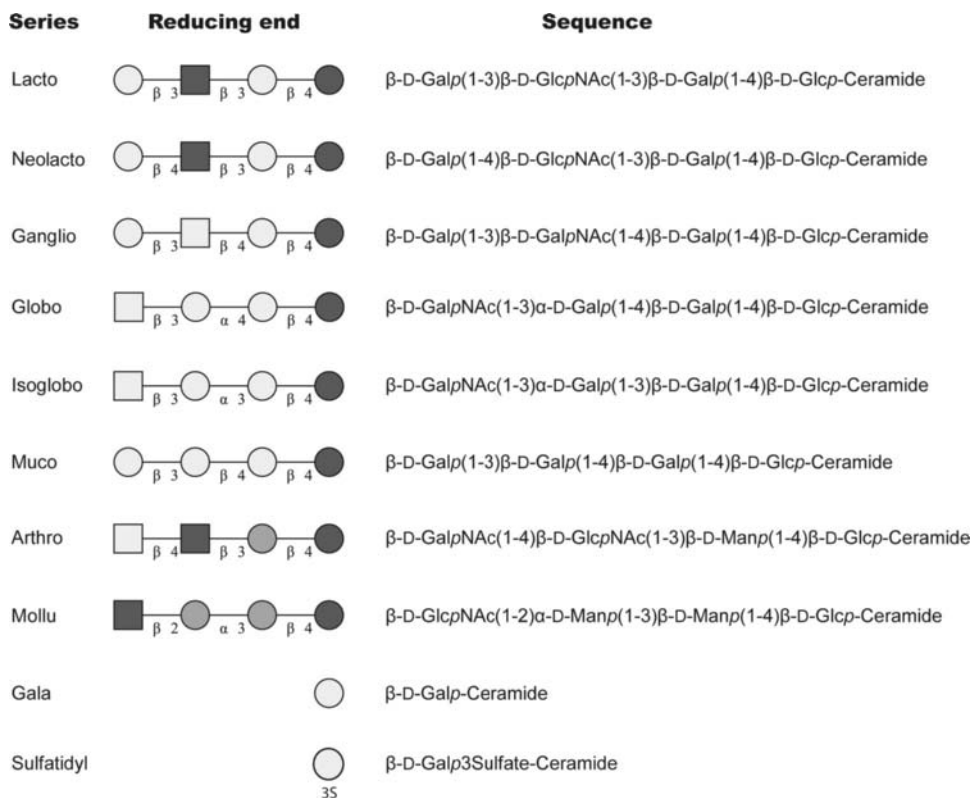


Figure 2.11 Common structural motifs constitute the main classification criterion for glycosphingolipids.

wall component of plants, glycogen as an energy storage saccharide, and chitin as the main exoskeleton material of insects.

2.3.3.1 Proteoglycans and Glycosaminoglycans. Proteoglycans (PGs) [22, 32] are a ubiquitous family of biomolecules that are composed of a core protein and one or more covalently attached sulfated glycosaminoglycan (GAG) [33] chains. GAGs can also appear unconjugated. GAGs are linear polymers built up from disaccharide building blocks, containing, as their name suggests, amino sugars and uronic acids (Figure 2.13). The degree and position of sulfation and epimerization (GlcA to IdoA) are extremely variable in GAGs. They constitute a very heterogeneous class of substances, as sulfation and epimerization reactions occur in a non-stoichiometric manner depending on the tissue/cellular/metabolic context.

2.3.3.2 Bacterial Carbohydrate Structures. Bacteria are capable of synthesizing complex polysaccharides with an enormous structural variety. These polysaccharides are mainly associated with the different layers of the cell wall. The more complex cell wall of Gram-negative bacteria consists of two membrane layers, which are separated by the periplasmic space (Figure 2.14a). In this intermembrane cavity, the two main saccharide classes of glucans and peptidoglycans have been described (Figure 2.14c, 14d). The outer leaflet

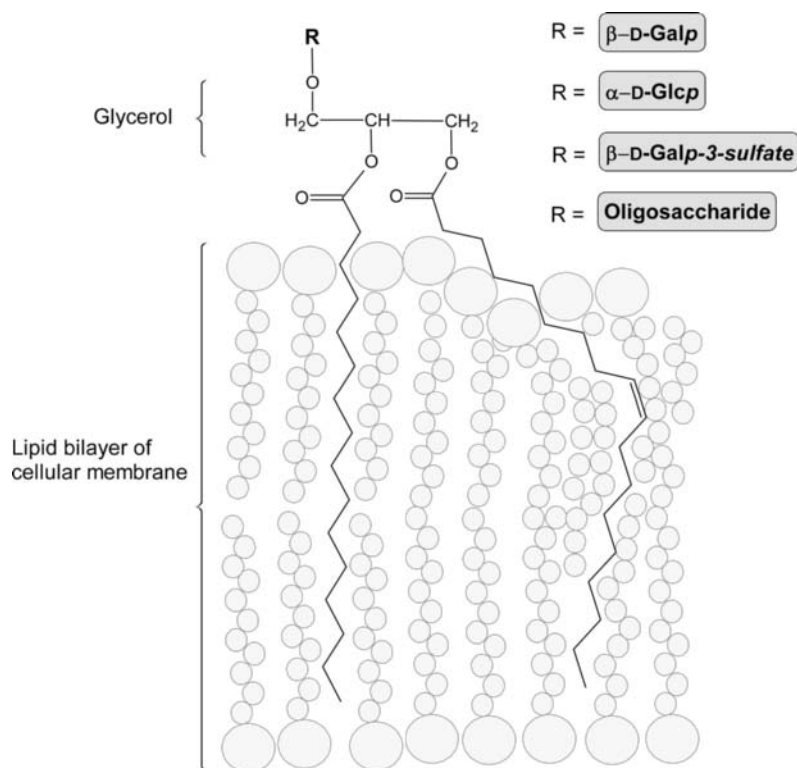


Figure 2.12 An example of a glycosylglycerolipid attached to a cellular membrane. It is constituted of carbohydrate moieties and a membrane-anchoring part, here a diacylglycerol.

of the outer membrane contains the lipopolysaccharide (LPS), a membrane-anchored polysaccharide with a distinct structure (Figure 2.14b). Depending on bacterial species and serotype, the LPS can be modified with a diverse range of substituents and rare monosaccharides. Gram-positive bacteria have a simpler cell wall, which consists of one membrane and a thicker peptidoglycan layer. Teichoic acids (polyribitol- or polyglycerolphosphates) conjugated to peptidoglycans or the membrane are distinct marker substances for Gram-positive bacteria. Frequently, bacteria have the potential to ensheath themselves in an additional layer of complex capsular polysaccharides (K-antigens as opposed to O-antigens from LPS and F-antigens from their mobility apparatus). These capsule polysaccharides can be coarsely divided into groups according to their monosaccharide content. There is now also evidence suggesting a glycosylation machinery for proteins in certain bacterial species [34].

An analysis of digitally available bacterial carbohydrate sequences reveals marked differences in disaccharide patterns compared with the mammalian situation [13, 35]. Most of the frequently occurring disaccharide patterns can be attributed to the specialized polysaccharides of the bacterial cell as discussed above.

2.3.3.3 Carbohydrate Structures of Plants. Apart from particularities in their *N*- and *O*-glycosylation machinery, which we will not discuss here (for more information, see

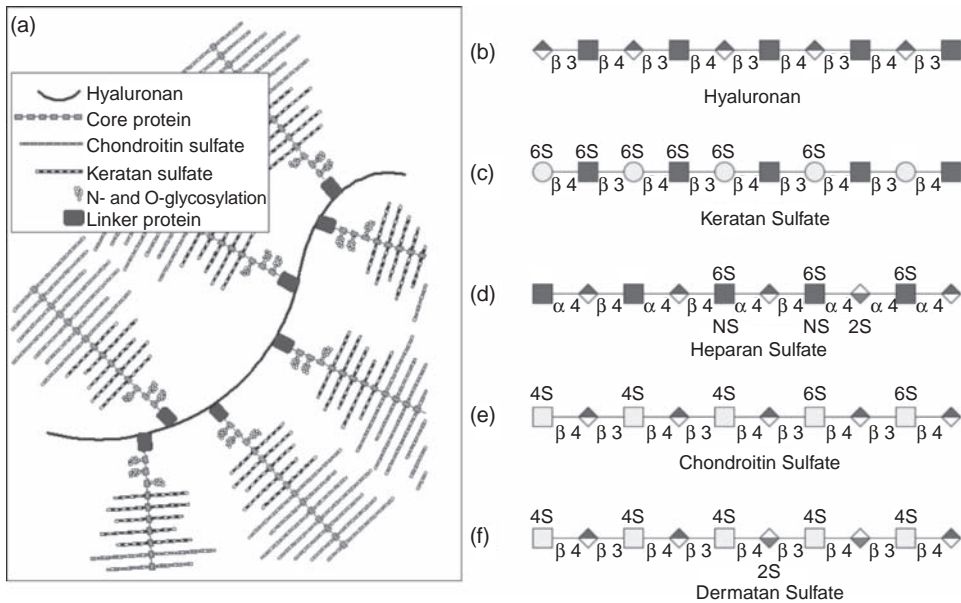


Figure 2.13 Proteoglycans and glycosaminoglycans. (a) Typical schema of a proteoglycan. A central protein is bound to a hyaluronan chain. The central protein is heavily glycosylated with both *N*- and *O*-glycans and glycosaminoglycans. (b)–(f) Different glycosaminoglycans. (b) Hyaluronan is not covalently attached to a protein, but directly secreted to the ECM. A polymer can contain up to 10 000 disaccharide building blocks. (c) Keratan sulfate is a sulfated glycosaminoglycan that contains a polylactosamine building block. Keratan sulfate can be linked both to *N*- and common *O*-glycan cores. (d) Heparan sulfate is linked via a xylose to a serine of the core protein. (e), (f) Chondroitin sulfate and dermatan sulfate are linked in the same way.

[36]), plants exhibit in their cell wall a variety of polymeric saccharides [37, 38]. The most abundant biogenic substance is cellulose, a linear homopolysaccharide built up by several hundred to over 10 000 $\beta(1-4)$ -linked glucose residues, with a defined secondary structure. Other cell wall components are hemicelluloses, a smaller branched heteropolysaccharide class with glucose, mannose, galactose, xylose, and arabinose content. Rhamnogalacturonans (RG-I) form another diverse class of plant polysaccharides with a $-2)-\alpha$ -L-Rhap-(1-4)- α -D-GalpA-(1- backbone to which side branches containing arabinose, galactose, xylose, and fucose are connected. The data for plants in carbohydrate sequence databases are sparse.

2.3.4 Motifs

Another means of grouping saccharides is the identification of sequence motifs, which can be found in glycolipids and *N*- and *O*-glycans. Regulated biosynthetic processes determine the precise nature of the terminal modifications of the core structures. Importantly, terminal mono- and oligosaccharides are those best accessible for receptor interactions. The frequently found extensions have received trivial names, with the primary examples being those of the human blood groups, which correspond to certain

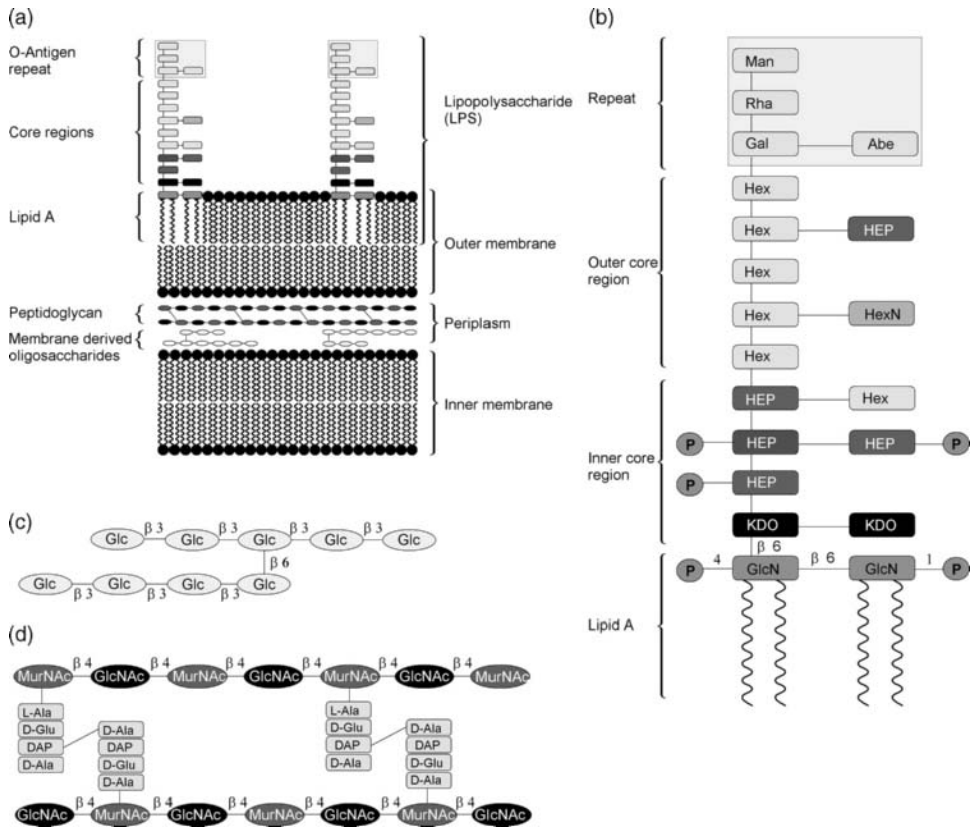


Figure 2.14 Selected cell wall components of Gram-negative bacteria with a carbohydrate content. (a) Schematic drawing of an idealized cell wall of a Gram-negative bacterium. Two membrane layers flank a periplasmic space, in which peptidoglycans and membrane-derived oligosaccharides (glucans) reside. The outer membrane contains lipopolysaccharides (LPS). (b) The general structure of lipopolysaccharides. Two glycosidically linked glucosamine residues are conjugated to membrane-anchoring lipophils; this substructure is called lipid A or endotoxin. An inner core region rich in ketodeoxyoctulosonic acid (KDO) and heptoses is attached to the lipid A, followed by an outer diverse core region. The polymeric *O*-antigen terminates the LPS. (c) β -Glucan structures are frequently found in the periplasmic space. (d) The main structure of peptidoglycan. A disaccharide repeating unit builds up long polymers, which are interconnected by a peptide bond between adjacent tetrapeptides conjugated to muramic acid.

small oligosaccharide fragments located on the erythrocytes (Figure 2.15: blood group antigens). Other oligosaccharide extensions found on erythrocyte membranes are the Lewis antigen with its different forms (Figure 2.15: Lewis A-Y, sialyl Lewis A, sialyl Lewis X), and the lactosamine extension with its different subtypes (Figure 2.15: lactosamine, neo-lactosamine, LacDiNAc). The polylactosamine motif is a disaccharide repeating unit found on *N*-glycans, *O*-glycans, and glycolipids (Figure 2.15: polylactosamine) [3]. The human P-antigens are described on glycosphingolipids (Figure 2.15: P-, Pk- and P1-antigens). CAD- and Sda-antigens are also found as blood group determinants.

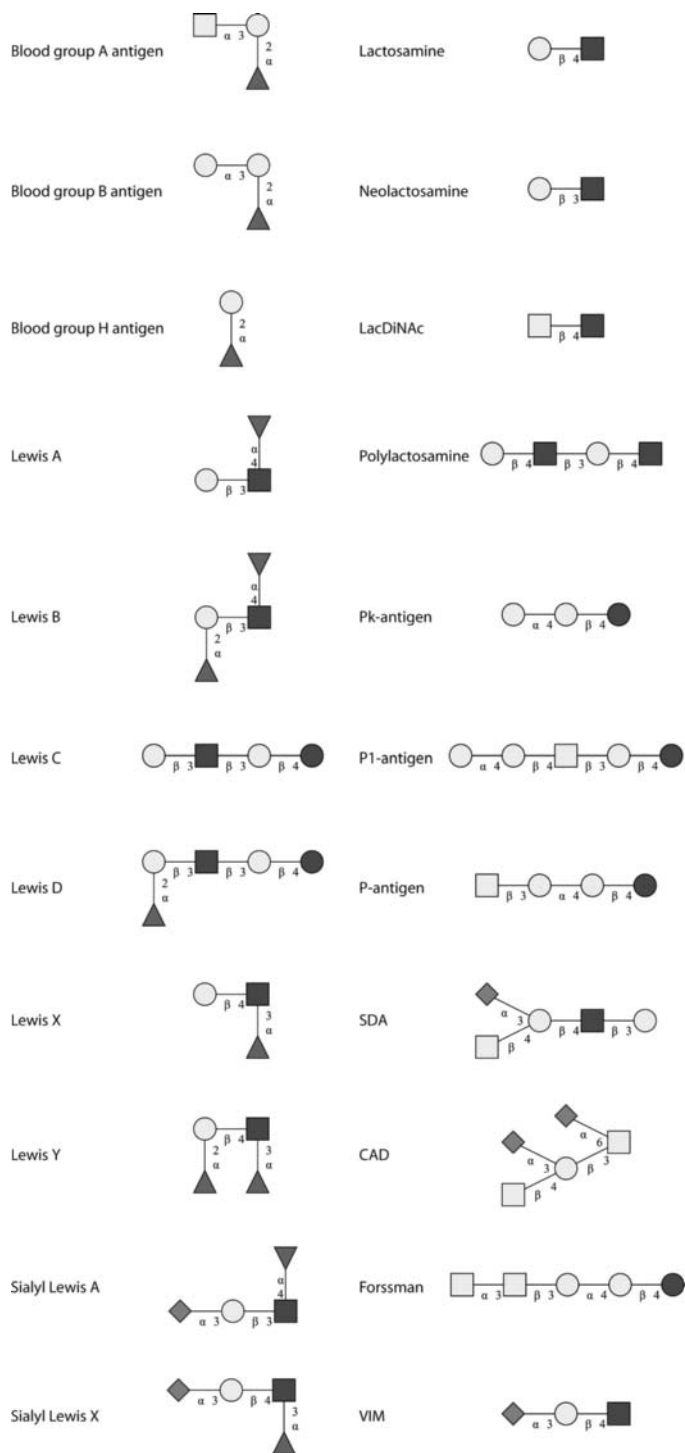


Figure 2.15 Named carbohydrate motifs. For explanation, see text.

2.4 Conclusion

A large variety of carbohydrates exist, which are often specific to certain species. Nevertheless, natural organisms do not populate the large number of potential isomers that mono- and oligosaccharides could provide. Each living organism uses a small set of monosaccharide building blocks determined ultimately by their genetically determined repertoire of glyco-related enzymes to produce the monosaccharides and also to link them together. Through statistical analysis of the carbohydrate structures deposited in databases [13, 39], one can derive an overview of the structural diversity of experimentally determined mammalian and bacterial “glycospace.”

It has been recognized that the information available in carbohydrate sequence databases is not complete, and large “glycome” projects – where all carbohydrates produced by an organism are systematically determined experimentally – may not be feasible. Therefore, bioinformatic approaches may be used to estimate the size of the glycome of an organism by analyzing the so-called “glycogenes.” Bioinformatic methods have already been used to find all glycosyltransferase genes in humans [40, 41]. Generally, the glyco-related pathways in mammals are now well-investigated regulatory networks [15]. However, although the 10 “mammalian” monosaccharides (Figure 2.5) and about 180 human glycogenes (<http://riodb.ibase.aist.go.jp/rcmg/ggdb/>) [40], most of which produce a specific linkage, are fairly well known and characterized, it is very difficult to give a realistic estimate of how many glycan structures can really be found in mammalian tissues. Because the biosynthesis of carbohydrates is not under direct genetic control, dozens of different enzymes are involved in the synthesis of the sugar chains attached to proteins or lipids. Depending on which of these enzymes are expressed and which donors are available in the cell that synthesizes the carbohydrates, various different glycan chains will be produced.

Recently, a mathematical model for the *N*-glycan pathway, the best understood biosyntheses process in glycobiology, was developed to simulate possible extension of the *N*-glycan core structure encoding the reaction of 11 different enzymes. Although several simplifications were included, the model generates 7565 *N*-glycans in a network of 22 871 reactions [42]. The number of distinct *N*-glycan structures currently deposited in publicly available databases totals about 4000 [9]. In another approach to estimating possible *N*-glycan structures, a knowledge-based approach to generate sets of possible *N*-glycans was implemented in the Cartoonist program, used to assist rapid interpretation of mass spectra. Experts have to manually include so-called archetype cartoons. These are essentially known basic substructures occurring for a specific class of glycans. Based on these archetype cartoons and sets of rules, the program generates a library of possible *N*-glycan structures. The latest version of Cartoonist generates about 145 000 different *N*-glycan structures [43, 44]. Although many details of the *N*-glycosylation pathway are known, prediction of the structures actually synthesized is further complicated by the fact that the *N*-glycan branching is ultra-sensitive to increases in the UDP-GlcNAc donor levels. Consequently, glycoprotein receptors have evolved with low or high numbers of *N*-glycan sites so they may take advantage of differential *N*-glycan branching to regulate the strength of their association with a cell-surface lattice [45, 46]. As a conclusion, systems biology approaches [47, 48] to estimate the number and structures of *N*-glycans synthesized in a specific tissue will have to take into account all of the above-outlined influences on the level of available enzymes and also donor molecules.

References

1. Weerapana E, Imperiali B: Asparagine-linked protein glycosylation: from eukaryotic to prokaryotic systems. *Glycobiology* 2006, **16**:91R–101R.
2. Gagneux P, Varki A: Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology* 1999, **9**:747–755.
3. Varki A, Cummings RD, Esko JD, Freeze H, Hart G (eds): *Essentials of Glycobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2008.
4. Brooks SA, Dwek MV, Schuhmacher U: *Functional and Molecular Glycobiology*. Oxford: Bios Scientific Publishers; 2002.
5. Kamerling JP (ed.): *Comprehensive Glycoscience – from Chemistry to Systems Biology*. Oxford: Elsevier; 2007.
6. Lindhorst TK: *Essentials of Carbohydrate Chemistry and Biochemistry*, 3rd edn. Weinheim: Wiley-VCH Verlag GmbH; 2007.
7. Ernst B, Hart GW, Sinaÿ P (eds): *Carbohydrates in Chemistry and Biology*. Weinheim: Wiley-VCH Verlag GmbH; 2000.
8. Herget S, Ranzinger R, Maass K, von der Lieth C-W: GlycoCT – a unifying sequence format for carbohydrates. *Carbohydr Res* 2008, **343**:2162–2171.
9. Ranzinger R, Herget S, Wetter T, von der Lieth C-W: GlycomeDB – integration of open-access carbohydrate structure databases. *BMC Bioinformatics* 2008, **9**:384.
10. McNaught AD: Nomenclature of carbohydrates (recommendations 1996). *Adv Carbohydr Chem Biochem* 1997, **52**:43–177.
11. Lowe JB, Marth JD: A genetic approach to mammalian glycan function. *Annu Rev Biochem* 2003, **72**:643–691.
12. Ohtsubo K, Marth JD: Glycosylation in cellular mechanisms of health and disease. *Cell* 2006, **126**:855–867.
13. Herget S, Toukach PV, Ranzinger R, Hull WE, Knirel YA, von der Lieth C-W: Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Struct Biol* 2008, **8**:35.
14. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M: Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* 2005, **21**:3976–3982.
15. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M: KEGG as a glycome informatics resource. *Glycobiology* 2006, **16**:63R–70R.
16. Kornfeld R, Kornfeld S: Assembly of asparagine-linked oligosaccharides. *Annu Rev Biochem* 1985, **54**:631–664.
17. Helenius A, Aebi M: Intracellular functions of N-linked glycans. *Science* 2001, **291**:2364–2369.
18. Apweiler R, Hermjakob H, Sharon N: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999, **1473**:4–8.
19. Bause E: Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem J* 1983, **209**:331–336.
20. Chavan M, Lennarz W: The molecular basis of coupling of translocation and N-glycosylation. *Trends Biochem Sci* 2006, **31**:17–20.
21. Lehle L, Strahl S, Tanner W: Protein glycosylation, conserved from yeast to man: a model organism helps elucidate congenital human diseases. *Angew Chem Int Ed* 2006, **45**:6802–6818.
22. Lamoureux F, Baud'huin M, Duplomb L, Heymann D, Redini F: Proteoglycans: key partners in bone cell biology. *BioEssays* 2007, **29**:758–771.
23. Wells L, Hart GW: O-GlcNAc turns twenty: functional implications for post-translational modification of nuclear and cytosolic proteins with a sugar. *FEBS Lett* 2003, **546**:154–158.
24. Peter-Katalinić J: Methods in enzymology: O-glycosylation of proteins. *Methods Enzymol* 2005, **405**:139–171.

25. Merrill AH Jr, Wang MD, Park M, Sullards MC: (Glyco)sphingolipidology: an amazing challenge and opportunity for systems biology. *Trends Biochem Sci* 2007, **32**:457–468.
26. Hanzal-Bayer MF, Hancock JF: Lipid rafts and membrane traffic. *FEBS Lett* 2007, **581**:2098–2104.
27. Chester MA: IUPAC–IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature of glycolipids – recommendations 1997. *Eur J Biochem* 1998, **257**:293–298.
28. Svennerholm L: Chromatographic separation of human brain gangliosides. *J Neurochem* 1963, **10**:613–623.
29. Murthy PP: Structure and nomenclature of inositol phosphates, phosphoinositides, and glycosylphosphatidylinositols. *Subcell Biochem* 2006, **39**:1–19.
30. Jones DR, Varela-Nieto I: The role of glycosyl-phosphatidylinositol in signal transduction. *Int J Biochem Cell Biol* 1998, **30**:313–326.
31. Holzl G, Dormann P: Structure and function of glycoglycerolipids in plants and bacteria. *Prog Lipid Res* 2007, **46**:225–243.
32. Hardingham TE, Fosang AJ: Proteoglycans: many forms and many functions. *FASEB J* 1992, **6**:861–870.
33. Sasisekharan R, Venkataraman G: Heparin and heparan sulfate: biosynthesis, structure and function. *Curr Opin Chem Biol* 2000, **4**:626–631.
34. Abu-Qarn M, Eichler J, Sharon N: Not just for Eukarya anymore: protein glycosylation in Bacteria and Archaea. *Curr Opin Struct Biol* 2008, **18**:544–550.
35. Toukach FV, Knirel YA: New database of bacterial carbohydrate structures. *Glycoconj J* 2005, **22**:216–217.
36. Reiter WD: Biosynthesis and properties of the plant cell wall. *Curr Opin Plant Biol* 2002, **5**:536–542.
37. Joshi CP, Mansfield SD: The cellulose paradox – simple molecule, complex biosynthesis. *Curr Opin Plant Biol* 2007, **10**:220–226.
38. Lytovchenko A, Sonnewald U, Fernie AR: The complex network of non-cellulosic carbohydrate metabolism. *Curr Opin Plant Biol* 2007, **10**:227–235.
39. Werz DB, Ranzinger R, Herget S, Adibekian A, von der Lieth C-W, Seeberger PH: Exploring the structural diversity of mammalian carbohydrates (“glycospace”) by statistical databank analysis. *ACS Chem Biol* 2007, **2**:685–691.
40. Narimatsu H: Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj J* 2004, **21**:17–24.
41. Kikuchi N, Narimatsu H: Bioinformatics for comprehensive finding and analysis of glycosyltransferases. *Biochim Biophys Acta* 2006, **1760**:578–583.
42. Krambeck FJ, Betenbaugh MJ: A mathematical model of N-linked glycosylation. *Biotechnol Bioeng* 2005, **92**:711–728.
43. Goldberg D, Sutton-Smith M, Paulson J, Dell A: Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* 2005, **5**:865–875.
44. Goldberg D, Bern M, Parry S, Sutton-Smith M, Panico M, Morris HR, Dell A: Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J Proteome Res* 2007, **6**:3995–4005.
45. Stanley P: A method to the madness of N-glycan complexity? *Cell* 2007, **129**:27–29.
46. Lau KS, Partridge EA, Grigorian A, Silvescu CI, Reinhold VN, Demetriou M, Dennis JW: Complex N-glycan number and degree of branching cooperate to regulate cell proliferation and differentiation. *Cell* 2007, **129**:123–134.
47. Ingolia NT, Weissman JS: Systems biology: reverse engineering the cell. *Nature* 2008, **454**:1059–1062.
48. Murrell MP, Yarema KJ, Levchenko A: The systems biology of glycosylation. *ChemBioChem* 2004, **5**:1334–1347.

3 Digital Representations of Oligo- and Polysaccharides

Stephan Herget and Claus-Wilhelm von der Lieth

*Central Spectroscopic Unit, Deutsches Krebsforschungszentrum
(German Cancer Research Center), 69120 Heidelberg, Germany*

3.1 Introduction

Glycoscience is an interdisciplinary field of research and scientists from various communities are involved in elucidating the biological roles of carbohydrates from diverse perspectives, using a wide variety of experimental techniques. Each of these communities has its preferred way to represent carbohydrate structures. Whereas biomedically oriented scientists tend to use coarse descriptions such as composition and symbolic representations, biochemically oriented groups often prefer a residue-based alphanumeric description (mainly one of the various IUPAC depictions). In contrast, chemists clearly favor graphical representations, which show all atomic details including stereochemistry. X-ray crystallographers, NMR spectroscopists and molecular modelers prefer 3D representations. This chapter will present an overview of the sequence formats used in glycobioinformatics.

3.1.1 An Abstract View of Carbohydrate Sequences

Sequences of DNA or protein can be handled bioinformatically as simple linear strings, whereas carbohydrate sequences contain special informatic challenges caused by the property of branching. They can probably be best described in computational terms as graphs, with the monosaccharide residues as the vertices (nodes) and the glycosidic linkages as edges (lines) (Figure 3.1). As carbohydrate sequences contain a preferred direction from the reducing end to the non-reducing end, the graphs can be viewed as directed (digraphs). The existence of potential multiple connections between two residues (e.g. sialyl glycoside lactonization) can degenerate the graph to a multigraph. The rare cyclization of carbohydrate structures (e.g. cyclodextrins) can lead to cyclic graphs. To avoid a combinatorial expansion of identical substructures, so-called repeating units are frequently encoded as special entities. Limited analytical techniques resulting in partial structure elucidation can produce uncertainties in the sequences, especially regarding the location of terminal residues (capping units). Some secondary modifications (e.g. sulfation) are present only on a fraction of the residues (nodes) of repeating units, leading to non-stoichiometric modification

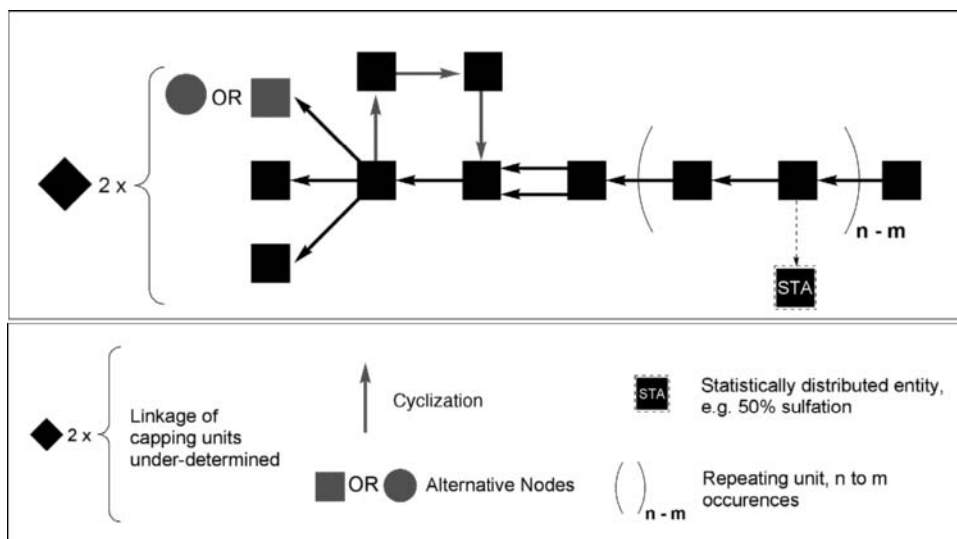


Figure 3.1 A generic schema of a complex carbohydrate sequence. Squares represent monosaccharide residues (graph vertices), arrows symbolize glycosidic linkages (directed edges). Carbohydrate sequences may contain repeating units with non-stoichiometric modifications (STA), multiple connections between vertices, cyclic subgraphs, alternative residue declarations, and fuzzily defined capping unit locations.

patterns. Finally, alternative residue names at a certain position vertex can be the result of an impartial structure elucidation.

3.1.2 Historical Overview of Carbohydrate Sequence Formats and Storage Capabilities

An intuitive way of storing carbohydrate sequence topologies is a simple two-dimensional sketch analogous to that shown in Figure 3.1a. ASCII 2D plots, with the residues and linkage data inserted as text in close resemblance to IUPAC recommendations [1], were used by the first initiative for carbohydrate sequence storage, the Complex Carbohydrate Structure Database (CCSD) [2, 3] (see Section 3.2.2, Figure 3.5). Subsequent initiatives using relational databases opted for storing the carbohydrate sequences as strings, similar to those used in protein or DNA databases. These strings were obtained by an ordered traversal of the carbohydrate sequences and could thus serve as primary keys in database systems (e.g. canonicalized string representations such as used in LINUCS [4], GlycoSuite [5, 6], LinearCode [7], and BCSDB [8]). Later, the connectivity information in carbohydrate sequences was stored using connection table-like representations (KCF [9], GlycoCT [10]). These connection tables can be naturally expressed in XML encodings (Glyde [11], CabosML [12]). The different formats used in glycoinformatics have different capabilities to store the complex information potentially present in carbohydrate sequences (Table 3.1).

Table 3.1 A comparison of the structural information storage capabilities of the different sequence formats used in glycoinformatics.

Sequence format	Multiple connections	Cyclization	Repeating units	Capping unit under-determined	Non-stoichiometric modification	Alternative residues
CCSD	+	–	+	–	–	–
LINUXS	–	+	+	–	–	–
BCSDB	–	–	+	–	–	+
LinearCode	–	+	–	+	–	+
KCF	+	+	+	–	–	–
CabosML	+	+	+	–	–	–
Glyde-II	+	+	+	+	+	–
GlycoCT	+	+	+	+	+	+

The following sections will provide more detailed insights into the different sequence formats used in glycoinformatics, focusing mainly on the topological descriptors developed so far.

3.2 Sequence Formats

3.2.1 The Chemical Standard IUPAC–IUBMB

The standardized nomenclature of carbohydrates is the recommendation of the joint commission of the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Biochemistry and Molecular Biology (IUBMB) [1]. Since the IUPAC definitions are to some extent used by most of the existing databases, the main agreements and rules will be summarized here with respect to digital encoding of complex carbohydrate topologies. See Chapter 2 for an introduction to monosaccharide naming conventions.

3.2.1.1 Topological Descriptors According to IUPAC. Since the IUPAC recommendations are able to describe most of the structural features of naturally occurring oligosaccharides, most of their basic definitions have been adopted by the publicly available databases:

- The conventional depiction has the reducing sugar on the right and the non-reducing end on the left. Also, when there is a glycosyl linkage to a non-carbohydrate moiety (e.g. protein, peptide or lipid), the glycosyl residue involved should appear on the left.
- The linkage of two monosaccharide units is described by the locant of the anomeric carbon atom, an arrow or a hyphen, and the locant of the connecting oxygen of the next monosaccharide unit which are set in parentheses between the names of the residues concerned [e.g. Gal-(1→4)-Glc]. The anomeric centers are described either within the linkage or on the residue level.

IUPAC provides several ways to write the abbreviated forms of oligosaccharide sequences. Examples are given for three encoding schemata for the *N*-glycan core in Figure 3.2.

The extended and condensed forms according to IUPAC are two-dimensional graphical representations of oligosaccharide structures; in the condensed form the more frequent

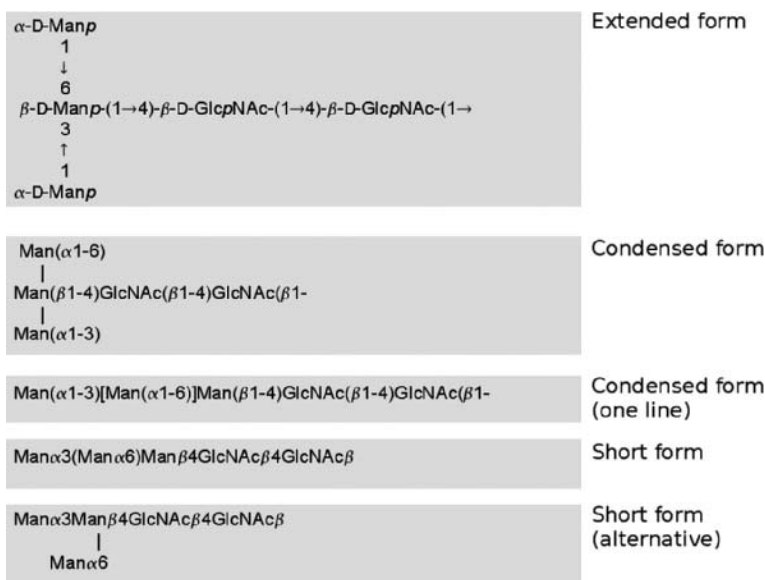


Figure 3.2 IUPAC defines several alternatives for the description of oligosaccharide sequences. Some databases use variants of the condensed form, adding a sorting algorithm to provide a unique key for relational databases.

configuration (D or L) and ring size can be omitted. The recommendation contains, furthermore, an example of a pure string representation for an oligosaccharide [condensed form (one line)], which places branching chains in square brackets. The short form reduces the sequence even more, defining the C1 atom of aldoses and the C2 atom of ketoses as the default connection, which can be left out. The recommendation, however, defines no precedence rules for sorting and therefore does not result in unique sequences for branched oligosaccharides.

3.2.2 Complex Carbohydrate Structure Database (CCSD)

The Complex Carbohydrate Structure Database (CCSD) – often called CarbBank with reference to the retrieval software used to access the data – was developed and maintained by the Complex Carbohydrate Research Center of the University of Georgia (USA) [2, 3]. An online version of the database is still available [13]. It was the largest endeavor during the 1990s to collect glycan structures, mainly through retrospective manual extraction from the literature. The main aim of the CCSD was to store all publications in which specific carbohydrate structures were reported. However, when the funding stopped during the second half of the 1990s, CarbBank was not developed further and the CCSD was no longer updated. Nevertheless, with about 50 000 entries of about 20 000 different structures, the CCSD is still one of the largest repositories of glycan-related data. Figure 3.3 depicts the CarbBank encoding scheme for monosaccharides.

CarbBank defined the first sequence format for oligosaccharide structures, which was designed for computational purposes, basically to facilitate retrieval operations. It stored the sequence information using the extended IUPAC description as 2D graphical

(a)

①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬
1,6An-	D-	gro-	a-	L-	5-en	6-deoxy-	araHex	3ulo	p	A	subst	-ol

① anhydro
 ② ③ For sugars with > 6 stereogenic C-atoms, additional stereocenters are designated with an identifier here
 ④ ⑤ anomeric and configurational information
 ⑥ ⑦ double bonds (en) and deoxygenations (deoxy)
 ⑧ basetype of sugar
 ⑨ ulo indicates keto-function
 ⑩ ring form
 ⑪ Uronic acid
 ⑫ position and identity of substituents
 ⑬ -ol for alditols, -aric for glycaric acids, -onic for glyconic acids

(b)

b-D-Glcp	β-D-glucose, pyranose form
b-D-GlcpA	β-D-glucuronic acid
b-D-GalpNAc	β-D-galactose, position 2: N-acetylamino
D-Glcp-onic	D-gluconic acid
a-L-6-deoxy-Talp	α-L-talose, pyranose form, C6 deoxygenated
D-Gro-a-D-manHepp	D-glycero-α-D-manno-heptose, pyranose form
b-D-3en-eryHexpA	β-D-erythro-hexose, uronic acid, 3,4 C-C double bond

Figure 3.3 (a) CarbBank introduced a defined encoding scheme for monosaccharides, which resembles a condensed and ordered IUPAC string. Standard positions for common substitutions and trivial names were used. In simplification of IUPAC representations, α and β were represented by a and b, respectively, and the font size of the configurational descriptors D and L (usually small capitals) was made the same as that of the normal text. (b) Examples.

representations (Figures 3.4 and 3.5). Extensions of the notation allowed it to encode cyclic structures, repeating units and ambiguities on both linkage and building block level. No hierarchy rules for the ordering of branches of the oligosaccharide were explicitly defined. However, by convention, the branch with the highest locant was positioned at the top.

Although the CarbBank rules are in principle able to provide a unique description of a monosaccharide, they have unfortunately not been consistently applied to all CCSD entries. The inconsistency in monosaccharide encoding used within CarbBank results from the fact that different people have entered structures and that no automatic check using a controlled vocabulary was implemented. Additionally, the scientific curators had to include modifications of carbohydrate structures, which were not completely covered by the CarbBank encoding schema. Unfortunately, only a few copies of the manual are still circulating and it seems that not all conventions – especially not the latest ones – applied to the database are completely documented.

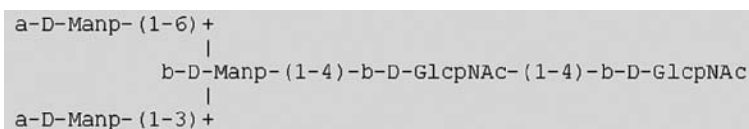


Figure 3.4 CarbBank used 2D graphical representations to store the oligosaccharide sequences.

<pre> ; start of record ; db=ccsd29 CC: CCSD:17587 AU: Priem B; Gitti R; Bush CA; Gross KC TI: Structure of ten free N-glycans in ripening tomato fruit CT: Plant Physiol (1993) 102: 445-458 BS: (GS) Lycopersicon esculentum, (CN) tomato MT: N-linked glycoprotein DA: 23-11-1993 ----- structure: a-D-Manp-(1-6)+ b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)+ a-L-Araf-(1-2)+ a-L-Fucp-(1-3)+ D-GlcNAc D-GlcNAc =====end of record- </pre>	<pre> Start flag Database revision Complex Carbohydrate ID Authors Title of publication Citation source Biological source Metatype Date of entry ASCII 2 D graph of structure End flag </pre>
---	---

Figure 3.5 Excerpt from CarbBank for CCSD-ID 17587. The sequences were stored in a flat-file, which contained different sections for additional information and annotation for each entry.

CarbBank provided options for exact and substructure searches, but did not check if a newly entered structure was already contained in the database, as the major aim of this database was literature tracking. Since the CCSD data have been included in nearly all active projects to some extent, the ideas and conventions introduced by CarbBank are still of importance.

3.2.3 Linear Notation for Unique Description of Carbohydrate Sequences (LINUCS)

The GLYCOSCIENCES.de portal [14, 15] – the former SweetDB [16] – was established at the end of the 1990s at the German Cancer Research Center (DKFZ) and includes most of the sequences from the CCSD and Sugabase [17, 18] (a carbohydrate NMR database that combines CarbBank data with proton and carbon chemical shift values). Additionally, the database is updated with manually selected structures and NMR spectra. The CarbBank nomenclature is used as the encoding scheme for the monosaccharides. For the sequence format, the GLYCOSCIENCES.de portal uses a newly defined sequence format called LINUCS (LInear Notation for Unique description of Carbohydrate Sequences) [4], which produces unique strings for branched oligosaccharides by using the linkage information to establish a hierarchy of branches (Figure 3.6). The sequences start at the reducing end and are sorted alphanumerically by the linkage information. Each residue is followed by obligatory curly brackets (braces), which contain possible sub-branches. The format allows implicitly for ambiguities on residue and linkage level to occur: the unknown information is represented by the metacharacter “?”. Repeating elements and cyclizations have been foreseen by the developers of this format.

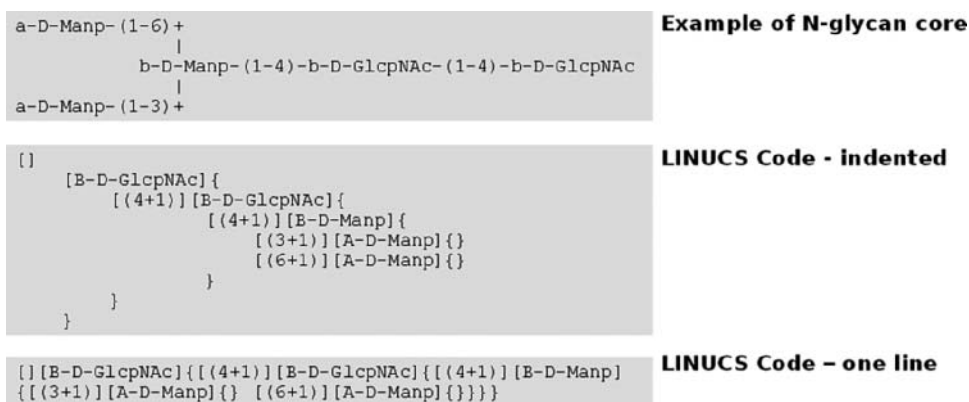


Figure 3.6 Linear notation, which is used in the GLYCOSCIENCES.de portal, exemplified for the *N*-glycan core. The resulting code is shown in indented and one-line variants.

3.2.4 Bacterial Carbohydrate Structure Database (BCSDB)

The aim of the Bacterial Carbohydrate Structure Database (BCSDB) [19] is to provide a database of structural, bibliographic, and related information on bacterial carbohydrate structures [8, 20]. The BCSDB uses a monosaccharide encoding, which is in some aspects similar to the extended IUPAC format. Each residue name is composed of several identifiers following each other without any separators: anomeric configuration, configurational symbol, carbohydrate stem type, ring size, modification position(s) and modifiers. In contrast to IUPAC conventions, some monovalent substituents (e.g. Me = methyl, Ac = acetyl) are described as separate residues, e.g. aDGal(1-3)bDGlcNAc is recorded as aDGal(1-3)[Ac(1-2)]bDGlcN. Sequences are encoded in a linear fashion using brackets. The BCSDB has a set of rules which results in unique representations for oligosaccharides by ordering the carbohydrate sequences. A special focus is on repeating units, which are common in bacterial polysaccharides. Examples of the BCSDB formats are shown in Figure 3.7; the indented form is included for readability.

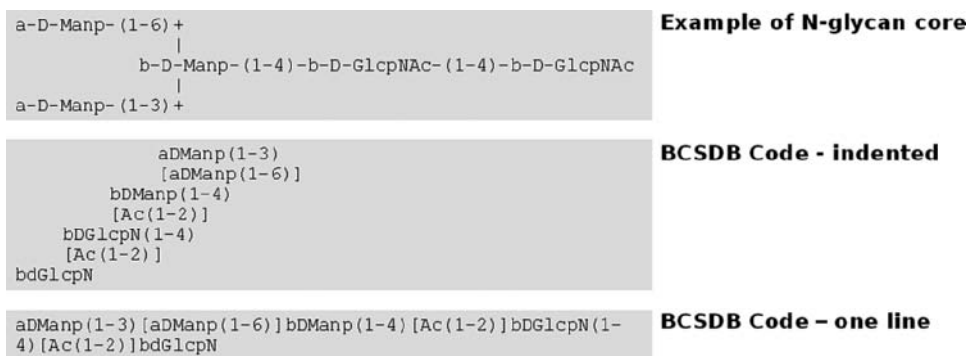


Figure 3.7 The BCSDB encoding exemplified for the *N*-glycan core. Acetyl is treated as a separate residue, in contrast to other encoding schemes.

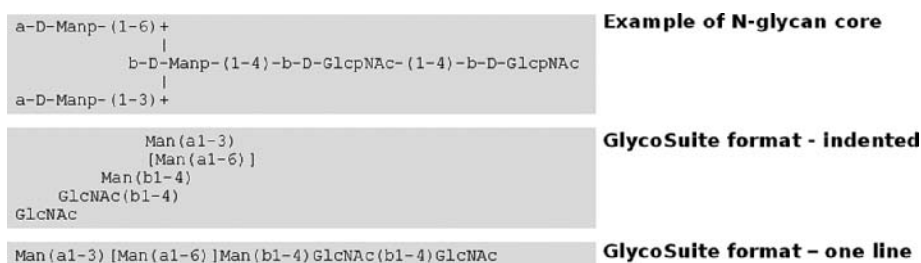


Figure 3.8 GlycoSuite’s format for representing oligosaccharide sequences.

3.2.5 GlycoSuiteDB – Sequence Format

Established by the Australian company Proteome Systems and derived from the precursor database BOLD [21], the GlycoSuite database [5, 6] aims to cover all published, structurally completely defined, mammalian *O*- and *N*-glycan structures. GlycoSuite [22], which is a commercial, manually curated database, uses its own sequence format. The sequence format is based on the IUPAC condensed form with extensions for repeating units and ambiguous structures. Sequences are made unique by ordering them by longest chain, linkage precedence, and finally alphanumeric sorting. Examples of the condensed and one-line GlycoSuite formats are shown in Figure 3.8.

3.2.6 Glycominds – LinearCode

The company Glycominds has developed an encoding scheme for sugar structures which differs considerably from other approaches [23]. In LinearCode, the monosaccharide building blocks are encoded in an abbreviated form, with the most common forms being represented by a single-letter code (Table 3.2) [7]. The one-letter code *M*, for example, stands for D-mannose in pyranose form. This approach results in very short names for common monosaccharides (see Figure 3.9). The monosaccharide building blocks are given a hierarchical order defined by their position in the list. The ordered controlled list of base-type monosaccharides (Table 3.2a) serves both as a vocabulary and as a sorting criterion for the algorithm to define the main chain in branched oligosaccharides. An introduction to the LinearCode is available at the company’s website [24].

The short codes for the monosaccharide stem types can be modified by a series of operators to encode structural changes. The letters ‘a’ and ‘b’ following the monosaccharide symbol are used to encode the anomeric configuration (see Figure 3.9). Substituents on the common core sugars are encoded as suffixes originating from a controlled vocabulary (Table 3.2b), for example, a sulfate at position 3 of D-glucose in pyranose form is written as “G[3S]”. Glycoconjugates are divided in the subclasses of amino acids, lipid moieties, and other chemical entities.

The sequences of this format are grouped along the main chain, which is defined by two rules:

1. If residues at branch points are identical, the lower numerical linkage defines the main chain.
2. If residues at a branch point are not identical, the hierarchy of the dictionary defines the main chain.

Table 3.2 (a) Monosaccharide translation table for the LinearCode. The order of this list is crucial for the sorting algorithm which provides unique sequences. (b) Controlled vocabulary for substituents. (c) Examples for modified monosaccharides. The standard definition as found in part (a) is modified with suffixes for ring size and configuration using the operators caret (^), apostrophe ('), and tilde (~). The caret symbol indicates a change of ring size (pyranose/furanose) compared with the definition in part (a), an apostrophe is used to indicate the corresponding stereoisomer (D/L), and the tilde changes both ring size and absolute configuration.

(a) Basic monosaccharides		(b) Substituents		(c) Examples of modifications	
Monosaccharide	Code	Substituent	Code	Monosaccharide	Code
D-Glcp	G	Acid	A	D-Araf'	R'
D-Galp	A	N-Methylcarbamoyl	ECO	D-Galf	A^
D-GlcpNAc	GN	Pentyl	EE	D-Xylf	X^
D-GlcpN	GQ	Octyl	EH	D-Rhaf'	H~
D-GalpNAc	AN	Ethyl	ET	D-GlcpNGc	GJ
D-GalpN	AQ	Inositol	IN		
D-Manp	M	N-Glycolyl	J		
D-ManpNAc	MN	Methyl	ME		
Neu5Ac	NN	Hydroxyl	OH		
Neu5Gc	NJ	Phosphate	P		
Neu	N	Phosphocholine	PC		
Kdn	K	Phosphoethanolamine	PE		
Kdo	W	Pyruvate acetal	PYR		
D-GalpA	L	Amine	Q		
L-IdopA	I	N-Sulfate	QS		
L-Rhap	H	Sulfate	S		
L-Fucp	F	O-Acetyl	T		
D-Xylp	X	Deoxy	Y		
D-Ribp	B				
L-Araf	R				
D-GlcpA	U				
D-Allp	O				
D-Apif	P				
D-Tagp	T				
Abe	Q				
D-Xulf	D				
D-Fruf	E				

LinearCode also contains rules to encode repeating units and circular structures. Furthermore, a number of ambiguities can also be encoded (Table 3.3).

3.2.7 The Format of the Kyoto Encyclopedia of Genes and Genomes (KCF)

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [25] has developed the KCF (KEGG Chemical Function) description for storage and encoding of chemical compounds [9]. This format has been adapted for saccharide structures. It represents the first published sequence format for saccharides which uses a connection table approach (Figure 3.10). All other previous attempts to encode glycan structures linearized the tree-like structures to flat strings. The KEGG encoding scheme for building blocks follows the established CarbBank style. The NODE section of the format describes the building blocks (monosaccharides)

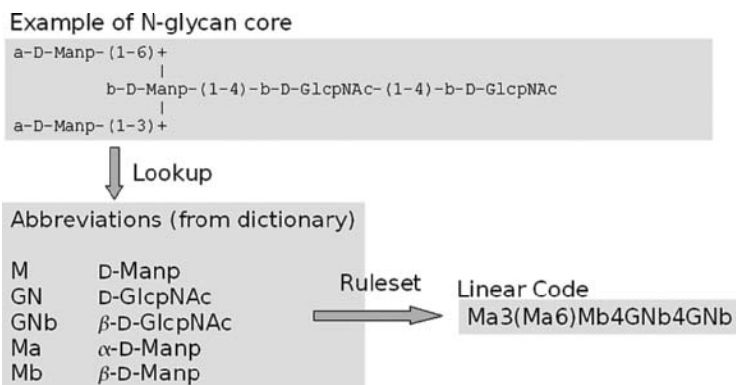


Figure 3.9 LinearCode developed by Glycominds. A compact linear representation results for common oligosaccharides.

sequentially. The list is numbered and its members constitute the nodes of a graph. The EDGE section describes the linkages (glycosidic bonds) between the entries defined in the NODE section. The resulting sequences are not necessarily ordered. Consequently, more than one valid KCF description for a given glycan can be generated.

The KCF description includes X,Y coordinates for each node, used for drawing purposes. An extension of the format allows repeating structures to be encoded, using an additional BRACKET section, which introduces X,Y coordinates of parentheses indicating the repeating unit in the graphical output. The graphical output can be generated by the KegDraw software, which is a stand-alone Java application for editing and visualizing oligosaccharide structures and chemical molecules [26].

3.2.8 Extensible Markup Language (XML)-based Approaches

As demonstrated in the previous sections, a variety of encoding schemes and sequence formats for glycan structures exist. Essentially, nearly every database project has developed its proprietary way to encode glycan structures. The current situation of databases in glycobiology/glycomics can be characterized by the existence of multiple disconnected and incompatible islands of experimental data, data resources, and specific applications.

Table 3.3 Coding of ambiguities in LinearCode.

Ambiguity	Operator
Anomer	?
Substituent position	?
Unknown sugar	*
Undefined linkage	List of linkages, delimiter is “/”
Sugar identity	List of possibilities, delimiter is “/ . . . /”
Unknown position of terminal residues	“ . . . =2% . . . =1% ” to define a list of terminal residues, “2%”, “1%”, etc., to mark the possible positions of these residues in the structure

<pre> a-D-Manp-(1-6)+ b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-GlcpNAc a-D-Manp-(1-3)+ </pre>																																																																																			
KCF																																																																																			
	<table border="0"> <thead> <tr> <th colspan="2"></th> <th>ENTRY</th> <th colspan="2">Glycan</th> <th></th> </tr> <tr> <th colspan="2"></th> <th>NODE</th> <th colspan="2"></th> <th>Total number of nodes</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>Number of node in sequence, unordered</td> <td>5</td> <td colspan="2"></td> <td rowspan="5">} Node (Residue) list including coordinates for graphical display</td> </tr> <tr> <td>GlcNAc</td> <td>Residue name</td> <td>1</td> <td>GlcNAc</td> <td>3.0 0.0</td> </tr> <tr> <td>3.0 0.0</td> <td>XY-coordinates (for image display)</td> <td>2</td> <td>Man</td> <td>-7.0 0.0</td> </tr> <tr> <td></td> <td></td> <td>3</td> <td>Man</td> <td>-16.0 5.0</td> </tr> <tr> <td></td> <td></td> <td>4</td> <td>Man</td> <td>-16.0 -5.0</td> </tr> <tr> <td></td> <td></td> <td>5</td> <td>GlcNAc</td> <td>16.0 0.0</td> <td></td> </tr> <tr> <td></td> <td></td> <th>EDGE</th> <th colspan="2"></th> <th>Total number of edges</th> </tr> <tr> <td>4</td> <td>Number of linkage</td> <td>4</td> <td colspan="2"></td> <td rowspan="4">} Edge (linkage) list</td> </tr> <tr> <td>1:b1</td> <td>Residue 1 is connected via C-atom 1 (beta anomeric) to</td> <td>1</td> <td>2:b1</td> <td>1:4</td> </tr> <tr> <td></td> <td></td> <td>2</td> <td>3:a1</td> <td>2:6</td> </tr> <tr> <td></td> <td></td> <td>3</td> <td>4:a1</td> <td>2:3</td> </tr> <tr> <td>5:4</td> <td>Residue 5 via C-atom 4</td> <td>4</td> <td>1:b1</td> <td>5:4</td> </tr> <tr> <td></td> <td></td> <td>///</td> <td colspan="2"></td> <td></td> </tr> </tbody> </table>			ENTRY	Glycan					NODE			Total number of nodes	1	Number of node in sequence, unordered	5			} Node (Residue) list including coordinates for graphical display	GlcNAc	Residue name	1	GlcNAc	3.0 0.0	3.0 0.0	XY-coordinates (for image display)	2	Man	-7.0 0.0			3	Man	-16.0 5.0			4	Man	-16.0 -5.0			5	GlcNAc	16.0 0.0				EDGE			Total number of edges	4	Number of linkage	4			} Edge (linkage) list	1:b1	Residue 1 is connected via C-atom 1 (beta anomeric) to	1	2:b1	1:4			2	3:a1	2:6			3	4:a1	2:3	5:4	Residue 5 via C-atom 4	4	1:b1	5:4			///			
		ENTRY	Glycan																																																																																
		NODE			Total number of nodes																																																																														
1	Number of node in sequence, unordered	5			} Node (Residue) list including coordinates for graphical display																																																																														
GlcNAc	Residue name	1	GlcNAc	3.0 0.0																																																																															
3.0 0.0	XY-coordinates (for image display)	2	Man	-7.0 0.0																																																																															
		3	Man	-16.0 5.0																																																																															
		4	Man	-16.0 -5.0																																																																															
		5	GlcNAc	16.0 0.0																																																																															
		EDGE			Total number of edges																																																																														
4	Number of linkage	4			} Edge (linkage) list																																																																														
1:b1	Residue 1 is connected via C-atom 1 (beta anomeric) to	1	2:b1	1:4																																																																															
		2	3:a1	2:6																																																																															
		3	4:a1	2:3																																																																															
5:4	Residue 5 via C-atom 4	4	1:b1	5:4																																																																															
		///																																																																																	

Figure 3.10 The format of the Japanese initiative *Kyoto Encyclopedia of Genes and Genomes*.

Therefore, there are several ongoing projects aiming at overcoming this unfavorable situation, which mainly concentrate on providing an XML-based sequence format for glycan structures. XML provides a text-based means to describe and apply a tree-based structure to information. At its base level, all information manifests as text, interspersed with markup that indicates the information's separation into a hierarchy of *character data*, container-like *elements*, and *attributes* of those elements. There are currently two XML descriptions for sugar structures published in the literature.

3.2.8.1 Glycan Data Exchange (GLYDE) Format. As a part of the Integrated Technology Resource for Biomedical Glycomics, established by the National Center for Research Resources, a team from the Complex Carbohydrate Research Center (CCRC), at the University of Georgia, has proposed GLYDE (**Glycan data exchange format**) as an XML-based representation format to foster interoperability and exchange of glycomics data.

Version 1.1 of GLYDE, which is based on a tree-like representation of glycan structures (Figure 3.11), using IUPAC-like naming of monosaccharide residues, was published in 2005 [11].

As a result of an international collaboration, GLYDE-II [27] (Figure 3.12), the successor of GLYDE, based on a connection table approach, has been suggested, to overcome the limitations of GLYDE. GLYDE-II uses a controlled vocabulary for glycans based on the conventions of GlycoCT (discussed in Section 3.4.2). GLYDE-II is capable of encoding incompletely or ambiguously defined oligosaccharides, and can handle repeating units. An atom replacement formalism models the atomistic details of the bonds between the different entities.

3.2.8.2 CabosML. As a part of the Japanese Glycogene Project carried out at the National Institute of Advanced Industrial Science and Technology (AIST), a bioinformatics system for the comprehensive identification and *in silico* cloning of human genes which encode for carbohydrate-active enzymes was created. To cope with the complexity of glycan

```

a-D-Manp-(1-6)+
  |
  b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc
  |
a-D-Manp-(1-3)+

```

```

GLYDE
<Glycan>
<residue>
  <residue anomeric_carbon="1" anomer="b" chirality="D" monosaccharide="GlcNAc" ringform="p" >
    <residue link="4" anomeric_carbon="1" anomer="b" chirality="D" monosaccharide="GlcNAc" ringform="p" >
      <residue link="4" anomeric_carbon="1" anomer="b" chirality="D" monosaccharide="Man" ringform="p" >
        <residue link="3" anomeric_carbon="1" anomer="a" chirality="D" monosaccharide="Man" ringform="p" >
          </residue>
        <residue link="6" anomeric_carbon="1" anomer="a" chirality="D" monosaccharide="Man" ringform="p" >
          </residue>
        </residue>
      </residue>
    </residue>
  </residue>
</Glycan>

```

Figure 3.11 GLYDE in its first published version. Tree-like oligosaccharides can be encoded using a parent-child hierarchy of XML elements.

```

a-D-Manp-(1-6)+
  |
  b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc
  |
a-D-Manp-(1-3)+

```

```

<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE GlydeII SYSTEM "GLYDE-II_v2.14.DTD"[
  <!ENTITY mDBget "ref=http://www.MonoSaccharideDB.org/GLYDE-II.jsp?G">
]>
<GlydeII xmlns:glydeII="http://glycomics.ccrcc.uga.edu/GLYDE-II_v2.14">
  <structure type="molecule" id="molecule_1" name="pentaglycoside">
    <part type="residue" subtype=substituent partid="1" ref="&mDBget;=nac"/>
    <part type="residue" subtype=base_type partid="3" ref="&mDBget;=b-dglc-hex-1:5"/>
    <part type="residue" subtype=base_type partid="4" ref="&mDBget;=b-dglc-hex-1:5"/>
    <part type="residue" subtype=base_type partid="5" ref="&mDBget;=b-dman-hex-1:5"/>
    <part type="residue" subtype=base_type partid="6" ref="&mDBget;=a-dman-hex-1:5"/>
    <part type="residue" subtype=base_type partid="7" ref="&mDBget;=a-dman-hex-1:5"/>
    <link from="1" to="3">
      <link from="N1" to="C2" from_replaces="O2" bond_order="1"/>
    </link>
    <link from="2" to="4">
      <link from="N1" to="C2" from_replaces="O2" bond_order="1"/>
    </link>
    <link from="4" to="3">
      <link from="C1" to="O4" to_replaces="O1" bond_order="1"/>
    </link>
    <link from="5" to="4">
      <link from="C1" to="O4" to_replaces="O1" bond_order="1"/>
    </link>
    <link from="6" to="5">
      <link from="C1" to="O3" to_replaces="O1" bond_order="1"/>
    </link>
    <link from="7" to="5">
      <link from="C1" to="O6" to_replaces="O1" bond_order="1"/>
    </link>
  </structure>
</GlydeII>

```

Figure 3.12 GLYDE-II encoding of the *N*-glycan core. GLYDE-II is based on a connection table (CT) formalism using XML syntax. Residues as defined by the MonoSaccharideDB (see Section 3.4.1) are either base-type or substituents. All links between residues have to be exactly described, indicating the atoms which are connected and those which will be replaced in the creation of the linkage. For example, the monosaccharide *N*-acetyl-D-glucosamine is encoded with two residues: its stem type *D-Glcp*, (*dglc-hex-1:5*), and the substituent, *N*-acetylamino (*nac*). A bond is formed between the N of the *N*-acetylamino group and C2 of *D-Glcp* by replacing the O2 of the stem residue (atom replacement formalism).



Figure 3.13 CabosML encoding of the N-glycan core. Topologies are reflected by the parent–child relationships of the element *g:MS*.

structures, a **carbohydrate sequence markup language** (CabosML), an XML description of carbohydrate structures, was developed [12] (Figure 3.13). This encoding is also used in their non-public structural database [12]. The format is capable of encoding cyclic and repeating structures.

3.3 Other Descriptors and Tools

3.3.1 Symbolic Representations

Symbolic representations of complex glycans consist of a series of geometric symbols, one for each type of monosaccharide, which are connected by a direct line to indicate the glycosidic linkage. They are intensively used in many biomedically oriented publications. Recently, a standardized set of symbols to represent monosaccharide residues (see the inside cover of this book) was proposed by the Consortium for Functional Glycomics (CFG) [28]. It was adapted, and modified, from the symbol set used in the book *Essentials of Glycobiology* [29, 30]. Another set of symbols, and a linkage representation, have been proposed by the Oxford Glycobiology Institute [31] (Figure 3.14a). Two different ways of depicting the anomeric configuration and the glycosidic linkage are in use:

1. Using characters added to the linkage line to indicate the stereochemistry and connection points of a glycosidic linkage.
2. Using different line types (full lines or dashed lines) to indicate the anomeric configuration, and associating the angle of the line linking to the adjacent residue with the linkage position. The same angle always represents the same linkage position. In this way, the linkage position is clearly shown without the need to add additional numbers.

The CFG system is probably the most prominent example of the first approach. Here, the anomeric configuration is designated by an α or β symbol, and a number indicates the position of the carbon on the residue that is receiving the glycosidic bond (Figure 3.14b).

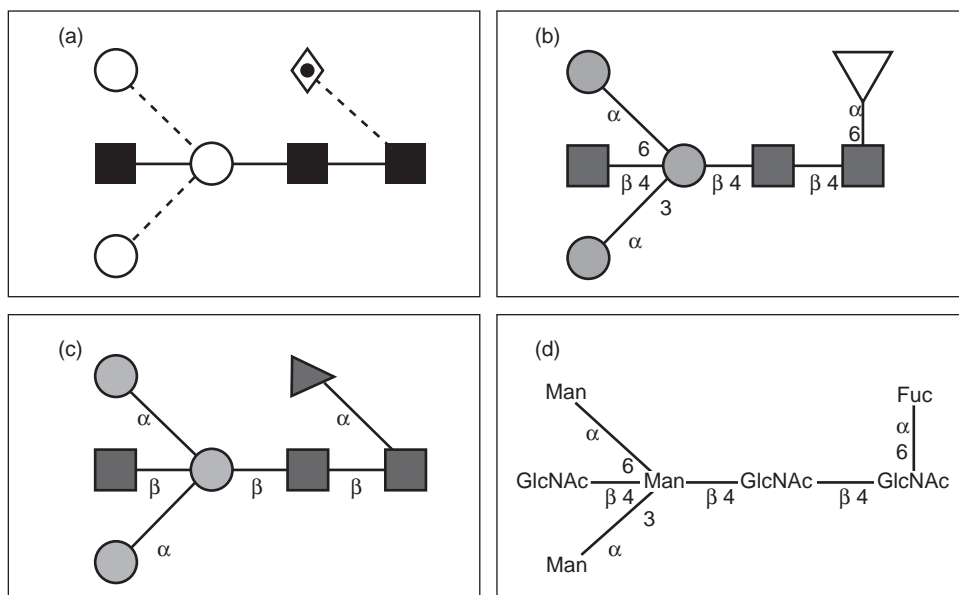


Figure 3.14 Representation of an oligosaccharide structure using symbols. For the meaning of the monosaccharide residue symbols, see the inside cover of this book. (a) Oxford notation. (b) CFG black-and-white symbolic notation. (c) CFG symbolic notation with angle-specific indications of linkages as used in the Oxford notation. (d) ASCII 2D plot. The IUPAC three-letter codes are used to describe the monosaccharides. All images in this figure were created using *GlycanBuilder* [33, 34] (see Chapter 13).

The second approach to representing anomeric configuration and the linkage (Figure 3.14a), which was initially suggested by Kamerling and Vliegthart in 1992 [32], has been adopted by the Oxford Glycobiology Institute.

3.3.2 Encoding Monosaccharides Using Small Molecule Descriptors

Since the building blocks of oligosaccharides are small molecules with typically 5–11 C-atoms, in principle the molecular descriptors in use for small organic molecules could also be used for the encoding of complex carbohydrates. A prerequisite is, of course, that the format allows an encoding of the stereochemistry. Such encoding schemes have been used to represent carbohydrates in small-molecule libraries. However, none of the existing glyco-related databases uses an atomistic encoding of monosaccharides. Equally important is a consistent handling of the non-carbohydrate portions of the oligosaccharides, for which simple text identifiers are in use.

3.3.2.1 SMILES. SMILES is a general-purpose chemical nomenclature and data exchange format, for which different variations exist [35] (Figure 3.15). The variants include stereospecific (isomeric) and unique SMILES. The cheminformatics community uses this format extensively and most of the structurally related chemical databases are capable of generating this format as an output, for example, PubChem and ChEBI.



Figure 3.15 Encodings for small molecules can be used especially on the aglycone level in database applications.

3.3.2.2 InChI. A more recent development is the IUPAC NIST Chemical Identifier (InChI) [36, 37] (Figure 3.15). It is a digital equivalent of the IUPAC name, containing composition, connectivity, charge, stereochemistry, isotopic, and tautomerism layers, which are separated by a slash, “/”. Its primary aim is to provide a unique identifier for chemical compounds. This identifier is also broadly used in databases for small molecules.

3.3.2.3 Stereocodes. A direct numerical representation of the stereocenters of the monosaccharide core has been utilized, for example, to detect carbohydrates in data of the Protein Data Bank (PDB) [38, 39]. Generally, each carbon atom of the main chain is assigned a number which reflects its stereochemical status (Figure 3.16). The stereocode encodes the orientation of the hydroxyl group [or the substituent(s) by which it is replaced] in relation to the plane spanned by the carbon to which it is attached and the two neighboring

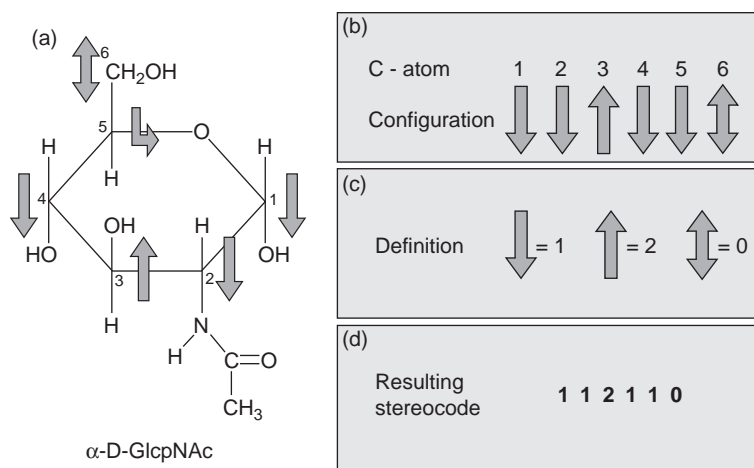


Figure 3.16 Numerical representation of the stereocenters of a monosaccharide. (a) *N*-Acetyl- α -D-glucosamine with stereovectors. The stereovectors indicate the orientation of the hydroxyl groups/substituents, so at C5 the orientation of the ring oxygen and not that of the CH₂OH group is taken into account. (b), (c) The position of the stereogenerating non-hydrogen substituents is used to determine a configurational numerical code. (d) The resulting stereocode can be used to identify uniquely monosaccharide core structures.

carbons. This approach facilitates computational handling of monosaccharides and maintenance of databases.

3.4 Ongoing and Future Developments

The current situation in glycobioinformatics shows clearly a lack of standardization on both monosaccharide and sequence levels, as several “island” databases with different encodings coexist, which are only weakly interconnected. Furthermore, no generally accepted sequence exchange format exists, and none of the existing sequence formats is truly able to handle all relevant experimentally derived structural evidence in a consistent manner. Unpublished analysis has shown that all existing databases in the glycobioinformatics area contain errors on the monosaccharide level, as the sugar names are typically treated as free-text entries, which easily leads to errors even by careful and skilled curators. A short overview of two promising solutions to these challenges is given in this section.

3.4.1 MonosaccharideDB

The MonosaccharideDB, which is currently under development within the EuroCarbDB project [40], is intended to be a unified resource to validate monosaccharide names. Its primary target is the generation of a unique identifier for each monosaccharide, along with two- and three-dimensional standard representations, stereocodes, and different textual

MonoSaccharideDB

home

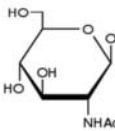
notation

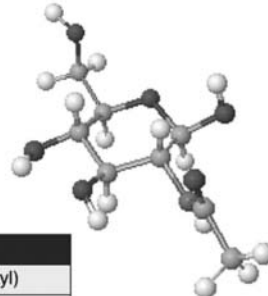
query

* exact search
* fuzzy search

Monosaccharide: β -D-GlcpNAc

Properties:	
ID:	1
Name:	β -D-GlcpNAc
BaseType:	GLC (Glucose)
RootType:	GLC
Size:	6
Anomeric:	b
Anomeric Carbon:	C1
Abs. Config:	D
Ring Type:	p
Stereocode:	212110





Modifications:	
Name:	NAc (N-Acetyl)
Position:	2

Figure 3.17 MonosaccharideDB prototype web frontend – entry for β -D-GlcpNAc. The current implementation of the MonosaccharideDB is a prototype, which handles CarbBank nomenclature and performs normalization on this encoding scheme. The long-term goal of this project is a freely accessible standardizing and translating web service and dictionary for the glycobiology community.

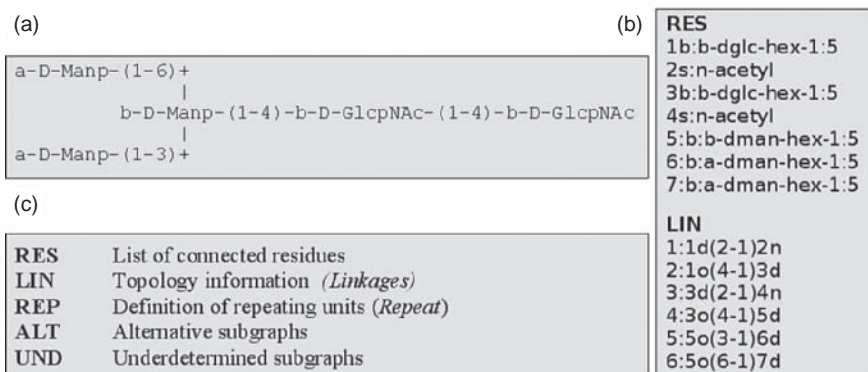


Figure 3.18 Encoding of the *N*-glycan core (a), in the GlycoCT format (b). The RES section contains the monosaccharide units, the vertices of the carbohydrate sequence graph, the LIN section enumerates the linkages, the connecting edges of the sequence graph. An atom replacement formalism on the linkage level [e.g. in 1:1d(2-1)2n, d = deoxy] reflects the changes in the monosaccharides due to the linking process. The format contains several distinct sections (c).

synonyms currently in use (Figure 3.17). A mid-term goal of this project is a name parsing and generating routine for further notation schemes such as BCSDB, KCF, or GLYDE.

3.4.2 GlycoCT

The EuroCarbDB project is a design study to implement a distributed database of primary experimental data and glycan sequences. For this project, a new sequence format has been suggested, which aims to provide a consistent solution for carbohydrate sequences. The general concept of the sequence format is a connection table approach (Figure 3.18). A systematic machine-readable encoding scheme for monosaccharides is part of its definition (Figure 3.19). The full documentation of this format is available online [10].

3.5 Summary and Outlook

As demonstrated in this chapter, the current situation of digital representations for carbohydrates is characterized by the existence of various encoding schemata, which have been developed by different consortia, institutions, or local groups. Most of the representations developed so far are based on a tree-like linear encoding of glycan structures. However, more recently, several groups have independently proposed to describe the connectivity of complex carbohydrates using a connection table approach. This change in conception reflects, on the one hand, the fact that the encoding of branched carbohydrates is more similar to the digital description of small organic molecules, where connection table-based descriptions are commonly used. On the other hand, this change reflects the demand to develop descriptions which can reliably encode all types of existing carbohydrates, including non-stoichiometric substitutions, probabilistic attachment of terminal residues, and specific topologies such as cyclic structures and complex repeating of basic units.

In many cases, the currently existing representations have been designed to fulfill mainly the specific needs of a particular scientific purpose. The encoding schemata have not

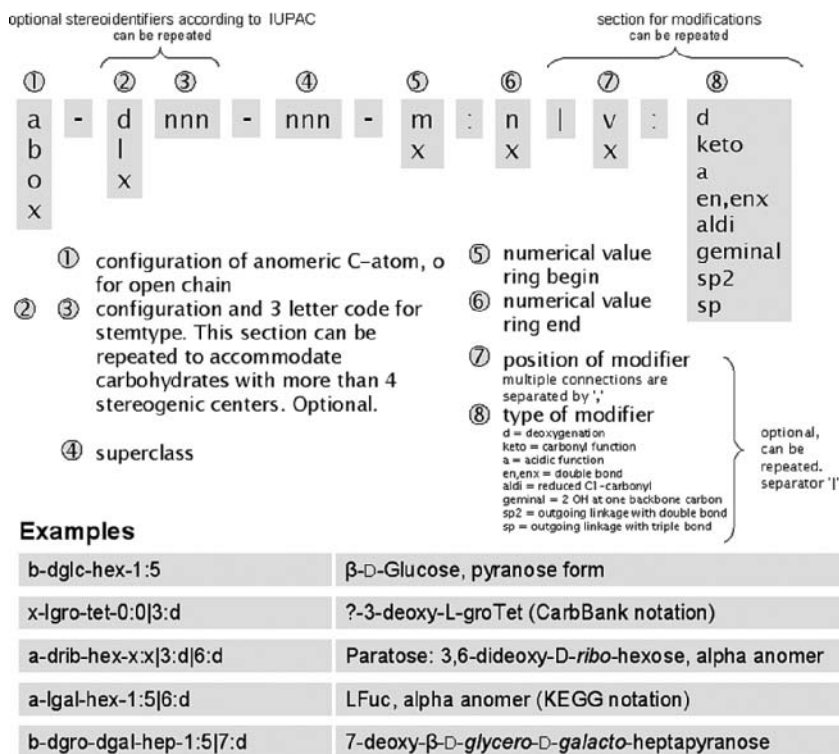


Figure 3.19 The monosaccharide notation of GlycoCT is both machine-readable and human-interpretable.

necessarily been conceived to provide communication mechanisms. Nevertheless, approaches to link distributed data have been conceptually worked out and implemented, for example, a link between the GLYCOSCIENCES.de portal and the Bacterial Carbohydrate Structure Database (BCSDB) [20]. In addition, translators are being developed for the different encoding schemes and sequence formats, including a “universal translator” [41], to allow cross-referencing and data integration.

Prerequisites for successful and efficient communication between distributed data resources are standard representations of both the glycan structures identified and the associated experimental data. It can be regarded as essential progress in the development of glycan exchange formats that in 2006 agreement was reached between the major carbohydrate database initiatives in the United States, Germany, and Japan to use GLYDE-II as a common structural data exchange format. This agreement signifies the urgent need to establish generally accepted XML-based exchange formats for glycan structures. The controlled vocabulary for monosaccharides and the ability to describe all types of topologies, defined by GlycoCT, is an integral part of GLYDE-II.

Currently, only a limited amount of software related to glycomics is freely available to be shared by various projects. The largest bottleneck up to now has been the lack of a common language to exchange glycan structures and related data. With the agreement to accept GLYDE-II as the central format to exchange structural data, a central prerequisite

for the development of glyco-related software for glycomics has been achieved. It is of paramount importance that software implementations using the GLYDE-II format will become publicly available in the near future.

References

1. McNaught AD: Nomenclature of Carbohydrates (Recommendations 1996). *Carbohydr Res* 1997, **297**:1–92.
2. Doubet S, Bock K, Smith D, *et al.*: The Complex Carbohydrate Structure Database. *Trends Biochem Sci* 1989, **14**:475–477.
3. Doubet S, Albersheim P: CarbBank. *Glycobiology* 1992, **2**:505.
4. Bohne-Lang A, Lang E, Forster T, von der Lieth C-W: LINUCS: LInear Notation for Unique description of Carbohydrate Sequences. *Carbohydr Res* 2001, **336**:1–11.
5. Cooper CA, Harrison MJ, Wilkins MR, Packer NH: GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res* 2001, **29**:332–335.
6. Cooper CA, Joshi HJ, Harrison MJ, *et al.*: GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res* 2003, **31**:511–513.
7. Banin E, Neuberger Y, Altshuler Y, *et al.*: A novel LinearCode[®] nomenclature for complex carbohydrates. *Trends Glycosci Glycotechnol* 2002, **14**:127–137.
8. Toukach FV, Knirel YA: New database of bacterial carbohydrate structures. *Glycoconj J* 2005, **22**:216–217.
9. Aoki KF, Yamaguchi A, Ueda N, *et al.*: KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res* 2004, **32**:W267–W272.
10. Encoding glycan structures and their biological occurrence: Glyco-CT format; MonosaccharideDB (EUROCarbDB): <http://www.eurocarbdb.org/recommendations/encoding/>.
11. Sahoo SS, Thomas C, Sheth A, *et al.*: GLYDE – an expressive XML standard for the representation of glycan structure. *Carbohydr Res* 2005, **340**:2802–2807.
12. Kikuchi N, Kameyama A, Nakaya S, *et al.*: The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics* 2005, **21**:1717–1718.
13. SUGABASE: <http://boc.chem.uu.nl/sugabase/sugabase.html>.
14. Lütteke T, Bohne-Lang A, Loss A, *et al.*: GLYCOCIENCES.de: an internet portal to support glycomics and glycobiology research. *Glycobiology* 2006, **16**:71R–81R.
15. GLYCOSCIENCES.de portal: <http://www.glycosciences.de>.
16. Loss A, Bunsmann P, Bohne A, *et al.*: SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res* 2002, **30**:405–408.
17. van Kuik JA, Hard K, Vliegthart JFG: A ¹H NMR database computer program for the analysis of the primary structure of complex carbohydrates. *Carbohydr Res* 1992, **235**:53–68.
18. van Kuik JA, Vliegthart JF: Databases of complex carbohydrates. *Trends Biotechnol* 1992, **10**:182–185.
19. Bacterial Carbohydrate Structure DataBase (BCSDB): <http://www.glyco.ac.ru/bcsdb/start.shtml>.
20. Toukach P, Joshi HJ, Ranzinger R, *et al.*: Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the Bacterial Carbohydrate Structure DataBase and GLY-COSCIENCES.de. *Nucleic Acids Res* 2007, **35**:D280–D286.
21. Cooper CA, Wilkins MR, Williams KL, Packer NH: – a biological O-linked glycan database. *Electrophoresis* 1999, **20**:3589–3598.
22. GlycoSuite (Proteome Systems): <http://www.glycosuite.com>.
23. Glycominds on World Wide Web URL: <http://www.glycominds.com>.

24. Glycomics database and LinearCode[®] (Glycominds) (link to Syntax at bottom of page): <http://www.glycominds.com/index.asp?menu=Research&page=glycoit#>.
25. Kanehisa M, Goto S, Hattori M, *et al.*: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006, **34**:D354–D357.
26. KegDraw (Java application for drawing compound and glycan structures): <http://www.genome.jp/download>.
27. GLYDE-II: <http://glycomics.ccruc.uga.edu/GLYDE-II>.
28. Symbol and text nomenclature for representation of glycan structure (Nomenclature Committee, Consortium for Functional Glycomics: <http://glycomics.scripps.edu/CFGnomenclature.pdf>).
29. Symbol nomenclature for glycans (from *Essentials of Glycobiology* [30]): <http://grtc.ucsd.edu/symbol.html>.
30. Varki A, Cummings R, Esko J, *et al.*: *Essentials of Glycobiology*, 1st edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1999.
31. Royle L, Dwek RA, Rudd PM: Determining the structure of oligosaccharides N- and O-linked to glycoproteins. In *Current Protocols in Protein Science* (eds J Coligan, B Dunn, D Speicher, P Wingfield). New York: John Wiley & Sons, Inc., 2006, Suppl 43, pp.12.16.11–12.16.45.
32. Kamerling JP, Vliegenthart JFG: High-resolution ¹H-nuclear magnetic resonance spectroscopy of oligosaccharide-alditols released from mucin-type O-glycoproteins. In *Carbohydrates and Nucleic acids (Biological Magnetic Resonance)* (ed. L Berliner, J Reuben). New York: Plenum Press, 1992 (*Biol. Magn. Reson.* Vol **10**), pp. 1–194.
33. Ceroni A, Dell A, Haslam SM: The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* 2007, **2**:3.
34. Tools to normalize and convert glycan structures. GlycanBuilder: a rapid and intuitive glycan structure editor (EUROCarbDB): <http://www.eurocarbdb.org/applications/structure-tools>.
35. Weininger D: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988, **28**:31–36.
36. Stein SE, Heller SR, Tchekhovskoi D: An open standard for chemical structure representation: the IUPAC Chemical Identifier. In *International Chemical Information Conference; Nimes: 2003*: 131–143.
37. The IUPAC International Chemical Identifier (InChITM): <http://www.iupac.org/inchi>.
38. Lütteke T, von der Lieth CW: pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics* 2004, **5**:69.
39. Lütteke T, Frank M, von der Lieth CW: Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr Res* 2004, **339**:1015–1020.
40. EUROCarbDB: <http://www.eurocarbdb.org>.
41. GlycomeDB: <http://www.glycome-db.org>.

Evolutionary Considerations in Studying the Sialome: Sialic Acids and the Host–Pathogen Interface

Amanda L. Lewis and Ajit Varki

Glycobiology Research and Training Center, Departments of Medicine, Biological Sciences and Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA 92093-0687, USA

4.1 Introduction and Summary

The apparent variation in cell surface oligosaccharide (glycan) structures within and between species has long been an interesting and yet puzzling aspect of glycobiology. It has been suggested that this diversity reflects the often-conflicting pressures of evading pathogens, while simultaneously maintaining endogenous functions [1–4]. Since most pathogens replicate much faster than their hosts, they can rapidly evolve different ways to target or mimic structures that are critical for host processes – a feature called the “Red Queen” effect¹, that may be especially relevant to glycans [2–5]. Two categories of pathogen molecules are relevant to protein–glycan interactions at the host–pathogen interface: (1) the pathogen receptors/toxins that recognize and *bind to* host glycans, and (2) pathogen surface molecules that *mimic* host glycans. A great number and variety of pathogens use host glycan structures for targeted adherence, invasion, or cytotoxicity. In fact, the vast majority of known glycan-binding lectins are those used by pathogens [6, 7]. Host population heterogeneity of the targeted glycan structures may also ensure survival of some individuals and reduce the chance of epidemic spread, a phenomenon referred to in other contexts as “herd immunity” [2, 8, 9]. The extent of population heterogeneity (of a targeted glycan structure) is constrained by the existence and stringency of internal host functions in which the same structure participates. In a similar yet opposing pathogenic mechanism, many microbes decorate themselves with glycans that are similar to or identical with structures expressed in the host (i.e. “molecular mimicry”) [10, 11]. Microbial hijacking of host lectins via molecular mimicry is a virulence mechanism likely broader in scope than currently recognized.

Despite frequent pathogen targeting and mimicry of carbohydrate structures, *all* cells in nature are covered with a dense coating of glycans [12]. This remarkable “rule”

¹ The “Red Queen” effect in evolutionary processes is based on the observation to Alice by the Red Queen in Lewis Carroll’s *Through the Looking Glass* – that “it takes all the running you can do, to keep in the same place.”

suggests that glycans afford the most flexible way to adapt a cell surface away from pathogen recognition, and/or that certain cell surface glycans are required for non-dispensable host functions. Broadening our understanding of host–pathogen co-evolution requires further study of glycans and lectins on *both* sides of the host–pathogen equation. This chapter focuses mainly on the sialic acids (Sias), providing examples of ongoing studies in the area of Sia-dependent host–pathogen interactions. To place these interactions in a broader context, we consider the diversity, distribution, biosynthesis, and evolution of Sias. We also review common glycan analysis techniques that can result in loss of Sias or Sia modifications and finally suggest a “Sialome”² project for archiving information about Sias in nature.

4.2 Pathogens that Target or Mimic Host Sialic Acids

Sialic acids are a diverse family of sugars that often occupy the non-reducing outermost ends of glycan chains in many animals [13–17]. Due in part to this terminal location, Sias are the glycan receptors most frequently targeted for recognition by pathogens. There are numerous documented examples of pathogens that use their Sia-binding proteins (called agglutinins in viruses, adhesins in bacteria, and lectins in protozoa and fungi) to adhere to, or gain entry into, host cells. There are also a number of bacterial toxins that bind Sias to effect their toxicity on target cells. Figure 4.1 provides an incomplete listing of some common pathogens that target Sia residues [see ref 17 for a listing of ~100 Sia-binding pathogens/toxins in nature].

Many common and sometimes fatal illnesses are caused by Sia-binding pathogens. For example, *Plasmodium falciparum* is a common agent of malaria, the leading cause of illness (300–500 million per year) and death (>1 million per year) worldwide (World Health Organization estimates). *Plasmodium falciparum* merozoites typically infect red blood cells in a Sia-dependent manner, leading to fever, chills, and flu-like symptoms, and if not treated, kidney failure, seizures, coma, and death. Influenza virus A (the causative agent of the “flu”) is also a Sia-binding pathogen responsible for >100 000 hospitalizations, >30 000 deaths, and ~\$15 billion costs in the United States each year. Yet another is *Helicobacter pylori*, commonly involved in the formation of gastric and duodenal ulcers, a condition experienced by as many as 5 million persons at any one time in the United States (estimate from Digestive Diseases in the United States: Epidemiology and Impact, NIH Publication No. 94-1447, 1994).

Among the Sia-expressing pathogens, *Escherichia coli* (K1 and K92), *Neisseria meningitidis* (Groups B, C, Y, and W135), and Group B Streptococci all express capsular Sias and cause sepsis and meningitis, particularly in young children [10]. Other important Sia-expressing pathogens include *Haemophilus influenzae*, the most common cause of childhood ear infections [18], and some strains of *Escherichia coli* that cause gastrointestinal and urinary tract infections [10]. Indeed, the list of Sia-binding and Sia-expressing pathogens continues to expand, and recently available genomic data suggest that this may be even more common than previously realized [19].

² The term “Sialome” was recently coined [5] to denote “the total complement of Sia types and linkages and their modes of presentation on a particular organelle, cell, tissue, organ or organism – as found at a particular time and under specific conditions.”

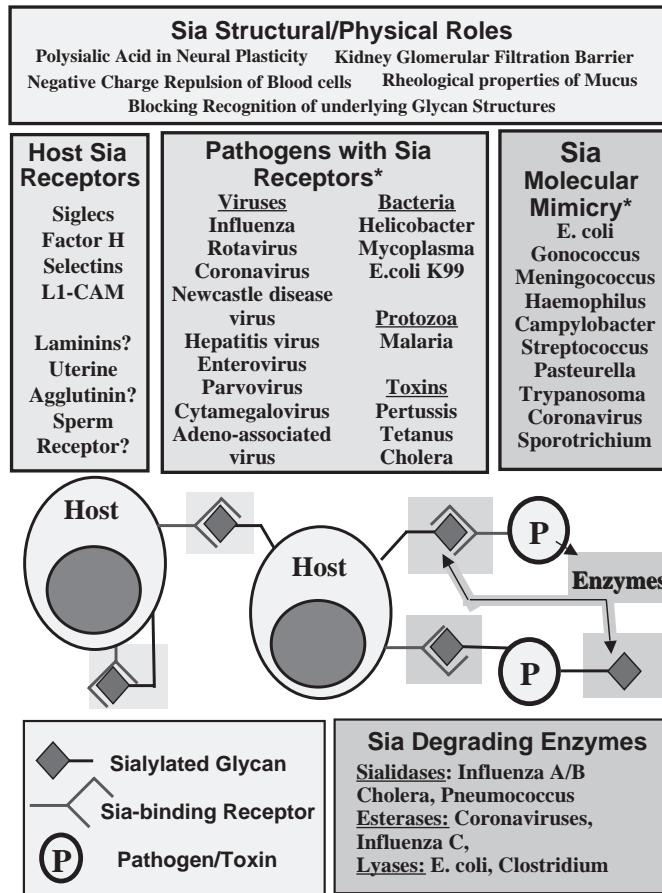


Figure 4.1 Biological roles of sialic acids. The diverse biological roles of Sias include structural/physical functions, or functions that require Sia-recognizing proteins. In the latter category, intrinsic (host) Sia-recognizing proteins are typically involved in endogenous functions, while extrinsic (pathogen) Sia-recognizing proteins mediate mechanisms of host cell adherence or entry. Microbial pathogens can also decorate themselves with Sia to hijack host processes. In addition, some microbes express enzymes that degrade Sias. See the text and the cited literature for details about the biological functions of Sias depicted in this figure. Note: most pathogens express Sias or Sia-binding proteins in a strain-specific manner; hence these properties do not apply to all members of the indicated pathogen classes. The asterisks indicate that, this figure represents an incomplete listing of most of the categories above. See [17] for details. A full-color version of this figure is included in the Plate section of this book.

In the light of the importance of Sias in host–pathogen interactions, the biology and evolution of this family of sugars are especially relevant. Why have so many pathogens come to rely on Sia binding or mimicry for colonization and/or invasion of host tissues? Has the host immune system adapted to meet these widespread pathogenic strategies? Can we use a bioinformatic approach to studying the sialome, and to orchestrate scientific research better in this important area of public health?

4.3 Diversity and Biology of Sialic Acids in Nature

4.3.1 Biological Importance of Sialic Acids

N-Acetylneuraminic acid (Neu5Ac) (Figure 4.2) is the most common Sia in mammals, and serves as a “core” structure that can be modified and presented as per the enzymatic repertoire of a particular cell (see Section 4.3.2). In mammals, the physical properties of Sias are known to be involved in processes such as nervous system plasticity and learning,

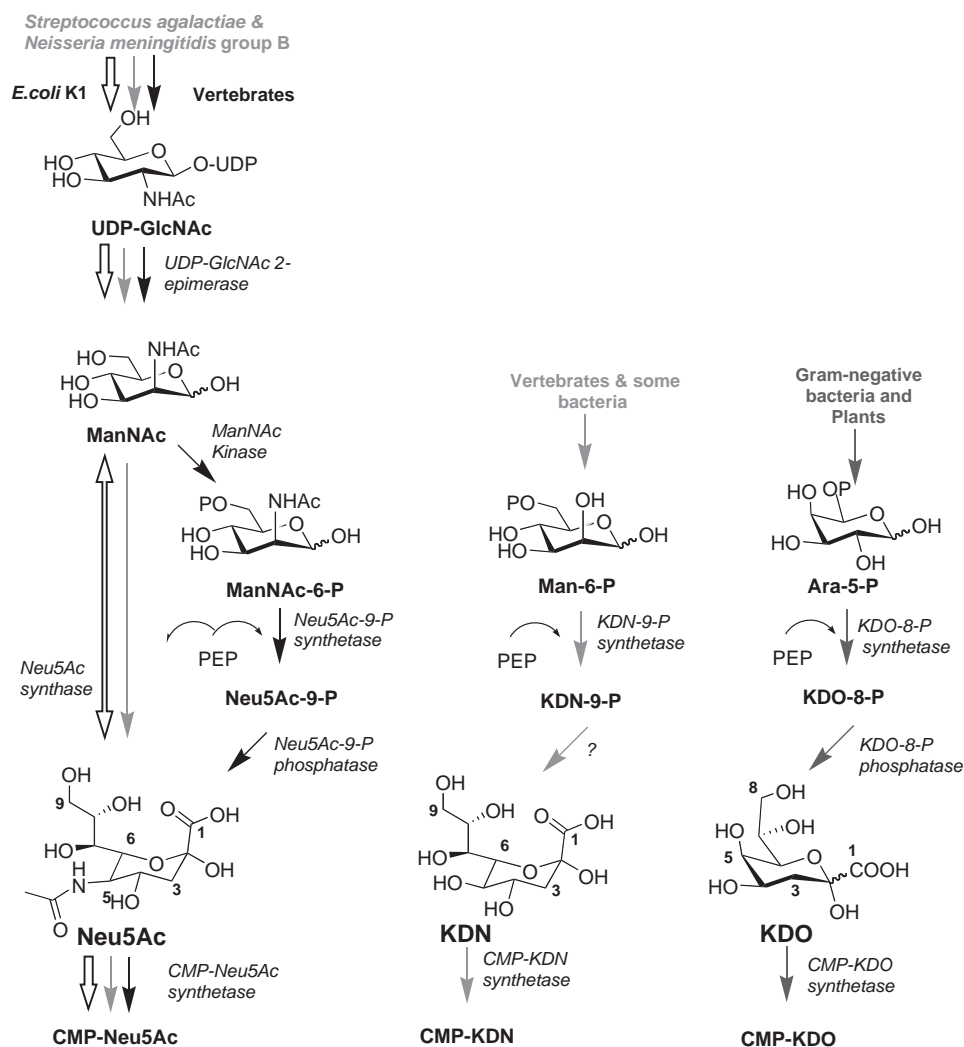


Figure 4.2 Biosynthetic pathways of sialic acids and sialic acid-like molecules. Shown on the left is the Neu5Ac biosynthesis pathway in vertebrates, which differs slightly from the pathways employed by bacteria. Biosynthetic pathways for Kdn and Kdo are shown for comparison. Refer to text for known evolutionary relationships.

kidney glomerular filtration, and repulsion between circulating blood cells [20–22]. Well-established immunological functions mediated by Sia-binding proteins include leukocyte trafficking by the selectins [23] and regulation of the alternative complement pathway activation by Factor H [24]. The Siglecs are a more recently discovered family of Sia-binding proteins that appear to mediate regulatory and/or phagocytic immune functions. [5, 12, 25]. Sialic acids are crucial for mammalian development, as evidenced by embryonic lethality of mice that lack a key biosynthetic enzyme [26]. Sialyltransferases can be regulated in a tissue-specific manner [27], and by stress [28, 29], infection [30], or malignancy [31]. Mammalian sialidases (enzymes that remove Sias) are also carefully regulated in normal and malignant mammalian cells [32, 33]. Sias are found in diverse structural contexts in Nature. This diversity exists on multiple levels, including modifications of core Sia structures, the linkage of Sia to an underlying sugar, the identity and arrangement of underlying sugars, structural attributes of the glycosylated molecule, and the higher level cellular and organismal milieu. Changes in Sia structural context can regulate Sia-dependent events by altering or preventing recognition by Sia-binding lectins and sialidases. The following sections briefly describe Sia diversity and distribution as they relate to Sia modifications/types and their representation among different species.

4.3.2 Sialic Acid Diversity in Nature

Sialic acids comprise a naturally occurring family of over 50 structures, many of unknown biological importance [16, 17]. *N*-Acetylneuraminic acid (Neu5Ac) and *N*-glycolylneuraminic acid (Neu5Gc) are the most common Sias in mammals, and serve as “core” structures that can be extensively modified. The other core Sia is Kdn (3-keto-2-deoxy-*D*-manno-nonulosonic acid). Although Kdo (2-keto-3-deoxy-*D*-manno-octulosonic acid) is not defined as a sialic acid, we have recently emphasized [17] that it is closely related both in structure and biosynthetic mechanism to Neu5Ac (discussed further in Section 4.6). CMP-Neu5Gc is derived from CMP-Neu5Ac via the enzymatic addition of an oxygen atom to the *N*-acetyl group [34]. Humans do not express Neu5Gc due to an inactivating mutation in the hydroxylase gene that mediates this step [35]. Neuraminic acid (Neu) is assumed (although not proven) to be derived from glycosidically-bound Neu5Ac by deacetylation of the *N*-acetyl group [36, 37]. Enzymatic alteration of Neu5Ac, Neu5Gc, Kdn or Neu can occur at various carbon positions [e.g. *O*-methylation (at C8), or esterification with acetyl (C4, C7, C8, or C9), lactyl (C9), sulfate (C8), or phosphate (C9) groups]. Further variability is derived by lactonization or lactamization of Sia structures [see [16, 17] for attempts at a comprehensive listing of chemical names, abbreviations, and reference publications of known Sia structures].

Sialic acid *O*-acetylation is a modification described in many biological contexts. While the biosynthetic mechanisms are still being elucidated, Sia *O*-acetylation is implicated in processes such as apoptotic regulation, colonic development and cancer, alternative complement pathway regulation, bacterial polysaccharide immunogenicity, and visceral leishmaniasis [38–43]. A mechanistic understanding of these and other *O*-acetylation-dependent processes has been hampered by the many failed attempts to clone the mammalian enzyme(s) responsible for this modification. Even in bacteria, definitive biochemical and genetic identification of Sia *O*-acetyltransferases was achieved only recently [19, 44]. Unfortunately for those trying to clone the mammalian enzymes, vertebrate genomes do not appear to

have any obvious homologs of these bacterial Sia *O*-acetyltransferases. Pre-existing work [16, 17] and unpublished data from our laboratory indicate extensive inter-species diversity in the expression profiles of vertebrate Sia *O*-acetylation. Thus, the biological roles of this modification are probably species specific in some instances.

4.3.3 Species Distribution of Sialic Acids

Understanding the distribution of Sias in nature requires two types of descriptions for different organisms, tissues or cells: (a) the presence of “core” Sia biosynthetic ability and (b) the presence of Sia-modifying enzymes. Echinoderms such as sea urchins and starfish were once thought to express the broadest diversity of Sias (i.e. the most Sia-modifying enzymes), with humans being the simplest [13]. With the advent of more sensitive techniques, it is now known that even human cells express a large variety of modified Sias, albeit in smaller quantities [45, 46]. Warren’s pioneering work in developing and using the thiobarbituric acid assay to detect and quantify Sias suggested that the presence of “core” Sia biosynthetic ability was limited to certain species [47]. Specifically, Warren detected Sias in the deuterostome lineage of animals (vertebrates, and certain “higher” invertebrates such as echinoderms that literally have “two mouths” during development), but not in the protostomes (insects, mollusks, etc.). With the advent of very sensitive detection techniques and the ability to search genomes for genes involved in Sia synthesis, we now know that Sias have a much broader distribution than was originally thought [17]. They are expressed in octopus and squid (mollusks) [48], some insects (arthropods) [49, 50], and possibly even in some plants [51] – although the last finding has not been replicated by others [52, 53]. What is clear is that Sias are not ubiquitous among living organisms and certainly do not have the same biological roles in all contexts. This is particularly evident in the case of Sia-expressing pathogens, which apparently use this decoration to “hijack” a ride through host tissues.

4.4 Host Mechanisms for Evading Glycan-binding Pathogens

4.4.1 Is Sialic Acid Diversity a Reflection of Host Defense Against Past Invaders?

When pathogens exploit critical structures such as Sias, there are serious difficulties for the host. Does the structural diversity of Sia presentation reflect past host attempts at evading pathogens that bind to these structures? Consider the case of the Influenza virus hemagglutinin, which interacts with host cells in a Sia-dependent manner [54, 55]. The context of Sias varies widely between different animals, and dictates the specificity of Influenza virus binding. The host range of a particular Influenza virus is partly determined by Sia linkages (α 2–3 or α 2–6) and also by modifications (Neu5Ac, Neu5Gc, Neu5,9Ac₂) recognized by the virus hemagglutinin [55]. The critical importance of specific Sia recognition for viral evolution suggests that mammalian Sia structures have diverged (at least in part) to avoid host infection by Influenza viruses, harbored by other animals living in close proximity.

Another potential example of host evasion from pathogens by altering Sia expression involves Neu5Gc, a common Sia in nature that is not expressed by humans (see section on 4.3.2). Pathogens such as *E. coli* K99 can cause potentially lethal diarrheal infections by targeting host intestinal Neu5Gc residues [56]. One possibility in human evolution is that

a Neu5Gc-binding pathogen such as K99 eliminated all but a few individuals (who did not express Neu5Gc) in an early hominid population. Unfortunately, as in many such examples, it is nearly impossible to prove that a host species altered Sia linkages or modifications in order to escape viral or bacterial infection or limit cross-species transmission. Nevertheless, such discussion arouses interest about the effects of Sia diversity on pathogenic tropism for various species. Indeed, this tropism affects the “fitness” of both pathogens and hosts, and is broadly relevant to public health.

4.4.2 Mucins and Milk Oligosaccharides Can Prevent Infections

Mammals have evolved mucosal defense mechanisms against glycan-binding pathogens. Mucins and free oligosaccharides (OS) produced by mucosal tissues have long been hypothesized to have antibacterial properties. All mucosal surfaces exposed to microbes secrete copious amounts of mucus, the main component of which is a family of heterogeneously and heavily glycosylated mucin molecules [57]. By providing multivalent arrays of glycan binding sites for pathogens, these act as “decoys”, to prevent pathogens and toxins from reaching the cell surface, where they can initiate invasion. Such bound and trapped pathogens can also be eliminated mechanically, by expelling the mucus. Also, OS-based mucosal immunity transferred from mother to infant via breast milk appears to be important for the prevention of infant gastrointestinal infections. Human milk OS concentrations are estimated at $\sim 10 \text{ g l}^{-1}$ [58] and mass spectrometric studies indicate >900 different OS species [59]. Compositional analyses of human breast milk and infant stool indicate that (uncharged) milk OS not only survive the infant digestive tract, but are actually present in higher concentrations than in mothers’ milk [60]. This finding suggests that (neutral) milk OS may not have a primarily nutritive role; rather, they are in the right place at the right time to inhibit binding of potential pathogens. It has been argued that the considerable maternal resources expended on milk OS production must have an important biological role [61]. Certainly, the benefits of breast-feeding for protection from infant infection are well documented, but how much of the effect can we attribute to oligosaccharides? Is the expression of milk glycans an evolutionary response to the common pathogen strategy of glycan-binding? In this regard, it is interesting that the most complex mixtures of milk OS are typically found in large-brained social animals such as humans and elephants, which have relatively immature young, needing prolonged care before they become independent [62].

Interestingly, most of the OS in human milk have terminal fucose and/or Sia residues, both of which are commonly used by various pathogens for targeted entry. Studies indicate that the levels of $\alpha 1-2$ -linked fucosylated OS in milk correlate inversely with the risk of diarrheal infection caused by fucose-binding pathogens such as *Campylobacter* species, *Vibrio cholerae*, *E. coli* stable toxin, and calciviruses to target cells [63]. Similarly to fucosylated milk OS, some studies suggest that sialylated milk OS and mucins may inhibit Sia-dependent interactions between mammalian cells and pathogens. For example, sialidase treatment of mucins dramatically reduced their inhibitory activity against *Helicobacter pylori* [64] and *Haemophilus influenzae* [65]. Another study describes a Sia-dependent resistance to adenoviral gene transfer mediated by the mucin MUC1 [66]. Indeed, mucins are upregulated during infectious challenges of both murine [67] and human cells [68]. Furthermore, binding and replication of rotavirus, the major cause of severe

dehydrating diarrhea in young children, is inhibited in a Sia-dependent manner by a particular mucin in human milk [69]. Sialylated gangliosides and OS in milk also inhibit binding of several *E. coli* strains (or enterotoxins) that cause diarrhea and urinary tract infections [70, 71]. *Vibrio cholerae* and *E. coli* Sia-binding toxins are also inhibited by a Sia-rich ganglioside fraction of human milk [71]. The prevalence of Sia-binding pathogens and the abundance of sialylated glycans in breast-milk and mucins suggest that these glycans may have an even broader protective role against Sia-binding pathogens than is currently recognized.

4.5 Pathogens Exchange Glycan Biosynthesis Genes, Allowing Molecular Mimicry and Hijacking of Host Lectins

4.5.1 The Diversity of Bacterial Polysaccharides

Bacterial surface glycans exhibit remarkable diversity and in many cases influence pathogenicity. Bacterial glycans come in the form of cell-wall peptidoglycans, polysaccharide capsules, glycoproteins, or glycolipids. Capsules (sometimes referred to as “K-antigens”) can be elaborated by both Gram-positive and Gram-negative bacteria and are comprised of single-monosaccharide homopolymers or repeating units of many monosaccharide types. In contrast, bacterial glycolipids are only found in the outer membranes of Gram-negative bacteria and are either referred to as lipopolysaccharides (LPS) or lipooligosaccharides (LOS). The core structure of bacterial glycolipids is often modified with outer “O-antigens,” which can vary widely. As an example, *Escherichia coli* has well over 70 capsular polysaccharides and over 170 known O-antigen immunotypes. Likewise, *Streptococcus pneumoniae* can express over 90 different strain-specific capsular polysaccharides [72]. Pathogenic bacteria are thus adept at evolving or acquiring the machinery for synthesizing host-like structures such as sialic acids. Below we compare Sia biosynthesis pathways in vertebrates and bacteria and discuss known and potential mechanisms that make Sia decoration a successful pathogenic mechanism.

4.5.2 Vertebrates and Bacteria Use Phylogenetically-related Sialic Acid Biosynthesis Pathways

The biosynthesis of free Neu5Ac varies slightly depending on whether it is happening in a vertebrate or bacterial cell [17] (Figure 4.2). In vertebrates, UDP-GlcNAc is converted to ManNAc by UDP-GlcNAc 2-epimerase/ManNAc kinase, a bifunctional enzyme, which also phosphorylates ManNAc to give ManNAc-6-phosphate [73, 74], which is then converted into Neu5Ac-9-phosphate by Neu5Ac-9-phosphate synthetase [75, 76]. Bacteria obtain ManNAc in the same way as vertebrates, by epimerization of UDP-GlcNAc [77, 78]. *Neisseria meningitidis* was postulated in one study to obtain ManNAc by epimerization of GlcNAc-6-phosphate (via gene product SiaA), followed by dephosphorylation [79]. A subsequent study, however, demonstrated that *N. meningitidis* SiaA catalyzes the epimerization of UDP-GlcNAc to ManNAc [78]. In contrast to animals, bacteria synthesize Neu5Ac via a Neu5Ac synthetase directly from ManNAc, rather than ManNAc-6-phosphate [80, 81]. In both bacteria and animals, activation of Neu5Ac is accomplished by converting Neu5Ac to CMP-Neu5Ac using CTP (cytidine 5'-triphosphate) and a CMP-Neu5Ac synthetase [82].

Despite the biochemical differences in Sia biosynthesis between vertebrates and bacteria, genes encoding the responsible enzymes are evolutionarily related. Moreover, phylogenetic analysis of Neu5Ac and CMP-Neu5Ac synthetases suggests that there have been multiple horizontal transfers of Sia biosynthesis genes among bacteria [17]. Unlike the enzymes involved in Sia biosynthesis, bacterial sialyltransferases do not appear to be related to those of animals and were likely “reinvented”, on multiple occasions. Alternatively, pathogens without complete biosynthetic pathways have devised multiple ways to “steal” sialic acids from their hosts [10], employing truncated Sia pathways that allow scavenging of host Sias for incorporation into their polysaccharides [10].

4.5.3 Microbial Sialic Acid Hijacks Host Factor H

The presence of Sias in microbial polysaccharides is often associated with pathogenicity in humans. Microbial expression of Sia enhances serum-resistance due to inactivation of the alternative complement pathway [83, 84], a tightly regulated proteolytic cascade that mediates an immediate antibody-independent response to foreign entities, and constitutes one of the frontline defenses against many invading pathogens [24]. Factor H is a regulatory component of the alternative complement pathway, which binds to a number of “self” components including sialic acid, and down-regulates activation of the pathway [85]. In this way, Factor H serves a critical function by preventing self-reactivity (i.e. “friendly fire”). Unfortunately, the otherwise useful molecular assumption that “Sia = self” becomes a liability when sialylated pathogens invade the body. The common mammalian terminal trisaccharide Neu5Ac α -2-3Gal β 1-(3/4)GlcNAc appears to be very common among pathogenic bacterial polysaccharide capsules and lipooligosaccharides (LOS) and may reflect bacterial optimization for binding to Factor H or other host Sia-binding proteins, such as Siglecs (see below) [11].

4.5.4 Is Siglec Hijacking Responsible for the Rapid Evolution of This Protein Family?

Siglecs are a large family of Sia-binding proteins that are expressed, with some exceptions, on cells of the immune system [5, 86, 87]. The precise contexts in which many Siglecs function remain largely unknown, although some clearly have important endogenous functions such as regulation of B-cell stimulation [88] and the maintenance of myelin [89]. A group of closely related Siglecs (known as the CD33-related Siglecs) are capable of regulating immune cells *in vitro* [90–92], as well as *in vivo* [93].

The CD33-related Siglecs are undergoing rapid evolution that is particularly evident in their amino-terminal V-set Ig-like Sia-binding domains [94]. This evolution likely reflects multiple selective pressures, including the human-specific loss of Neu5Gc. Indeed, it appears that the Sia-binding preference of at least one human Siglec (Siglec-9) has shifted from Neu5Gc towards Neu5Ac, when compared with the chimpanzee ortholog [95]. The loss of Neu5Gc may (partly) explain accelerated evolution in the human lineage, but the fact that Siglecs are evolving rapidly in multiple lineages requires further explanation. We have suggested that Sia-binding and Sia-expressing pathogens increased the rate of evolution among Siglecs in distinct, but interdependent, ways [5, 94].

Sia-binding and Sia-expressing mechanisms of pathogenesis not only apply to humans, but appear to be fairly ancient. Indeed, Sia binding and Sia expression are used by pathogens that infect a wide range of vertebrate host species. Pathogenic Sia expression may result in direct interactions with Siglecs, possibly to hijack inhibitory roles of these proteins in the host [12]. Through many generations and/or epidemics, Siglecs may have diverged from recognizing particular contexts of pathogen-expressed Sias – a theoretical example of the “Red Queen effect.” In the case of Sia-binding pathogens, particular host Sias may have diverged to avoid pathogen recognition (see Section 4.4.1). Such changes in host Sia structure(s) may have created the need for Siglec divergence in order to preserve endogenous function(s). This hypothetical situation embodies a concept recently referred to as a “Secondary Red Queen effect” [5].

In a variation on this theme, two porcine viruses [reproductive and respiratory syndrome virus (RRSR) and arterivirus], apparently exploit the host Sia-recognizing lectin (Siglec-1/sialoadhesin) for invasion of macrophages, via recognition of the Sias on the viral cell membranes [96, 97].

4.5.5 Enzymes That Release, Destroy, or Alter Sialic Acids

There are many sialidases encoded by both host and pathogen that remove Sia molecules. In addition, enzymes such as lyases and esterases destroy Sias or Sia modifications, respectively (Figure 4.1). Microbial removal of host Sia is often related to survival in the host. For example, *Streptococcus pneumonia* (pneumococcus) is a pneumonia-causing agent and a leading cause of fatal infections in the very young and the very old. Pneumococcus expresses a sialidase (neuraminidase) that exposes underlying Gal β 1–4GlcNAc residues and increases bacterial adhesion [98, 99]. Another example of a neuraminidase-expressing pathogen is Influenza virus, which has both a Sia-binding hemagglutinin and a neuraminidase, both of which are critical for replication and spread of the virus. Some bacteria are also able to use Sias as an energy source, and may encode sialidases to release nutritional value that is ‘locked up’ on the host cell surface [100–102]. Although it appears that host and pathogen sialidases share a common evolutionary origin, the roles of endogenous (host) sialidases are not as well understood as those of their microbial counterparts [103, 104]. Just as has been shown for Sia-binding lectins, many sialidases prefer particular Sia structures; the most commonly described situation is to find that sialidase action is hindered by *O*-acetylation [105, 106].

4.6 Evolution of Sialic Acids

Evolutionary relationships between biosynthetic pathways are often inferred by phylogenetic, structural, and/or mechanistic studies of their component enzymes. Some such studies are lacking for particular Sia or Sia-like molecules. Nonetheless, preliminary analyses suggest the biosynthetic pathways of Neu5Ac, Kdn, Kdo, and other ‘sialic acid-like’ molecules are at least partly homologous [17].

4.6.1 Neu5Ac versus Kdo

A bacterial sugar called Kdo (2-keto-3-deoxy-D-manno-octulosonic acid) bears a close resemblance to Sias (structures and biosynthetic pathways are shown in Figure 4.2). Kdo

is found as part of the “core” structure of Gram-negative bacterial lipopolysaccharides (LPS) and also in plant cell walls. Unlike the differences in Neu5Ac biosynthesis between bacteria and vertebrates (see Figure 4.2 and text above), it appears that the Kdo pathway in Gram-negative bacteria and plants does not differ substantially. Indeed, a Gram-negative mutant deficient in Kdo-8-phosphate synthetase was complemented by the orthologous enzyme from a pea plant [107]. Phylogenetic examination of CMP-Kdo synthetases indicates that this gene underwent horizontal gene transfer from bacteria to plants prior to or shortly after the divergence of plants from other eukaryotes [108]. There is also significant sequence similarity between CMP-Neu5Ac and CMP-Kdo synthetases, confirming their homologous relationship. These synthetases likely diverged by ancient gene duplication, as evidenced by their separate clustering in a phylogenetic tree [17]. Unfortunately, the exact timing of the presumed gene duplication leading to Neu5Ac and Kdo CMP-synthetases is difficult to infer. Analysis of the biosynthetic step prior to CMP activation indicates that Neu5Ac-9-P and Kdo-8-P synthetases do *not* show much sequence similarity; however, both reactions proceed by a similar mechanism that employs a TIM barrel fold, and is also shared by DHAP synthase (3-deoxy-D-arabino-heptulosonate-7-phosphate [109]. Recently, the structure of a bacterial Neu5Ac synthetase was also shown to have the TIM barrel fold [110]. Of course, the common TIM barrel fold shared by these enzymes is a very stable fold that is used by a large number of enzymes that perform many different functions [111].

4.6.2 Biosynthetic Machinery for Kdn, Legionaminic Acid, and Pseudaminic Acid

Although the enzymes of Kdn biosynthesis have not been cloned and characterized, biochemical studies indicate that the pathway intermediates are analogous to those of Neu5Ac biosynthesis, but are mostly catalyzed by distinct enzymes [112]. The essential difference between Kdn and Neu5Ac biosynthesis is that the Kdn pathway begins with mannose-6-phosphate (Man-6-P) rather than ManNAc. Interestingly, both human and *Drosophila* Neu5Ac-9-P synthetase enzymes can accept Man-6-P to generate Kdn-9-P [113, 114]. Studies of rat Neu5Ac-9-P synthetase, however, indicate that it does not accept Man-6-P [115]. Discovery of the genetic basis for Kdn biosynthesis will allow further characterization of both the evolutionary history and biological importance of this sialic acid.

Other Sia-like molecules produced by microbes include pseudaminic acid (5,7-diamino-3,5,7,9-tetradecyloxy-L-glycero-L-manno-nonulosonic acid) and legionaminic acid (5,7-diamino-3,5,7,9-tetradecyloxy-D-glycero-D-galacto-nonulosonic acid). Whereas Kdo is found in all Gram-negative bacteria and plants, pseudaminic and legionaminic acids have so far been chemically identified in only a relatively small number of bacteria [19]. Surprisingly, bacterial synthetases (whether Neu5Ac, legionaminic acid, or pseudaminic acid) share 30–35% identity at the amino acid level when compared with the human Neu5Ac synthetase. All-in-all, although the enzymes of Sia biosynthesis appear to be related, an understanding of their ancient history will likely require a phylogenetic approach based on structure or structural modeling, not just on sequence identity [116]. Recent advances in our understanding of pseudaminic acid biosynthesis will be of particular interest towards this end [117].

4.7 Current Status and Future Directions of Sialic Acid Bioinformatics

4.7.1 Existing Glycan Databases Often Misrepresent Sialic Acids

Sialic acid linkages and modifications are somewhat delicate (i.e. susceptible to hydrolysis) compared with other carbohydrates. Methods for the release and analysis of glycans often damage Sias or their modifications (see Table 4.1); thus, Sias are often inaccurately depicted in glycan databases. In practice, there has been an unfortunate tendency to assume that any Sia that has been removed by mild acid or by a sialidase, or detected by virtue of its negative charge, must be *N*-acetylneuraminic acid. In fact, in most instances where the literature or databases indicate the presence of Neu5Ac (or other acronyms such as NeuAc, NeuNAc, or NANA), the actual native Sia structure remains unknown. For example, Sia modifications such as *O*-acetylation, *O*-lactylation, and sulfation are sensitive to pH and are often inadvertently removed during glycan release, purification, or analysis. Also, Sia *O*-acetyl and *N*-glycolyl groups are sensitive to the conditions employed during hydrazinolysis and methylation analysis (Table 4.1). *O*-Acetyl esters at the C7 position are known to migrate along the Sia side-chain to C9, even under physiological conditions [118, 119]. Hence the detection of a 9-*O*-acetylated Sia should be cautiously interpreted as “9(7)-*O*-acetylated” until further characterization reveals the native location.

Overall, it is our recommendation that the term “Sia” (rather than Neu5Ac) be used whenever the type of Sia at a particular position on a glycan has not been definitively determined. All existing databases need to be “cleaned up” following the same convention. Table 4.1 lists several glycan analysis techniques that employ conditions likely to damage Sias or the Sia modifications mentioned above; it also provides some practical suggestions for retaining native Sia structures.

4.7.2 The Need for a “Sialome” Project

The myriad glycan structures found in nature represent a vast and expanding biological frontier. As elaborated throughout this chapter, Sias are an extreme example of glycan diversity, both in structure and in function. We have recently defined the term “Sialome” as “the total complement of Sia types and linkages and their modes of presentation on a particular organelle, cell, tissue, organ, or organism – as found at a particular time and under specific conditions” [5].

In other words, a Sialome can be thought of as reflecting at least six levels of complexity:

1. Different possible “core” Sia structural variations at the 5-position of Neu, Neu5Ac, Neu5Gc, or Kdn.
2. Various modifications of the above, sometimes in combinations (see Section 4.3.2).
3. Differing linkages of the Sia to the underlying glycan (mostly α 2–3 or α 2–6 to various sugars, or α 2–8 or α 2–9 to underlying Sias).
4. The precise identity and arrangement of glycans immediately below the Sia.
5. Structural attributes of the underlying glycan (e.g. *N*-linked or *O*-linked to protein, lipid- or GPI-anchored, cell-associated, or secreted).
6. The higher level cellular, organismal, and environmental milieu (each of these ideally deserve their own category).

Table 4.1 Examples of methods for studying glycans and their negative impact on sialic acids.

Method	Effect on Sia	Recommendation*
Endoglycosidase release, e.g. PNGase F (peptide: <i>N</i> -glycosidase F) release of <i>N</i> -glycans. This enzyme exhibits broad specificity and is very commonly used in <i>N</i> -glycan analysis	Sia <i>O</i> -acetylation susceptible to migration and/or hydrolysis at the pH optimum of this enzyme (8.6), especially during prolonged incubations	Use larger amounts of enzyme at pH 7.5 for short time periods to retain <i>O</i> -acetylation
Hydrazinolysis for release of <i>N</i> - and <i>O</i> -linked glycans. More commonly used for <i>O</i> -glycans, since there are no known broad-spectrum. Peptide: <i>O</i> -glycanases	Results in loss of both <i>N</i> - and <i>O</i> -acetyl groups (and likely some others)	Products of this reaction are often subject to re- <i>N</i> -acetylation without regard for the possibility that <i>N</i> -glycolyl could have been the native structure or that <i>O</i> -modifications are not replaced
Use of ~0.05–0.1% TFA (trifluoroacetic acid) in applications such as dialysis or column elution	Can result in desialylation and may damage some Sia modifications	If necessary for separation from contaminants, freeze-dry immediately and interpret cautiously
Base treatment often used to release <i>O</i> -glycans, and sometimes to release GPI-anchored glycans or bacterial polysaccharides, e.g. Group B streptococcal (GBS) polysaccharide, linked through phosphodiester to cell-wall peptidoglycan	Harsh basic conditions are certain to destroy some Sia modifications which go unreported, e.g. <i>O</i> -acetylation of Sias on GBS capsule went unreported for ~25 years	Use alternatives to base-treatment to retain native structure or interpret cautiously
Thiobarbituric acid (TBA) assay for Sia quantitation	Assay relies on periodate-oxidation of the Sia side-chain, a reaction that is impeded by the presence of <i>O</i> -acetylation and other modifications of Sia at C9	The TBA assay does not quantitate modified Sias unless modifications are first removed (i.e. remove <i>O</i> -acetylation by base treatment)
Fluorescent derivatization of monosaccharides (DMB-sialic acid) and reducing termini of oligosaccharides (2-AB and 2-AMAC) for HPLC-fluorescence resolution and detection of different Sia species	Derivatization often employed under acidic conditions, which could result in some loss or migration of Sia modifications. Not systematically studied	Keep reactions cold after derivatization and analyze as soon as possible
Various	The <i>O</i> -acetyl ester modification of Neu5Ac migrates from C7 to C9 under physiological and experimental conditions (e.g. below pH 3–4 and above pH 8, especially for prolonged periods and/or high temperatures)	Avoid conditions under which migration is accelerated. Dry samples and freeze for storage or use immediately for best results. When position unknown, use “9(7)- <i>O</i> -acetylation”

*If there is any concern of damage during release or analysis, indicate “Sia” rather than “Neu5Ac.”

Indeed, there are many factors that influence the presentation of Sias on cell surfaces, including normal developmental processes and pathological states such as inflammation, infection, and cancer – and many of these are cell-type and/or species-specific. Carefully constructed Sia databases will likely shed light on mechanisms of pathogen tropism for various tissues or species and may provide further insight into pathogenic mechanisms of invasion and evasion mediated by Sias. Given the diverse and extensive involvement of Sias in biological processes and disease states, these resources would no doubt provide insights into many other areas of interest. Despite the clear benefits of doing so, there has not been a concerted effort to collect Sia information into a comprehensive user-friendly database. Certainly, building databases that contain detailed information about glycans is a tedious task that must reflect methodological limitations of the past and present. We suggest that the field of Sia research needs two databases of Sia information in relation to particular biological and species contexts: one which would allow cataloguing of Sia types (points 1 and 2 above) and a second that would provide an overall idea of Sia linkages and modes of presentation (points 3–5 above). Information could be input and searched based on methodology or biological context. Such a “Sialome project” could have great predictive value in biological studies aimed at unraveling the functional significance and evolutionary history of this curious group of molecules.

References

1. Varki A: Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* 1993, **3**:97–130.
2. Gagneux P, Varki A: Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology* 1999, **9**:747–755.
3. Varki A: Nothing in glycobiology makes sense, except in the light of evolution. *Cell* 2006, **126**:841–845.
4. Bishop JR, Gagneux P: Evolution of carbohydrate antigens – microbial forces shaping host glycomes? *Glycobiology* 2007, **17**:23R–34R.
5. Varki A, Angata T: Siglecs – the major subfamily of I-type lectins. *Glycobiology* 2006, **16**:1R–27R.
6. Mammen M, Choi S-K, Whitesides GM: Polyvalent Interactions in Biological Systems: Implications for Design and Use of Multivalent Ligands and Inhibitors. *Angew Chem Int Ed* 1998, **37**:2754–2794.
7. Ilver D, Johansson P, Miller-Podraza H, Nyholm PG, Teneberg S, Karlsson KA: Bacterium–host protein–carbohydrate interactions. *Methods Enzymol* 2003, **363**:134–157.
8. Wills C, Green DR: A genetic herd-immunity model for the maintenance of MHC polymorphism. *Immunol Rev* 1995, **143**:263–292.
9. Lindesmith L, Moe C, Marionneau S, Ruvoen N, Jiang X, Lindblad L, Stewart P, LePendu J, Baric R: Human susceptibility and resistance to Norwalk virus infection. *Nat Med* 2003, **9**:548–553.
10. Vimr E, Lichtensteiger C: To sialylate, or not to sialylate: that is the question. *Trends Microbiol* 2002, **10**:254–257.
11. Carlin AF, Lewis AL, Varki A, Nizet V: Group B streptococcal capsular sialic acids interact with Siglecs (immunoglobulin-like lectins) on human leukocytes. *J Bacteriol* 2007, **89**:1231–1237.
12. Varki A: Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. *Nature* 2007, **446**:1023–1029.

13. Schauer R: Chemistry, metabolism, and biological functions of sialic acids. *Adv Carbohydr Chem Biochem* 1982, **40**:131–234.
14. Varki A: Diversity in the sialic acids. *Glycobiology* 1992, **2**:25–40.
15. Troy FA: Polysialylation: from bacteria to brains. *Glycobiology* 1992, **2**:5–23.
16. Kelm S, Schauer R: Sialic acids in molecular and cellular interactions. *Int Rev Cytol* 1997, **175**:137–240.
17. Angata T, Varki A: Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective. *Chem Rev* 2002, **102**:439–469.
18. Bouchet V, Hood DW, Li J, Brisson JR, Randle GA, Martin A, Li Z, Goldstein R, Schweda EK, Pelton SI, *et al.*: Host-derived sialic acid is incorporated into *Haemophilus influenzae* lipopolysaccharide and is a major virulence factor in experimental otitis media. *Proc Natl Acad Sci USA* 2003, **100**:8898–8903.
19. Lewis AL, Desa N, Hansen EE, Knirel Y, Gordon JI, Gagneux P, Nizet V, Varki A: Innovations in host and microbial sialic acid biosynthesis revealed by genomic and phylogenetic prediction of nonulosonic acid structure. *Proc Natl Acad Sci USA* 2009. (in press).
20. Kerjaschki D, Vernillo AT, Farquhar MG: Reduced sialylation of podocalyxin – the major sialoprotein of the rat kidney glomerulus – in aminonucleoside nephrosis. *Am J Pathol* 1985, **118**:343–349.
21. Fujimoto I, Bruses JL, Rutishauser U: Regulation of cell adhesion by polysialic acid. *J Biol Chem* 2001, **276**:31745–31751.
22. Ronn LC, Berezin V, Bock E: The neural cell adhesion molecule in synaptic plasticity and ageing. *Int J Dev Neurosci* 2000, **18**:193–199.
23. Ley K: The role of selectins in inflammation and disease. *Trends Mol Med* 2003, **9**:263–268.
24. Pangburn MK: Host recognition and target differentiation by factor H, a regulator of the alternative pathway of complement. *Immunopharmacology* 2000, **49**:149–157.
25. Crocker PR, Paulson JC, Varki A: Siglecs and their roles in the immune system. *Nat Rev Immunol* 2007, **7**:255–266.
26. Schwarzkopf M, Knobeloch KP, Rohde E, Hinderlich S, Wiechens N, Lucka L, Horak I, Reutter W, Horstkorte R: Sialylation is essential for early development in mice. *Proc Natl Acad Sci USA* 2002, **99**:5267–5270.
27. Martin LT, Marth JD, Varki A, Varki NM: Genetically altered mice with different sialyltransferase deficiencies show tissue-specific alterations in sialylation and sialic acid 9-*O*-acetylation. *J Biol Chem* 2002, **277**:32930–32938.
28. Jamieson JC, McCaffrey G, Harder PG: Sialyltransferase: a novel acute-phase reactant. *Comp Biochem Physiol B* 1993, **105**:29–33.
29. Dabelic S, Flogel M, Maravic G, Lauc G: Stress causes tissue-specific changes in the sialyltransferase activity. *Z Naturforsch, Teil C* 2004, **59**:276–280.
30. Knight PA, Pemberton AD, Robertson KA, Roy DJ, Wright SH, Miller HR: Expression profiling reveals novel innate and inflammatory responses in the jejunal epithelial compartment during infection with *Trichinella spiralis*. *Infect Immun* 2004, **72**:6076–6086.
31. Fukushima K, Hara-Kuge S, Seko A, Ikehara Y, Yamashita K: Elevation of alpha2→6 sialyltransferase and alpha1→2 fucosyltransferase activities in human choriocarcinoma. *Cancer Res* 1998, **58**:4301–4306.
32. Stamatou NM, Liang F, Nan X, Landry K, Cross AS, Wang LX, Sshezhetsky AV: Differential expression of endogenous sialidases of human monocytes during cellular differentiation into macrophages. *FEBS J* 2005, **272**:2545–2556.
33. Miyagi T, Wada T, Yamaguchi K, Hata K: Sialidase and malignancy: a minireview. *Glycoconj J* 2004, **20**:189–198.
34. Kawano T, Koyama S, Takematsu H, Kozutsumi Y, Kawasaki H, Kawashima S, Kawasaki T, Suzuki A: Molecular cloning of cytidine monophospho-*N*-acetylneuraminic acid hydroxylase.

- Regulation of species- and tissue-specific expression of *N*-glycolylneuraminic acid. *J Biol Chem* 1995, **270**:16458–16463.
35. Hayakawa T, Satta Y, Gagneux P, Varki A, Takahata N: Alu-mediated inactivation of the human CMP-*N*-acetylneuraminic acid hydroxylase gene. *Proc Natl Acad Sci USA* 2001, **98**:11399–11404.
 36. Manzi AE, Sjoberg ER, Diaz S, Varki A: Biosynthesis and turnover of *O*-acetyl and *N*-acetyl groups in the gangliosides of human melanoma cells. *J Biol Chem* 1990, **265**:13091–13103.
 37. Sonnenburg JL, Van Halbeek H, Varki A: Characterization of the acid stability of glycosidically linked neuraminic acid – use in detecting de-*N*-acetyl-gangliosides in human melanoma. *J Biol Chem* 2002, **277**:17502–17510.
 38. Chen HY, Varki A: *O*-Acetylation of GD3: an enigmatic modification regulating apoptosis? *J Exp Med* 2002, **196**: 1529–1533.
 39. Shen Y, Tiralongo J, Kohla G, Schauer R: Regulation of sialic acid *O*-acetylation in human colon mucosa. *Biol Chem* 2004, **385**: 145–152.
 40. Shen Y, Kohla G, Lrhorfi AL, Sipos B, Kalthoff H, Gerwig GJ, Kamerling JP, Schauer R, Tiralongo J: *O*-Acetylation and de-*O*-acetylation of sialic acids in human colorectal carcinoma. *Eur J Biochem* 2004, **271**:281–290.
 41. Shi WX, Chammas R, Varki NM, Powell L, Varki A: Sialic acid 9-*O*-acetylation on murine erythro leukemia cells affects complement activation, binding to I-type lectins, and tissue homing. *J Biol Chem* 1996, **271**:31526–31532.
 42. Berry DS, Lynn F, Lee CH, Frasch CE, Bash MC: Effect of *O*-acetylation of *Neisseria meningitidis* serogroup A capsular polysaccharide on development of functional immune responses. *Infect Immun* 2002, **70**:3707–3713.
 43. Chava AK, Bandyopadhyay S, Chatterjee M, Mandal C: Sialoglycans in protozoal diseases: their detection, modes of acquisition and emerging biological roles. *Glycoconj J* 2004, **20**: 199–206.
 44. Deszo EL, Steenbergen SM, Freedberg DI, Vimr ER: *Escherichia coli* K1 polysialic acid *O*-acetyltransferase gene, neuO, and the mechanism of capsule form variation involving a mobile contingency locus. *Proc Natl Acad Sci USA* 2005, **102**:5564–5569.
 45. Zanetta JP, Pons A, Iwersen M, Mariller C, Leroy Y, Timmerman P, Schauer R: Diversity of sialic acids revealed using gas chromatography/mass spectrometry of heptafluorobutyrate derivatives. *Glycobiology* 2001, **11**:663–676.
 46. Bulai T, Bratosin D, Pons A, Montreuil J, Zanetta JP: Diversity of the human erythrocyte membrane sialic acids in relation with blood groups. *FEBS Lett* 2003, **534**:185–189.
 47. Warren L: The distribution of sialic acids in Nature. *Comp Biochem Physiol* 1963, **10**:153–171.
 48. Saito M, Kitamura H, Sugiyama K: Occurrence of gangliosides in the common squid and Pacific octopus among protostomia. *Biochim Biophys Acta Bio-Membr* 2001, **1511**:271–280.
 49. Roth J, Kempf A, Reuter G, Schauer R, Gehring WJ: Occurrence of sialic acids in *Drosophila melanogaster*. *Science* 1992, **256**:673–675.
 50. Malykh YN, Krisch B, Gerardy-Schahn R, Lapina EB, Shaw L, Schauer R: The presence of *N*-acetylneuraminic acid in Malpighian tubules of larvae of the cicada *Philaenus spumarius*. *Glycoconj J* 1999, **16**:731–739.
 51. Shah MM, Fujiyama K, Flynn CR, Joshi L: Sialylated endogenous glycoconjugates in plant cells. *Nat Biotechnol* 2003, **21**:1470–1471.
 52. Seveno M, Bardor M, Paccalet T, Gomord V, Lerouge P, Faye L: Glycoprotein sialylation in plants? *Nat Biotechnol* 2004, **22**:1351–1352.
 53. Zeleny R, Kolarich D, Strasser R, Altmann F: Sialic acid concentrations in plants are in the range of inadvertent contamination. *Planta* 2006, **224**:222–227.
 54. Couceiro JNSS, Paulson JC, Baum LG: Influenza virus strains selectively recognize sialyloligosaccharides on human respiratory epithelium; the role of the host cell in selection of hemagglutinin receptor specificity. *Virus Res* 1993, **29**:155–165.

55. Suzuki Y: Sialobiology of influenza: molecular mechanism of host range variation of influenza viruses. *Biol Pharm Bull* 2005, **28**:399–408.
56. Kyogashima M, Ginsburg V, Krivan HC: *Escherichia coli* K99 binds to *N*-glycolylsialoparagloboside and *N*-glycolyl-GM3 found in piglet small intestine. *Arch Biochem Biophys* 1989, **270**:391–397.
57. Hollingsworth MA, Swanson BJ: Mucins in cancer: protection and control of the cell surface. *Nat Rev Cancer* 2004, **4**:45–60.
58. Newburg DS, Ruiz-Palacios GM, Altaye M, Chaturvedi P, Meinzen-Derr J, Guerrero MM, Morrow AL: Innate protection conferred by fucosylated oligosaccharides of human milk against diarrhea in breastfed infants. *Glycobiology* 2004, **14**:253–263.
59. Stahl B, Thurl S, Zeng J, Karas M, Hillenkamp F, Steup M, Sawatzki G: Oligosaccharides from human milk as revealed by matrix-assisted laser desorption/ionization mass spectrometry. *Anal Biochem* 1994, **223**:218–226.
60. Chaturvedi P, Warren CD, Buescher CR, Pickering LK, Newburg DS: Survival of human milk oligosaccharides in the intestine of infants. *Adv Exp Med Biol* 2001, **501**:315–323.
61. Newburg DS: Innate immunity and human milk. *J Nutr* 2005, **135**:1308–1312.
62. Urashima T, Saito T, Nakamura T, Messer M: Oligosaccharides of milk and colostrum in non-human mammals. *Glycoconj J* 2001, **18**:357–371.
63. Morrow AL, Ruiz-Palacios GM, Jiang X, Newburg DS: Human-milk glycans that inhibit pathogen binding protect breast-feeding infants against infectious diarrhea. *J Nutr* 2005, **135**:1304–1307.
64. Hirno S, Kelm S, Iwersen M, Hotta K, Goso Y, Ishihara K, Suguri T, Morita M, Wadström T, Schauer R: Inhibition of *Helicobacter pylori* sialic acid-specific haemagglutination by human gastrointestinal mucins and milk glycoproteins. *FEMS Immunol Med Microbiol* 1998, **20**:275–281.
65. Solzbacher D, Hanisch FG, van Alphen L, Gilsdorf JR, Schrotten H: Mucin in middle ear effusions inhibits attachment of *Haemophilus influenzae* to mucosal epithelial cells. *Eur Arch Otorhinolaryngol* 2003, **260**:141–147.
66. Arcasoy SM, Latoche J, Gondor M, Watkins SC, Henderson RA, Hughey R, Finn OJ, Pilewski JM: MUC1 and other sialoglycoconjugates inhibit adenovirus-mediated gene transfer to epithelial cells. *Am J Respir Cell Mol Biol* 1997, **17**:422–435.
67. Holmen JM, Olson FJ, Karlsson H, Hansson GC: Two glycosylation alterations of mouse intestinal mucins due to infection caused by the parasite *Nippostrongylus brasiliensis*. *Glycoconj J* 2002, **19**:67–75.
68. McNamara N, Basbaum C: Signaling networks controlling mucin production in response to Gram-positive and Gram-negative bacteria. *Glycoconj J* 2001, **18**:715–722.
69. Yolken RH, Peterson JA, Vonderfecht SL, Fouts ET, Midthun K, Newburg DS: Human milk mucin inhibits rotavirus replication and prevents experimental gastroenteritis. *J Clin Invest* 1992, **90**:1984–1991.
70. Martín-Sosa S, Martín MJ, Hueso P: The sialylated fraction of milk oligosaccharides is partially responsible for binding to enterotoxigenic and uropathogenic *Escherichia coli* human strains. *J Nutr* 2002, **132**:3067–3072.
71. Otnaess AB, Laegreid A, Ertresvag K: Inhibition of enterotoxin from *Escherichia coli* and *Vibrio cholerae* by gangliosides from human milk. *Infect Immun* 1983, **40**:563–569.
72. Weintraub A: Immunology of bacterial polysaccharide antigens. *Carbohydr Res* 2003, **338**:2539–2547.
73. Hinderlich S, Stäsche R, Zeitler R, Reutter W: A bifunctional enzyme catalyzes the first two steps in *N*-acetylneuraminic acid biosynthesis of rat liver – purification and characterization of UDP-*N*-acetylglucosamine 2-epimerase/*N*-acetylmannosamine kinase. *J Biol Chem* 1997, **272**:24313–24318.

74. Stäsche R, Hinderlich S, Weise C, Effertz K, Lucka L, Moormann P, Reutter W: A bifunctional enzyme catalyzes the first two steps in *N*-acetylneuraminic acid biosynthesis of rat liver – molecular cloning and functional expression of UDP-*N*-acetyl-glucosamine 2-epimerase/*N*-acetylmannosamine kinase. *J Biol Chem* 1997, **272**:24319–24324.
75. Lawrence SM, Huddleston KA, Pitts LR, Nguyen N, Lee YC, Vann WF, Coleman TA, Betenbaugh MJ: Cloning and expression of the human *N*-acetylneuraminic acid phosphate synthase gene with 2-keto-3-deoxy-D-glycero-D-galacto-nononic acid biosynthetic ability. *J Biol Chem* 2000, **275**:17869–17877.
76. Nakata D, Close BE, Colley KJ, Matsuda T, Kitajima K: Molecular cloning and expression of the mouse *N*-acetylneuraminic acid 9-phosphate synthase which does not have deaminoneuraminic acid (KDN) 9-phosphate synthase activity. *Biochem Biophys Res Commun* 2000, **273**:642–648.
77. Vann WF, Daines DA, Murkin AS, Tanner ME, Chaffin DO, Rubens CE, Vionnet J, Silver RP: The NeuC protein of *Escherichia coli* K1 is a UDP *N*-acetylglucosamine 2-epimerase. *J Bacteriol* 2004, **186**:706–712.
78. Murkin AS, Chou WK, Wakarchuk WW, Tanner ME: Identification and mechanism of a bacterial hydrolyzing UDP-*N*-acetylglucosamine 2-epimerase. *Biochemistry* 2004, **43**:14290–14298.
79. Petersen M, Fessner W, Frosch M, Luneberg E: The *siaA* gene involved in capsule polysaccharide biosynthesis of *Neisseria meningitidis* B codes for *N*-acetylglucosamine-6-phosphate 2-epimerase activity. *FEMS Microbiol Lett* 2000, **184**:161–164.
80. Vann WF, Tavarez JJ, Crowley J, Vimr E, Silver RP: Purification and characterization of the *Escherichia coli* K1 neuB gene product *N*-acetylneuraminic acid synthetase. *Glycobiology* 1997, **7**:697–701.
81. Ringenberg MA, Steenbergen SM, Vimr ER: The first committed step in the biosynthesis of sialic acid by *Escherichia coli* K1 does not involve a phosphorylated *N*-acetylmannosamine intermediate. *Mol Microbiol* 2003, **50**:961–975.
82. Rodríguez-Aparicio LB, Luengo JM, González-Clemente C, Reglero A: Purification and characterization of the nuclear cytidine 5'-monophosphate *N*-acetylneuraminic acid synthetase from rat liver. *J Biol Chem* 1992, **267**:9257–9263.
83. Jarvis GA, Vedros NA: Sialic acid of group B *Neisseria meningitidis* regulates alternative complement pathway activation. *Infect Immun* 1987, **55**:174–180.
84. Marques MB, Kasper DL, Pangburn MK, Wessels MR: Prevention of C3 deposition by capsular polysaccharide is a virulence mechanism of type III group B streptococci. *Infect Immun* 1992, **60**:3986–3993.
85. Meri S, Pangburn MK: Discrimination between activators and nonactivators of the alternative pathway of complement: regulation via a sialic acid /polyanion binding site on factor H. *Proc Natl Acad Sci USA* 1990, **87**:3982–3986.
86. Crocker PR, Varki A: Siglecs, sialic acids and innate immunity. *Trends Immunol* 2001, **22**:337–342.
87. Crocker PR: Siglecs in innate immunity. *Curr Opin Pharmacol* 2005, **5**:431–437.
88. Cornall RJ, Goodnow CC, Cyster JG: Regulation of B cell antigen receptor signaling by the Lyn/CD22/SHP1 pathway. *Curr Top Microbiol Immunol* 1999, **244**:57–68.
89. Pan B, Fromholt SE, Hess EJ, Crawford TO, Griffin JW, Sheikh KA, Schnaar RL: Myelin-associated glycoprotein and complementary axonal ligands, gangliosides, mediate axon stability in the CNS and PNS: neuropathology and behavioral deficits in single- and double-null mice. *Exp Neurol* 2005, **195**:208–217.
90. Avril T, Floyd H, Lopez F, Vivier E, Crocker PR: The membrane-proximal immunoreceptor tyrosine-based inhibitory motif is critical for the inhibitory signaling mediated by Siglecs-7 and -9, CD33-related Siglecs expressed on human monocytes and NK cells. *J Immunol* 2004, **173**:6841–6849.
91. Ikehara Y, Ikehara SK, Paulson JC: Negative regulation of T cell receptor signaling by Siglec-7 (p70/AIRM) and Siglec-9. *J Biol Chem* 2004, **279**:43117–43125.

92. Lajaunias F, Dayer JM, Chizzolini C: Constitutive repressor activity of CD33 on human monocytes requires sialic acid recognition and phosphoinositide 3-kinase-mediated intracellular signaling. *Eur J Immunol* 2004, **35**:243–251.
93. Zhang M, Angata T, Cho JY, Miller M, Broide DH, Varki A: Defining the in vivo function of Siglec-F, a CD33-related Siglec expressed on mouse eosinophils. *Blood* 2007, **109**:4280–4287.
94. Angata T, Margulies EH, Green ED, Varki A: Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc Natl Acad Sci USA* 2004, **101**:13251–13256.
95. Sonnenburg JL, Altheide TK, Varki A: A uniquely human consequence of domain-specific functional adaptation in a sialic acid-binding receptor. *Glycobiology* 2004, **14**:339–346.
96. Delputte PL, Nauwynck HJ: Porcine arterivirus infection of alveolar macrophages is mediated by sialic acid on the virus. *J Virol* 2004, **78**:8094–8101.
97. Vanderheijden N, Delputte PL, Favoreel HW, Vandekerckhove J, Van Damme J, Van Woensel PA, Nauwynck HJ: Involvement of sialoadhesin in entry of porcine reproductive and respiratory syndrome virus into porcine alveolar macrophages. *J Virol* 2003, **77**:8207–8215.
98. Barthelson R, Mobasser A, Zopf D, Simon P: Adherence of *Streptococcus pneumoniae* to respiratory epithelial cells is inhibited by sialylated oligosaccharides. *Infect Immun* 1998, **66**:1439–1444.
99. Tong HH, Liu X, Chen Y, James M, Demaria T: Effect of neuraminidase on receptor-mediated adherence of *Streptococcus pneumoniae* to chinchilla tracheal epithelium. *Acta Otolaryngol* 2002, **122**:413–419.
100. Corfield AP, Wagner SA, O'Donnell LJD, Durdey P, Mountford RA, Clamp JR: The roles of enteric bacterial sialidase, sialate *O*-acetyl esterase and glycosulfatase in the degradation of human colonic mucin. *Glycoconj J* 1993, **10**:72–81.
101. Vimr ER, Kalivoda KA, Deszo EL, Steenbergen SM: Diversity of microbial sialic acid metabolism. *Microbiol Mol Biol Rev* 2004, **68**:132–153.
102. Sonnenburg JL, Xu J, Leip DD, Chen CH, Westover BP, Weatherford J, Buhler JD, Gordon JI: Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* 2005, **307**:1955–1959.
103. Roggentin P, Schauer R, Hoyer LL, Vimr ER: The sialidase superfamily and its spread by horizontal gene transfer. *Mol Microbiol* 1993, **9**:915–921.
104. Taylor G: Sialidases: structures, biological significance and therapeutic potential. *Curr Opin Struct Biol* 1996, **6**:830–837.
105. Varki A, Diaz S: A neuraminidase from *Streptococcus sanguis* that can release *O*-acetylated sialic acids. *J Biol Chem* 1983, **258**:12465–12471.
106. Zenz KI, Roggentin P, Schauer R: Isolation and properties of the natural and the recombinant sialidase from *Clostridium septicum* NC 0054714. *Glycoconj J* 1993, **10**:50–56.
107. Brabetz W, Wolter FP, Brade H: A cDNA encoding 3-deoxy-D-manno-oct-2-ulosonate-8-phosphate synthase of *Pisum sativum* L. (pea) functionally complements a kdsA mutant of the Gram-negative bacterium *Salmonella enterica*. *Planta* 2000, **212**:136–143.
108. Royo J, Gimez E, Hueros G: CMP-KDO synthetase: a plant gene borrowed from Gram-negative eubacteria. *Trends Genet* 2000, **16**:432–433.
109. Tanner ME: The enzymes of sialic acid biosynthesis. *Bioorg Chem* 2005, **33**:216–228.
110. Gunawan J, Simard D, Gilbert M, Lovering AL, Wakarchuk WW, Tanner ME, Strynadka NC: Structural and mechanistic analysis of sialic acid synthase NeuB from *Neisseria meningitidis* in complex with Mn²⁺, phosphoenolpyruvate, and *N*-acetylmannosaminitol. *J Biol Chem* 2004, **280**:3555–3563.
111. Wise EL, Rayment I: Understanding the importance of protein structure to Nature's routes for divergent evolution in TIM barrel enzymes. *Acc Chem Res* 2004, **37**:149–158.

112. Angata T, Nakata D, Matsuda T, Kitajima K, Troy FAII: Biosynthesis of KDN (2-keto-3-deoxy-D-glycero-D-galacto-nononic acid) – identification and characterization of a KDN-9-phosphate synthetase activity from trout testis. *J Biol Chem* 1999, **274**:22949–22956.
113. Lawrence SM, Huddleston KA, Tomiya N, Nguyen N, Lee YC, Vann WF, Coleman TA, Betenbaugh MJ: Cloning and expression of human sialic acid pathway genes to generate CMP-sialic acids in insect cells. *Glycoconj J* 2001, **18**:205–213.
114. Kim K, Lawrence SM, Park J, Pitts L, Vann WF, Betenbaugh MJ, Palter KB: Expression of a functional *Drosophila melanogaster* *N*-acetylneuraminic acid (Neu5Ac) phosphate synthase gene: evidence for endogenous sialic acid biosynthetic ability in insects. *Glycobiology* 2002, **12**:73–83.
115. Chen H, Blume A, Zimmermann-Kordmann M, Reutter W, Hinderlich S: Purification and characterization of *N*-acetylneuraminic acid-9-phosphate synthase from rat liver. *Glycobiology* 2002, **12**:65–71.
116. Pirun M, Babnigg G, Stevens FJ: Template-based recognition of protein fold within the midnight and twilight zones of protein sequence similarity. *J Mol Recognit* 2005, **18**:203–212.
117. Schoenhofen IC, McNally DJ, Brisson JR, Logan SM: Elucidation of the CMP-pseudaminic acid pathway in *Helicobacter pylori*: synthesis from UDP-*N*-acetylglucosamine by a single enzymatic reaction. *Glycobiology* 2006, **16**:8C–14C.
118. Varki A, Diaz S: The release and purification of sialic acids from glycoconjugates: methods to minimize the loss and migration of *O*-acetyl groups. *Anal Biochem* 1984, **137**:236–247.
119. Kamerling JP, Schauer R, Shukla AK, Stoll S, van HH, Vliegthart JFG: Migration of *O*-acetyl groups in *N,O*-acetylneuraminic acids. *Eur J Biochem* 1987, **162**:601–607.

**Section 3:
Carbohydrate-active
Enzymes and Glycosylation**

5 Carbohydrate-active Enzymes

Database: Principles and Classification of Glycosyltransferases

**Pedro M. Coutinho¹, Corinne Rancurel¹, Mark Stam¹,
Thomas Bernard¹, Francisco M. Couto², Etienne G. J. Danchin¹
and Bernard Henrissat¹**

¹*Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités d'Aix-Marseille I & II, 13402 Marseille cedex 20, France*

²*Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal*

5.1 Introduction

Carbohydrates, and to a lesser extent glycoconjugates, sequester most of the carbon immobilized by living organisms. Glycosidic bonds link different monosaccharides, forming oligo- and polysaccharides, but may also link glycans to a variety of other compounds as in glycoproteins and glycolipids and in an enormous variety of other glycoconjugates. Glycosyltransferases (GTs) are the enzymes responsible for the biosynthesis of these glycosidic bonds in living organisms, using phospho-activated sugar donors as substrates. The resulting bonds and compounds may be degraded or modified by other carbohydrate-active enzymes. These enzymes are often very selective as glycans can present varied ring shapes, types of glycosidic bonds, and the possibility of branching [1].

As exemplified in the Chapter 2, glycans found in Nature exhibit a broad variability of structures. Their diversity is harnessed by living organisms to achieve a multiplicity of roles, ranging from the “more basic” structural role in the cell walls and membranes to carbon and energy reserves, from control of protein folding [2, 3]. To a “higher level,” carbohydrates are involved in a variety of intra- and intercellular communication events both within the organism [4] but also in the interactions between organisms, with specific emphasis in defense and pathogenesis [5, 6].

The diversity of carbohydrate and glycoconjugate structures results from the action of a plethora of GTs whose action depends on the enzymes' ability to recognize both an activated-sugar donor and the respective acceptor and to form a new glycosidic bond between the sugar and the acceptor. The donors are phospho-activated sugars where the sugar reducing end is bound to one or two phosphate groups that may be linked to different nucleosides (forming nucleotide mono- and diphosphate sugars designated here as NMP- and NDP-sugars, respectively) and to lipid groups based on dolichol-related lipid

groups (dolichyl-mono- and dolichyl-diphosphate-sugars, designated here as DMP- and DDP-sugars, respectively, and other dolichol-like prenols). This list is complemented by reducing-end phosphorylated sugars (as in sugar-1-phosphates). The acceptors are diverse and ultimately could correspond to any molecule containing simple reactive groups such as hydroxyl and amino groups. Typical acceptors include carbohydrates, proteins, lipids, and a variety of compounds from the basal and secondary metabolism. The importance and the diverse roles of carbohydrates and glycoconjugates are strictly dependent on the selective glycosylation step catalyzed by GTs, and on their selective cleavage or rearrangement achieved by other carbohydrate-active enzymes.

In this chapter, a review of the basis of classification of GTs and related carbohydrate-active enzymes found in CAZy (Carbohydrate-Active Enzymes database, <http://www.cazy.org/>) is given. This database provides an enzyme classification based on the conservation of sequence and structural features and covers different enzyme activity classes dealing with the formation and breakdown of glycosidic bonds and associated activities. Within limits, it provides means to correlate structural and enzyme mechanistic data within enzyme families based on biochemically characterized members, but contrasts with the traditional enzyme classifications by the fact that families are not centered on enzyme specificity. When applied to the classification of GTs, it allows researchers to establish relationships between different catalytic modules present in enzyme sequences, the nature of their substrates and the nature of the glycosidic bonds resulting from their action.

5.2 Classifications of Carbohydrate-active Enzymes in CAZy

In the early 1990s, a family classification system was developed based on protein sequence and structure similarities, which now covers different enzyme classes catalyzing the synthesis, degradation, and modification of glycoconjugates in general. Gradually having been made available online since 1998 and currently available in the CAZy database, these classifications emphasize enzyme mechanisms and protein fold rather than enzyme specificity, and are much adapted to the classification of proteins lacking biochemical characterization when sequence similarities to characterized enzyme exist. At present, CAZy covers the following classes of enzyme activities:

1. Glycoside hydrolases (GHs), including all glycosidases and transglycosidases [7–10], are responsible for the hydrolysis and/or transglycosylation of glycosidic bonds. They are very abundant and are currently classified into 106 families. Most GHs act by two well-known and established mechanisms [11], yielding products that invert or retain the configuration of the anomeric carbon of the cleaved glycosidic bond of the substrate. Only enzymes acting by anomeric retention are known to act as transglycosidases. Otherwise, a single catalytic water molecule is necessary for the action of most retaining and inverting hydrolases. A catalytic mechanism requiring NAD^+ as a cofactor has been unveiled recently [12], but so far it appears restricted to a limited number of families (eg. GH4). GHs are present in most free-living organisms, and correspond to almost half of the enzymes classified in CAZy. Given their widespread biotechnological use, GHs constitute so far the best biochemically characterized set of enzymes present in the database. They have been the subject of several reviews [13, 14].

2. Glycosyltransferases (GTs) are responsible for the biosynthesis of glycosidic bonds [15, 16]. These enzymes are currently classified into 84 families and are described in more detail later in this chapter. GTs are present in virtually every single organism and are as abundant as GHs.
3. Polysaccharide lyases (PLs) cleave the glycosidic bonds of uronic acid-containing polysaccharides by a β -elimination mechanism [12]. They are currently found in 18 families in CAZy [17], corresponding to about 1.5% of CAZy content. Many PLs have biotechnological and biomedical applications and are among the enzymes having the most biochemically characterized examples present in the database.
4. Carbohydrate esterases (CEs) remove some ester-based modifications present in poly- and oligosaccharides and may facilitate the action of GHs. Currently described in 14 families [17], CEs represent roughly 5% of CAZy entries. As the barrier between carbohydrate esterase and other esterase activities is low, it is probable that some likely enzymes present in the classification have limited action on carbohydrate modifications.
5. Carbohydrate-binding modules (CBMs) have no enzymatic activity *per se* but are known to potentiate the activity of many enzymes described above by targeting specific forms of the substrate. CBMs are most often associated with other carbohydrate-active enzyme catalytic modules in the same polypeptide and can target different substrate forms depending on different structural characteristics [18, 19]. However, occasionally they can be present in isolated or tandem forms not coupled with an enzyme. They are currently classified in 45 families in CAZy [17], and are likely to increase with better enzyme characterization.

5.2.1 CAZy Enzyme Families in Contrast with the More Traditional EC Classification

The CAZy enzyme families contrast with the more traditional EC classification at different levels:

1. Characterized enzymes are not classified by EC activity but solely according to similarity at sequence level, which correlates with the similarity of their 3D structure;
2. Proteins not yet biochemically characterized can be assigned to a family, allowing the inference of some mechanistic, structural, and/or even specificity or functional properties;
3. A single enzyme family may contain enzymes of different specificities, bringing limits to the inference of specificity.

As the age of genomics results in the accumulation of sequence data in databases, putative proteins found in new genomes can be analyzed *in silico* and classified. If significant sequence similarities are found with catalytic and ancillary non-catalytic modules present in already classified proteins, a new protein can therefore be added to the classification in the absence of biochemical characterization.

The individual but related classifications described here are regularly updated with new sequences, structures, and biochemical information and new families are frequently added

and reflect the advances in experimental characterization. New families are exclusively created based on the availability of at least one biochemically characterized member for which a sequence is available. This sequence will serve as a seed for the family that is extended with the addition of related sequences identified by diverse bioinformatics methods. Updated lists of families of carbohydrate-active enzymes can be found in the CAZy database site.

5.3 Bioinformatics: from Sequences and Structures to Biochemistry and Genomes

The principles of sequence classification in CAZy are described in this section. The methodology used to update and maintain the sequences and structure information is described along with the continuous curation efforts needed to integrate biochemical and other characterization data from the literature and the analysis of full sets of protein sequences present in a single genome.

Carbohydrate-active enzymes can exhibit a modular structure, where a module can be defined as a structural and functional unit [9, 10, 17]. Each family in CAZy is dependent on the definition of a common segment in each full sequence that ultimately contains the catalytic module. The definition of the limits within the sequence of the composing modules depends on available information issued from different approaches:

- protein three-dimensional structures
- reported deletion studies
- protein sequence analysis and comparisons.

Different sequence comparison tools have been used to define enzyme families. Informatics approaches used to populate families have evolved with the creation of new and improved biocomputing tools. In the 1990s, when the number of sequences available in the public sequence and structural databases was limited, hydrophobic cluster analysis (HCA) was often performed to identify the conserved globular protein segments that could be grouped to form families [20]. This analysis permits the identification not only of the limits of catalytic and ancillary modules but also of the contiguous unstructured elements such as linkers or variable anchoring segments, of signal peptides and transmembrane anchors, and so on. HCA allows a careful pairwise comparison of protein sequences, but relies on trained users. The alignment scores between sequences are difficult to obtain, and HCA plot interpretations can be subjective. The major drawback of HCA, however, was its low throughput so that it was not adapted to the massive release of new sequence data resulting from genome sequencing. In the late 1990s, the management of thousands of protein sequences, sequence and structure accessions codes, and modular organization descriptions became a challenge for an ever-growing database. This problem led to a major rearrangement of the CAZy management and annotation strategy in 1999 into a new evolving system.

The present layers of CAZy structure and management are described in Figure 5.1. The core of the system is a relational database that allows the management of individual proteins, their biochemical and modular annotations, and the management of families in a taxonomic context. Over 46 000 non-redundant protein entries and 84 000 single individual

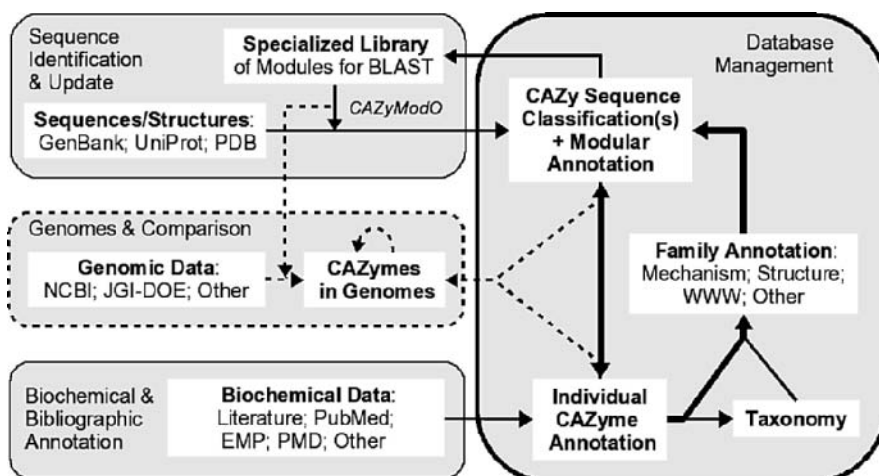


Figure 5.1 The CAZy integrated database and annotation system. A database management layer manages individual carbohydrate active enzyme entries, comprising sequence and biochemical information, its modular description, the overall family description, and taxonomic distribution. A sequence identification layer, relying on the expert modular annotation system CAZyModO, analyzes new sequence information to assign it to new entries or for addition to already known entries. A biochemical and bibliographic annotation layer correlates individual CAZy entries with biochemical information, and provides the means to gather bibliographic data from different sources including manual. These tools are providing the support for the identification and annotation of carbohydrate active enzymes in newly sequenced genomes and their comparison in what constitutes a new genome and genome comparison layer.

modules were present in CAZy in April 2006, the overall size doubling approximately every 3 years. In CAZy, sequence and structure accessions corresponding to the same protein in a given organism are grouped and form a single entry. Using the relational database, the information may be organized at the protein level but also according to biochemical function of a family, organisms, and even full taxonomic groups. This database is continuously fed with new sequences and biochemical data by two additional independent modules.

An essential layer of sequence comparison and initial modular annotation allows the incorporation of newly released sequences into CAZy (see the Sequence Identification and Update layer in Figure 5.1). With the increased throughput in sequence release by public databases observed in the late 1990s, following genome sequencing, new annotation approaches faster than HCA have been implemented based on gapped BLAST [21]. All the sequences corresponding to the catalytic modules (and ancillary CBMs) of carbohydrate-active enzymes are excised from the full protein sequence and grouped in a BLAST library. This library integrates an evolving annotation system designated CAZyModO (Carbohydrate-Active Enzyme Modular Organization annotation) that is used internally for the analysis of newly released sequences from public databases [17]. Positive hits against this “high-quality” library are entered into the database by trained curators following manual checking on a daily basis. The definition and annotation of the modular organization for the new entries rely on a semi-automatic setup based on comparison with previous annotations,

BLAST-based analysis against family specific libraries [17], and family-derived Hidden Markov models [22].

In parallel with sequence management, biochemical information on carbohydrate-active enzymes is integrated and continuously updated (see the Biochemical and Bibliographic Annotation layer in Figure 5.1). Biochemical characterization of new proteins is essential not only for the identification and creation of new protein families, but also for the annotation of individual protein entries, for activity assignment to previously identified genes, and to update family descriptions [17]. The accumulated information on biochemical characterization found in CAZy helps in genome annotation. The annotation of protein or enzyme activities is a parallel effort to that made by major sequence and structure databases. Annotations in these databases often suffer from bad practices, and lack of enforcement of common annotation standards. Even though rigorous efforts to annotate enzyme activities are made by Swiss-Prot [23], high-throughput sequence release depends strongly on automated sequence annotation. In the majority of newly released sequences, enzyme activities are derived from a small number of biochemically characterized cases and a large number of cases that were themselves inferred following various and disparate criteria. Most of the errors contaminating annotations result from the under- and over-interpretation of similarity analysis, from the insufficient accounting for modularity in many annotation pipelines, and from the use of ambiguous, inappropriate, and/or erroneous nomenclature [24]. Furthermore, errors are often propagated [25] and the identification of the initial error is often impossible, making corrections in public databases very difficult. To avoid similar problems, the practice of annotation in CAZy relies on the systematic removal of all predicted functional annotation from all newly released genomic entries and careful analysis for the other cases. Doubtful cases are checked manually, often ignored, or simply tagged for confirmation by later data, as missing or omitted characterization information is preferred to erroneous annotation.

Enzyme activity annotations currently found in CAZy arise mostly from literature survey and information from collaborators and from the glycobiology community in general. This information is complemented with information resulting from cross-analysis between CAZy entries and entries from other curated databases. In particular, biochemical data extracted from the literature following enzyme and mutant characterization have been made by the Enzymes and Metabolic Pathways database [26] and the Protein Mutant database [27], respectively. Inclusion of reference data compiled by communities centered on model organisms is under consideration for the future. Bibliographic references are included in CAZy with a specific layer that includes over 13 000 different bibliographic references annotated to protein entries. These references were either extracted automatically from individual accessions using the Protein Functional Annotation through Literature tool [28] or entered manually (over 4 000 entries).

A new layer dealing with the analysis of whole protein sets issued from genomes has been evolving more recently (see the Genomes and Comparison layer in Figure 5.1). Modular annotation has in fact been applied to genome data released by the NCBI, with over 300 genomes analyzed. Approximately 2–5% of the proteins encoded by a typical genome are covered by CAZy. More recently, annotation of proteins in recently sequenced genomes prior to full release has been performed by the CAZy team in collaboration with scientists from the Joint Genome Institute. Such annotation benefits from both the CAZyModO system described earlier, but equally from ongoing function prediction efforts relying on subfamily analysis [22]. This annotation method attempts to overcome the common pitfalls in “traditional” genome annotation mentioned earlier.

5.4 Classification and Description of Glycosyltransferases

GTs were originally described in the Enzyme Classification (EC) system [29], a classification system started in the late 1950s that provides a catalog and description of the diversity of enzyme reactions described in the literature. In this system, the enzymes that catalyze the formation of glycosidic bonds are classified as hexosyltransferases (EC 2.4.1.x), pentosyltransferases (EC 2.4.2.x), or other glycosyltransferases (EC 2.4.99.x), depending on the nature of the transferred sugar. Because the EC classification predated structural and detailed enzyme characterizations, the different classes were defined independently of the sugar activator and underlying enzyme mechanism: only donor, acceptor, and product specificity were taken into account, the last with varying degrees of tolerance. This reference classification system therefore presents some shortcomings for the discrimination and compartmentalization of GTs and of carbohydrate-active enzymes in general. The situation is worsened by “accepted” imprecisions in the terminology used by the biological community. For GTs:

1. Only the “macroscopic” aspect of reactions are retained by the EC classification; analogous enzymes are grouped together independently of their structure, molecular mechanism, and evolutionary history;
2. Enzymes not dependent on activated sugars and that “simply” perform transglycosylation, the two-step cleavage and recombination of oligo- and polysaccharides, share similar EC numbers. Up to 20 EC 2.4.1.x numbers correspond to this case and common examples include branching enzymes (EC 2.4.1.8), cyclomaltodextrin glucanotransferases (EC 2.4.1.19), and xyloglucan:xyloglucosyl transferases (EC 2.4.1.207).
3. 1-Phospho-sugar transferases (EC 2.7.8.-) are often misnamed glycosyltransferases, although they do not form new glycosidic bonds.

To simplify their categorization, GTs may be simply assigned into one of two groups according to the change in the relative orientation of the anomeric carbon between the sugar donor and the resulting product [11, 15, 16, 30]. This categorization is based on the pyranose sugar donor depicted in the 4C_1 conformation. The newly formed glycosidic bond can therefore retain or invert the configuration of the anomeric carbon – typically designated α or β but here addressed simply as axial (ax) or equatorial (eq) for systematization purposes – of the original donor sugar-phosphate glycosidic bond, therefore allowing the assignment of the GT catalyzing the transfer reaction into an inverting or a retaining catalytic mechanism, respectively. The nature of the phospho-activated sugar may vary more substantially:

- axial-linked nucleotide-diphospho-sugars (NDP-sugars), where N represents a variety of nucleotide bases as in ADP-, UDP-, GDP-sugars, and so on;
- equatorial-linked nucleotide-monophospho-sugars (NMP-sugars), where N is usually C as for CMP-sugars;
- equatorial-linked dolichyl-diphospho-sugars (DDP-sugars);
- equatorial-linked dolichyl-monophospho-sugars (DMP-sugars), where the dolichyl moiety may be changed to another prenyl such as undecaprenyl or dodecaprenyl groups;
- axial-linked lipid-diphosphate-sugars (LDP-sugars), where the lipid is typically undecaprenol;
- axial-linked sugar-1-phosphates.

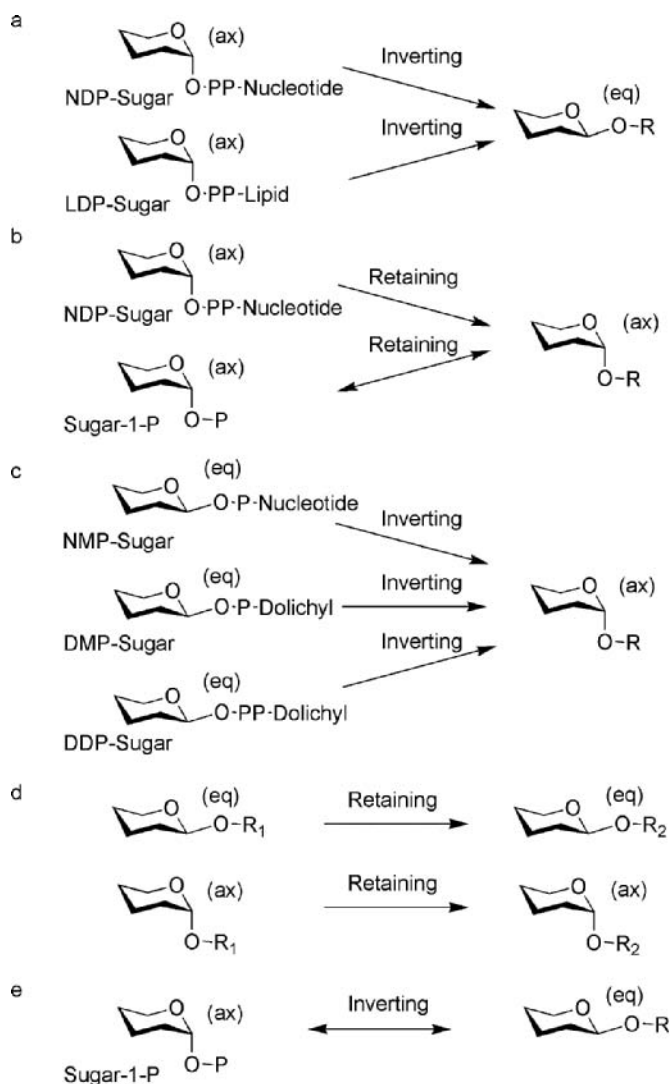


Figure 5.2 Stereochemical outcome of the action of glycosyltransferases, transglycosidases, and related phosphorylases on axial (ax) or equatorial (eq) pyranose-containing substrates and products. Action of: (a) inverting glycosyltransferases using NDP- and LDP-sugar donors; (b) retaining glycosyltransferases and related sugar phosphorylases on NDP-sugar and sugar-1-phosphate donors; (c) inverting glycosyltransferases on NMP-, DMP-, and DDP-sugar donors; (d) retaining transglycosidases on axial and equatorial glycosides; and (e) inverting sugar phosphorylases. Similar relationships may be expected for furanose-containing substrates and products. Note that this classification is based on pyranose sugars in the 4C_1 conformation.

The stereochemical outcome of the reaction catalyzed by GTs and related phosphorylases is shown in Figure 5.2a–c and the observed changes for GHs and related phosphorylases can be seen in Figure 5.2d and e. Unfortunately, transglycosylating enzymes, described in Figure 5.2d, are often designated “glycosyltransferases” and are designated EC 2.4.1.x, and likewise GTs that make use of phospho-activated sugars. Transglycosidases behave very much like retaining glycosidases, a subset of EC 3.2.1.x, the water molecule participating

in catalysis in the latter being replaced by the hydroxyl group of a new carbohydrate in the former [11]. Transglycosidases are in fact known to share structural features with glycosidases and may also exhibit varying levels of hydrolytic activity. Beyond their mechanistic description in Figure 5.2d, these enzymes will not be described in this chapter. Although they share the ability to form new glycosidic bonds, these are obtained after cleaving other glycans, and not by using phospho-activated-sugar donors.

The reaction mechanisms exhibited by known GTs may be constrained by the axial or equatorial nature of the glycosyl-phosphate bond in the activated sugar. The configuration of the anomeric carbon of phospho-activated sugar is axial for NDP-, and LDP-sugars and for sugar-1-phosphate donors whereas for NMP-, DMP-, and DDP-sugar donors it is equatorial when the sugar ring is in the 4C_1 conformation). Not all sugars adopt the 4C_1 conformation; for example, in CMP-Neu5Ac the bond to the CMP is axial in the “natural” 2C_5 ring form (equivalent to 1C_4), where it is viewed as equatorial only if the sugar is drawn in the 5C_2 form as shown in Figure 5.2c (see Figure 5.2a–c). For axial-type NDP- and LDP-sugar donors, both inverting and retaining GT activities have been described (see Figure 5.2a and b), resulting in the formation of new equatorial and axial glycosidic bonds, respectively. However, for the other activated sugar forms, only an inverting mechanism has been described. This is the case with the equatorial-linked NMP-, DMP-, and DDP-sugar donors, resulting in the exclusive formation of axial-type glycosidic bonds (see Figure 5.2c). Lack of information on the nature of the sugar donor or of the configuration of the newly formed glycosidic bond may render difficult the identification of the GT as inverting or retaining. The nature of the sugar donor may be used as a second level of mechanistic classification of GTs.

5.5 Sequence/Structure Classification of Glycosyltransferases

The basis of the sequence-based classification of GTs was established in 1997, following the analysis and classification of NDP-hexosyltransferases into 27 families [15]. The GT classification was subsequently extended to include other phospho-activated pentosyl and other transferases [16]. A total of 84 GT families were found in CAZy as of April 2006, but new GT families are still regularly created. The seed sequences often arise from the set of unclassified GTs already present in CAZy, inclusion of which often awaits complementary characterization data from the literature. Most GT families exhibit more than one EC activity. However, the inclusion of a new, non-biochemically characterized protein in a GT family may already have predictive power. Sequences belonging to a given GT family have:

- a conserved unique fold, very often GT-A or GT-B [16]
- a similar active site architecture for each family with potential similarities between closely related families [15, 16]
- a conserved nature of the activated sugar donor (NDP-, NMP-, DDP-, DMP-, LDP-, or phosphate-1-linked), that often extends to a preferred nucleotide in a given family.

Furthermore, a conserved stereochemical outcome of the reaction, inverting or retaining, is usually observed within a family [15, 16]. However, the structural distinctiveness associated with inverting versus retaining mechanisms observed for GHs [14] appears less strict among GTs. Studies of the closely related families GT2 and GT78 [31] show that very similar active site architectures result in an inverting and a retaining mechanism, respectively, a feature that

may be more general given the low level of biochemically characterized enzymes currently observed. We need, however, to keep in mind that about 7% of all known sequences in the GT classification in CAZy have an attributed EC number, a proportion that decreases rapidly with the flow of uncharacterized data arising from genome sequencing.

GTs are known for a limited structural diversity, as the majority of the structures obtained for different GT families are variations on two folds, GT-A and GT-B [15], exemplified in Figure 5.3a and b. These two folds are both composed of two $\beta/\alpha/\beta$ domains. In fold GT-A these two domains are tightly associated and the nucleotide binding is N-terminal, whereas in fold GT-B the two domains are less tightly associated, and nucleotide-binding takes place in the C-terminal module. Various sequence- and structure-based analyses suggest that many of the GT families still without structural representatives fit into one of these two folds and that at least one more fold of integral membrane GTs is likely to exist [32–34]. This situation contrasts strongly with the rich structural diversity found in other CAZy family classifications, namely in GHs, where more than a dozen folds are known [13, 14].

This chapter resulted from an analysis performed in April 2006. Since then several new families and enzyme activities where added to the different classifications described, including that of glycosyltransferases. For updates, please consult the CAZy database at <http://www.cazy.org>.

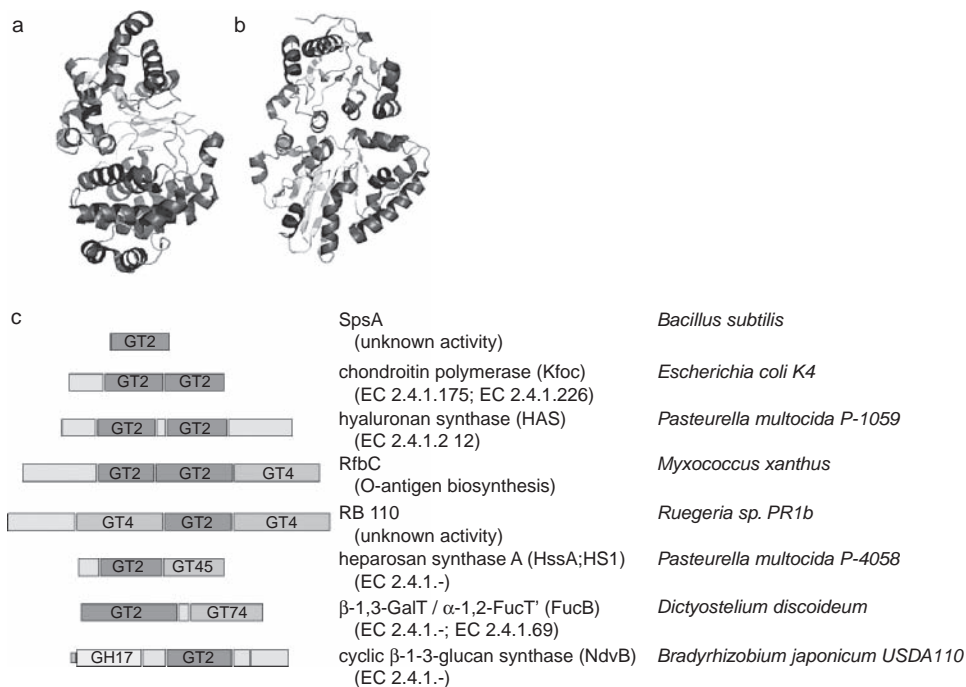


Figure 5.3 Representative folds of glycosyltransferases. (a) A GT-A representative from family GT78: *Rhodothermus marinus* mannosylglycerate synthase [31]. (b) A GT-B representative from family GT80: *Pasteurella multocida* α -2,3-sialyltransferase (Ni *et al.*, 2006). (c) Examples of modular combinations found in glycosyltransferases from family GT2. Individual catalytic modules are identified as boxes. Legend for identified modules: glycosyltransferase, GT; glycoside hydrolase, GH; carbohydrate-binding module, CBM; transmembrane segment, TM. Multiple catalytic GT(+GH) modules often correspond to unique enzymes whose global activity depends on the concerted activity of the individual catalytic domains.

As mentioned earlier, sequence integration in CAZy is centered on modular descriptions. When applied to all GT families, a large number of multi-modular GTs have been unveiled. Even though the full coverage of the GT modular variability exceeds the scope of this chapter, a few examples of GT modular annotation for several GT2 containing proteins are given in Figure 5.3c. GT2 is the largest GT family (see Table 5.1) and was among the first GT families to have a three-dimensional structure determined: the single module *Bacillus subtilis* SpsA enzyme [35]. This structure helped to refine the limits of GT2 modules defined earlier [15]. When applied to the few thousand proteins now present in family GT2, several singular examples of multi-modular enzymes have been revealed. Each catalytic module having a singular GT, activity is often combined with other independent GT modules to produce complex oligo- or polysaccharides in a given pathway. These modules can also be appended to other modules in order to combine their activities. The combination of two different GT2 modules in *Escherichia coli* K4 chondroitin polymerase (EC 2.4.1.175; EC 2.4.1.226) [36] and in *Pasteurella multocida* P-1059 hyaluronan synthase (EC 2.4.1.212) [37] yields enzymes able to produce alternating polymers. Chondroitin and hyaluronan both exhibit a disaccharide repeating unit, whose biosynthesis depends on single-step additions by the two cooperating GT2 modules. Tandem pairs of GT modules can contain modules from different families as in *Pasteurella multocida* P-4058 heparosan synthase A [38], and in the bifunctional β -1,3-galactosyltransferase/ α -1,2-fucosyltransferase (EC 2.4.1.-; EC 2.4.1.69) from *Dictyostelium discoideum* [39], combining a GT2 module with GT45 and GT74 (see Figure 5.3c) modules, respectively. Combinations containing GT2 modules are regularly found in newly sequenced bacterial genomes. A few cases of combination of three GT modules have also been found, such as the *Myxococcus xanthus* O-antigen protein RfbC [40] or the *Ruegeria* sp. *PR1b* ORF RB110, but unfortunately none has yet been biochemically characterized. Other modular combinations are possible where GT modules can be combined with a GH module. An example is given for *Bradyrhizobium japonicum* USDA110 cyclic β -1–3-glucan synthase (EC 2.4.1.-) [41], where a β -1–3-glucan-synthesizing GT2 module is found adjacent to a likely transglycosylating GH17 module that would promote cyclization. Other combinations of GT modules may be found in other CAZy GT families, integrating not only other GTs and GHs, but also carbohydrate-binding modules (CBMs), carbohydrate esterases (CEs), myosin motors, and many other modules.

5.6 Comparing Glycosyltransferase Families

The nature of the activated sugar and the stereochemical outcome of the reaction are known for the majority of GT families. The information on GT families can be found in Table 5.1. This table lists all currently known activities of each family and also its taxonomic and biochemical characterization coverage. This table is too extensive for a complete interpretation in this chapter, but includes the major findings available for each GT family. Major highlights from its analysis are described here.

The first observation is that most GT families use NDP-sugar donors, 41 families presenting an inverting mechanism and only 24 a retaining mechanism. The corresponding families contain approximately 51% and 35% of the classified GTs, respectively. Only seven families (GT29, GT30, GT38, GT42, GT52, GT73, and GT80) rely on NMP-sugars, in fact always CMP-sugars to date, corresponding to roughly 3% of all GTs. All are inverting enzymes, but interestingly, whereas GT42 is structurally related to fold GT-A, the recently

Table 5.1 Description of the known enzymatic activities, nature of sugar donor, stereochemical outcome of reaction, taxonomic and biochemical characterization, coverage of glycosyltransferases, and sugar-phosphorylase families present in CAZy in April 2006. The taxonomic coverage indicates the members of each kingdom present in each family and for Eukaryota the subgroup is indicated only if four or less subgroups are present.

Family	Fold/clan from 3D	Reaction stereo-chemical outcome	Activated sugar (orientation)	Orientation of bond formed	Description (EC nomenclature)	Taxonomic range	Total entries	Entries with EC
Glycosyltransferase families								
GT1	GT-B	Inverting	NDP-sugar (ax)	eq	UDP-glucuronosyltransferase (EC 2.4.1.17); 2-hydroxyacylsphingosine 1- β -galactosyltransferase (EC 2.4.1.45); <i>N</i> -acylsphingosine galactosyltransferase (EC 2.4.1.47); flavonol 3- <i>O</i> -glucosyltransferase (EC 2.4.1.91); indole-3-acetate β -glucosyltransferase (EC 2.4.1.121); sterol glucosyltransferase (EC 2.4.1.173); ecdysteroid UDP-glucosyltransferase (EC 2.4.1.-); zeaxanthin glucosyltransferase (EC 2.4.1.-); zeatin <i>O</i> - β -glucosyltransferase (EC 2.4.1.203); zeatin <i>O</i> - β -xylosyltransferase (EC 2.4.2.40); limonoid glucosyltransferase (EC 2.4.1.210); sinapate 1-glucosyltransferase (EC 2.4.1.120); anthocyanin 3- <i>O</i> -galactosyltransferase (EC 2.4.1.-); anthocyanin 5- <i>O</i> -glucosyltransferase (EC 2.4.1.-); anthocyanidin 3- <i>O</i> -glucosyltransferase (EC 2.4.1.115); dTDP- β -2-deoxy-L-fucose α -L-2-deoxyfucosyltransferase (EC 2.4.1.-); UDP- β -L-rhamnose α -L-rhamnosyltransferase (EC 2.4.1.-); UDP-glucose 4-hydroxybenzoate 4- <i>O</i> - β -glucosyltransferase (EC 2.4.1.194); flavonol L-rhamnosyltransferase (EC 2.4.1.159)	Archaea (14); Bacteria (409); Eukaryota (855); Viruses (39); unclassified (1)	1318	242

GT2	GT-A	Inverting	NDP-sugar (ax)	eq	Cellulose synthase (EC 2.4.1.12); chitin synthase (EC 2.4.1.16); dolichyl-phosphate β -D-mannosyltransferase (EC 2.4.1.83); dolichyl-phosphate β -glucosyltransferase (EC 2.4.1.117); <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.-); <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.-); hyaluronan synthase (EC 2.4.1.212); chitin oligosaccharide synthase (EC 2.4.1.-); β -1,3-glucan synthase (EC 2.4.1.34); β -1,4-mannan synthase (EC 2.4.1.-); β -mannosylphosphodecaprenol manno oligosaccharide α -1,6-mannosyltransferase (EC 2.4.1.199); α -1,3-L-rhamnosyltransferase (EC 2.4.1.-) Glycogen synthase (EC 2.4.1.11)	Archaea (283); Bacteria (3653); Eukaryota (835); Viruses (31)	4802	234
GT3		Retaining	NDP-sugar (ax)	ax	α -1,6-mannosyltransferase (EC 2.4.1.199); α -1,3-L-rhamnosyltransferase (EC 2.4.1.-) Glycogen synthase (EC 2.4.1.11)	Archaea (2); Bacteria (5); Eukaryota (54) {Fungi; Metazoa; Viridiplantae}	61	11
GT4		Retaining	NDP-sugar (ax)	ax	Sucrose synthase (EC 2.4.1.13); sucrose-phosphate synthase (EC 2.4.1.14); α -glucosyltransferase (EC 2.4.1.52); lipopolysaccharide <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.56); GDP-mannose α -mannosyltransferase (EC 2.4.1.-); 1,2-diacylglycerol 3-glucosyltransferase (EC 2.4.1.157); diglucosyl diacylglycerol synthase (EC 2.4.1.208); digalactosyl diacylglycerol synthase (EC 2.4.1.141); trehalose phosphorylase (EC 2.4.1.64); phosphatidylinositol α -mannosyltransferase (EC 2.4.1.57); UDP-galactose α -galactosyltransferase (EC 2.4.1.-)	Archaea (279); Bacteria (3055); Eukaryota (301); Viruses (19); unclassified (1)	3655	115

(Continued)

Table 5.1 (Continued)

Family	Fold/clan from 3D	Reaction stereo-chemical outcome	Activated sugar (orientation) formed	Orientation of bond formed	Description (EC nomenclature)	Taxonomic range	Total entries	Entries with EC
GT5	GT-B	Retaining	NDP-sugar (ax)	ax	UDP-glucose-glycogen glucosyltransferase (EC 2.4.1.11); NDP-glucose-starch glucosyltransferase (EC 2.4.1.242); UDP-glucose α -1,3-glucan synthase (EC 2.4.1.183)	Archaea (12); Bacteria (171); Eukaryota (1023)	1206	66
GT6	GT-A	Retaining	NDP-sugar (ax)	ax	α -1,3-Galactosyltransferase (EC 2.4.1.87); α -1,3- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.40); α -galactosyltransferase (EC 2.4.1.37); globoside α - <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.88)	Eukaryota (82) {Metazoa}; Viruses (1)	83	11
GT7	GT-A	Inverting	NDP-sugar (ax)	eq	Lactose synthase (EC 2.4.1.22); β - <i>N</i> -acetylglucosaminyl-glycopeptide β -1,4-galactosyltransferase (EC 2.4.1.38); <i>N</i> -acetylglucosamine synthase (EC 2.4.1.90); β -1,4- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.-); xylosylprotein β -4-galactosyltransferase (EC 2.4.1.133)	Bacteria (1); Eukaryota (105) {Metazoa}	106	31
GT8	GT-A	Retaining	NDP-sugar (ax)	ax	Lipopolysaccharide galactosyltransferase (EC 2.4.1.44); lipopolysaccharide glucosyltransferase 1 (EC 2.4.1.58); glycogenin glucosyltransferase (EC 2.4.1.186); inositol 1- α -galactosyltransferase (galactinol synthase) (EC 2.4.1.123); homogalacturonan α -1,4-galacturonosyltransferase (EC 2.4.1.43)	Bacteria (169); Eukaryota (219) {Fungi; Metazoa; Viridiplantae}; Viruses (19)	407	34
GT9	GT-B	Inverting	NDP-sugar (ax)	eq	Lipopolysaccharide <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.56); heptosyltransferase (EC 2.4.-.-)	Archaea (3); Bacteria (454)	457	8

GT10	Inverting	NDP-sugar (ax)	eq	Galactoside α -1,3/1,4-L-fucosyltransferase (EC 2.4.1.65); galactoside α -1,3-L-fucosyltransferase (EC 2.4.1.152); glycoprotein α -1,3-L-fucosyltransferase (EC 2.4.1.214)	Bacteria (24); Eukaryota (184) {Metazoa; Mycetozoa; Viridiplantae}; Viruses (1)	209	34
GT11	Inverting	NDP-sugar (ax)	eq	Galactoside α -1,2-L-fucosyltransferase (EC 2.4.1.69)	Bacteria (40); Eukaryota (107) {Euglenozoa; Metazoa}; Viruses (2)	149	43
GT12	Inverting	NDP-sugar (ax)	eq	[<i>N</i> - Acetylneuraminyl]galactosylglucosylceramide <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.92)	Bacteria (3); Eukaryota (7)	10	4
GT13	Inverting	NDP-sugar (ax)	eq	α -1,3-Mannosyl-glycoprotein β -1,2- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.101)	Eukaryota (34) {Metazoa; Viridiplantae}	34	18
GT14	Inverting	NDP-sugar (ax)	eq	β -1,3-Galactosyl- <i>O</i> -glycosyl-glycoprotein β -1,6- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.102); <i>N</i> -acetylglucosaminide β -1,6- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.150); protein <i>O</i> - β -xylosyltransferase (EC 2.4.2.26)	Bacteria (20); Eukaryota (120); Viruses (7)	147	21
GT15	Retaining	NDP-sugar (ax)	ax	Glycolipid 2- α -mannosyltransferase (EC 2.4.1.131); GDP-mannose: α -1,2-mannosyltransferase (EC 2.4.1.-)	Eukaryota (69) {Fungi}	69	6
GT16	Inverting	NDP-sugar (ax)	eq	α -1,6-Mannosyl-glycoprotein β -1,2- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.143)	Eukaryota (17) {Metazoa; Viridiplantae}	17	5

(Continued)

Table 5.1 (Continued).

Family	Fold/clan from 3D	Reaction stereo-chemical outcome	Activated sugar (orientation)	Orientation of bond formed	Description (EC nomenclature)	Taxonomic range	Total entries	Entries with EC
GT17	Inverting	NDP-sugar (ax)	eq	β -1,4-Mannosyl-glycoprotein β -1,4- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.144)	Bacteria (1); Eukaryota (24)	25	3	
GT18	Inverting	NDP-sugar (ax)	eq	α -1,3(6)-Mannosylglycoprotein β -1,6- <i>N</i> -acetyl-glucosaminyltransferase (EC 2.4.1.155)	Eukaryota (8) {Metazoa}	8	6	
GT19	Inverting	NDP-sugar (ax)	eq	Lipid- α -disaccharide synthase (EC 2.4.1.182)	Bacteria (194); Eukaryota (2) {Viridiplantae}	196	7	
GT20	GT-B	NDP-sugar (ax)	ax	α , α -Trehalose-phosphate synthase [UDP-forming] (EC 2.4.1.15)	Archaea (12); Bacteria (98); Eukaryota (110)	220	30	
GT21	Inverting	NDP-sugar (ax)	eq	Ceramide glucosyltransferase (EC 2.4.1.80)	Bacteria (24); Eukaryota (32) {Fungi; Metazoa; Viridiplantae}	56	12	
GT22	Inverting	Dolichyl-P-sugar (eq)	ax	Dolichyl-phosphate-mannose α -mannosyltransferase (EC 2.4.1.-)	Bacteria (2); Eukaryota (85)	87	5	
GT23	Inverting	NDP-sugar (ax)	eq	<i>N</i> -Acetyl- β -D-glucosaminide α -1,6-fucosyltransferase (EC 2.4.1.68)	Bacteria (40); Eukaryota (22) {Metazoa}	62	9	
GT24	Retaining	NDP-sugar (ax)	ax	UDP-glucose glycoprotein α -glucosyltransferase (EC 2.4.1.-)	Eukaryota (35)	35	3	
GT25	Inverting	NDP-sugar (ax)	eq	Lipopolysaccharide biosynthesis protein; β -1,4-galactosyltransferase (EC 2.4.1.-)	Bacteria (168); Eukaryota (28) {Euglenozoa; Fungi; Metazoa}; Viruses (3)	199	6	

GT26	Inverting	NDP-sugar (ax)	eq	UDP-ManNAcA β - <i>N</i> -acetyl mannosaminuronic acid transferase (EC 2.4.1.-); UDP-Glc β -1,4-glucosyltransferase (EC 2.4.1.-)	Bacteria (173)	173	3
GT27	Retaining	NDP-sugar (ax)	ax	Polypeptide α - <i>N</i> -acetylgalactosaminyltransferase (EC 2.4.1.41)	Bacteria (2); Eukaryota {121} {Alveolata; Metazoa}	123	50
GT28	Inverting	NDP-sugar (ax)	eq	1,2-Diacylglycerol 3- β -galactosyltransferase (EC 2.4.1.46); 1,2-diacylglycerol 3- β -glucosyltransferase (EC 2.4.1.157); undecaprenyldiphospho-muramoylpentapeptide β - <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.227)	Archaea (1); Bacteria (328); Eukaryota (16) {Viridiplantae}	345	11
GT29	Inverting	NMP-sugar (eq)	ax	Sialyltransferase (EC 2.4.99.-); β -galactoside α -2,6-sialyltransferase (EC 2.4.99.1); α - <i>N</i> -acetylgalactosaminide α -2,6-sialyltransferase (EC 2.4.99.3); β -galactoside α -2, 3-sialyltransferase (EC 2.4.99.4); <i>N</i> -acetyllactosaminide α -2,3-sialyltransferase (EC 2.4.99.6); (α - <i>N</i> -acetylneuraminyl-2,3- β -galactosyl-1,3)- <i>N</i> -acetylgalactosaminide α -2, 6-sialyltransferase (EC 2.4.99.7); α - <i>N</i> -acetyl-neuraminide α -2, 8-sialyltransferase (EC 2.4.99.8); lactosylceramide α -2,3-sialyltransferase (EC 2.4.99.9)	Eukaryota (274) {Metazoa; Viridiplantae}; Viruses (2)	276	64
GT30	Inverting	NMP-sugar (eq)	ax	α -3-Deoxy-D-manno-octulosonic acid (KDO) transferase (EC 2.-.-.-)	Bacteria (194); Eukaryota (2) {Viridiplantae}	196	19

(Continued)

Table 5.1 (Continued)

Family	Fold/clan from 3D	Reaction stereo-chemical outcome	Activated sugar (orientation)	Orientation of bond formed	Description (EC nomenclature)	Taxonomic range	Total entries	Entries with EC
GT31	Inverting	NDP-sugar (ax)	eq	<i>N</i> -Acetyllactosaminide β-1,3- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.149); glycoprotein- <i>N</i> -acetylgalactosamine 3-β-galactosyltransferase (EC 2.4.1.122); fucose-specific β-1,3- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.-); globotriosylceramide β-1,3-GalNAc transferase (EC 2.4.1.79); chondroitin synthase (β-1,3-GlcUA) and β-1,4-GalNAc transferase (EC 2.4.1.-); chondroitin β-1,4- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.-)	Eukaryota (325)	325	45	
GT32	Retaining	NDP-sugar (ax)	ax	α-1,6-Mannosyltransferase (EC 2.4.1.-); α-1,4- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.-); α-1,4- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.-); GDP-mannose:chitobiosyldiphosphodolichol β-mannosyltransferase (EC 2.4.1.142) UDP-galactose:galactomannan α-1,6-galactosyltransferase (EC 2.4.1.-); UDP-xylose:xyloglucan α-1,6-xylosyltransferase (EC 2.4.2.39); α-1,2-galactosyltransferase (EC 2.4.1.-)	Bacteria (93); Eukaryota (88) {Fungi; Metazoa; Viridiplantae}; Viruses (1) Eukaryota (32) Eukaryota (62)	182	9	
GT33	Inverting	NDP-sugar (ax)	eq	Glycogen and starch phosphorylase (EC 2.4.1.1)	Archaea (52); Bacteria (232); Eukaryota (91)	375	50	
GT34	Retaining	NDP-sugar (ax)	ax	Renamed; now family GH94 (see below) Galactoside 2- <i>L</i> -fucosyltransferase (EC 2.4.1.69) Polysialyltransferase (EC 2.4.-.-)	Eukaryota (31) {Viridiplantae} Bacteria (16)	62	3	
GT35	Retaining	Phospho-sugar (ax)	ax					
GT36	Inverting	NDP-sugar (ax)	eq					
GT37	Inverting	NMP-sugar (eq)	ax					
GT38	Inverting	NMP-sugar (eq)	ax					

GT39	Inverting	DMP-sugar (eq)	ax	Dolichyl-phosphate-mannose-protein mannosyltransferase (EC 2.4.1.109)	Bacteria (27); Eukaryota (70) {Fungi; Metazoa}	97	15
GT40	Inverting	NDP-sugar (ax)	eq	β -1,3-Galactofuranosyltransferases (EC 2.4.1.-)	Eukaryota (11) {Euglenozoa}	11	2
GT41	Inverting	NDP-sugar (ax)	eq	UDP- <i>N</i> -acetylglucosamine:peptide <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.94)	Bacteria (80); Eukaryota (31) {Fungi; Metazoa; Viridiplantae}	111	4
GT42	Related to GT-A	NMP-sugar (eq)	ax	α -2,3-Sialyltransferase (EC 2.4.99.-)	Bacteria (31); Eukaryota (27)	58	15
GT43	Inverting	NDP-sugar (ax)	eq	β -Glucuronyltransferase (EC 2.4.1.135)	Eukaryota (125)	125	6
GT44	Retaining	NDP-sugar (ax)	ax	UDP-glucose glucosyltransferase (EC 2.4.1.-); UDP-GlcNAc GlcNAc-transferase (EC 2.4.1.-)	Bacteria (31)	31	5
GT45	Retaining	NDP-sugar (ax)	ax	α -GlcNAc transferase (EC 2.4.1.-)	Bacteria (8)	8	4
GT46	unknown	unknown		Putative glycosyltransferases	Bacteria (15)	15	0
GT47	Inverting	NDP-sugar (ax)	eq	Heparan β -glucuronyltransferase (EC 2.4.1.225); xyloglucan β -galactosyltransferase (EC 2.4.1.-); heparan synthase (EC 2.4.1.-); arabinan α -L-arabinosyltransferase (EC 2.4.2.-); 1,3- β -D-Glucan synthases (EC 2.4.1.34)	Eukaryota (113) {Fungi; Metazoa; Viridiplantae}	113	8
GT48	Inverting	NDP-sugar (ax)	eq	β -1,3- <i>N</i> -Acetylglucosaminyltransferases (EC 2.4.1.-)	Eukaryota (109) {Fungi; Viridiplantae}	109	11
GT49	Inverting	NDP-sugar (ax)	eq	β -1,3- <i>N</i> -Acetylglucosaminyltransferases (EC 2.4.1.-)	Eukaryota (30) {Metazoa; Mycetozoa}	30	1
GT50	Inverting	DMP-sugar (eq)	ax	Dolichyl-phosphate-mannose α -1,4-mannosyltransferase (EC 2.4.1.-)	Eukaryota (35)	35	2

(Continued)

Table 5.1 (Continued)

Family	Fold/clan from 3D outcome	Reaction stereo-chemical outcome	Activated sugar (orientation)	Orientation of bond formed	Description (EC nomenclature)	Taxonomic range	Total entries	Entries with EC
GT51	Inverting	LDP-sugar (ax)	eq	Murein polymerases (EC 2.4.1.129)	Bacteria (1008); Eukaryota (1) {Viridiplantae}	1009	9	
GT52	Inverting	NMP-sugar (eq)	ax	α -2,3-Sialyltransferase (EC 2.4.99.4); α -glucosyltransferase (EC 2.4.1.-)	Bacteria (46)	46	5	
GT53	Inverting	NDP-sugar (ax)	eq	UDP-L-arabinose:arabinosyltransferase (EC 2.4.2.-)	Bacteria (41)	41	4	
GT54	Inverting	NDP-sugar (ax)	eq	UDP- <i>N</i> -acetylglucosamine: α -1,3-D-mannoside β -1,4- <i>N</i> -acetylglucosaminyltransferase (EC 2.4.1.145)	Eukaryota (27) {Metazoa}	27	4	
GT55	Retaining	NDP-sugar (ax)	ax	GDP-mannose:mannosyl-3-phosphoglycerate synthase (EC 2.4.1.217)	Archaea (11); Bacteria (5); Eukaryota (2) {Fungi}	18	6	
GT56	Inverting	NDP-sugar (ax)	eq	TDP-Fuc4NAc:lipid II Fuc4NAc transferase (EC 2.4.1.-)	Bacteria (27)	27	1	
GT57	Inverting	DMP-sugar (eq)	ax	Dolichyl-phosphate-glucose α -1,3-glucosyltransferase (EC 2.4.1.-)	Eukaryota (50)	50	3	
GT58	Inverting	DMP-sugar (eq)	ax	Dol-P-mannose:dolichol pyrophosphate-mannose α -1,3-mannosyltransferase (EC 2.4.1.130); Dol-P-mannose:dolichol pyrophosphate-Man5GlcNAc2 α -1,2-mannosyltransferase (EC 2.4.1.130)	Eukaryota (25)	25	3	
GT59	Inverting	DMP-sugar (eq)	ax	Dolichyl-phosphate-glucose α -1,2-glucosyltransferase (EC 2.4.1.-)	Eukaryota (21)	21	1	
GT60	Retaining	NDP-sugar (ax)	ax	UDP-GlcNAc:hydroxyproline polypeptide α -GlcNAc-transferase (EC 2.4.1.-)	Bacteria (10); Eukaryota (6) {Euglenozoa; Mycetozoa}	16	1	

GT61	Inverting	NDP-sugar (ax)	eq	β -1,2-Xylosyltransferase (EC 2.4.2.38)	Eukaryota (95) {Fungi; Metazoa; Viridiplantae}	95	4
GT62	Retaining	NDP-sugar (ax)	ax	α -1,2-Mannosyltransferase (EC 2.4.1.-); α -1,6-mannosyltransferase (EC 2.4.1.-)	Eukaryota (39) {Fungi}	39	1
GT63	Inverting	NDP-sugar (ax)	eq	DNA β -glucosyltransferase (EC 2.4.1.27)	Viruses (1)	1	1
GT64	Retaining	NDP-sugar (ax)	ax	Heparan α -N-acetylhexosaminyltransferase (EC 2.4.1.224)	Eukaryota (44) {Metazoa; Viridiplantae}	44	7
GT65	Inverting	NDP-sugar (ax)	eq	GDP-fucose:protein <i>O</i> - α -fucosyltransferase (EC 2.4.1.-)	Eukaryota (29) {Metazoa; Viridiplantae}	29	6
GT66	Inverting	DDP-sugar (eq)	ax	Oligosaccharyl transferase (EC 2.4.1.119)	Archaea (34); Bacteria (11); Eukaryota (49)	94	0
GT67	Inverting	NDP-sugar (ax)	eq	Phosphoglycan β -1,3-galactosyltransferase (EC 2.4.1.-)	Eukaryota (33) {Euglenozoa}	33	0
GT68	Inverting	NDP-sugar (ax)	eq	GDP-fucose:protein <i>O</i> - α -fucosyltransferase (EC 2.4.1.-)	Eukaryota (26) {Alveolata; Metazoa; Viridiplantae}	26	1
GT69	Retaining	NDP-sugar (ax)	ax	GDP-mannose: α -1,3-mannosyltransferase (EC 2.4.1.-)	Eukaryota (37) {Fungi; Euglenozoa}	37	1
GT70	Inverting	NDP-sugar (ax)	eq	UDP-GlcA β -glucuronosyltransferase (EC 2.4.1.-)	Bacteria (10)	10	1
GT71	Retaining	NDP-sugar (ax)	ax	α -Mannosyltransferase (EC 2.4.1.-)	Eukaryota (65) {Fungi}	65	6
GT72	Retaining	NDP-sugar (ax)	ax	DNA α -glucosyltransferase (EC 2.4.1.26)	Viruses (2)	2	1
GT73	Inverting	NMP-sugar (eq)	ax	α -3-Deoxy-D-manno-octulosonic-acid (KDO) transferase (EC 2.-.-)	Bacteria (17)	17	2
GT74	Inverting (inferred)	NDP-sugar (ax)	eq	α -1,2- <i>L</i> -Fucosyltransferase (EC 2.4.1.69)	Bacteria (1); Eukaryota (1) {Mycetozoa}	2	1

(Continued)

Table 5.1 (Continued)

Family	Fold/clan from 3D	Reaction stereo-chemical outcome	Activated sugar (orientation)	Orientation of bond formed	Description (EC nomenclature)	Taxonomic range	Total entries	Entries with EC
GT75		Inverting	NDP-sugar (ax)	eq	Self-glucosylating UDP-glucose β -glucosyltransferase (EC 2.4.1.-)	Archaea (2); Bacteria (2); Eukaryota (25) {Viridiplantae}	29	1
GT76		Inverting	DMP-sugar (eq)	ax	Dolichyl-phosphate-mannose α -1,6-mannosyltransferase (EC 2.4.1.-)	Eukaryota (34)	38	2
GT77		Retaining	NDP-sugar (ax)	ax	α -Xylosyltransferase (EC 2.4.2.39); α -1,3-galactosyltransferase (EC 2.4.1.37)	{Viridiplantae}	34	3
GT78	GT-A	Retaining	NDP-sugar (ax)	ax	GDP-mannose α -mannosyltransferase (mannosylglycerate synthase) (EC 2.4.1.-)	Bacteria (1); Eukaryota (1) {Rhodophyta}	2	1
GT79		Retaining	NDP-sugar (ax)	ax	GDP- α -D-arabinose phosphoglycan α -1,2-arabinopyranosyltransferase I (EC 2.4.2.-)	Eukaryota (3) {Euglenozoa}	3	1
GT80	GT-B	Inverting	NMP-sugar (eq)	ax	β -Galactoside α -2,6-sialyltransferase (EC 2.4.99.1); β -galactoside α -2,3-sialyltransferase (EC 2.4.99.4)	Bacteria (3)	3	2
GT81		Retaining	NDP-sugar (ax)	eq	GDP-Glc: α -glucosyl-3-phosphoglycerate synthase (EC 2.4.1.-)	Archaea (4); Bacteria (13)	17	1
GT82		Inverting	NDP-sugar (ax)	eq	UDP-GalNAc β -1,4- <i>N</i> -acetylgalactosaminyltransferase (EC 2.4.1.-)	Bacteria (14); unclassified (8)	22	1
GT83		Inverting	DMP-sugar (eq)	ax	Undecaprenyl phosphate-L-Ara4N:4-amino-4-deoxy- β -L-arabinosyltransferase (EC 2.4.2.-); dodecaprenyl phosphate- β -galacturonic acid:lipopolysaccharide core α -galacturonosyl transferase (EC 2.4.1.-)	Bacteria (130)	130	7
GT84		Inverting	NDP-sugar (ax)	eq	Cyclic β -1,2-glucan synthase (EC 2.4.1.-)	Bacteria (23)	23	2
GT*	-	-	-	-	Unclassified glycosyltransferases awaiting further characterization and/or other seeding sequences. Some sequences may exhibit weak similarity to established GT families	Archaea (14); Bacteria (109); Eukaryota (15); Virus (1)	138	3

Phosphorylase-containing GH families								
GH13	GH-H	Retaining	Sugar (ax); phospho- sugar (ax)	ax	α -Amylase (EC 3.2.1.1); pullulanase (EC 3.2.1.41); cyclomaltodextrin glucanotransferase (EC 2.4.1.19); cyclomaltodextrinase (EC 3.2.1.54); trehalose-6-phosphate hydrolase (EC 3.2.1.93); oligo- α -glucosidase (EC 3.2.1.10); maltogenic amylase (EC 3.2.1.133); neopullulanase (EC 3.2.1.135); α -glucosidase (EC 3.2.1.20); maltotetraose-forming α -amylase (EC 3.2.1.60); isoamylase (EC 3.2.1.68); glucodextranase (EC 3.2.1.70); maltohexaose-forming α -amylase (EC 3.2.1.98); branching enzyme (EC 2.4.1.18); trehalose synthase (EC 5.4.99.16); 4- α -glucanotransferase (EC 2.4.1.25); maltopentaose-forming α -amylase (EC 3.2.1.-); amylosucrase (EC 2.4.1.4); sucrose phosphorylase (EC 2.4.1.7); malto-oligosyltrehalose trehalohydrolase (EC 3.2.1.141); isomaltulose synthase (EC 5.4.99.11)	Archaea (65); Bacteria (1541); Eukaryota (979); unclassified (6)	2591	542
GH65	GH-L	Inverting	Sugar (ax); phospho- sugar (eq)	ax	Trehalase (EC 3.2.1.28); maltose phosphorylase (EC 2.4.1.8); trehalose phosphorylase (EC 2.4.1.64); kojibiose phosphorylase (EC 2.4.1.230)	Archaea (1); Bacteria (80); Eukaryota (25) {Euglenozoa; Fungi; Metazoa}	106	13
GH94	Related to GH-L	Inverting	Phospho- sugar (ax)	eq	Cellobiose phosphorylase (EC 2.4.1.20); cellodextrin phosphorylase (EC 2.4.1.49); chitobiose phosphorylase (EC 2.4.1.-); cyclic β -1,2-glucan synthase (EC 2.4.1.-)	Bacteria (64); Eukaryota (2) {Fungi}	66	14

characterized family GT80 clearly belongs to fold GT-B. No NMP-sugar families have yet been discovered in *Archaea*.

Among the phospho-lipid-activated GTs, seven families (GT22, GT39, GT50, GT57, GT58, GT59, and GT76) are known to use equatorial-type DMP-sugars, one (GT66) employs equatorial-type DDP-sugars, and one (GT51) uses axial-type LDP-sugars. They contain approximately 2, 0.5, and 5% of all classified GTs, respectively. DMP-sugars are mostly found in *Eukaryota* but families GT22 and GT39 have bacterial proteins, whereas family GT83 is exclusively bacterial. Interestingly, family GT83 uses undecaprenyl- or dodecaprenyl-activated sugars and not dolichyl-sugars as apparently found elsewhere. The single DDP-sugar-dependent family GT66 is found in all kingdoms, except viruses. The LDP-sugar-dependent murein polymerase GT51 is found only in bacteria.

The single phospho-sugar-retaining GT35 family, whose mechanism is similar to that of other GTs in Figure 5.2b, represents 2% of all classified GTs as it is present in most organisms relying on glycogen or starch. Interestingly, other phosphate-sugar-acting phosphorylases are present in other CAZy families, namely in GH families GH13, GH65, and GH94 (also shown in Table 5.1). Family GH94, previously named family GT36, was moved to the GH classification based on obvious structural and mechanistic similarities with other GH families [42, 43], and is the only currently known family exhibiting the mechanism described in Figure 5.2e.

Among the GT families having at least one structural representative, the GT-A fold is represented by more than 10 families present in Table 5.1. All characterized members from families presenting this fold make use of NDP-sugars, but four of these families are inverting and seven are retaining (see Figure 5.2a and b). At the taxonomy level, we can observe that the only archaeal proteins in a family containing a structure presenting the GT-A fold are found in family GT2.

The GT-B fold has been described for almost 10 other structurally characterized families. Interestingly, four families are retaining and five inverting. Moreover, most of these families rely on NDP-sugars, but two exceptions are found: NMP-sugars for inverting family GT80, and phospho-sugars for retaining family GT35, which have in common the formation of an axial bond (see Figure 5.1b and c). Structural and mechanistic similarities between families GT35 and GT20 have been already described [44].

The fact that both retaining and inverting enzymes are found within the same fold, and donors of different nature are found within a fold (as in GT-B described above), suggests an ancient and complex evolutionary history for these enzymes. It is likely that GTs originated from a few independent folds at different moments in time, that diverged to present-day families. Unlike what has been observed for the GHs, where mechanistic shifts from retaining to inverting anomeric configuration have rarely been observed [43], it appears that such shifts among GTs have been more frequent and this probably reflects (i) fundamental differences between the molecular mechanisms of GHs and of GTs and (ii) that it probably takes only a minor adjustment of the positioning of the acceptor to transform a retaining into an inverting GT.

5.7 Final Remarks

Only 7% of GTs have an associated EC number in CAZy. This small proportion results from the practical difficulties in determining GT activities and from our strict policy

of not extrapolating EC descriptions based solely on sequence similarity. Biochemical characterization of GTs presents different degrees of difficulty depending on the nature of the enzyme, cellular location, and type of organism. Only a few small families whose scope is taxonomically limited have higher levels of biochemical characterization. In most cases, the lack of characterization comes from practical problems, for instance, the simultaneous need to define an activated sugar donor and an acceptor. If on the one hand bacterial and archaeal proteins may be simpler to produce and characterize, difficulties exist in defining the role of these intracellular proteins in the metabolism, and therefore in the choice of the corresponding donors and acceptors for testing. This problem is usually worse in Eukaryotes, where multiple isoforms, specific cellular location in organelles, and membrane anchoring or integration are more prevalent. The compartmentalization of GTs, particularly observed in Eukaryotes, may be at the origin of *in vitro* characterization limitations. In practice, GT specificity *in vivo* may depend on its cellular localization, and on a corresponding limited set of donor and acceptor molecules, rather than only on actual active site architecture. *In vitro* studies carried out with MGS synthase, the only known example from family GT78, show that the enzyme displays a fairly wide donor and acceptor specificity that was evidenced by the use of high-throughput parallel assays and that would have perhaps escaped more traditional assays [31].

These facts suggest that GT specificity prediction, either sequence or structure-based, is and will remain a hot issue in the years to come. Enzyme specificity will present different meanings in the context of physiological or *in vivo* characterization of individual enzymes and pathways that is likely to contrast with *in vitro* or recombinant usage in a biotechnological context.

A large number of uncharacterized GTs are now known. As the characterization of all proteins is likely to remain a challenging issue in the near future, one can foresee that the hypothetical number of different GT activities will exceed that of GHs. Furthermore, the resulting large variety of donors and acceptors may indeed create problems in establishing new EC numbers in the future. To prepare for the biochemical characterization challenges of the future, the conservative CAZy annotation system, which relies on sequence conservation rather than on specificity, is part of the answer. CAZy is at present the most complete resource on GTs, gathering the interests of “independent” sub-communities within glycobiology. Hopefully, it will accompany a sorely needed standardization in family, mechanistic, and genomic nomenclature in glycobiology.

Abbreviations

BLAST	Basic Local Alignment Search Tool
CAZy	Carbohydrate-Active Enzymes database
CAZyModO	Carbohydrate-Active enzyme modular organization annotation
CBM	carbohydrate-binding module
CE	carbohydrate esterase
GH	glycoside hydrolase
DMP, DDP	dolichyl mono- and diphosphate-sugars
EC	Enzyme Classification
EMP	Enzymes and Metabolic Pathways database
GT	glycosyltransferase

HCA	hydrophobic cluster analysis
NCBI	National Center for Biotechnology Information
NMP, NDP	nucleotide mono- and diphosphate-sugars designated
PL	polysaccharide lyase
PDB	Protein Databank
PMD	Protein Mutant Database

Acknowledgments

EGJD, PMC and BH wish to acknowledge financial support from the European Commission (STREP FungWall grant, contract: LSHB-CT-2004-511952) and the French Ministry of Research (programme ACI-BCMS, Enzywall). FMC and PMC thank the program Actions Universitaires Intégrées Luso-françaises for financial support. Yves Bourne is thanked for reviewing the manuscript.

References

1. Laine R: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05×10^{12} structures for a reducing hexasaccharide: the isomer barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* 1994, **4**:759–767.
2. Helenius A, Aebi M: Intracellular functions of N-linked glycans. *Science* 2001, **291**:2364–2369.
3. Bertozzi C, Kiessling L: Chemical glycobiology. *Science* 2001, **291**:2357–2364.
4. Crocker P, Feizi T: Carbohydrate recognition systems: functional triads in cell–cell interactions. *Curr Opin Struct Biol* 1996, **6**:679–691.
5. Gabius H: Glycobiology of host defense mechanisms. In *Glycociences* (eds H Gabius, S Gabius). London: Chapman & Hall; 1997: pp. 497–506.
6. Rudd P, Elliott T, Cresswell P, Wilson I, Dwek R: Glycosylation and the immune system. *Science* 2001, **291**:2370–2376.
7. Henrissat B: A classification of glycosyl hydrolases based on amino-acid sequence similarities. *Biochem J* 1991, **280**:309–316.
8. Henrissat B, Bairoch A: New families in the classification of glycosyl hydrolases based on amino-acid sequence similarities. *Biochem J* 1993, **293**:781–788.
9. Henrissat B, Bairoch A: Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* 1996, **316**:695–696.
10. Henrissat B, Bork P: On the classification of modular proteins. *Protein Eng* 1996, **9**:725–726.
11. Lairson L, Withers S: Mechanistic analogies amongst carbohydrate modifying enzymes. *Chem Commun* 2004, 2243–2248.
12. Yip V, Withers S: Breakdown of oligosaccharides by the process of elimination. *Curr Opin Chem Biol* 2006, **10**:147–155.
13. Davies G, Henrissat B: Structures and mechanisms of glycosyl hydrolases. *Structure* 1995, **3**:853–859.
14. Henrissat B, Davies G: Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol* 1997, **7**:637–644.
15. Campbell J, Davies G, Bulone V, Henrissat B: A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J* 1997, **326**:929–939.
16. Coutinho P, Deleury E, Davies G, Henrissat B: An evolving hierarchical family classification for glycosyltransferases. *J Mol Biol* 2003, **328**:307–317.

17. Coutinho P, Henrissat B: Carbohydrate-active enzymes: an integrated database approach. In *Recent Advances in Carbohydrate Bioengineering* (eds H Gilbert, G Davies, B Henrissat, B Svensson). Cambridge: Royal Society of Chemistry; 1999, pp. 3–12.
18. Boraston A, Bolam D, Gilbert H, Davies G: Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* 2004, **382**:769–781.
19. Tomme P, Warren R, Miller R Jr, Kilburn D, Gilkes N: Cellulose-binding domains: classification and properties. In *Enzymatic Degradation of Insoluble Polysaccharides* (eds J Saddler, M Penner). Washington, DC: American Chemical Society; 1995: pp. 142–163.
20. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon J: Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* 1997, **53**:621–645.
21. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, **25**:3389–3402.
22. Stam M, Danchin E, Rancurel C, Coutinho P, Henrissat B: Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng Des Sel* 2006, **19**:555–562.
23. Bairoch A, Boeckmann B, Ferro S, Gasteiger E: Swiss-Prot: juggling between evolution and stability. *Brief Bioinformatics* 2004, **5**:39–55.
24. Doerks T, Bairoch A, Bork P: Protein annotation: detective work for function prediction. *Trends Genet* 1998, **14**:248–250.
25. Gilks W, Audit B, De Angelis D, Tsoka S, Ouzounis C: Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* 2002, **18**:1641–1649.
26. Selkov E, Basmanova S, Gaasterland T, Goryanin I, Gretchkin Y, Maltsev N, Nenashev V, Overbeek R, Panyushkina E, Pronevitch L, *et al.*: The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucleic Acid Res* 1996, **24**:26–28.
27. Kawabata T, Ota M, Nishikawa K: The Protein Mutant Database. *Nucleic Acids Res* 1999, **27**:355–357.
28. Couto F, Silva M, Coutinho P: ProFAL: Protein Functional Annotation through Literature. In *VIII Conference on Software Engineering and Databases (JISBD)*, Alicante, Spain: 2003.
29. NC-IUBMB: *Enzyme Nomenclature. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. London: Academic Press; 1992.
30. Davies MJ, Hounsell EF: HPLC and HPAEC of oligosaccharides and glycopeptides. *Methods Mol Biol* 1998, **76**:79–100.
31. Flint J, Taylor E, Yang M, Bolam D, Tailford L, Martinez-Fleites C, Dodson E, Davis B, Gilbert H, Davies G: Structural dissection and high-throughput screening of mannosylglycerate synthase. *Nat Struct Mol Biol*. 2005, **12**:608–614.
32. Franco O, Rigden D: Fold recognition analysis of glycosyltransferase families: further members of structural superfamilies. *Glycobiology* 2003, **13**:707–712.
33. Liu J, Mushegian A: Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci* 2003, **12**:1418–1431.
34. Rosen M, Edman M, Sjoström M, Wieslander A: Recognition of fold and sugar linkage for glycosyltransferases by multivariate sequence analysis. *J Biol Chem* 2004, **279**:38683–38692.
35. Charnock S, Davies G: Structure of the nucleotide-diphospho-sugar transferase, SpsA from *Bacillus subtilis*, in native and nucleotide-complexed forms. *Biochemistry* 1999, **38**:6380–6385.
36. Ninomiya T, Sugiura N, Tawada A, Sugimoto K, Watanabe H, Kimata K: Molecular cloning and characterization of chondroitin polymerase from *Escherichia coli* strain K4. *J Biol Chem* 2002, **277**:21567–21575.
37. Williams K, Halkes K, Kamerling J, Deangelis P: Critical elements of oligosaccharide acceptor substrates for the *Pasteurella multocida* hyaluronan synthase. *J Biol Chem* 2006, **281**:5391–5397.

38. DeAngelis P, White C: Identification and molecular cloning of a heparosan synthase from *Pasteurella multocida* type D. *J Biol Chem* 2002, **277**:7209–7213.
39. Van Der Wel H, Fisher S, West C: A bifunctional diglycosyltransferase forms the Fuc-a-1,2-Gal- β -1, 3-disaccharide on Skp1 in the cytoplasm of dictyostelium. *J Biol Chem* 2002, **277**:46527–46534.
40. Guo D, Bowden M, Pershad R, Kaplan H: The *Myxococcus xanthus* rfbABC operon encodes an ATP-binding cassette transporter homolog required for O-antigen biosynthesis and multicellular development. *J Bacteriol* 1996, **178**:1631–1639.
41. Chen R, Bhagwat A, Keister D: A motility revertant of the ndvB mutant of *Bradyrhizobium japonicum*. *Curr Microbiol* 2003, **47**:431–433.
42. Honda Y, Kitaoka M, Hayashi K: Reaction mechanism of chitobiose phosphorylase from *Vibrio proteolyticus*: identification of family 36 glycosyltransferase in *Vibrio*. *Biochem J*. 2004, **377**:225–232.
43. Stam M, Blanc E, Coutinho P, Henrissat B: Evolutionary and mechanistic relationships between glycosidases acting on alpha- and beta-bonds. *Carbohydr. Res.* 2005, **340**:2728–2734.
44. Gibson R, Turkenburg J, Charnock S, Lloyd R, Davies G: Insights into trehalose synthesis provided by the structure of the retaining glucosyltransferase OtsA. 6. *Chem Biol* 2002, **9**:1337–1346.

6 Other Databases Providing Glycoenzyme Data

Thomas Lütteke¹ and Claus-Wilhelm von der Lieth²

¹*Faculty of Veterinary Medicine, Institute of Biochemistry and Endocrinology, Justus-Liebig University Gießen, 35392 Gießen, Germany*

²*Formerly at the Central Spectroscopic Unit, Deutsches Krebsforschungszentrum (German Cancer Research Center), 69120 Heidelberg, Germany*

In addition to CAZy [1] (see Chapter 5), there are several other databases containing information on glycoenzymes available on the Internet. These can be divided into two categories: general protein databases and databases that specialize in glycoenzymes or contain an individual subsection on glycoenzymes. In this chapter, these databases are briefly discussed.

6.1 GlycoGene DataBase

The GlycoGene DataBase [2] (GGDB; for URLs, see Table 6.1) provides information merely on enzymes involved in carbohydrate biosynthesis. In contrast to CAZy, no hydrolases are included. The database can be queried by carbohydrate structures, by protein sequence, or using a keyword search. A search by tissue type is also offered. In addition, the database can be navigated by enzyme families such as fucosyltransferases, sulfotransferases, or enzymes involved in GAG biosynthesis. For the single enzymes, information such as gene and protein name, synonyms, substrate specificity, tissue distribution, mRNA and protein sequences, and links to external sources are given. Sequence data and species-specific links cover human enzymes only. The respective data for other species (mouse, rat, *Drosophila melanogaster*, *Caenorhabditis elegans*, and yeast) are given in a separate “homologous gene” menu.

6.2 KEGG

The *Kyoto Encyclopedia of Genes and Genomes* [3] (KEGG) (see Chapter 7) is not focused on glycoenzyme data exclusively. Nevertheless, some subcategories contain merely glycoenzymes. In addition to enzymes involved in glycan biosynthesis, KEGG also includes enzymes that participate in monosaccharide metabolism. In addition to general data such as

Table 6.1 Overview of the databases described in this chapter. CAZy is described in detail in Chapter 5 and is added to this table for the sake of completeness.

Database	Comment	URL	Ref.
Databases specializing in glycoenzymes or with glycoenzyme section			
CAZy	Data on glycosyltransferases and hydrolases	www.cazy.org	[1]
GlycoGene DB	Glycosyltransferase data	riodb.ibase.aist.go.jp/rcmg/ggdb/	[2]
KEGG	Mainly focused on reaction pathways.	www.genome.jp/kegg/	[3]
CFG GlycoEnzyme	Mainly focused on reaction pathways.	www.functionalglycomics.org (→ cfg databases → glycosyltransferases)	[11]
CFG Microarray	Gene expression from microarray experiments	www.functionalglycomics.org (→ cfg data → gene microarray)	[4]
General databases			
Swiss-Prot/TrEMBL	General information on the proteins	www.expasy.org/sprot/	[6]
ENZYME	Nomenclature of enzyme families, catalyzed reactions	www.expasy.org/enzyme/	[7]
BRENDA	Notation, catalyzed reactions, experimental properties	www.brenda-enzymes.org/	[8]
NCBI databases	Sequence databases	www.ncbi.nlm.nih.gov	[9]
PDB	Protein 3D structures	www.pdb.org	[10]

gene and protein names, sequences and links to external databases, the entries in KEGG also contain some links to KEGG-specific resources such as pathway or gene orthology data.

Glycoenzyme data in KEGG can be accessed in several ways. In the “Carbohydrate Metabolism” section of KEGG Orthology (KO), entries are categorized in groups such as citrate cycle or galactose metabolism, whereas the “Glycan Biosynthesis and Metabolism” section covers glycosyltransferases grouped by the different glycan types such as *N*-glycan biosynthesis, heparan sulfate biosynthesis, or lipopolysaccharide biosynthesis. A keyword search and searches by amino acid or nucleotide sequence using BLAST or FASTA are also provided. Further, KEGG Pathway offers an easy means to survey the possibility of accessing the enzymes involved in a certain step of glycan biosynthesis. The categories by which the pathways are arranged are the same as in KEGG Orthology.

6.3 CFG GlycoEnzyme

Recently, the Consortium for Functional Glycomics (CFG) also launched a pathway-based database of glycosyltransferases, CFG GlycoEnzyme. The pathways differ from the KEGG Pathway in such a way that they are divided into core pathways and extension pathways. This reflects the fact that glycan families such as *N*-glycan structures, mucin-type *O*-glycan structures and glycosphingolipids each share a fixed core structure, to which various extensions can be added. This concept makes the single pathways simpler and more

flexible than those present in KEGG. On the other hand, to assign the enzymes that are involved in the biosynthesis of an oligosaccharide structure, often several pathways have to be combined.

From the pathway maps or using a keyword search, the molecule pages of the single glycosyltransferases can be accessed. In addition to general data such as amino acid and nucleotide sequences or links to external resources, links to all pathways in which an enzyme is involved and to other CFG data are provided.

6.4 CFG Microarray Data

Another glycoenzyme resource provided by the Consortium for Functional Glycomics is CFG Microarray Data [4]. The CFG glycogene chips include human and murine genes encoding for proteins involved in glycan biosynthesis and degradation and also for glycan-binding proteins. A variety of human and murine samples of diverse tissues have been analyzed using the glycogene chips. The results of these experiments can be downloaded from the CFG website (for URL, see Table 6.1). Microarray readout data can be searched to display the results for specific genes or results matching specific constraints only.

The CFG Microarray Data resource enables users to determine which glycogenes are expressed in which tissues or in combination with which diseases. Recently, a combination of these data with the pathway information provided by KEGG was used successfully to predict glycan structures [5].

6.5 Swiss-Prot/TrEMBL

Swiss-Prot is a manually curated protein database. The information on the proteins is extracted by experts from the literature or from other databases. The entries in TrEMBL (Translated EMBL), in contrast, are automatically assigned and cover mainly those proteins for which no Swiss-Prot entry is available yet [6].

Swiss-Prot provides detailed data on the proteins. These include the protein and gene names and a list of synonyms used for the protein name. The species from which the protein originates is given together with the taxonomy information. The list of references includes a brief comment for each publication indicating the kind of information that can be extracted from that reference. Only the amino acid sequence is given in Swiss-Prot; the nucleotide sequence can be easily accessed using links to various gene databases, which are provided together with cross-references to diverse other external resources.

Information on the function, catalytic activity, and subcellular location is also given, provided that these data can be found in the literature. The same applies to information on post-translational modifications or known mutations. As these data are assigned manually, they are not available in TrEMBL but only in Swiss-Prot.

Glycoenzymes can only be searched using a keyword search. Because most databases described in this chapter provide links to the respective Swiss-Prot/TrEMBL entries, the more specialized databases can also be used to search for glycoenzyme data in Swiss-Prot and TrEMBL.

6.6 ExPASy ENZYME

The ExPASy ENZYME [7] database represents a hierarchical classification of enzyme families. The primary focus is information related to the nomenclature of enzymes. Entries are marked by an identifier composed of four numbers, the so-called EC number. In higher hierarchy levels, the numbers of non-determined levels are replaced by dashes. To glycosyltransferases, the EC number 2.4.-.- is assigned. A database query using this number reveals a list of all entries starting with 2.4., that is, all entries that are classified as glycosyltransferases. Glycosylases are classified with EC number 3.2.-.-.

The single entries provide information on the enzyme class such as official name according to IUBMB recommendations and also synonyms, the reaction catalyzed, comments, and crosslinks to other databases, including a list of Swiss-Prot entries relating to the enzyme family. Protein-specific data such as amino acid sequences are not available in ENZYME, as this is a database on enzyme families.

Apart from using the EC number to access the database, keyword searches for enzyme names, chemical compounds, cofactors, and comment lines are also provided.

6.7 BRENDA

Entries in the BRENDA [8] enzyme database are classified using the EC numbers provided by ENZYME (see above). Therefore, newly discovered proteins, to which no EC number has yet been assigned, are not included. BRENDA provides more search options than ExPASy ENZYME. In addition to various keyword searches, which can be combined with functional parameters such as pH optimum and K_i value, enzymes can be found using the pathway classification provided by KEGG (see above) or by gene ontology. The latter options provide a convenient way to find glycoenzymes.

The information provided in BRENDA entries is extracted from the literature. References are given together with the data. Entries in BRENDA also contain enzyme properties that are of interest for researchers working with these proteins, such as pH/temperature stability or information on references that describe purification methods for the protein.

BRENDA can be used free of charge by academic institutions. Commercial use or inclusion of BRENDA data in other databases requires a license.

6.8 NCBI Databases

The US National Center for Biotechnology Information (NCBI) hosts diverse sequence databases [9]. NCBI Nucleotide and NCBI Protein provide DNA/RNA or amino acid sequences, respectively, together with the reference in which the sequence was published. Entrez Gene gives data on the gene location within the chromosome, which can also be graphically displayed using Map Viewer. Sequence information in other databases is often taken from NCBI Nucleotide or NCBI Protein.

Similarly to Swiss-Prot, glycoenzyme data can only be retrieved using keyword searches, so that here also crosslinks from the databases, which are focused on glycoenzymes, can be used specifically to retrieve the respective data from the NCBI databases.

6.9 Protein Databank (PDB)

The PDB [10] is a resource of biomolecular 3D structures, most of which are resolved by X-ray crystallography or NMR spectroscopy (see Section 20.2). Glycoenzymes can only be searched for by keywords, but similarly to Swiss-Prot and the NCBI databases, PDB entries are crosslinked from most other databases mentioned in this chapter.

6.10 Conclusion

There are several freely available resources where glycoenzymes are collected and classified and their functions are described. Many of the data are derived from large public resources such as Swiss-Prot, NCBI or the scientific literature. Therefore, some basic information such as nomenclature or amino acid sequences is available in virtually all of the databases. However, several features are unique to particular databases. These features comprise both data and possibilities to search for specific enzymes. Crosslinks between the databases in many cases facilitate using the search options that are unique to one database and still easily access the information that is available only in other databases.

References

1. Coutinho P, Henrissat B: Carbohydrate-active enzymes: an integrated database approach. In *Recent Advances in Carbohydrate Bioengineering* (eds H Gilbert, G Davies, B Henrissat, B Svensson). Cambridge: Royal Society of Chemistry; 1999: pp. 3–12.
2. Kikuchi N, Narimatsu H: Bioinformatics for comprehensive finding and analysis of glycosyltransferases. *Biochim Biophys Acta* 2006, **1760**: 578–583.
3. Hashimoto K, Kawano S, Aoki-Kinoshita K, Ueda N, Hamajima M, Kawasaki T, Kanehisa M: KEGG as a glycome informatics resource. *Glycobiology* 2006, **16**: 63R–70R.
4. Comelli E, Head S, Gilmartin T, Whisenant T, Haslam S, North S, Wong N, Kudo T, Narimatsu H, Esko J, *et al.*: A focused microarray approach to functional glycomics: transcriptional regulation of the glycome. *Glycobiology* 2006, **16** (2): 117–131.
5. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M: Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* 2005, **21**: 3976–3982.
6. Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin M, Michoud K, O'Donovan C, Phan I, *et al.*: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003, **31**: 365–370.
7. Bairoch A: The ENZYME database in 2000. *Nucleic Acids Res* 2000, **28**: 304–305.
8. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D: BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004, **32**: D431–433.
9. Maglott D, Ostell J, Pruitt K, Tatusova T: Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005, **2005**: D54–D58.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res* 2000, **28**: 235–242.
11. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R: Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology* 2006, **16**: 82R–90R.

7 Bioinformatics Analysis of Glycan Structures from a Genomic Perspective

Kiyoko F. Aoki-Kinoshita¹ and Minoru Kanehisa²

¹*Department of Bioinformatics, Faculty of Engineering, Soka University, Tokyo 192-8577, Japan*

²*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan*

7.1 Introduction

The analysis of glycans entails the study not only of their structures but also of their functions, which are intertwined with the rest of the biological system such as interacting proteins and chemical compounds. Bioinformatics approaches to these types of analyses have been implemented by various groups. In this chapter, we present a variety of analytical techniques based on the data in KEGG (*Kyoto Encyclopedia of Genes and Genomes*), which includes a comprehensive glycan data resource called KEGG GLYCAN. KEGG encompasses all aspects of the biological system (i.e. all major components of biological processes), incorporating genomic information integrated with pathways and reactions and also chemical compounds [1]. We first describe this data resource before delving into the bioinformatics approaches for glycan structure analysis.

7.2 KEGG GLYCAN

The KEGG resources at <http://www.genome.jp> provide an integrated knowledge base of protein networks, the gene universe, and the chemical universe [1]. KEGG consists of many databases of pathways (PATHWAY), orthologs (KO, or KEGG ORTHOLOGY), reactions (REACTION), chemical compounds (COMPOUND), enzymes (ENZYME) and glycans (GLYCAN). KEGG GLYCAN is currently available at <http://www.genome.jp/kegg/glycan/>. It consists of glycan structures originally derived from the 45 000 carbohydrate structures in CarbBank [2]. However, since many of these were redundant, the entire database was inspected and cleaned so as to obtain non-redundant entries. In particular, many of the same molecules that were originally represented by different symbols were unified, and entries for redundant structures were found using KCaM (described later) and merged into single entries with links back to the originals. As a part

of KEGG, many records in GLYCAN have cross-references to the other databases both within and outside KEGG, such as PubMed [3].

The KEGG GLYCAN database provides many tools for glycan analysis. First, a tool for comparing glycan structures was developed to allow for efficient queries of the database. This tool is called KCaM and utilizes a verified algorithm for accurate comparisons of glycan tree structures. Second, a freely downloadable Java application called KegDraw is available to draw, save, and search glycan structures. Third, the Composite Structure Map (CMM), described later in this section, provides a tool to analyze glycan structures within the entire scope of all possible glycan structures.

7.2.1 *KCaM: Structure Search Tool*

The structure comparison program used in all the analyses in this section is based on an algorithm called KCaM, for KEGG Carbohydrate Matcher [4], which implements a dynamic programming technique and a theoretically proven efficient algorithm for finding the maximum common subtree between two trees [5]. This tool is currently available both online and through KegDraw for users to perform queries against the KEGG GLYCAN database. That is, by specifying a glycan structure as a query structure, the most similar structures (alignments) from the KEGG GLYCAN database are returned, just as a BLAST query searches for similar protein sequences.

KCaM consists of two main variations, an approximate matching algorithm and an exact matching algorithm. The former aligns monosaccharides while allowing gaps in the alignment, whereas the latter aligns linkages and disallows any gaps, resulting in a stricter criterion for alignment. Both variations provide local and global options. The local approximate matching algorithm does not penalize the gaps for unaligned regions whereas the global version does. Thus only conserved regions can be found using the local approximate matching algorithm. The local exact matching algorithm simply finds the first largest matching subtree to the query, whereas the global version attempts to find as many matching subtrees as possible. As a guideline, local exact matching should be sufficient for queries using specific structures. In contrast, local approximate matching can be used for more general queries when detailed information such as linkage conformation is unknown.

7.2.2 *KegDraw: Glycan Structure Drawing Tool*

Users who are interested in using the KEGG GLYCAN data are encouraged to download KegDraw, a freely available software tool for drawing and querying chemical structures. Although there are several applications already available for drawing chemical compounds, few are available for drawing glycans. KegDraw is a Java application, so it runs locally in a platform-independent manner. It consists of two drawing modes:

1. Compound mode, for drawing chemical compounds in a similar manner to the widely used ChemDraw software (www.cambridgesoft.com/software/ChemDraw/), and
2. Glycan mode, for drawing glycans with monosaccharide units.

In Glycan mode, glycan structures can be drawn in a variety of ways. The simplest method is to select monosaccharides and linkage conformations individually from popup

menus. Convenient functionalities such as cut-and-paste and predefined template structures are also available. KegDraw currently handles files in KEGG Chemical Function (KCF) format [6] (see Chapter 3), so glycan structures can be input by specifying a data file in this format. Conversely, a glycan structure input on to the canvas can be transferred into KCF as output data. Furthermore, structures drawn can be queried against the KEGG GLYCAN database to retrieve the most similar structures.

Finally, an Application Programming Interface (API) is available to allow programmers to access all data from KEGG, including glycan information, over the Internet. This API library can be used by anyone who wishes to develop their own programs to access the KEGG resources for personal data analysis.

7.3 Composite Maps

The main bioinformatics approach for glycans that we introduce in this section is based on the concept of glycan Composite Maps, which provide a global view of glycans based on their structures. Different types of views can be obtained by modifying the methodology used. We present two different methods used for generating a Composite Map:

1. Composite Reaction Map (CRM): a map derived from differences in linkages, corresponding to glycosyltransferases. All of the possible pathways concerning glycans can be seen on this map.
2. Composite Structure Map (CSM): a map based on superimposing whole glycan structures to obtain a comprehensive map illustrating all possible structures of glycans.

7.3.1 Composite Reaction Map

The development of the CRM came from the fact that, by taking two known glycan structures that differ by only one linkage (i.e. by one extra monosaccharide residue, hereafter called “one link”), a glycosyltransferase reaction that adds this one link to the smaller (source) structure to result in the larger (target) structure may be found. If this process is repeated with increasingly larger structures differing by one link, then by connecting the source and target structures, a global map may be constructed. An illustration of this process for two paths is given in Figure 7.1.

During the development of the CRM, the KEGG GLYCAN database contained a total of 10 385 glycan structures, from which redundant structures and those containing cycles of monosaccharide chains were removed, resulting in 7891 structures with which to construct the CRM. Starting from a single linkage (hereafter referred to as the root linkage), the local exact matching algorithm of KCaM was utilized to find those structures that contained the same link and an additional linkage. That is, structures of three monosaccharides (two linkages) were retrieved, where one linkage matched the root linkage exactly. These retrieved structures were then “connected” to the root structure. This process was subsequently repeated for each of these retrieved structures to obtain structures of three linkages where two of the linkages match the previous structure exactly. Again, these newly retrieved structures were then connected to the corresponding matching structure. This would thus result in a branched “tree” structure of connected glycan structures. Figure 7.1 shows this process for two different paths.

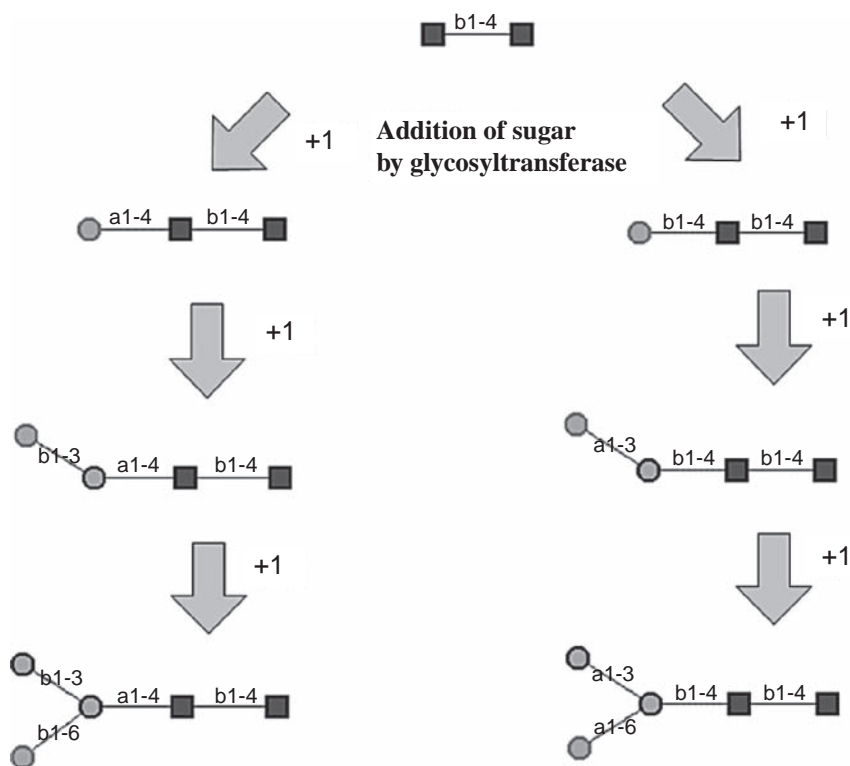


Figure 7.1 Process of comparing glycan structures that differ by one link.

As a result, in order to form a global CRM, all possibly related structures were connected together into a tree structure, with nodes represented by glycans and edges by the link that connected them. Figure 7.2 is the single global CRM, within which the different glycan classes can be clearly delineated. This procedure can be specialized by starting with different root structures; such a specialized sub-map is given in Figure 7.3 for sphingolipids, which is a tree with Gal(β 1-4)Glc at the root.

Figure 7.4 displays the tree for *N*-glycans, with the number of one-links demarcated. The largest one-link-connected structure that could be obtained contained 18 links, meaning that the biosynthesis of the single largest structure that could be traced from the root consisted of 18 reactions (disregarding pruning). Therefore, given this CRM for *N*-glycans, we considered how the structures actually compared with the *N*-glycan biosynthesis pathway. Figure 7.5 illustrates this comparison. It is clear how one branch (biosynthesis starting from PP-dol) could be traced to its end, and then backtracked for pruning, eventually to branch off into different paths corresponding to different *N*-glycan subtypes.

Using this CRM, we can assess the coverage of the knowledge of all the known glycan structures in KEGG. That is, if all glycan structures in KEGG can be connected by one link in the CRM, then we can assume that all structures are accounted for in the database. However, we found that this CRM contained 4713 structures, corresponding to a coverage of just 61.2%. The next question was to ask what it would take to connect all structures. Note

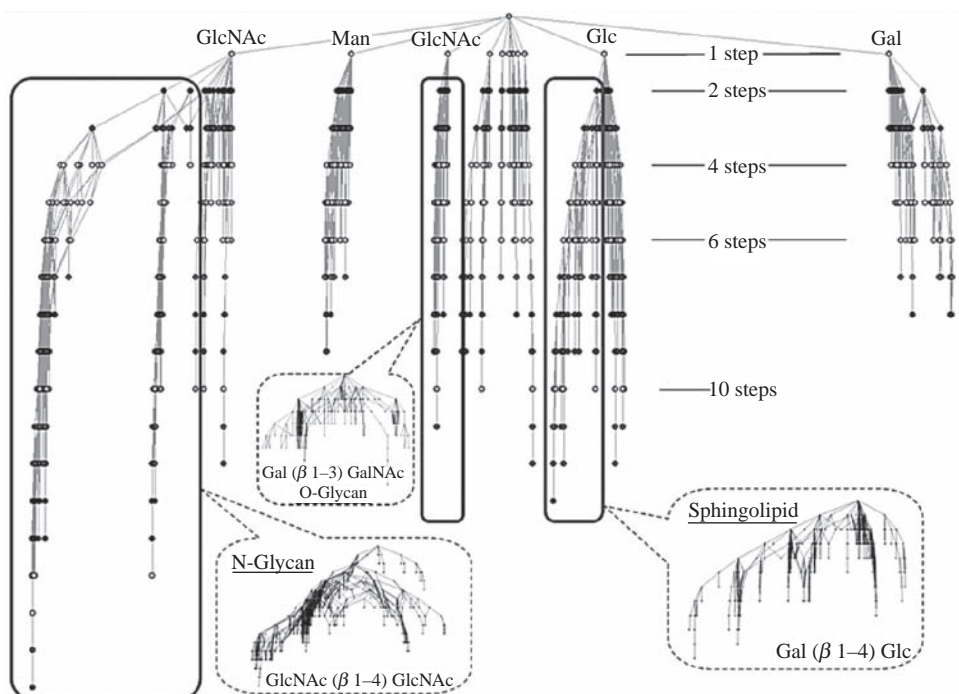


Figure 7.2 The final Composite Reaction Map, with areas circled corresponding to known glycan classes.

that this map was created using a very strict limitation of one link to connect structures. We can loosen this limitation by connecting structures that are connected by up to two links. That is, we can apply the same KCaM algorithm to retrieve iteratively structures that match a query structure for all but two linkages, resulting in a different CRM. Then we may repeat this for even looser matches producing more CRMs. In fact, we generated CRMs for up to seven links; Table 7.1 lists the amount of coverage achieved for all these maps. At seven links, the coverage increased to almost 98%. We then did not see a change in coverage until we loosened the requirement to 21 links. That is, only when we reached differences of 21 links could we finally account for almost all structures in the database. Considering that it takes four link differences to obtain even 90% coverage, we may estimate that there are many missing structures yet to be reported and entered into the database.

7.3.2 Composite Structure Map (CSM)

We now take a look at a different method of constructing a Composite Map based on the superimposition of whole structures as opposed to looking at links. CSM is a static representation of all possible variations of glycan structures in a tree format [7]. That is, CSM is not limited to those structures that differ by one link; it superimposes structures together to obtain the union, such that all possibilities of the given structures can be seen on the map. As will be described later, despite the wide variety of glycosidic

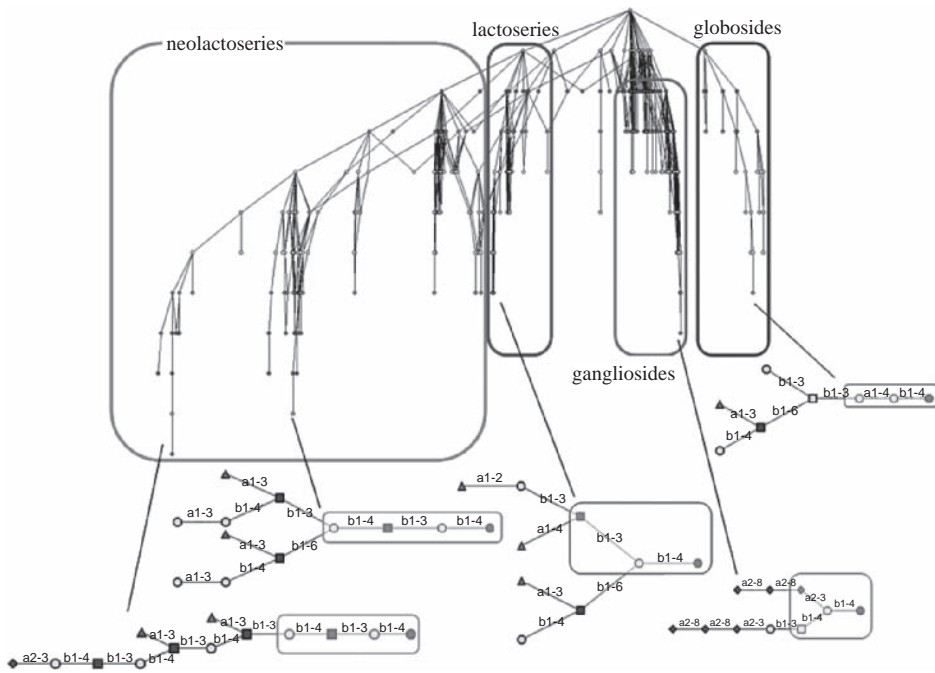


Figure 7.3 CRM for sphingolipids with Gal(β 1-4)Glc at the root.

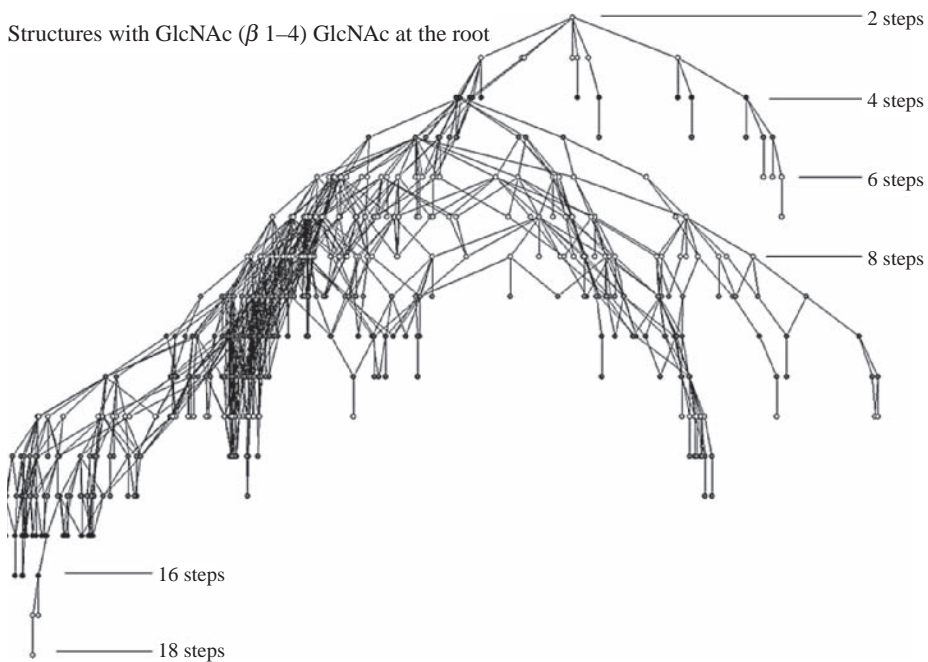


Figure 7.4 N-Glycan CRM indicating that the largest one-link-connected N-glycan structure that could be obtained contained 18 links.

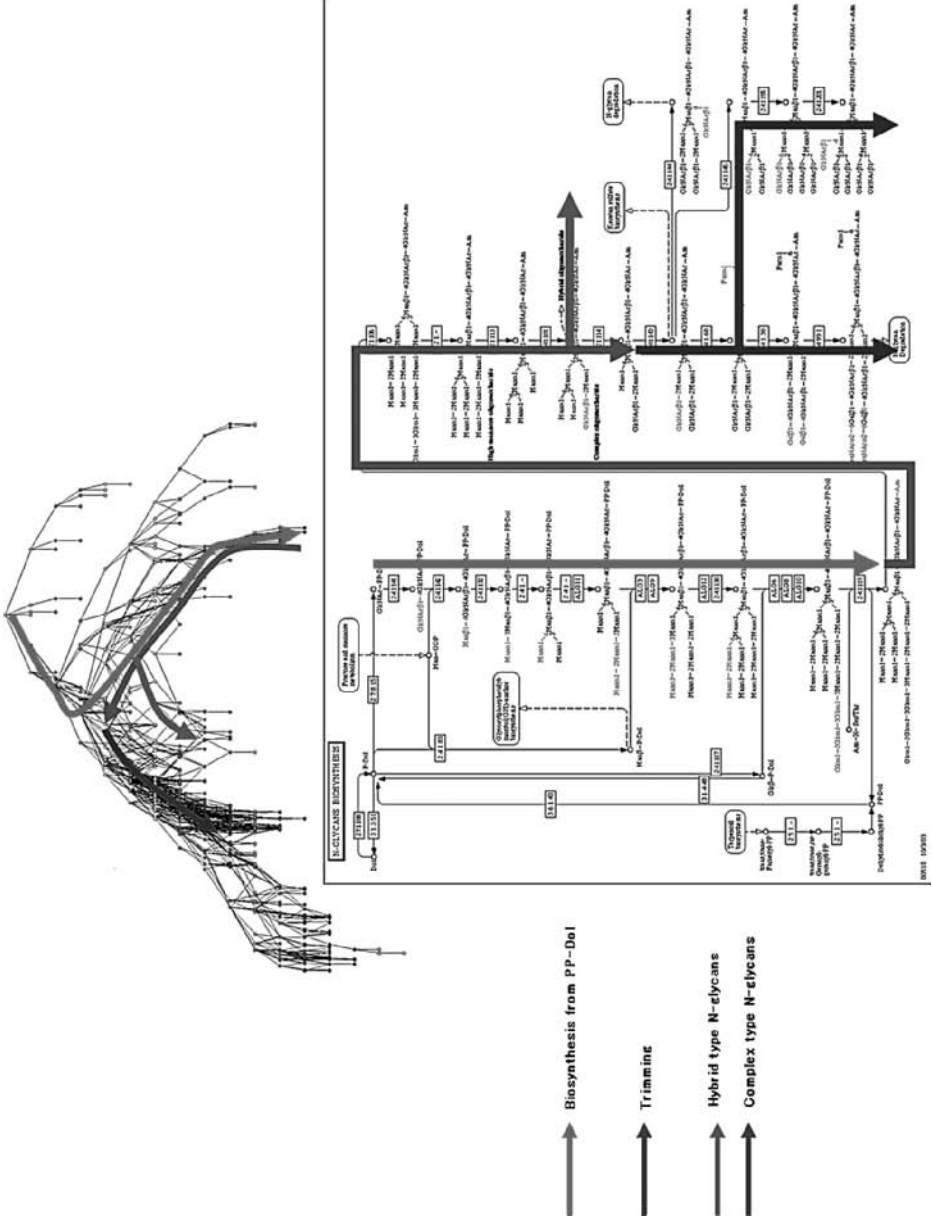


Figure 7.5 Correspondence of N-glycan biosynthesis pathway with the CRM. A full-color version of this figure is included in the Plate section of this book.

Table 7.1 Coverage of glycan structures contained in the CRM differentiated by the number of links used to construct the CRM.

	No. of links							
	1	2	3	4	5	6	7	21
No. of structures	4713	6177	6903	7210	7383	7465	7517	7666
Coverage (%)	61.2	80.3	89.7	93.7	95.9	97.0	97.7	99.6

linkages theoretically possible that may result in an uncountable number of plausible glycan structures, this map illustrates that the variety of structures currently studied is rather limited. Since a tool to visualize all carbohydrate structures in this way had never been offered before, we have made CSM available online at http://www.genome.jp/kegg-bin/draw_csm.

CSM was constructed by taking all the glycan structures containing a particular monosaccharide at its root and superimposing them into one unified structure by a tree structure alignment program (a modified version of KCaM). Note that glycan entries having bonds between two reducing ends were disregarded and that modification molecules attached to monosaccharides after glycan biosynthesis were removed. Thus, among the remaining 7891 non-redundant structures, pairs of glycan structures would be compared, and any differences would then be added to eventually obtain a massive map covering matches with all structures containing the same root monosaccharide. In particular, the following steps were taken:

1. Take the path of glycosidic linkages from the root monosaccharide to the end of every branch and store it as a linear structure.
2. Remove any duplicate linear structures.
3. For all the linear branches for all of the glycan structures containing this root monosaccharide, merge them together, overlapping the same sub-paths and keeping different paths distinct.

This results in a single tree with many branches. Note that every glycosidic linkage conformation was distinguished, so different glycosidic linkages were represented by different edges even if the attached monosaccharide is the same. That is, CSM would contain different edges for “Gal β 1–3 Glc” and “Gal β 1–4 Glc.”

A massive map for every major monosaccharide was thus generated and made available as a clickable tool, illustrated in Figure 7.6. This figure shows a CSM having Glc subsequently linked to Gal as the root. This is the tree containing all variations of glycan structures whose root monosaccharides are Glc linked to Gal. A different tree can be displayed for any combination of up to three monosaccharides composing the root by specifying them in the appropriate textbox at the top. Each monosaccharide in the tree is represented by symbols in accordance with the standards set forth by the Consortium for Functional Glycomics (CFG) (<http://www.functionalglycomics.org/>).

Any node on CSM corresponds to a list of glycan structures that are located on the single path from the root to that node. Correspondingly, each node on the online map is hyperlinked to its corresponding list, as illustrated at the top right of Figure 7.6. This list

contains figures of each structure and also hyperlinks to relevant information. Further detail for each structure in the resulting list is also available via a hyperlink to each corresponding GLYCAN entry.

Other online functionality includes organism-specific visualization which is available from the pull-down menu at the top. When a specific organism is selected, each edge is colored and hyperlinked to the KEGG GENES entry for the glycosyltransferase that catalyzes the biosynthesis of the particular linkage (see bottom right of Figure 7.6). Each GENES entry contains all available related information such as involved reactions, pathways, and literature links. On the other hand, when “All organisms in KEGG” is selected, each edge is hyperlinked to the ortholog group (called the KO, or KEGG Orthology) of the glycosyltransferase to which it corresponds (across all organisms). This provides an idea of the orthologs of the given glycosyltransferase calculated manually based on KEGG pathway information, providing a connection between chemical compound information and gene information.

It is also possible to specify colors for the edges corresponding to a specification of colors and genes stored in a local file. Thus, a variety of analyses can be performed using this tool. For example, microarray gene expression data can be displayed in this tree by color coding the up or down regulation of genes involved in glycan synthesis, linking gene expressions to expressed glycan structures. That is, a list of glycan structures of interest may be color-coded on to CSM to view their structural relationships from a bird’s eye perspective.

7.4 Glycan Structure Prediction from Glycosyltransferase Expression Data

In contrast to the use of KEGG for directly analyzing a dataset of glycan structures, we can also take advantage of additional experimental data to develop methods for glycan structure prediction. The use of microarrays is common in bioinformatics, and there are many sources of gene expression data for glycosyltransferases. Therefore, we consider the fact that, given a repertoire of glycosyltransferases and glycosidases and also a set of donor monosaccharides, it should be plausible to predict the possible glycan structures in an organism or at a particular stage in the development and functioning of the cell, as captured by microarray expression data [8] That is, by capturing patterns of bonds existing in current glycans, new glycans can be predicted from a set of bonds. This first requires a statistical computation of the structures in the glycan database.

7.4.1 Reaction Library Construction

To begin, a library consisting of bond-formation patterns of glycosyltransferase reactions was first constructed. Human glycosyltransferase genes were collected from KEGG GLYCAN based on gene annotation information. Reaction specificity was determined according to the published literature and was characterized by the following three features: (1) the acceptor monosaccharide in the glycan chain, (2) the donor monosaccharide, and (3) the glycosidic bond between the acceptor and donor. A total of 146 glycosyltransferase genes in the human genome were collected, and 52 were found to be involved in the synthesis of glycosidic bonds (reduced due to paralogs). Among these 52 genes, 41 attached a sugar

donor to a sugar acceptor. Other reaction patterns catalyzed by other glycosyltransferases such as those that attach sugars to proteins or lipids were not considered in this analysis. We also limited this library only to those structures containing the following nine monosaccharides: glucose, galactose, mannose, *N*-acetylglucosamine, *N*-acetylgalactosamine, fucose, xylose, glucuronic acid, and *N*-acetylneuraminic acid. As a result, a total of 4107 human glycan structures from KEGG GLYCAN containing only these nine monosaccharides were extracted.

7.4.2 Co-occurrence Score

Because many of the same glycosidic linkages are found within the same classes of glycans, a co-occurrence score was developed to use in our prediction method. All glycan structures in the database were broken down into “reaction pattern” components consisting of the donor, acceptor, and linkage conformation. These were used to construct a reaction pattern matrix of correlation coefficients to measure the co-occurrence of each pair of reaction patterns. The matrix containing the cosine coefficient values for every pair of reaction patterns was then used to calculate co-occurrence scores. A preliminary test on the utility of this score was performed by clustering the negative score values using the Ward method [9]. Six clusters resulted, with each cluster corresponding to different glycan classes: *N*-glycans, *O*-glycans and glycolipids, proteoglycans, xyloglucans, and galactomannans. The effectiveness of the co-occurrence score was evident from the clear delineations of each glycan class within different clusters.

7.4.3 Prediction Method

The prediction method consisted of taking the list of reaction patterns for a glycan structure (used as the query) and calculating a score for each candidate glycan structure. This score is calculated as the sum of the co-occurrence scores for every combination of reaction patterns between the query and candidate normalized by the number of reaction patterns in the candidate. That is, the score S_c was calculated as

$$S_c = \frac{1}{m} \sum \sum S_0(q_i, b_j)$$

where S_0 is the co-occurrence score between the query reaction pattern (q) and the reaction pattern in the candidate glycan structure (b), and m is the number of reaction patterns in the candidate glycan structure. This method was tested on the KEGG GLYCAN database and compared against prediction using random data. Given a particular glycan structure, we used its list of reaction patterns and calculated its score against our glycan structure library, ranking the scores from highest to lowest. This was repeated for all entries in the library. As a result, 72% of all entries could be predicted as among the top 10 highest scoring glycans. This method was improved by introducing a penalty score; a value of one was subtracted from each co-occurrence score being summed if the particular reaction pattern pair did not exist in the query but did exist in the candidate. That is,

$$S_c = \frac{1}{m} \sum \sum [S_0(q_i, b_j) - p_j]$$

where $p_j = 0$ if b_j exists in the query and 1 if it does not. With this new method, 81% of all entries could be predicted among the top 10 highest scoring glycans.

Among the falsely predicted glycans, many structures contained repeated sequences of reaction patterns. However, almost every glycan predicted as the top score belonged to the same class as the query glycan. Furthermore, because various glycans are often expressed in the same cell regardless of the set of expressed glycosyltransferases, these prediction results can be seen as favorable.

7.4.4 Microarray Data Analysis

To determine the usefulness of this new prediction method on actual microarray data, DNA microarray expression profiles of human cell lines were obtained from the Consortium for Functional Glycomics at <http://www.functionalglycomics.org/static/consortium/organization/sciCores/coree.shtml>. Human data from five experiments such as lung, leukemia, and carcinoma cell lines were collected; 80 glycosyltransferase genes were mounted on these arrays, corresponding to 37 reaction patterns. Genes were assumed to be positively expressed if the data indicated such (i.e. the Affymetrix data indicated that the gene expression value exceeded a certain threshold such that it was determined as being "Present") in the majority of the same experimental conditions.

Because of the lack of data regarding glycosyltransferases, there is the possibility that the penalty score used in this method may be falsely used for unidentified glycosyltransferases. However, it is assumed in this experiment that all glycosyltransferases corresponding to reaction patterns on the microarray have been identified since homologous enzymes may catalyze the same reaction and can be easily found using a homology search on genome data.

The human carcinoma U937 cell line was analyzed using this prediction method. The top 10 predicted structures could be divided into two groups: hybrid *N*-glycans and gangliosides. Although these classes are known to have different core structures, they share the same terminal structure at the non-reducing end: sialic acid (Neu5Ac) and fucose, and the internal *N*-acetylglucosamine structure (Gal β 1–4 GlcNAc). The sialyl Lewis X epitope [Neu5Ac α 2–3 Gal β 1 (LFuc α 1–3) 4 GlcNAc] was also found in some of the resulting glycans. Because it has been reported that carcinoma cells tend to over-produce the sialyl Lewis X epitope and also terminal sialic acid [10], these prediction results were promising.

This methodology is just the first step in predicting glycan structure from glycosyltransferase expression data. However, it is not limited to DNA expression profiles, but may also be applied to other forms such as protein expression profiles. Further improvement of the methodology may be gained from the incorporation of such information as the relative position of reaction patterns to capture more properties of glycan structures.

7.5 Mining Glycan Data

One of the earliest methods in bioinformatics for retrieving patterns in sequence data was the hidden Markov model (HMM), which was then applied to bioinformatics in the form of the profile HMM [11], which allowed large amounts of protein sequence data to be

mined for “profiles” in the data. These profiles were computed probabilistically such that good alignments could be made between large amounts of data very efficiently. These profiles were then collected and are now available as the Pfam database [12]. Applying this model to glycans, there were several aspects of glycan structures to take into consideration: (1) the branching structure and (2) the relationship between “siblings,” meaning those monosaccharides that have glycosidic bonds with a common monosaccharide. That is, the conformations of these “sibling” monosaccharides need to be taken into consideration in the model. As a result, probabilistic models for tree structures have been developed, including the probabilistic sibling-dependent tree Markov model (PSTMM) [13] and the ordered tree Markov model (OTMM) [14]. Following on the latest versions of these models, a profile PSTMM model which is capable of extracting the actual tree-structure profiles from the input data was developed [15]. As a result, profiles similar to sequence logos [16] could be generated from groups of glycan structures, allowing the extraction of glycan motifs. Further details regarding these methods are beyond the scope of this chapter, but the interested reader is encouraged to refer to the relevant literature.

Another technique for mining patterns in abundant amounts of data is the kernel method, which takes as input groups (or classes) of structures and, depending on the model, determines which characteristics in the data most identify and differentiate between the classes. Recently, two kernel methods for extracting glycan motifs in certain types of cells have been developed. The first was developed in particular for leukemic cells [17]. It used the concept of *layers*, where the layer of a monosaccharide was defined as the number of monosaccharides with which it was linked in order from the root. Moreover, the input glycans were broken down into trimers, such that all combinations of three linked monosaccharide structures were used to represent a single glycan. Each trimer was weighted according to its layer, defined as the layer of the monosaccharide in the trimer of the smallest value. Thus, given two classes of glycan structures, the kernel would calculate all trimers and their layers and determine which trimer at which layer most distinguishes the two groups from one another. As a result, this layered-trimer kernel extracted the glycan structure α -D-Neup5Ac-(2-3)- β -D-Galp-(1-4)-D-GlcpNAc as being characteristic of leukemic cells.

Another kernel which expanded on the layered-trimer kernel is the gram distribution kernel [18], which not only broke down glycans into trimers but also dimers, quadrimers, up to 6-mers. This not only allowed the kernel to extract trimer motifs, but any motif which would most characterize a group of glycans, regardless of size. As a result, the gram distribution kernel was able to extract not only the same structure as the layered trimer kernel for leukemic cells, but also a 6-sulfated GlcNAc structure as being characteristic of cystic fibrosis in comparison with other respiratory and bronchial mucins. The score of this structure, however, was much lower than what was extracted for leukemic cells, indicating that more data are required to extract better motifs.

7.6 A Genomic Perspective of Glycan Structures

Considering the thousands of unique structures already in the KEGG GLYCAN database, there seems to be a wide variety of glycan structures in Nature. The complexity of glycan structures is caused by the combination of different monosaccharides with different

glycosidic linkages, which are also related to the complexity of biosynthetic pathways and responsible genes. However, we find that the majority of structures include core and common structures such as the root of *N*-glycans, *O*-glycans, and glycolipids. That is, the structures are not constructed randomly, and their variety is rather limited. This reflects the architecture of the biosynthetic pathways, consisting of conserved portions and terminal variations as shown, for example, in the KEGG pathway maps of “*N*-glycan biosynthesis” and “High-mannose type *N*-glycan biosynthesis.” This, in turn, reflects the inventory of genes in the genome, containing orthologous genes and paralogous genes, respectively, for the universality and diversity of such pathways.

The Composite Maps are tools for the integrated analysis of both structural and genomic information. By considering the one-link differences between structures, the minimal global tree called CRM could be obtained to illustrate the current set of known glycosyltransferase reactions. This could then be mapped to the glycan pathways to assess the coverage of glycan structures. Alternatively, using a different approach of superimposing entire structures, the current set of known glycan structures can be represented compactly in CSM. In addition to structural variations, CSM also illustrates the relationship between a glycosidic linkage and the glycosyltransferase that catalyzes it. Currently, 59 known ortholog groups in KEGG, which include 176 organisms, are assigned to the edges in CSM as shown in Table 7.2 (see also the updated list at <http://www.genome.jp/kegg/glycan/GT.html>). However, this corresponds to only 24% of the edges that are present in CSM, suggesting that there are still a large number of genes that are unknown or whose functions have not yet been fully characterized.

Hence the development of methodologies to predict structures may and should be utilized in order to assist in filling these gaps. By predicting structures from current data, potentially new structures may then be iteratively added back to the original computation of the Composite Map, which may then again be used to predict even more structures. These predictions are vital as they provide the generation of plausible structures that may not have been considered otherwise. Moreover, as more experimental data are added to the database, these predictions may be confirmed, as in the use of microarray expression data.

Although this chapter focused on analyses and prediction of glycan structures, this is only the first step. Once a better understanding of glycan structures has been gained, the functions of these glycans need to be studied, in combination with the proteins with which they interact. Thus, as research progresses, the current methods need to be developed continuously and new ideas generated for other bioinformatics approaches for glycan structure and function analysis. It is safe to say that there is much room for growth in bioinformatics approaches for glycan analysis in terms of both structure and function.

Abbreviations

API	Application Programming Interface
CFG	Consortium for Functional Glycomics
CRM	Composite Reaction Map
CSM	Composite Structure Map
HMM	hidden Markov model
KCaM	KEGG Carbohydrate Matcher
KCF	KEGG Chemical Function

Table 7.2 List of glycosyltransferases used in CSM.

Glycosidic linkage	KO	EC number	Human gene	CAZy
Gal α 1–3 Gal	K00707	2.4.1.37	ABO	GT6
	K00743	2.4.1.87		GT6
Gal α 1–4 Gal	K01988	2.4.1.228	A4GALT	
Gal α 1–6 Gal		2.4.1.67		
Gal β 1–3 Gal	K00734	2.4.1.134	B3GALT6	
Gal β 1–3 GalNAc	K03877	2.4.1.-	B3GALT5	GT31
	K00731	2.4.1.122	C1GALT1	GT31
	K00715	2.4.1.62	B3GALT4	
Gal β 1–4 GlcNAc	K00708	2.4.1.38	B4GALT1,2,3,5	GT7
	K00718	2.4.1.69	FUT1,2	GT11 GT37
	K00733	2.4.1.133	B4GALT7	GT7
GalNAc α 1–3 Gal	K00709	2.4.1.40	ABO	GT6
GalNAc α 1–3 GalNAc	K00722	2.4.1.88	GBGT1	GT6
GalNAc β 1–3 Gal	K00719	2.4.1.79	B3GALT3	GT31
GalNAc β 1–4 Gal	K00725	2.4.1.92	GALGT	GT12
GalNAc β 1–4 GlcA	K00746	2.4.1.174		
	K00747	2.4.1.175		
Glc α 1–2 Glc	K03850	2.4.1.-	ALG10	GT59
Glc α 1–3 Glc	K03849	2.4.1.-	ALG8	GT57
Glc α 1–3 Man	K03848	2.4.1.-	ALG6	GT57
GlcA β 1–3 Gal	K00735	2.4.1.135	B3GAT1,2,3	GT43
GlcA β 1–3 GalNAc	K03419	2.4.1.226		
GlcA β 1–4 GlcNAc	K03420	2.4.1.225		
GlcNAc α 1–4 GlcA	K02368	2.4.1.224	EXTL1	GT47 GT64
	K02369	2.4.1.223	EXTL2	GT64
	K02370	2.4.1.223 2.4.1.224	EXTL3	GT47 GT64
GlcNAc β 1–2 Man	K00726	2.4.1.101	MGAT1	GT13
	K00736	2.4.1.143	MGAT2	GT16
GlcNAc β 1–3 Gal	K00741	2.4.1.149	B3GNT1,6	GT31
GlcNAc β 1–3 GalNAc	K00739	2.4.1.147	B3Gn-T6	
GlcNAc β 1–4 Man	K00737	2.4.1.144	MGAT3	GT17
	K00738	2.4.1.145	MGAT4	GT54
GlcNAc β 1–4 MurNAc	K02563	2.4.1.227	MURG	GT28
GlcNAc β 1–6 Gal	K00742	2.4.1.150	GCNT2	GT14
GlcNAc β 1–6 GalNAc	K00727	2.4.1.102	GCNT1	GT14
	K00740	2.4.1.148		
GlcNAc β 1–6 Man	K00744	2.4.1.155	MGAT5	GT18
Kdo α 2–4 Kdo	K02527	2.-.-.-	KDTA	GT30
LFuc α 1–2 Gal	K00718	2.4.1.69	FUT1,2	GT11 GT37
LFuc α 1–3 GlcNAc	K03663	2.4.1.152	FUT9	GT10
LFuc α 1–4 GlcNAc	K00716	2.4.1.65	FUT3,5,6	GT10
LFuc α 1–6 GlcNAc	K00717	2.4.1.68	FUT8	GT23
Man α 1–2 Man	K03844	2.4.1.-	ALG11	GT4
	K03846	2.4.1.130	ALG9	GT22
Man α 1–3 Man	K03843	2.4.1.132	ALG2	GT4
	K03845	2.4.1.130	ALG3	GT58
Man α 1–6 Man	K03847	2.4.1.130	ALG12	GT22
Man β 1–4 GlcNAc	K03842	2.4.1.142	ALG1	GT33

(continued)

Table 7.2 (Continued)

Glycosidic linkage	KO	EC number	Human gene	CAZy
Neu5Ac α 2–3 Gal	K03792	2.4.99.10	ST3GAL6	
	K00781	2.4.99.6	ST3GAL3	GT29
	K03370	2.4.99.9	B3GALT5	GT29
	K00780	2.4.99.4	SIAT4A	GT29
	K03368	2.4.99.4	SIAT4B	GT29
Neu5Ac α 2–6 Gal	K00778	2.4.99.1	ST6GAL1	GT29
Neu5Ac α 2–6 GalNAc	K03479	2.4.99.3	SIAT7A	GT29
	K03373	2.4.99.-	SIAT7C	GT29
	K03375	2.4.99.-	SIAT7E	GT29
	K03376	2.4.99.-	SIAT7F	GT29
Neu5Ac α 2–8 Neu5Ac	K03371	2.4.99.8	SIAT8A	GT29
	K03369	2.4.99.-	SIAT8E	GT29

KEGG *Kyoto Encyclopedia of Genes and Genomes*

KO KEGG Orthology

OTMM ordered tree Markov model

PSTMM probabilistic sibling-dependent tree Markov model

References

1. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006, **34**:D354–D357.
2. Doubet S, Bock K, Smith D, Darvill A, Albersheim P: The Complex Carbohydrate Structure Database. *Trends Biochem Sci* 1989, **14**:475–477.
3. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M: LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res* 2002, **30**:402–404.
4. Aoki K, Yamaguchi A, Ueda N, Akutsu T, Mamitsuka H, Goto S, Kanehisa M: KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res* 2004, **32**:W267–272.
5. Aoki K, Yamaguchi A, Okuno Y, Akutsu T, Ueda N, Kanehisa M, Mamitsuka H: Efficient tree-matching methods for accurate carbohydrate database queries. *Genome Informatics* 2003, **14**:134–143.
6. Hattori M, Okuno Y, Goto S, Kanehisa M: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* 2003, **125**:11853–11865.
7. Hashimoto K, S G, Kawano S, Aoki-Kinoshita K, Ueda N, Hamajima M, Kawasaki T, Kanehisa M: KEGG as a glycome informatics resource. *Glycobiology* 2006, **16**:63R–70R.
8. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M: Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* 2005, **21**:3976–3982.
9. Ward J: Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 1963, **58**:236–244.
10. Kim Y, Varki A: Perspectives on the significance of altered glycosylation of glycoproteins in cancer. *Glycoconj J* 1997, **14**:569–576.
11. Eddy S: Profile hidden Markov models. *Bioinformatics* 1998, **14**:755–763.

12. Finn R, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, *et al.*: Pfam: clans, web tools and services. *Nucleic Acids Res* 2006, Database Issue **34**:D247–D251.
13. Aoki K, Ueda N, Yamaguchi A, Kanehisa M, Akutsu T, Mamitsuka H: Application of a new probabilistic model for recognizing complex patterns in glycans. *Bioinformatics* 2004, **20**:I6–I14.
14. Hashimoto K, Aoki-Kinoshita K, Ueda N, Kanehisa M, Mamitsuka H: A new efficient probabilistic model for mining labeled ordered trees. *Proc KDD* 2006, 177–186.
15. Aoki-Kinoshita K, Ueda N, Mamitsuka H, Kanehisa M: ProfilePSTMM: capturing tree-structure motifs in carbohydrate sugar chains. *Bioinformatics* 2006, **22**:e25–e34.
16. Schneider T, Stephens R: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990, **18**:6097–6100.
17. Hizukuri Y, Yamanishi Y, Nakamura O, Yagi F, Goto S, Kanehisa M: Extraction of leukemia-specific glycan motifs in human by computational glycomics. *Carbohydr Res* 2005, **340**:2270–2278.
18. Kuboyama T, Hirata K, Aoki-Kinoshita K, Kashima H, Yasuda H: A gram distribution kernel applied to glycan classification and motif extraction. *Genome Informatics* 2006, **17**:25–34.

8 Glycosylation of Proteins

Claus-Wilhelm von der Lieth¹ and Thomas Lütteke²

¹Formerly at the Central Spectroscopic Unit, Deutsches Krebsforschungszentrum (German Cancer Research Center), 69120 Heidelberg, Germany

²Faculty of Veterinary Medicine, Institute of Biochemistry and Endocrinology, Justus-Liebig University Gießen, 35392 Gießen, Germany

8.1 Introduction

8.1.1 Biological Relevance of Glycosylation

Among the co- and post-translational modifications of proteins, glycosylation is by far the most common and the most complex. It differs from most other covalent protein modifications such as phosphorylation, acetylation, and formylation with respect to the size and the complexity of the added group, and also the magnitude of the cellular machinery devoted to synthesis and modulation [1, 2]. An analysis of the Swiss-Prot database has led to the estimation that about 50% of all proteins are glycosylated [3].

Simply because of their size and hydrophilicity, the glycan chains can alter the physico-chemical properties of glycoproteins, increasing their solubility and making them more stable by reducing backbone flexibility or protecting them from proteolysis [1, 4]. *N*-Linked glycans also play an important role in the protein folding process (see Section 8.1.3.2). Furthermore, glycan chains can serve as molecular addresses, marking the intra- or extracellular destination of a glycoprotein, and thus play an important role in protein trafficking. For example, the selective targeting of lysosomal hydrolases from the trans-Golgi network to endosomes and lysosomes is mediated by mannose-6-phosphate receptors [5]. In addition, some glycans are used as markers to select proteins that are to be cleared from the circulatory system. Glycosylated proteins on the cellular surfaces are implicated in a number of cell–cell and cell–matrix recognition events, ranging from fertilization and cellular development to pathological infections and immune responses. These events are based on protein–carbohydrate interactions, which are discussed in Chapters 21 and 22.

Glycosylation is – in contrast to protein or DNA/RNA synthesis – not a template-driven process. Although all cells in an organism possess the same glycosyltransferase genes, not all of them are expressed all the time. Depending on the developmental age

of the organism, the tissue, and the state of the cell, different patterns of transferases are expressed, leading to different glycan structures. Therefore, some characteristic glycan chains can be used as diagnostic markers, and as potential therapeutic targets, in diseases such as cancer [6, 7]. Even within one cell the glycosylation patterns of proteins may vary, so that glycoproteins generally exist as heterogeneous populations of glycosylated variants, the so-called glycoforms [8]. This capacity to produce varied glycan structures, allows fine tuning of protein functions.

8.1.2 Glycosylation Types

Protein glycosylation is classified by the type of atom to which the glycan chain is attached. The different glycan families are discussed in Chapter 2, so only a short overview of this topic is given here. In *N*-glycosylation, a carbohydrate chain is covalently linked to the N_δ2 atom of an Asn side-chain. Apart from very few exceptions [2], all *N*-glycan chains contain the common GlcNAc₂Man₃ pentasaccharide core structure Man- α -(1,6)-[Man- α -(1,3)]-Man- β -(1,4)-GlcNAc- β -(1,4)-GlcNAc- β -(1,N)-Asn. Depending on the residues outside the core, the structures are classified as complex, high mannose or hybrid [9].

An Asn residue can only be glycosylated if it is part of the so-called “sequon”, an Asn-Xxx-Ser/Thr sequence motif, where Xxx can be any amino acid except Pro [10]. During the transfer of the dolichol-linked GlcNAc₂Man₉Glc₃ tetradecasaccharide precursor to an Asn side-chain (see Section 8.1.3.1) the hydroxyamino acid at position +2 (Ser/Thr) serves as a hydrogen bond donor that is necessary for the function of the enzyme oligosaccharyltransferase (OST) [11, 12]. Not all sequons that are present in a protein chain are glycosylated, so the presence of a sequon is a necessary, but not a sufficient, criterion for *N*-glycosylation [3].

In *O*-glycosylation, carbohydrate chains are linked to the oxygen atoms of the hydroxyl groups present at the terminal ends of Ser, Thr or – to a lesser extent – Tyr, hydroxyproline, or hydroxylysine side-chains. In contrast to *N*-glycosylation, where there is only one ER-located enzyme – the oligosaccharyltransferase – responsible for the initial transfer of the precursor oligosaccharide to the Asn side-chain, there are a number of different glycosyltransferases known to catalyze the initial step of *O*-glycosylation, the connection of a monosaccharide to an amino acid. Within mucin-type *O*-glycosylation, for example, 15 different mammalian UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferases (pp-GalNAc-Ts), the enzymes catalyzing the initial attachment of GalNAc to Ser or Thr in mucin-type *O*-glycans, have been identified, which have overlapping but different specificities [13]. This might be the reason why so far no distinct sequence motif comparable to the *N*-glycosylation sequon has been reported for *O*-glycosylation [13] (see also Chapter 9). Even if each of the different enzymes that are responsible for the initial step of *O*-glycosylation, for example mucin-type *O*-glycosylation, might preferentially recognize a sequence motif, this becomes indistinct in a data set of *O*-glycosylated proteins, because the sequence motifs for the different glycosyltransferases would overlay each other and therefore be difficult to determine.

A third type of glycosylation is *C*-mannosylation. In this case an α -Man residue is linked via a C–C bond to the indole C2 of a Trp side-chain [2]. This modification is very rare and therefore is not discussed here.

8.1.3 N-Glycosylation

8.1.3.1 N-Glycan Biosynthesis. *N*-Glycosylation is the most frequent modification of secretory and membrane proteins in eukaryotic cells [14]. Proteins passing through the secretory pathway of the ER–Golgi conduit are transferred to the cell surface where they become exported or anchored to the plasma membrane, to the extracellular matrix, or to the cell wall. The carbohydrate moiety of these glycoproteins faces the outside of the cell and forms part of the glycocalyx, a dense and complex array of carbohydrates which covers all cells in nature [15].

All eukaryotic cells produce *N*-glycans and have conserved the earliest biosynthetic steps. The assembly of *N*-linked glycans can be divided into four stages: (a) formation of a lipid-linked precursor oligosaccharide, (b) *en bloc* transfer to the polypeptide chain, (c) initial processing, and (d) final trimming and elongation of the oligosaccharide. Whereas the first three steps take place in the ER (see Figure 8.1), the final trimming and elongation of the glycans is performed in the Golgi (see Figure 8.2). Each individual glycosyltransferase (GT) displays a strong preference towards a single oligosaccharide motif. This leads to a linear, stepwise biosynthetic pathway of the branched oligosaccharide.

The evolutionarily highly conserved glycosylation process is initiated in the ER, where the $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$ tetradecasaccharide – also referred to as the *N*-glycan precursor – is assembled on the lipid carrier dolichyl pyrophosphate and then transferred *en bloc* to selected asparagine residues of polypeptide chains (see Figure 8.1). The biosynthesis of the precursor starts in the cytoplasm and involves a translocation of a lipid-linked $\text{Man}_5\text{GlcNAc}_2$ heptasaccharide to the ER lumen [16]. There it is further elongated to the full-length lipid-linked tetradecasaccharide $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2\text{-PP-Dol}$. The transfer of the precursor oligosaccharide to the asparagine side-chain of a nascent protein is catalyzed by the oligosaccharyl transferase [17, 18]. The glycosylation occurs while the polypeptide is still unfolded and therefore can be classified as a co-translational protein modification.

Initial processing of the peptide-linked tetradecasaccharide involves the stepwise removal of the three glucose residues by two *exo*-glucosidases. These enzymes are part of the calnexin/calreticulin cycle, a quality control system that mediates the protein folding (see Section 8.1.3.2). Properly folded proteins leave this cycle and are transported from the ER to the Golgi apparatus. Beginning in the ER and continuing in the *cis* portion of the Golgi, some or all of the α 1,2-linked mannose residues are removed by a series of mannosidases. In the medial and *trans* Golgi apparatus, the glycan chains are further remodeled to yield complex or hybrid-type *N*-glycan structures (see Chapter 2), which can exhibit up to five antennae (see Figure 8.2). Biantennary glycans are most abundant, but tri- and tetraantennary glycans are also common.

8.1.3.2 Quality Control System That Monitors Proper Folding of Glycosylated Proteins. *N*-Glycans have an indirect role in the process of protein folding. This role is based on binding of the newly synthesized glycoproteins to lectins in the ER, the sub-cellular site where proteins following the secretory pathway acquire their proper tertiary structures. In this process, the glycans serve as sorting signals that the cell modifies to reflect the folding status of the protein. Proteins that are not yet properly folded are prevented from exiting the ER to the Golgi apparatus [19, 20].

The final step in the formation of the initial lipid-bound oligosaccharide in the *N*-glycan biosynthetic pathway is the addition of a terminal α -1,2-linked glucose residue to form

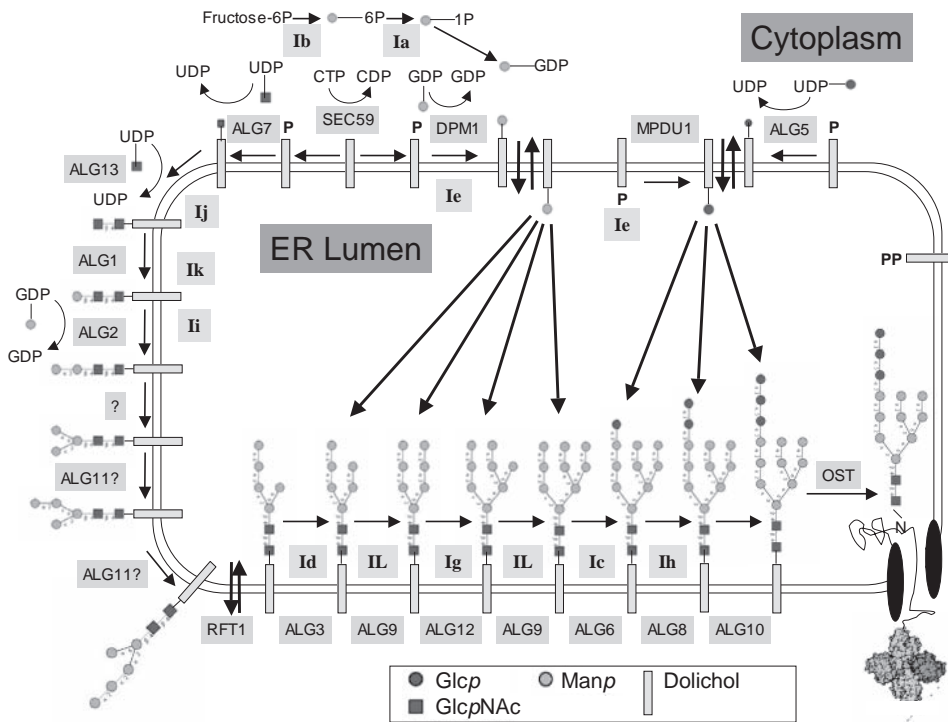


Figure 8.1 *N*-Glycan biosynthetic pathway – biosynthesis of lipid-bound oligosaccharide and transfer to a nascent polypeptide in the ER. The evolutionarily conserved *ALG* (asparagine-linked glycosylation) enzymes and other yeast loci involved in this pathway are displayed. The congenital disorders of glycosylation (CDG) disease classification (Ia–IL, see [38]) assigned to an enzyme's malfunction is depicted. Synthesis starts at the cytoplasmic face of the ER with UDP-GlcNAc and GDP-Man as glycosyl donors, transferring sugar residues onto dolichol (Dol). The $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ is then transferred to the luminal side with the help of *Rft1*, and elongated to the full-length lipid-linked oligosaccharide $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2\text{-PP-Dol}$ using Dol-P-Man and Dol-P-Glc. The oligosaccharide is subsequently linked by the oligosaccharyl transferase (OST) to the side-chain amido group of an asparagine residue within the consensus sequence Asn–Xaa–Ser/Thr of nascent secretory proteins. The glycosylation occurs while the polypeptide is still unfolded. Hence it can be classified as a co-translational protein modification. Adapted from [23] with slight modification. Reprinted, with permission, from Annual Reviews of Biochemistry, Volume 73, © 2004 by Annual Reviews, www.annualreviews.org. A full-color version of this figure is included in the Plate section of this book.

$\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$ (Figure 8.1). This terminal glucose residue is an important determinant for substrate recognition by the oligosaccharyl transferase [21]. Immediately after addition of the core glycan to a growing polypeptide chain, glucosidase I removes the outermost of the three glucose residues ($\text{Glc}_2\text{Man}_9\text{GlcNAc}_2$ is formed), followed by removal of the middle glucose residue by glucosidase II to yield $\text{Glc}_1\text{Man}_9\text{GlcNAc}_2$ (Figure 8.2). Glycoproteins with the $\text{Glc}_1\text{Man}_9\text{GlcNAc}_2$ epitope are recognized by the lectin domains of two ER-resident lectin chaperones, membrane-bound calnexin (CNX), and its soluble homolog, calreticulin (CRT) [22] (Figure 8.3).

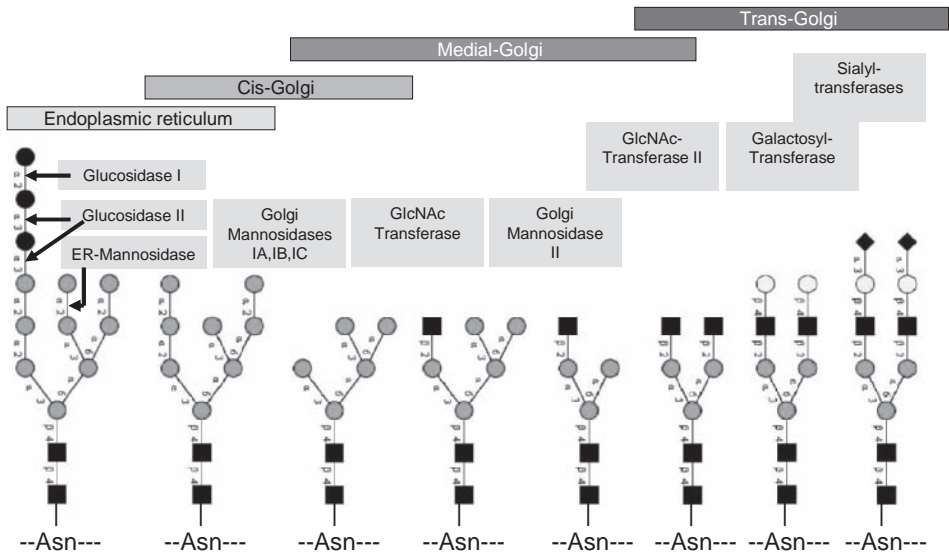
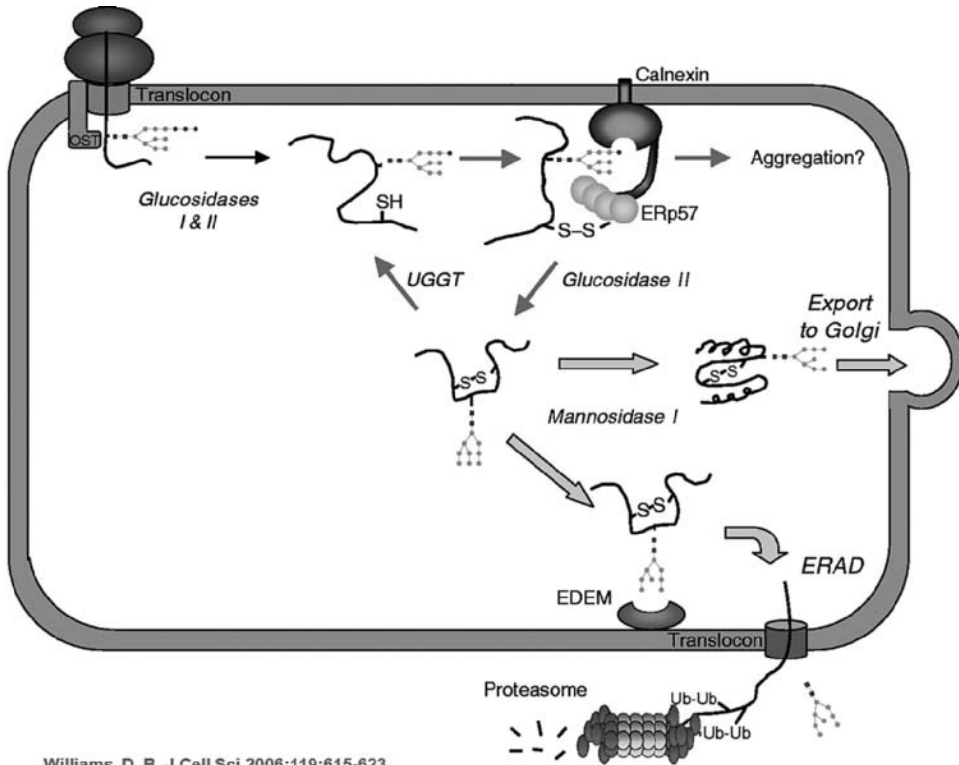


Figure 8.2 *N*-Glycan biosynthetic pathway – remodeling of the initial *N*-linked glycans to complex structures when transiting the ER and the Golgi apparatus. Following the transfer to the nascent polypeptide, the *N*-linked glycan is further processed, first by sequential removal of the terminal glucose residues by *exo*-glucosidases. Glucosidase I is responsible for the removal of the terminal glucose unit, and glucosidase II removes the two inner glucose residues, the last residue as part of the mediation of proper protein folding (see Figure 8.3). After it is properly folded, the glycosylated protein is ready for transport from the ER to the Golgi. The structure remaining after the action of the glucosidases is further truncated by a series of mannosidases that remove some or all of the α 1,2-linked mannose residues. This removal starts in the ER and continues in the *cis* portion of the Golgi apparatus. Further remodeling of the *N*-glycan core structure occurs in the medial and *trans* Golgi and results in complex (shown here) or hybrid type glycans which can exhibit up to five antennas. Although biantennary glycans are most abundant, tri- and tetraantennary glycans are also common.

In contrast to proteins in the cytosol, most proteins that fold in the ER acquire disulfide bonds through an oxidation process catalyzed by thiol–disulfide oxidoreductases. CNX and CRT form a complex with ERp57, a close homolog of the protein disulfide isomerase (PDI) having four thioredoxin-like domains, that catalyzes disulfide bond formation and isomerization [22]. When glucosidase II removes the remaining terminal glucose from the CNX/CRT-bound glycoprotein and $\text{Man}_9\text{GlcNAc}_2$ is formed, the glycoprotein dissociates from CNX/CRT. The glycoprotein is now able to leave the ER unless recognized by a soluble enzyme, UDP-Glc:glycoprotein glucosyltransferase (UGGT). UGGT only reglycosylates incompletely folded glycoproteins, and thus serves as a folding sensor in the cycle. If reglycosylated by UGGT, the glycoprotein rebinds to the lectins. A glycoprotein stays in the cycle until it is either properly folded and oligomerized or degraded.

It has been proven that when the *N*-glycosylation pathway is blocked, many polypeptides undergo improper or incomplete folding [1, 23]. Failing to reach the native conformation, they do not pass ER quality control. They are retained in the ER and eventually are degraded (Figure 8.3) [24, 25].



Williams, D. B. J Cell Sci 2006;119:615-623

Figure 8.3 Proposed mechanisms of the control system that mediates proper folding of glycosylated proteins. The polypeptide, as a nascent protein, enters the ER lumen via the translocon pore. The Asn–Xaa–Ser/Thr sequences may be recognized by OST and glycosylated with the preassembled $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$ tetradecasaccharide. The outer two glucose residues are then rapidly removed by glucosidases I and II to generate the $\text{Glc}_1\text{Man}_9\text{GlcNAc}_2$ oligosaccharide, which is recognized by calnexin (CNX) or its soluble homologue calreticulin (not shown). Cycles of glycoprotein release and rebinding are controlled solely by the removal and re-addition of the terminal glucose residue by glucosidase II and UDP-glucose:glycoprotein glucosyltransferase (UGGT), respectively. UGGT is the folding sensor because it only reglucosylates non-native glycoprotein conformers. Chaperone binding serves to retain non-native glycoproteins within the ER, and also recruits ERp57 to promote disulfide bond formation and isomerization. Folding takes place upon release from the chaperone, followed by further oligosaccharide trimming and export to the Golgi apparatus. For misfolded glycoproteins that remain for prolonged periods in the CNX cycle, trimming by mannosidase I generates a $\text{Man}_8\text{GlcNAc}_2$ structure that may be recognized by a putative lectin termed EDEM as part of a signal leading to retrotranslocation and proteasomal degradation [ER-associated degradation (ERAD)] [59]. Adapted with permission from [22] with slight modification.

8.1.4 O-Glycosylation

8.1.4.1 O-Glycan biosynthesis. In contrast to the biosynthesis of *N*-glycans, the initial step of which occurs co-translationally before the protein is folded (see above), *O*-glycan biosynthesis is a post-folding event. It takes place after folding and oligomerization of proteins in the late ER or in one of the Golgi compartments [7, 13]. As described above, *O*-glycosylation starts with the addition of a single monosaccharide to hydroxyamino acid,

predominantly Ser or Thr, of the protein chain. The type of the carbohydrate residue at the reducing end of the glycan chain (GalNAc, Xyl, GlcNAc, Gal, Man, Glc, or Fuc) is used to classify *O*-glycans [13]. In the following discussion, the major classes of *O*-glycans are described. For a comprehensive list of *O*-glycan classes and their occurrences in nature, the reader is referred to the literature [2, 13].

An α -linked GalNAc attached to a Ser or Thr side-chain, the so-called Tn antigen, forms the basis for mucin-type *O*-glycosylation. Depending on the sugar residue(s) directly attached to the GalNAc and the sugar linkage(s), eight mucin-type core structures are distinguished. Similarly to *N*-glycans, these core structures can be further extended. The glycosyltransferases that are involved in the elongation, branching, and termination of glycan chains, distant from the attachment site of the glycan to the protein, are regio- and stereospecific, that is, they add a particular monosaccharide with a certain linkage to a specific motif of the already present glycan chain. However, they are usually not specific for a certain core structure. Thus, the same enzymes that modify the terminal parts of *N*-glycans can also act on many *O*-glycan chains [13].

Glycosaminoglycans (GAGs) form another common type of *O*-glycans. GAGs are long, linear polysaccharides which contain a disaccharide repeating unit consisting of a GlcNAc or a GalNAc residue, together with a GlcA, IdoA, or Gal residue. Depending on the composition of the disaccharide repeating unit, three different classes of GAGs are distinguished in *O*-glycans: (1) dermatan sulfate and chondroitin sulfate (IdoA/GlcA + GalNAc), (2) heparin/heparan sulfate (GlcA/IdoA + GlcNAc) and (3) keratan sulfate (Gal + GlcNAc). Whereas keratan sulfate is linked to the core protein via either *N*- or core 2 *O*-glycans [26], the other GAGs are connected to Ser via a linker tetrasaccharide (GlcA β 1–3Gal β 1–3Gal β 1–4Xyl β 1–) [13, 27]. Hyaluronan, a fourth class of GAG (GlcA + GlcNAc), is not linked to proteins at all [27, 28]. The elongation of GAG chains is often followed by modification steps in which GlcA may be epimerized to IdoA, and sulfates are added. In the case of heparin/heparan sulfate, *N*-sulfation can occur at the C2 atom after *N*-deacetylation [29]. The product of one modification step is the substrate for the next step, leading to characteristic sulfation patterns. With the exception of some sulfotransferases, the transferases that are involved in GAG biosynthesis are specific for these pathways and are not involved in other glycoconjugate classes [13, 30].

O-Linked GlcNAc is a reversible type of *O*-glycosylation that involves only a single GlcNAc residue attached to a Ser or Thr side-chain. Unlike the other glycosylation classes, *O*-GlcNAc is not added during the secretory pathway in the ER or the Golgi apparatus. Instead, the enzymes that catalyze the addition (*O*-GlcNAc transferase, OGT) and removal (β -*N*-acetylglucosaminidase, *O*-GlcNAcase) of GlcNAc are found in the nucleus and cytosol, where they are involved in regulatory processes (see below) [31, 32].

O-Mannosylglycans were first discovered in yeast and for a long time were believed to be a specifically fungal type of glycosylation. However, they are also found on glycoproteins or glycopeptides in mammalian brain, nerves, and skeletal muscles [13, 33]. The biosynthetic pathway of yeast-type *O*-mannosylation differs from those of the other *O*-linked glycans described above. The latter mainly take place in the Golgi apparatus, require nucleotide sugars as sugar donors and thus resemble the extension processes of *N*-glycan chains. In contrast, *O*-mannosylation is initiated in the late ER. Similarly to the initial step of *N*-glycosylation, a dolichyl-phosphate activated sugar (Dol-P- β -Man) is transferred to the protein chain [33]. Recently, it could be shown that Dol-P- β -Man also serves as a donor in mammalian *O*-mannosylation [34, 35]. After the transfer of the first mannosyl moiety

to the protein, further extension of *O*-mannosylglycans is performed in the Golgi. Like the other Golgi-located glycosyltransferases, the enzymes that catalyze the elongation of *O*-mannosylglycans require nucleotide sugars as donors [33].

8.1.4.2 Functions of *O*-Linked Glycans. In addition to the functions that are common to *N*- and *O*-glycans, such as altering protein properties or providing recognition epitopes, an important role of mucin-type and GAG *O*-glycans is the binding of water. Mucins, which are proteins that are heavily *O*-glycosylated with mucin-type glycans, are frequently present at mucous membranes. The *O*-glycans often carry sialic acids, giving them a negative charge, which enables them to trap ions and bind large amounts of water. The mucins are cross-linked via intra- or intermolecular disulfide bonds and form a gelatinous matrix. This mucus forms a shield protecting the underlying epithelial cells, from, for example, pathogens and, in the case of the digestive tract, also from the body's own digestive enzymes [13, 28].

A second class of proteins that use *O*-glycan chains to bind water are proteoglycans. These are glycoproteins that carry GAG glycosylations. They carry between one and ~130 GAG chains. These chains can be relatively large (up to ~150 monosaccharides each), so that the molecular weight of the carbohydrate part often significantly exceeds that of the protein core. Proteoglycans are one of the major components of cartilage. The water that is bound by the GAG chains gives that tissue the ability to deform reversibly and thereby absorb pressure. Degradation of chondral proteoglycans is one of the major pathological features of arthritic diseases [28, 36]. Furthermore, GAGs in the extracellular matrix are involved in the assembly and stabilization of protein–protein complexes (e.g. the binding of growth factors to their receptors or growth factor oligomerization), or serve as co-receptors. Thereby they modulate various regulatory and signaling processes [27, 37]. In addition, heparan sulfate proteoglycans are involved in inflammatory processes after injuries by recruiting leukocytes into the damaged area [37]. Clinically, heparin/heparan sulfate is used as an anticoagulant [30].

O-GlcNAc modification resembles phosphorylation rather than “classical” glycosylation insofar as it consists of a single residue per modification site that is added or removed dynamically (see above). Both *O*-GlcNAc and phosphorylation are regulatory modifications, which sometimes work in a competitive way – either because a single amino acid is a target site for both modifications, or because *O*-GlcNAc blocks the ability of a kinase to phosphorylate residues in its neighborhood. The latter is the case, for example, in the C-terminal domain of RNA polymerase II. By modifying this polymerase, or transcription factors such as c-myc or sp1, *O*-GlcNAc is involved in transcriptional regulation [31, 32].

8.1.5 Congenital Disorders of Glycosylation

A protein modification that has been evolutionarily conserved from unicellular yeast to humans, in structure and biosynthetic pathway, is expected to have high biological importance. However, the specific biological function of glycosylation remained poorly understood for a long time. Only within the last decade did it become evident that if one of the glycan biosynthetic enzymes malfunctions then the cells in the body cannot glycosylate correctly. Several inherited human diseases, called congenital disorders of glycosylation (CDG), have been associated with deficiencies in the glycan biosynthetic pathways [38, 39]. However, the

impact on the body's structures and functions differs, and therefore the clinical symptoms vary, depending upon the altered enzyme.

CDGs constitute an instructive example of how the malfunction or absence of a specific glycosyltransferase, glycosidase, or nucleotide sugar transport protein (see Figure 8.1) can be directly associated with a disease. Most of the glycosylation disorders identified to date have been in the synthesis of *N*-linked oligosaccharides. In total there are six pathways (*N*-glycan, *O*-mannose, *O*-xylose, *O*-GalNAc, glycosphingolipid anchors, and glycosphingolipids) in which genetic disorders of glycosylation have been identified. So far, 19 types of CDG, each with its own unique defective enzyme, have been characterized [39]. A complete loss of glycans results in embryonic lethality [40]. Recent excellent reviews provide a comprehensive overview of the identified human diseases caused by genetic defects in the *N*- and *O*-glycosylation pathways in addition to the glycolipid synthesis pathway [38, 41].

Physicians observe even within each CDG type a wide variation of clinical symptoms, which may result from a combination of the severity of the mutations, genetic modifiers, and loss or substitution of specific glycans. Disease-causing defects in *O*-glycan biosynthesis occur primarily in the *O*-mannose and *O*-xylose pathways, and more patients are affected by these disorders than by *N*-glycan defects. Their clinical presentations are also usually distinct from those that are seen in *N*-glycosylation disorders: the *O*-mannose-based disorders all cause muscular dystrophy, an uncommon finding in *N*-glycan defects [41].

8.1.6 Sequence Dependence of Glycosylation

The reasons why some sequons or potential *O*-glycosylation sites are occupied and others are not are still not fully understood. Glycosylation is only found on the protein surface, but the initial step of *N*-glycosylation occurs co-translationally in the endoplasmic reticulum, before the protein folds to its tertiary structure. Hence the oligosaccharyltransferase that is responsible for this step does not necessarily “know” which part of the protein chain will become the surface in the folded state. And even among the potential glycosylation sites on the protein surface there are some that are occupied and others that are unoccupied. Therefore, the amino acids in the neighborhood of potential glycosylation sites probably affect the efficiency with which the site is glycosylated. The following section describes a statistical analysis of the amino acids around occupied glycosylation sites (Section 8.2), and the subsequent chapter describes tools to predict the status of potential glycosylation sites for a given protein sequence (Chapter 9).

8.2 GlySeq: Analysis of Experimentally Determined Occupied Glycosylation Sites

8.2.1 Data Sources for the Analysis of Protein Glycosylation

As many possible glycosylation sites are not occupied at all or are glycosylated inefficiently, there must be protein signals other than the glycosylation site itself that influence glycosylation [42]. To analyze statistically the influence of the amino acids in the neighborhood of

glycosylation sites on site occupancy, large data sets are needed. However, unambiguous information about which residues of a protein are glycosylated is hard to obtain. One data source that provides such data is the Protein Data Bank (PDB) [43] (<http://www.pdb.org>), which is the largest publicly available collection of biomolecular 3D structures. About 3.5% of the protein structures stored in the PDB feature covalently attached glycan chains, most of which are *N*-glycosidically linked (see Chapter 20.2). The fraction of glycosylated proteins in the PDB is significantly lower than in nature, where about 50% of all proteins are believed to be glycosylated [3]. The reasons for this difference are obvious. Many proteins in the PDB are bacterial proteins or recombinantly expressed in bacteria and therefore do not contain any glycans. If glycans are present, they may hamper crystal growth and hence are often removed enzymatically beforehand. Another problem that occurs with glycan structures is the fact that they are located on the protein surface and are highly flexible molecules. Hence they often do not provide sufficient electron density to resolve their 3D structures. For all these reasons, glycan chains are often not present in the PDB entries, even if the original protein is known to be glycosylated. This means that the presence of a glycan chain in a 3D structure unambiguously shows that the respective glycosylation site is occupied, but, in reverse, one cannot conclude that a potential glycosylation site is not occupied, if no attached glycan chain is present in the 3D structural data. The PDB data set is generated using the *pdb2linucs* software (see Chapter 20.2). These data are automatically updated every week.

Another valuable source for glycosylation data is the Swiss-Prot [44] database (<http://www.expasy.org/sprot/>), where information about occupied glycosylation sites is extracted from the literature by experts. As with the PDB, only data about occupied, and not about non-ambiguously unoccupied, glycosylation sites are given. The same applies to the third data source, the O-GLYCBASE [45] (<http://www.cbs.dtu.dk/databases/OGLYCBASE/>), which contains proteins with at least one experimentally verified *O*- or *C*-glycosylation site. Data from Swiss-Prot are extracted by searching that database for entries featuring the “FT <CARBOHYD>” tag. Glycosylation sites that are marked as “partial”, “potential”, “probable,” or “by similarity” are not entered into the data set.

8.2.2 *The GlySeq Software*

A statistical analysis of the glycoprotein data sets can be done using the GlySeq software. This program analyzes data sets of glycoprotein sequences and displays the occurrences of amino acids in the neighborhood of occupied glycosylation sites. Both absolute counts and deviations from natural abundances can be displayed. A web interface to GlySeq is available at the GLYCOSCIENCES.de web portal (www.glycosciences.de/tools/glyseq/). Through this web interface, six different data sets are accessible; a redundant and a non-redundant data set each for data retrieved from the PDB, Swiss-Prot, and a combination of both sources. In the non-redundant data sets, sequences with an identity of 95% or higher are excluded, so that identical or nearly identical sequences are omitted. This is of importance especially for the data derived from the PDB, since many protein structures are determined, for example, at different resolutions or with different ligands. This results in identical or, for example in cases where the influences of point mutations on the 3D structure are analyzed, nearly identical sequence entries. In the Swiss-Prot data set, only a few sequences are redundant at a 95% level.

For the sequence positions around occupied glycosylation sites, the diagrams presented in the following sections show the deviations of the observed occurrences of amino acids from the natural abundances of the residues. To calculate deviations, the frequency of an amino acid at each of the displayed positions is determined (F_{aa}). From this value, the natural abundance of the amino acid (A_{aa}) is subtracted and – for normalization purposes – the result is divided by the natural abundance. Multiplication by 100 gives a percentage value:

$$Dev_{aa} = \frac{F_{aa} - A_{aa}}{A_{aa}} \times 100$$

In the GlySeq output, position 0 marks the glycosylation site itself, -1 , -2 , and so on are the positions preceding the glycosylation site (towards the N-terminal end), and $+1$, $+2$, and so on the positions following it (towards the C-terminal end). Since the glycosylation site itself is always Asn in the case of *N*-glycans and Ser or Thr in the case of *O*-glycans, this position is omitted from the analysis. The same applies to position $+2$ of *N*-glycosylation sites, which is always Ser or Thr, or, in very few cases, Cys.

The following section presents some examples of results that can be obtained from glycoprotein sequence analysis using GlySeq. Recently published studies of *N*-glycosylation sites in the PDB [46] and Swiss-Prot [47] and *O*-glycosylation sites in O-GLYCBASE [48] revealed similar results.

8.2.3 Amino Acids in the Neighborhood of Glycosylation Sites

To investigate if the amino acids that are found in the neighborhood of glycosylation sites differ significantly from their natural abundances, the occurrences of different types of amino acids at the positions from -15 to $+15$ around *N*-glycosylation and *O*-glycosylation sites were examined. The results presented here are based on analyses of the non-redundant data set which contains data from both Swiss-Prot and the PDB. This data set (as of May 2007) contained 2759 occupied *N*-glycosylation sites. Much fewer data were available for *O*-glycosylation sites. In total, there were 934 occupied *O*-glycosylation sites (386 Ser + 548 Thr) in the data set used. In some PDB entries, carbohydrates are also linked to Asp or Glu. However, these are not real glycosylations but intermediate states of enzymatic reactions. Sequences that contain glycosylated Tyr residues are rarely found in the data sets and are therefore not discussed here.

Aromatic amino acids (Trp, Tyr, Phe) are over-represented around *N*-glycosylation sites. This effect is more pronounced in the positions preceding the glycosylation site than in those following it (Figure 8.4a). Around *O*-glycosylation sites, the occurrence of aromatic residues is below their natural abundance (Figure 8.4b). These results suggest that aromatic amino acids in the neighborhood of a potential glycosylation site may favor *N*-glycosylation, while they reduce the likelihood of *O*-glycosylation.

The opposite effect is observed with Pro. This amino acid is not found at position $+1$ of *N*-glycosylation sites, as noted in the description of the *N*-glycosylation sequon given above. (Strictly, there are three entries in the Swiss-Prot database where Pro is present at position $+1$ of an *N*-glycosylation site. However, one of these entries, IGHA2.HUMAN, features a variant form, the A2M(2) allotype, in which the Pro is replaced by a Ser and thus

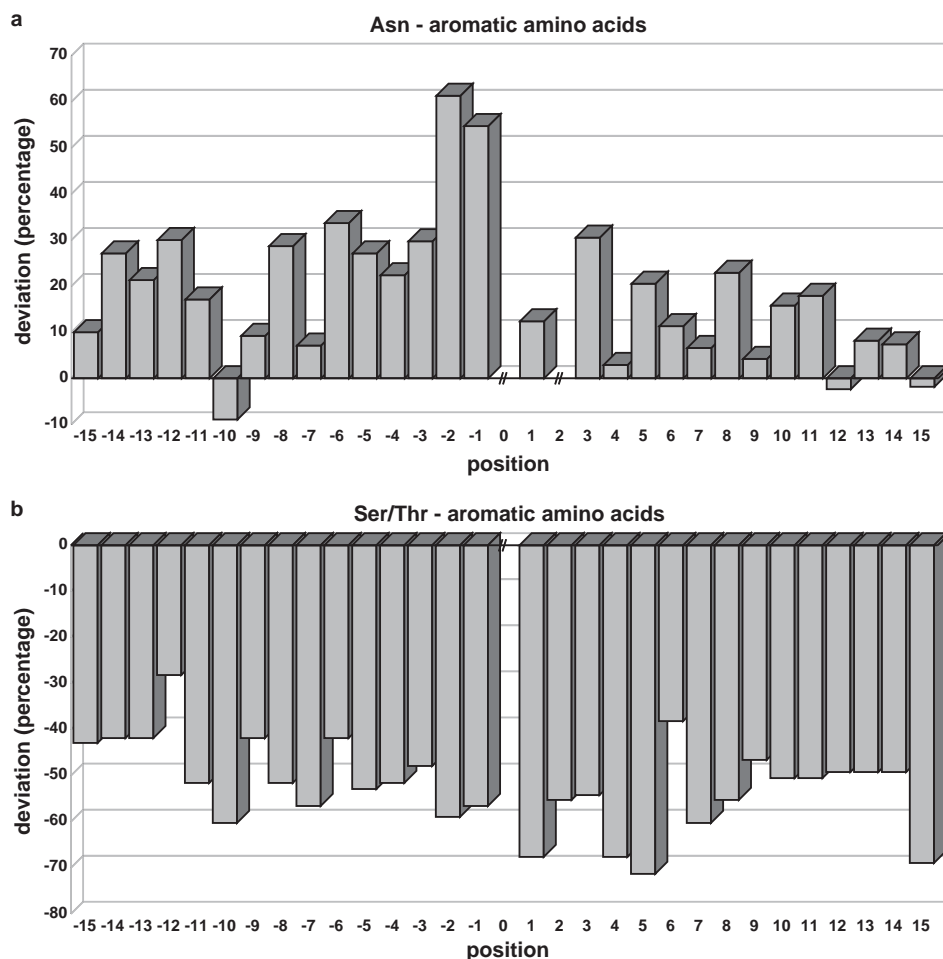


Figure 8.4 Frequency of aromatic amino acids (Trp, Tyr, Phe) in the neighborhood of glycosylation sites. Aromatic amino acids are over-represented in the environs of *N*-glycosylation sites (a), whereas they are under-represented around *O*-glycosylation sites (b).

reveals a correct sequon [49]. The second entry with Pro at position + 1 is NEPH_RAT, a glycopeptide of 21 amino acids that carries a trisaccharide of α -1–6-linked D-Glc_p residues instead of the normal *N*-glycan core structure [50]. This trisaccharide is not added by the oligosaccharyltransferase and therefore does not require the sequon. In the third entry with Pro following an Asn that is labeled as a glycosylation site, ITIH2_HUMAN, the respective position is explicitly mentioned not to be glycosylated in the related literature [51]. Therefore, this entry is wrongly assigned in the Swiss-Prot database.) At position +3 of an *N*-glycosylation site, the presence of Pro is not an exclusion criterion, but it is only seldom present (Figure 8.5a). This analysis supports the conclusion, found in previous studies [52–54], that Pro at this position significantly hampers *N*-glycosylation. The reason for this, and also for the fact that Pro at position +1 excludes glycosylation, is steric factors

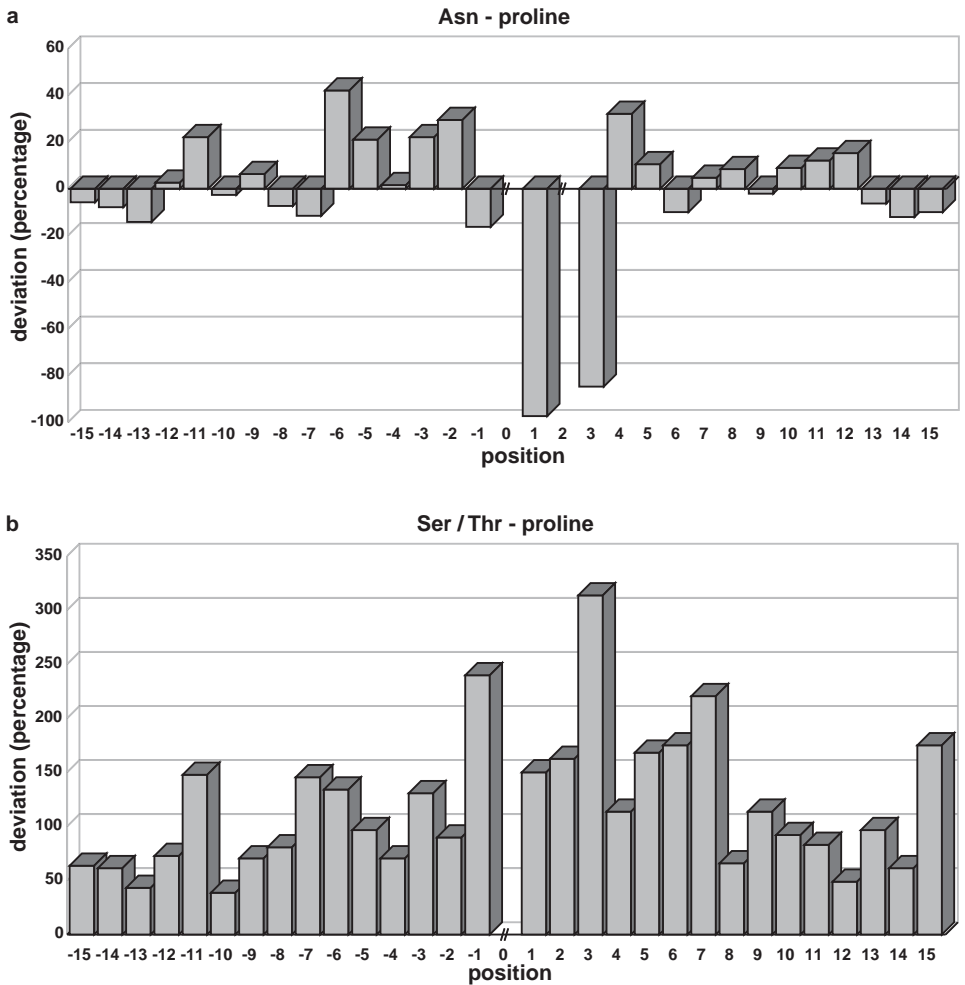


Figure 8.5 Frequency of Pro in the neighborhood of glycosylation sites. Pro is not found at position +1 of correctly assigned *N*-glycosylation sites. This residue is also seldom found at position +3 (a). For steric reasons, Pro present at those positions hampers *N*-glycosylation [55]. Around *O*-glycosylation sites, however, Pro is clearly over-represented (b).

[55]. More than one position away from the sequon, no clear trend can be observed for Pro, so it appears that this amino acid in these positions has only a slight effect on the efficiency of *N*-glycosylation.

In the neighborhood of *O*-glycosylation sites, Pro is clearly over-represented (Figure 8.5b). This leads to the conclusion that this residue has a positive influence on the occupancy of potential *O*-glycosylation sites.

Charged amino acids in the neighborhood of potential glycosylation sites appear to reduce the likelihood of the occupancy of both *N*- and *O*-glycosylation sites. Around occupied *O*-glycosylation sites, the frequency of basic amino acids (Lys, Arg, His) is about 45–75% below their natural abundances (Figure 8.6b). In the proximity of occupied *N*-glycosylation

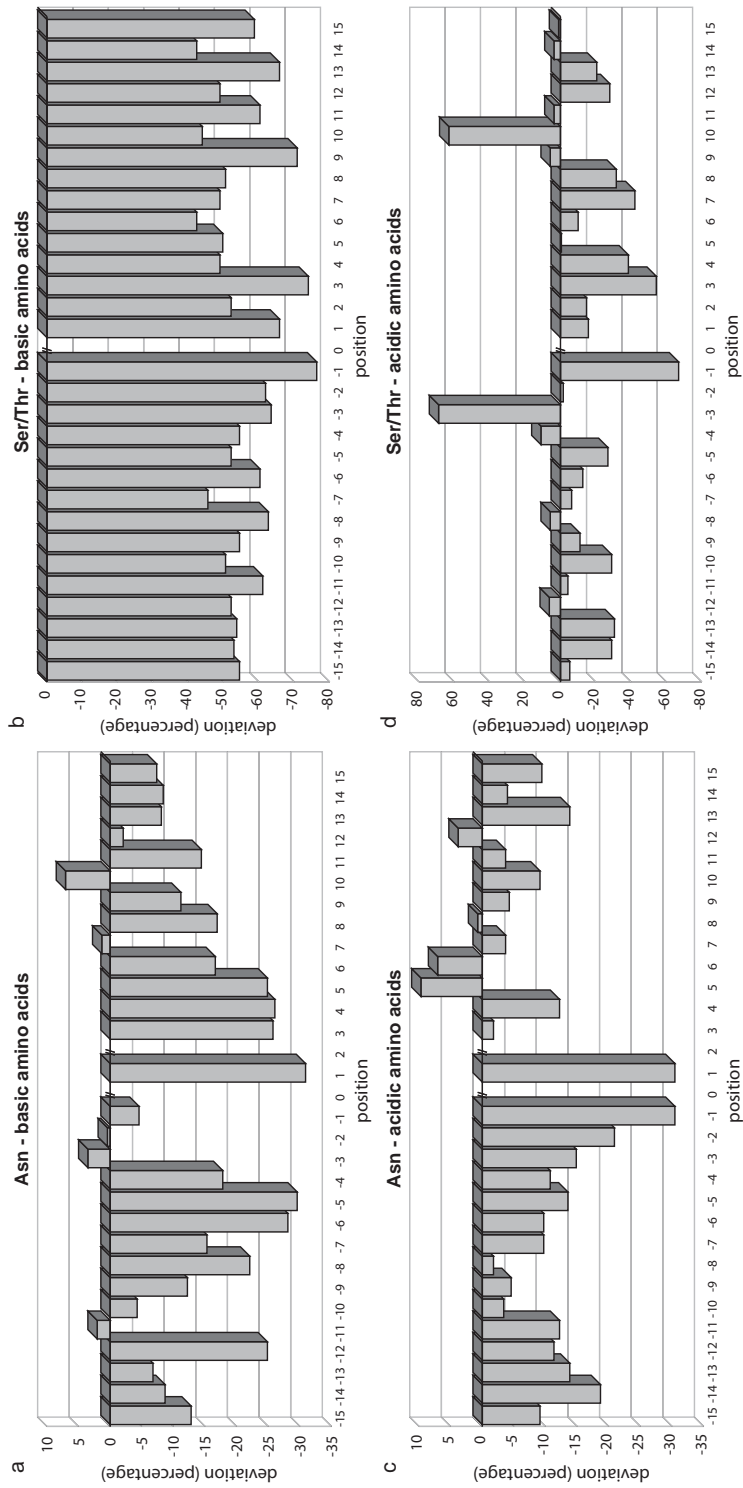


Figure 8.6 Frequencies of charged amino acids in the neighborhood of glycosylation sites. Basic amino acids are under-represented around both *N*-glycosylation (a) and *O*-glycosylation sites (b), with some exceptions for *N*-glycosylation sites. Acidic amino acids are also under-represented (c, d), but with exceptions for both *N*- and *O*-glycosylation sites.

sites, a similar effect is observed, albeit less pronounced (Figure 8.6a). Acidic amino acids (Asp, Glu) show a more complex pattern. In the positions preceding *N*-glycosylation sites, they are under-represented, whereas in the area following *N*-glycosylation sites, no clear trend can be observed for these residues (Figure 8.6c). Around *O*-glycosylation sites, acidic amino acids are under-represented at the majority of the positions (Figure 8.6d). At positions -3 and $+10$, however, these amino acids are clearly over-represented. The differences between the occurrences of basic and acidic amino acids are larger for *O*-glycosylation than for *N*-glycosylation sites.

In the average of all analyzed positions, charged amino acids are under-represented in the neighborhood of occupied glycosylation sites. This is surprising insofar as glycosylation is found on the protein surface, where charged amino acids are more frequent than in the protein core.

Ser or Thr residues in the neighborhood of occupied *O*-glycosylation sites are often themselves occupied. If only those Ser and Thr, which are not marked to be glycosylated in the data set, are counted, these residues are under-represented at most of the analyzed positions (Figure 8.7a). However, if both glycosylated and non-glycosylated Ser and Thr residues are investigated, they are clearly over-represented (Figure 8.7b). Between 50 and 75% of the Ser and Thr residues in positions -14 to $+14$ of occupied *O*-glycosylation sites are also glycosylated. This shows that, as reported in a previous study [48], *O*-glycosylation is often a bulk rather than a site-specific property. This observation can be explained by the fact that many GalNAc-transferases, which are responsible for the initial step of mucin-type *O*-glycosylation, possess a lectin domain. This domain recognizes GalNAc residues, which stabilizes the binding of the transferase to the protein chain in the neighborhood of already present *O*-glycan chains [56, 57].

The occupancy rate of potential *N*-glycosylation sites is influenced by the amino acid in position $+2$ (Ser or Thr). Depending on the data set, Asn-Xaa-Thr (NXT) sequons are between 50 and 75% more frequent than Asn-Xaa-Ser (NXS) sequons. The occupancy rate of NXT sequons is about 30% higher than that of NXS sequons. These findings are in agreement with the experimental result that a replacement of Ser by Thr in a sequon increases glycosylation efficiency [11, 58]. The frequencies of the amino acids at position $+1$ also differ between occupied NXS and NXT sequons. In sequons with Ser at position $+2$, Cys is highly over-represented at position $+1$ (Figure 8.8a). If position $+2$ of an *N*-glycosylation site is Thr, Cys is also over-represented, but less significantly so (Figure 8.8b). In general, small amino acids (Ala, Gly, Val, Ile) are frequently found at position $+1$. Of these, the smallest amino acids (Ala, Gly) are more frequent when Thr is found at position $+2$ than when Ser is present at that position.

8.3 Conclusion

Glycosylation is the most common and most complex modification of proteins. The glycan chains can change protein properties such as solubility or stability, are involved in the folding process, and serve as molecular addresses in protein trafficking. Furthermore, glycoproteins on the cell surface provide recognition epitopes for cell-cell and cell-matrix interactions. Malfunctions of enzymes that are involved in glycan biosynthesis result in inherited diseases called congenital disorders of glycosylation.

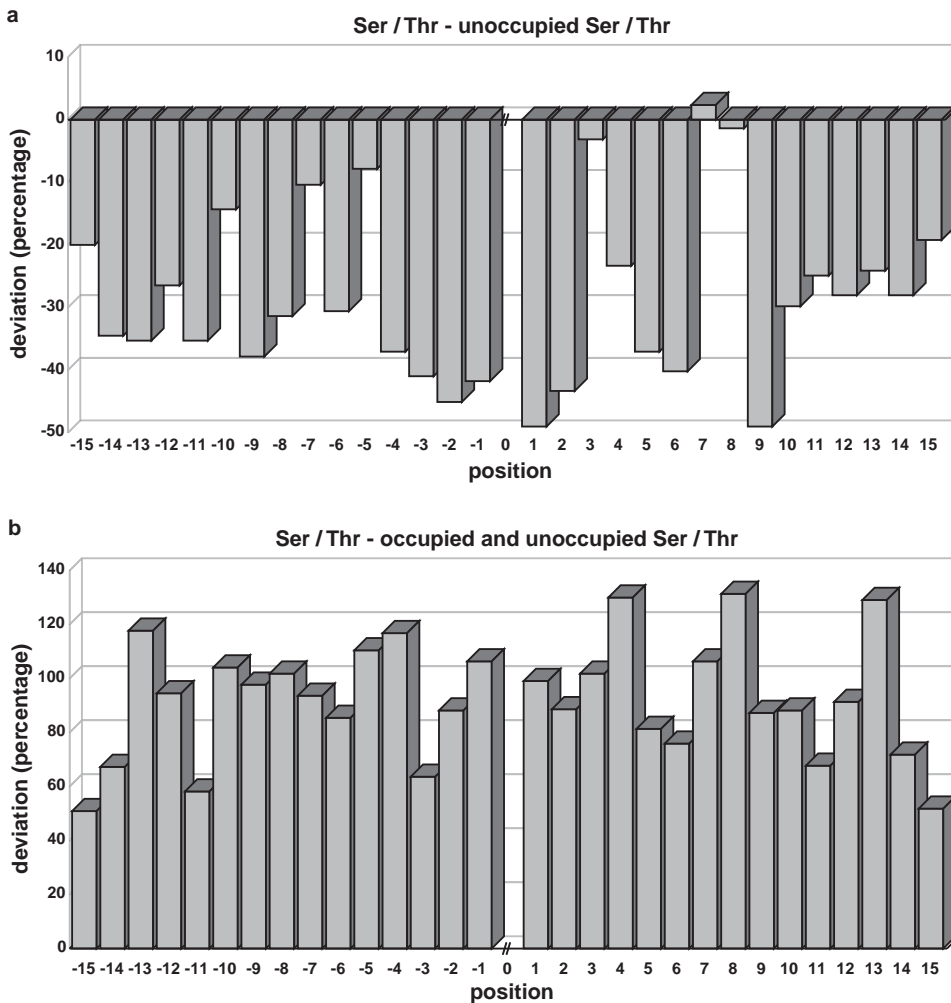


Figure 8.7 Occurrences of Ser and Thr residues in the neighborhood of occupied *O*-glycosylation sites. *O*-Glycosylation often occurs as a bulk property. If only unoccupied Ser and Thr residues are counted, these amino acids seem to be under-represented around *O*-glycosylation sites (a). If occupied Ser and Thr are also considered, these residues are over-represented (b). This shows that *O*-glycans often occur in clusters, which can be explained by the fact that many of the glycosyltransferases that catalyze the initial step of mucin-type *O*-glycosylation contain a lectin domain that recognizes *O*-glycan chains, which are already present [56, 57]. Thus, Ser or Thr residues close to an occupied *O*-glycosylation site are preferably glycosylated by those enzymes.

Nevertheless, not all potential glycosylation sites are occupied in nature. The analyses of occupied glycosylation sites presented in this chapter indicate that the amino acids in the neighborhood of potential glycosylation sites have an impact on the occupancy of these sites. This suggests that it should be possible to make predictions on the state of potential glycosylation sites from the sequence context. Such prediction methods are the subject of the following chapter.

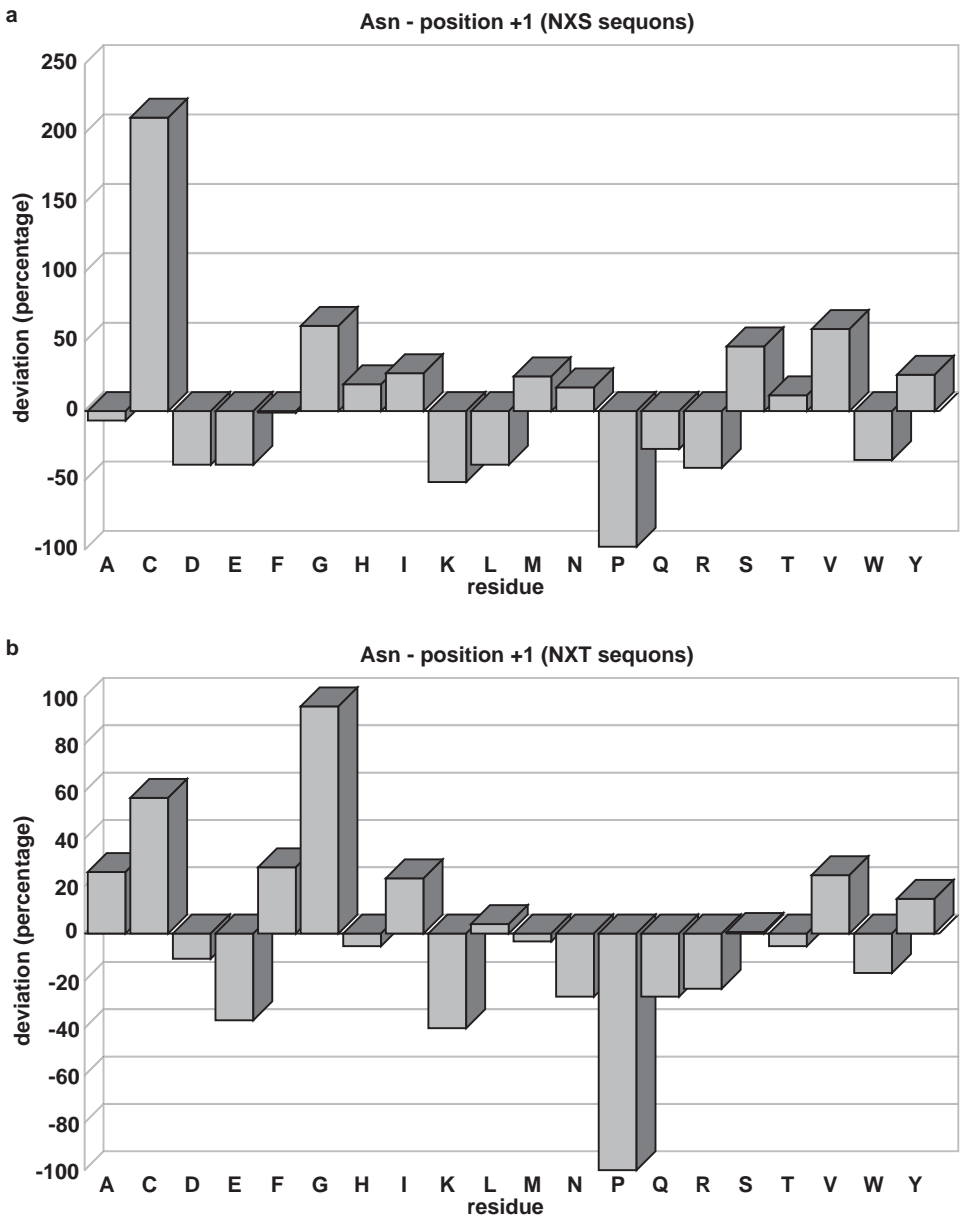


Figure 8.8 Frequencies of amino acids at position +1 of occupied *N*-glycosylation sites. The occurrences of the amino acids at position +1, the “Xaa” in the Asn–Xaa–Ser/Thr (NXS/T) sequon, differ depending on the residue at position +2 [(a) Ser; (b) Thr].

References

1. Helenius A, Aebi M: Intracellular functions of N-linked glycans. *Science* 2001, **291**:2364–2369.
2. Spiro RG: Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* 2002, **12**:43R–56R.
3. Apweiler R, Hermjakob H, Sharon N: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999, **1473**:4–8.
4. Wormald MR, Dwek RA: Glycoproteins: glycan presentation and protein-fold stability. *Structure Fold Des* 1999, **7**:R155–R160.
5. Le Borgne R, Hoflack B: Protein transport from the secretory to the endocytic pathway in mammalian cells. *Biochim Biophys Acta* 1998, **1404**:195–209.
6. Dube DH, Bertozzi CR: Glycans in cancer and inflammation – potential for therapeutics and diagnostics. *Nat Rev Drug Discov* 2005, **4**:477–488.
7. Brockhausen I: Mucin-type O-glycans in human colon and breast cancer: glycodynamics and functions. *EMBO Rep* 2006, **7**:599–604.
8. Rudd PM, Dwek RA: Glycosylation: heterogeneity and the 3D structure of proteins. *Crit Rev Biochem Mol Biol* 1997, **32**:1–100.
9. Dennis JW, Granovsky M, Warren CE: Protein glycosylation in development and disease. *BioEssays* 1999, **21**:412–421.
10. Marshall R: Glycoproteins. *Annu Rev Biochem* 1972, **41**:673–702.
11. Bause E, Legler G: The role of the hydroxy amino acid in the triplet sequence Asn–Xaa–Thr(Ser) for the N-glycosylation step during glycoprotein biosynthesis. *Biochem J* 1981, **195**:639–644.
12. Imperiali B, Hendrickson TL: Asparagine-linked glycosylation: specificity and function of oligosaccharyl transferase. *Bioorg Med Chem* 1995, **3**:1565–1578.
13. Wopereis S, Lefeber DJ, Morava E, Wevers RA: Mechanisms in protein O-glycan biosynthesis and clinical and molecular aspects of protein O-glycan biosynthesis defects: a review. *Clin Chem* 2006, **52**:574–600.
14. Aebi M, Hennet T: Congenital disorders of glycosylation: genetic model systems lead the way. *Trends Cell Biol* 2001, **11**:136–141.
15. Varki A: Nothing in glycobiology makes sense, except in the light of evolution. *Cell* 2006, **126**:841–845.
16. Helenius J, Ng DT, Marolda CL, Walter P, Valvano MA, Aebi M: Translocation of lipid-linked oligosaccharides across the ER membrane requires Rft1 protein. *Nature* 2002, **415**:382–383.
17. Kaplan HA, Welply JK, Lennarz WJ: Oligosaccharyl transferase: the central enzyme in the pathway of glycoprotein assembly. *Biochim Biophys Acta* 1987, **906**:161–173.
18. Dempsey RE, Imperiali B: Oligosaccharyl transferase: gatekeeper to the secretory pathway. *Curr Opin Chem Biol* 2002, **6**:844–850.
19. Parodi AJ: Protein glycosylation and its role in protein folding. *Annu Rev Biochem* 2000, **69**:69–93.
20. Lowe JB, Marth JD: A genetic approach to mammalian glycan function. *Annu Rev Biochem* 2003, **72**:643–691.
21. Burda P, Aebi M: The ALG10 locus of *Saccharomyces cerevisiae* encodes the alpha-1,2 glucosyltransferase of the endoplasmic reticulum: the terminal glucose of the lipid-linked oligosaccharide is required for efficient N-linked glycosylation. *Glycobiology* 1998, **8**:455–462.
22. Williams D: Beyond lectins: the calnexin/calreticulin chaperone system of the endoplasmic reticulum. *J Cell Sci* 2006, **119**:615–623.
23. Helenius A, Aebi M: Roles of N-linked glycans in the endoplasmic reticulum. *Annu Rev Biochem* 2004, **73**:1019–1049.
24. Ellgaard L, Helenius A: Quality control in the endoplasmic reticulum. *Nat Rev Mol Cell Biol* 2003, **4**:181–191.

25. Spiro RG: Role of N-linked polymannose oligosaccharides in targeting glycoproteins for endoplasmic reticulum-associated degradation. *Cell Mol Life Sci* 2004, **61**:1025–1041.
26. Funderburgh JL: Keratan sulfate: structure, biosynthesis, and function. *Glycobiology* 2000, **10**:951–958.
27. Raman R, Sasisekharan V, Sasisekharan R: Structural insights into biological roles of protein–glycosaminoglycan interactions. *Chem Biol* 2005, **12**:267–277.
28. Brooks SA, Dwek MV, Schumacher U: *Functional and Molecular Glycobiology*. Oxford: BIOS Scientific Publishers; 2002.
29. Aikawa J, Grobe K, Tsujimoto M, Esko JD: Multiple isozymes of heparan sulfate/heparin GlcNAc *N*-deacetylase/GlcNAc *N*-sulfotransferase. Structure and activity of the fourth member, NDST4. *J Biol Chem* 2001, **276**:5876–5882.
30. Liu J, Pedersen LC: Anticoagulant heparan sulfate: structural specificity and biosynthesis. *Appl Microbiol Biotechnol* 2007, **74**:263–272.
31. Wells L, Hart GW: *O*-GlcNAc turns twenty: functional implications for post-translational modification of nuclear and cytosolic proteins with a sugar. *FEBS Lett* 2003, **546**:154–158.
32. Vosseller K, Sakabe K, Wells L, Hart GW: Diverse regulation of protein function by *O*-GlcNAc: a nuclear and cytoplasmic carbohydrate post-translational modification. *Curr Opin Chem Biol* 2002, **6**:851–857.
33. Willer T, Valero MC, Tanner W, Cruces J, Strahl S: *O*-Mannosyl glycans: from yeast to novel associations with human disease. *Curr Opin Struct Biol* 2003, **13**:621–630.
34. Manya H, Chiba A, Yoshida A, Wang X, Chiba Y, Jigami Y, Margolis RU, Endo T: Demonstration of mammalian protein *O*-mannosyltransferase activity: coexpression of POMT1 and POMT2 required for enzymatic activity. *Proc Natl Acad Sci USA* 2004, **101**:500–505.
35. Akasaka-Manya K, Manya H, Nakajima A, Kawakita M, Endo T: Physical and functional association of human protein *O*-mannosyltransferases 1 and 2. *J Biol Chem* 2006, **281**:19339–19345.
36. Caterson B, Flannery CR, Hughes CE, Little CB: Mechanisms involved in cartilage proteoglycan catabolism. *Matrix Biol* 2000, **19**:333–344.
37. Bishop JR, Schuksz M, Esko JD: Heparan sulphate proteoglycans fine-tune mammalian physiology. *Nature* 2007, **446**:1030–1037.
38. Freeze HH: Genetic defects in the human glycome. *Nat Rev Genet* 2006, **7**:537–551.
39. Freeze HH, Aebi M: Altered glycan structures: the molecular basis of congenital disorders of glycosylation. *Curr Opin Struct Biol* 2005, **15**:490–498.
40. Marek KW, Vijay IK, Marth JD: A recessive deletion in the GlcNAc-1-phosphotransferase gene results in peri-implantation embryonic lethality. *Glycobiology* 1999, **9**:1263–1271.
41. Lehle L, Strahl S, Tanner W: Protein glycosylation, conserved from yeast to man: a model organism helps elucidate congenital human diseases. *Angew Chem Int Ed* 2006, **45**:6802–6818.
42. Shakin-Eshleman SH, Spitalnik SL, Kasturi L: The amino acid at the X position of an Asn–X–Ser sequon is an important determinant of N-linked core-glycosylation efficiency. *J Biol Chem* 1996, **271**:6363–6366.
43. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res* 2000, **28**:235–242.
44. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al.*: The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003, **31**:365–370.
45. Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE: O-GLYCBASE version 4.0: a revised database of *O*-glycosylated proteins. *Nucleic Acids Res* 1999, **27**:370–372.
46. Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR: Statistical analysis of the protein environment of *N*-glycosylation sites: implications for occupancy, structure and folding. *Glycobiology* 2004, **14**:103–114.
47. Ben-Dor S, Esterman N, Rubin E, Sharon N: Biases and complex patterns in the residues flanking protein *N*-glycosylation sites. *Glycobiology* 2004, **14**:95–101.

48. Julenius K, Molgaard A, Gupta R, Brunak S: Prediction, conservation analysis and structural characterization of mammalian mucin-type *O*-glycosylation sites. *Glycobiology* 2005, **15**:153–164.
49. Torano A, Putnam FW: Complete amino acid sequence of the alpha 2 heavy chain of a human IgA2 immunoglobulin of the A2m(2) allotype. *Proc Natl Acad Sci USA* 1978, **75**:966–969.
50. Shibata S, Takeda T, Natori Y: The structure of nephritogenoside. A nephritogenic glycopeptide with alpha-*N*-glycosidic linkage. *J Biol Chem* 1988, **263**:12483–12485.
51. Flahaut C, Capon C, Balduyck M, Ricart G, Sautiere P, Mizon J: Glycosylation pattern of human inter-alpha-inhibitor heavy chains. *Biochem J* 1998, **333**:749–756.
52. Mellquist JL, Kasturi L, Spitalnik SL, Shakin-Eshleman SH: The amino acid following an Asn—X—Ser/Thr sequon is an important determinant of N-linked core glycosylation efficiency. *Biochemistry* 1998, **37**:6833–6837.
53. Gavel Y, von Heijne G: Sequence differences between glycosylated and non-glycosylated Asn—X—Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng* 1990, **3**:433–442.
54. Roitsch T, Lehle L: Structural requirements for protein *N*-glycosylation. Influence of acceptor peptides on cotranslational glycosylation of yeast invertase and site-directed mutagenesis around a sequon sequence. *Eur J Biochem* 1989, **181**:525–529.
55. Bause E: Structural requirements of *N*-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem J* 1983, **209**:331–336.
56. Tenno M, Saeki A, Kezdy FJ, Elhammer AP, Kurosaka A: The lectin domain of UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase 1 is involved in *O*-glycosylation of a polypeptide with multiple acceptor sites. *J Biol Chem* 2002, **277**:47088–47096.
57. Wandall HH, Irazoqui F, Tarp MA, Bennett EP, Mandel U, Takeuchi H, Kato K, Irimura T, Suryanarayanan G, Hollingsworth MA, *et al.*: The lectin domains of polypeptide GalNAc-transferases exhibit carbohydrate-binding specificity for GalNAc: lectin binding to GalNAc-glycopeptide substrates is required for high density GalNAc-*O*-glycosylation. *Glycobiology* 2007, **17**:374–387.
58. Kasturi L, Eshleman JR, Wunner WH, Shakin-Eshleman SH: The hydroxy amino acid in an Asn—X—Ser/Thr sequon can influence N-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein. *J Biol Chem* 1995, **270**:14756–14761.
59. Ireland B, Niggemann M, Williams D, Imberty A: *In vitro* assays of the functions of calnexin and calreticulin, lectin chaperones of the endoplasmic reticulum. *Methods Mol Biol* 2006, **347**:331–342.

9 Prediction of Glycosylation Sites in Proteins

Karin Julenius^{1,2}, Morten B. Johansen³, Yu Zhang³, Søren Brunak³ and Ramneek Gupta³

¹Division of Matrix Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 171 77 Stockholm, Sweden

²Stockholm Bioinformatics Center, SCFAB, Stockholm University, 10691 Stockholm, Sweden

³Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, 2800 Lyngby, Denmark

9.1 Introduction

In the early days of molecular biology, the function and biochemical properties of a protein were usually known before its amino acid sequence. With access to exponentially growing sequence databases resulting from gene sequencing projects, this situation is now completely reversed. Almost as long as sequences have been accumulating in the databases, pattern recognition methods for their functional analysis have been designed. There is a long tradition of formulating motifs and consensus sequence methods for localization of features that may provide hints to the structure, function, and biochemical properties of a protein.

Protein glycosylation is more abundant and structurally diverse than all other types of post-translational modifications combined [1, 2]. Glycosylation is known to affect protein folding, localization, and trafficking, protein solubility, antigenicity, biological activity, and half-life, in addition to cell–cell interactions [3]. Prediction of glycosylation sites is a valuable tool when trying to characterize a new protein, for example, to help in interpreting mass spectrometry results. Methods to predict glycosylation sites in proteins have been shown to be important tools in the protein sequence-to-function predictor toolbox [4]. In addition, since glycosylation affects the structure of the protein and occurs primarily in surface-exposed regions, predicted glycosylation sites may be used to improve protein structure predictions. Prediction can also be useful in protein engineering, to engineer or abolish glycosylation sites and to design competitive inhibitors of glycosyltransferases [5]. Furthermore, knowing sites of glycosylation is often important in understanding and influencing antigenicity.

This chapter introduces ideas of pattern recognition for protein glycosylation site prediction from peptide sequence alone. A general introduction is provided to data-driven prediction methods for solving this problem, including a discussion on artificial neural networks. Also discussed are issues around performance evaluation and data collection. There is then a discussion about different linkages for glycosylation, and the prediction of their site specificities.

9.2 Data-driven Prediction Methods

Since the appearance of the very first DNA sequence, scientists have searched for patterns in biological sequences to explain and predict biochemical properties and phenomena. Early observations of simple conserved sequences led to the definition of consensus patterns. These consensus patterns were often a simplification that worked in a few cases (an example that works in most situations is “ATG” being the start codon), but were also inflexible and did not allow for any sequence substitutions. The next step was the introduction of weight matrices, where every position in an alignment of a certain pattern is scored according to the occurrence of that particular amino acid residue in that particular position in the sequence (or position in block alignment). This allowed for much more diverse patterns and provided the opportunity to score the hits according to how well they fit the pattern.

Stepping up even further in complexity, machine learning approaches were introduced to the field of biological sequence analysis. These included hidden Markov models (HMMs) and artificial neural networks (ANNs), which allowed for the classification of arbitrarily complex motifs. ANNs also have the advantage of being able to classify motifs containing correlations between different positions in the sequence. An example of a correlation would be that an unusually large amino acid could be accepted in a certain position if the adjacent amino acid is unusually small, or the other way round. Such cases can not be handled correctly by either weight matrices or HMMs. Although the sophisticated modern machine learning techniques usually are better suited than weight matrices at classifying highly complex sequence patterns, this comes at the cost of transparency. While it is fairly easy to infer the most important determinants of a functional site from a weight matrix, this becomes increasingly difficult when moving to a machine learning technique that takes into account correlations between positions. Hence, although the machine has learnt a pattern and can be shown to have good performance on an independent test set, it is not easy to ask the machine what features it has learnt and of what biological relevance these are.

9.2.1 *Neural Networks*

Artificial neural networks (ANNs) are capable of classifying highly complex and non-linear biological sequence patterns, where correlations between positions are important [6]. Not only does the network recognize the patterns seen during training, but it also retains the ability to generalize and recognize similar, though not identical, patterns. ANN algorithms have been used extensively in biological sequence analysis [7]. Most commonly used are the two-layered, feed-forward neural networks, trained with back-propagation. Other sophisticated machine learning methods used in biological sequence analysis are Support Vector machines and Kohonen self-organizing maps.

Training a neural network glycosylation site predictor involves presenting the network with the sequences around as many known glycosylation and non-glycosylation sites as possible (*training data*). For each glycosylation site, the weights of the neural network will be gradually adjusted to produce a network that will give a positive prediction (usually encoded as an output of 1.0). Correspondingly, for each non-glycosylation site, the weights will be adjusted to produce a negative prediction (usually encoded as an output of zero). The training examples are presented to the network many times until a maximum performance is reached. Care must be taken so that the network does not specifically learn only the exact training examples, which may lead to the loss of *generalization* ability (i.e. loss of the ability to recognize similar but not identical sites). This is usually referred to as *over-training* and may be the result of presenting the training examples too many times to the network and/or the choice of too complex network architecture. In general, making a model of more parameters can encode higher complexity, but also needs larger amounts of training data. Especially with glycosylation sites, high-quality training data (experimentally known glycosylation sites) are limited, so it is important to not introduce unnecessary complexity.

9.2.2 Evaluation Strategy

A relevant evaluation of the predictors is important not only for assessing performance and comparing different predictors, but also in the development of a predictor. During the development of a predictor, different window sizes in the input data (i.e. the number of residues around the glycosylation site that is presented to the network) and different network architectures (e.g. more hidden neurons means more complexity) are evaluated and the best one (or a set of best performing ones) is chosen. One also needs to evaluate how many times one should present the training data to the network to achieve maximum performance but to avoid over-training.

With a lot of biologically relevant problems, and glycosylation is no exception, the amount of experimentally verified data is very limited and therefore precious. In order to evaluate the predictor, one needs to divide the existing data into two groups – one for *training* and one for *testing*, with little sequence homology between the groups. However, the limited size of the available data makes it hard to reserve a large part of the data solely for network testing since this could mean sacrificing essential diversity in the training data. One commonly used method to tackle this difficulty is *cross-validation*, where the data are divided into a number of subsets, more or less equal in size. One subset is reserved for testing whereas the others are used for training (Figure 9.1). The predictive performance is recorded and the process is repeated with another subset as the test set (the previous test set is now included in the training set). The process is repeated until each subset has been the test set exactly once. The performances, recorded for each test set, are compiled and presented as the cross-validated performance of the neural network. It is very important to make the division into subsets so that the sequence similarity between subsets is low. Testing on data that are very similar to the training data will lead to over-estimation of the predictive performance and may indirectly lead to over-training.

In the development of a good predictor, one would like to maximize two parameters: *sensitivity* and *specificity*. One wants to identify as many true sites as possible (sensitivity), and on the other hand, ensure that those sites predicted as positives are in fact true (specificity). Predicting whether a particular site is modified or not is basically a classification task with two possible classes. In evaluating this, one needs to keep count of the number of

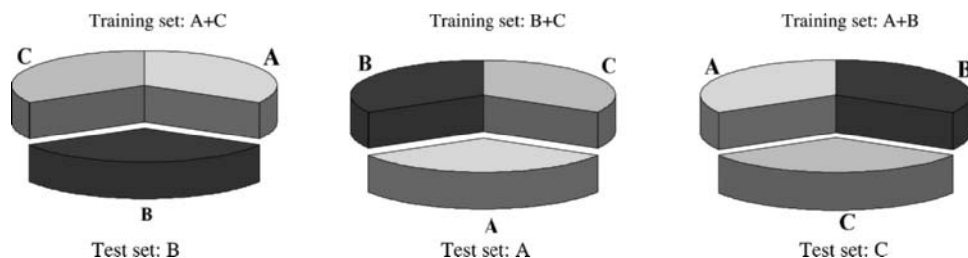


Figure 9.1 A cross-validation strategy for predictor evaluation. The data in this example are divided into three sets with low inter-set sequence similarity. The predictor is always evaluated on a test set not used for training. All data are used for training in this rotating scheme and the performance is calculated on the collected results of all test sets.

correct and incorrect predictions for each class, in total four possible outcomes: “true positives” (T_p) – experimentally verified modified sites that are also predicted to be modified; “true negatives” (T_n) – experimentally verified unmodified sites that are also predicted to be unmodified; “false positives” (F_p) – experimentally verified unmodified sites that are predicted (incorrectly) to be modified; and “false negatives” (F_n) – experimentally verified modified sites that are predicted (incorrectly) to be unmodified.

The sensitivity (S_n) of a method is defined as the proportion of positive sites that the method can correctly identify. Unfortunately, the definition of specificity (S_p) varies. In medical texts, the specificity (S_p^{med}) is usually defined as the proportion of negative sites correctly identified. However, another definition of specificity (S_p^{PPV}), more widespread in statistical texts, is the proportion of positive predictions that are in fact true [this is also known as the positive predictive value (PPV) of the method]. These measures are defined as follows:

$$S_n = \frac{T_p}{T_p + F_n}$$

$$S_p^{\text{med}} = \frac{T_n}{T_n + F_p}$$

$$S_p^{\text{PPV}} = \frac{T_p}{T_p + F_p}$$

As an unbiased overall performance estimator, the *Matthews correlation coefficient* (*CC*) [8] is widely used:

$$CC = \frac{T_p T_n - F_p F_n}{\sqrt{(T_n + F_n)(T_n + F_p)(T_p + F_n)(T_p + F_p)}}$$

The correlation coefficient takes both sensitivity and specificity into account. For example, a predictor that predicts every site to be positive, which is a predictor with 100% sensitivity but low specificity (S_p^{PPV}), would have a *CC* of zero. A predictor based on random guessing

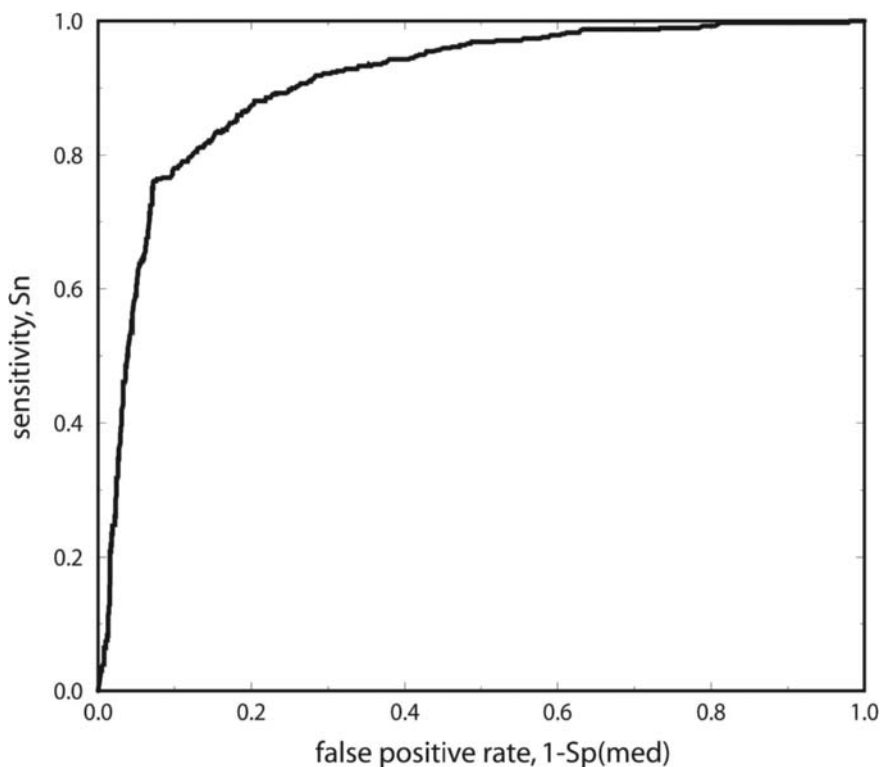


Figure 9.2 ROC curve for NetOGlyc 3.1 [9]. The sensitivity is the fraction of positive sites correctly predicted. The false positive rate is the fraction of negative sites wrongly predicted to be positive. A predictor making random guesses would perform along the diagonal and a perfect predictor along the y-axis.

would also have a CC of zero. A perfect predictor would have a CC of 1 and in the rare event of a predictor that is always inaccurate, this would have a CC of -1 .

Another common way of displaying and evaluating the performance of a predictor is the receiver operating characteristic (ROC) curve. The ROC is a type of curve that was developed in the 1950s as a by-product of research into making sense of radio signals contaminated by noise. More recently, it has become clear that ROC curves are remarkably useful in medical decision-making, since they can describe the quality of a classification method such as a predictor or a medical diagnostic tool. The ROC curve shows the trade-off between making many positive predictions, of which an increasing proportion are false, and making few predictions with higher quality and thereby missing some. An example of a ROC curve for a glycosylation predictor is shown in Figure 9.2 (this one is for NetOGlyc 3.1 [9]). The sensitivity is plotted against the *false positive rate* (the fraction of negative sites wrongly predicted to be positive = $1 - S_p^{\text{med}}$). A curve reaching far up into the upper left corner is to be preferred and completely random designation would perform along the diagonal.

When making two class predictions using neural networks that output a score between zero and one, a threshold is determined between zero and one that would obtain an

optimal separation between the two classes. For instance, in the case of glycosylated and non-glycosylated sites, a high threshold will result in few positive (“glycosylated site”) predictions, but a higher percentage of the predictions will in fact be true (in Figure 9.2, 40% of the positives can be found with only about 3% of the negatives being wrongly predicted to be positive). If a low threshold is used, a higher percentage of true positives are correctly predicted, but the prediction also generates more false positives (in Figure 9.2, 80% of the positives can be found while about 15% of the negative sites are wrongly predicted to be positive). Since non-glycosylated residues typically are much more common than glycosylated residues, it is normally preferred to keep the false positive rate as low as possible, as otherwise the specificity becomes very low. Available prediction methods usually use a default threshold that gives the maximum correlation coefficient, but often a raw output score can also be obtained and the users could define a threshold of their choice to suit their needs.

9.3 Data Collection

Our own experiences with databases include the curation of the glycosylation site database O-GlycBase, which was one of the first databases added to the EBI Sequence Retrieval System (SRS). SRS (<http://srs.ebi.ac.uk/>) was an early attempt to associate biological databases that existed mainly as large flat files (i.e. not in a relational database format). One of the very early parsers built into SRS was that of O-GlycBase. This was perhaps due to the free and easy availability of the entire database as a plain text file, or indeed due to the challenge in parsing plain text curated mostly by a molecular biologist. O-GlycBase (<http://www.cbs.dtu.dk/databases/OGLYCBASE/>) is a database of *O*- and *C*-glycosylated proteins where the *O*- and *C*-glycosylation sites have been experimentally mapped. This database was initially curated in concert with the construction of the NetOGlyc predictor [5, 9, 10], which relied on a substantial training set of sequence windows from glycosylated and non-glycosylated serines and threonines. The aim of this database was to collect site-mapped data with adequate reference so the data could be authenticated and glycosylation prediction reproduced if needed, but more importantly as a resource for other biologists.

In curating O-GlycBase, a valuable source of data is the Swiss-Prot database (<http://www.expasy.org/sprot/>). Swiss-Prot is an excellent resource for post-translational modifications and, since 2002, it categorizes glycosylation sites into the different types of attachments/linkages. Before inclusion in O-GlycBase, each Swiss-Prot entry with confirmed glycosylation sites is cross-checked against the original references. We also mine PubMed abstracts for data that Swiss-Prot does not contain, and communicate with the authors in ambiguous cases. Only proteins with experimentally verified glycosylation sites qualify for inclusion in O-GlycBase. Data found from PubMed, but not in Swiss-Prot, are usually communicated to Swiss-Prot for inclusion there. The most significant problem when compiling such a database is that there is almost never any negative data available. In other words, to know that a specific residue is *not* glycosylated can be extremely useful when making prediction methods. Unfortunately, such information is rarely published. To prove conclusively that a site is negative under all conditions is difficult, but to know that it is negative even in some contexts would be useful. Another problem while working with protein sequences from the literature is that database accession numbers are rarely mentioned. This leaves a lot of work for the database curator to track down the original

sequence and figure out the organism to which the protein belongs. Even when a database accession is found, the site numbering in the literature does not always correspond to the database entry, due to, for instance, signal peptides that are cleaved in the literature entry (and therefore not considered in counting amino acid position numbering in sequence).

To train a glycosylation predictor using a machine-learning approach, the non-glycosylation sites are as important as the glycosylation sites. Since they are rarely reported as such, their existence must be inferred from the original publications on the experimentally verified glycosylation sites. In many cases, the original work is a more or less exhaustive investigation of the glycosylation sites in a protein, or more often a part of a protein. In that case, the residues in the investigated part not reported to be glycosylated can be inferred to be most likely non-glycosylation sites. In other cases, the original work is clearly non-exhaustive, in which case nothing can be deduced about non-glycosylation sites in the sequence in question.

9.3.1 Sequence Logos

Glycosylation sites in this chapter have been illustrated using sequence logos, which are described here. In information theory, the Shannon entropy or information entropy is a measure of the uncertainty associated with a random variable. Here, Shannon information has been used to characterize sequence conservation around acceptor sites such as glycosylation and phosphorylation sites. Every “bit” of information represents the amount of information that can be obtained by answering one yes/no question. For a DNA sequence, every base position contains $\log_2 4 = 2$ bits of information, since it takes two yes/no questions to determine the base (e.g. “Is it an A or G?” and “Is it an A or C?”). For a protein sequence, the information content is $\log_2 20 \approx 4.32$ bits per position, since there are 20 possibilities in the form of amino acids.

A commonly used graphic representation of Shannon information content is a sequence logo [11] (see, e.g., Figure 9.3). Typically, sequence fragments are aligned at the acceptor

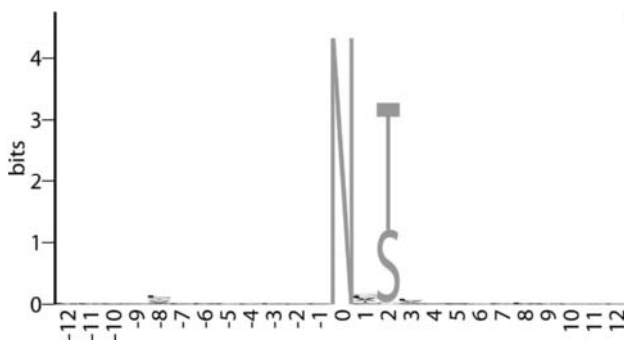


Figure 9.3 Shannon sequence logo for *N*-glycosylation sites. Position zero denotes the location of the glycosylated asparagine residue. Proline is disfavored at +1 and +3 relative to the asparagines. Position +2 is usually either serine or threonine (as the NXS/T motif dictates). Since no other sequence signals clearly stand out, a simple motif carries incomplete information for predicting *N*-glycosylation sites. A method that should have higher predictive capability than the NXS/T motif should consider correlations between sequence positions (not visible in such a sequence logo). A full-color version of this figure is included in the Plate section of this book.

site, and Shannon information is calculated for each position in a protein multiple alignment as

$$R = \log_2 20 - [H + e(n)]$$

where R is the amount of information in the position, $\log_2 20$ is the maximum information content for a protein sequence, $e(n)$ is a correction factor required when one only has a few (n) sample sequences, and H is the uncertainty, given by

$$H = - \sum_{aa=A}^Y f(aa) \log_2 f(aa)$$

where aa is one of the 20 amino acids and $f(aa)$ is the frequency of aa in that particular position.

Example 1: All amino acids are equally likely. The frequency for each is $1/20 = 0.05$.

$$H = -20 \times (0.05 \times \log_2 0.05) = -20 \times 0.05 \times (-4.32) = 4.32$$

$$R = \log_2 20 - 4.32 = 4.32 - 4.32 = 0$$

Example 2: Only one amino acid is represented. For this, the frequency is 1. For the other 19 amino acids, the frequency is 0.

$$H = -(1 \times \log_2 1) - 19 \times (0 \times \log_2 0) = -(1 \times 0) - 19 \times 0 = 0$$

$$R = \log_2 20 - 0 = 4.32$$

The information content in a position represents the importance of that particular position in the pattern described. This is indicated by the total height of the sequence logo in a particular position. Sequence logos are an *average* picture of a set of binding sites, which is why logos can have several letters at each position (in each stack). The amino acid residues are represented by their one-letter code and they are color-coded according to the following scheme: positively and negatively charged residues are shown in blue and red, respectively, uncharged polar residues are green, and hydrophobic residues are black. The size of each letter in a logo is determined by multiplying the frequency of the corresponding amino acid by the total information at that position [$f(aa) \times R$]. For more information on Shannon logos, the website <http://www-lmmb.ncifcrf.gov/~toms/> is recommended.

9.4 Linkage-specific Glycosylation

With the exception of glycation, which is a non-enzymatic reaction that will be treated in the last section of this chapter, glycosylation of protein occurs enzymatically in biological systems. A glycosyltransferase catalyzes the transfer of sugar from the donor and is thus responsible for the formation of the glycosidic linkage. A glycosyltransferase has specificities for a nucleotide sugar donor and an acceptor, which in the case of linking a glycan

Table 9.1 Summary of publicly available predictor services.

type	name	method	URL
N-glycosylation	PROSITE	consensus pattern	www.expasy.org/prosite/
mucin-type	NetNGlyc	neural network	www.cbs.dtu.dk/services/NetNGlyc/
	NetOGlyc 3.1	neural network	www.cbs.dtu.dk/services/NetOGlyc/
	Oglyc	support vector machines	www.biosino.org/Oglyc
O- α -GlcNAc	DictyOGlyc	neural network	www.cbs.dtu.dk/services/DictyOGlyc/
O- β -GlcNAc	YinOYang	neural network	www.cbs.dtu.dk/services/YinOYang/
proteoglycans	NetPGlyc 1.1	neural network	www.cbs.dtu.dk/services/NetPGlyc-1.1/
GPI-anchor	big-PI	linear function	mendel.imp.ac.at/gpi/gpi_server.html
	DGPI	rule based	129.194.185.165/dgpi/DGPI_demo_en.html
	GPI-SOM	Kohonen self-organized map	gpi.unibe.ch/
C-mannosylation Glycation	NetCGlyc 1.0	neural network	www.cbs.dtu.dk/services/NetCGlyc/
	NetGlycate	neural network	www.cbs.dtu.dk/services/NetGlycate/

to a protein is a polypeptide motif. Because of the strict donor and acceptor specificities of each glycosyltransferase, each enzyme can add only one type of sugar in a specific linkage. This concept is often referred to as the one enzyme, one linkage rule. Although each glycosyltransferase is highly specific, there are often several different glycosyltransferases with similar or overlapping specificities [12]. This is important to take into account when trying to predict glycosylation sites. One hopes for a predictor that has the same specificity as one glycosyltransferase or a group of highly related glycosyltransferases. Hoping that a predictor based on machine learning, would learn the specificities of all possible glycosyltransferases with a polypeptide acceptor motif is unrealistic. Table 9.1 provides a summary of publicly available predictor methods.

Protein glycosylation can be divided into four main categories, depending on the linkage between the amino acid and the sugar. These are *N*-linked glycosylation, *O*-linked glycosylation, glycosylphosphatidylinositol (GPI) anchor attachments and *C*-mannosylation. *N*-Glycosylation is the addition of a tetradecasaccharide (GlcNAc₂-Man₉-Glc₃) to the amino group (NH₂) of an asparagine. GPI anchors refer to glycosylphosphatidylinositol groups attached near the C-terminus of a protein chain that anchor the protein to the cell membrane. *C*-Mannosylation is the attachment of an α -mannopyranosyl residue to the indole C2 of a tryptophan residue via a C–C link.

In *O*-glycosylation, a sugar is attached to the hydroxyl group (OH) of a serine or threonine. *O*-Linked glycosylation reactions may happen at two cellular locations. Those taking place in the Golgi are initiated by the addition of various reducing terminal linkages such as *N*-acetylgalactosamine, *N*-acetylglucosamine, mannose, fucose, phosphodiester-linked *N*-acetylglucosamine, glucose, galactose or xylose to hydroxylated amino acids (usually serine or threonine). Recently, however, *O*-glycosylation has also been shown to occur in the nucleus and cytoplasm of cells [13]: this is characterized by the attachment of a monosaccharide, *N*-acetylglucosamine, to a serine or threonine residue. There is every reason to suspect that different types of *O*-glycosylation have different acceptor motifs, since

the glycosyltransferases involved in the recognition processes are different. Therefore, the different types of *O*-glycosylation will be dealt with separately, and only those for which there are available predictors are included.

9.4.1 *N*-Glycosylation

N-Linked glycosylation is perhaps the best studied and earliest known type of glycosylation. It is characterized by the transfer of a lipid-linked tetradecasaccharide (GlcNAc₂-Man₉-Glc₃) to an asparagine within a nascent polypeptide [14]. The process is catalyzed by a single enzyme, oligosaccharyl transferase (OST), and occurs co-translationally in the endoplasmic reticulum prior to protein folding. After the initial transfer of the triantennary oligosaccharide complex, diversification of the glycan is catalyzed by enzymes in the endoplasmic reticulum and Golgi apparatus during the maturation of the protein. The mature glycoforms fall into three main categories: high mannose, hybrid and complex. The process of linking the tetradecasaccharide to asparagine as catalyzed by OST seems to be conserved through eukaryote evolution [15]. However, the ability to form hybrid or complex *N*-glycans varies in different eukaryotic systems and during development within a given system or cell type.

Several congenital disorders causing aberrant *N*-glycosylation have been described [16]. These range from severe multisystemic disorders to disorders restricted to specific organs. Specific removal of an *N*-linked glycan by the site-directed mutagenesis of the asparagine residue in question can affect the secretion, function, stability, serum half-life, and/or trafficking properties of a protein [17–19]. Engineering additional *N*-glycosylation sites in a recombinant protein has been shown to increase the half-life *in vivo* and thus reduce the necessary dose of a protein-based pharmaceutical [20].

9.4.1.1 Proposed Sequence Motif. Contrary to widespread belief, acceptor sites for *N*-linked glycosylation on protein sequences are not well characterized. The sequence motif (*sequon*) Asn–Xaa–Ser/Thr (where Xaa is any amino acid except Pro) has been defined as a prerequisite for the modification [21, 22], but this is not sufficient to make an asparagine act as a glycosylation acceptor site. One- to two-thirds of these *N*-glycosylation sequons are estimated to be modified and the pattern by itself is thus not discriminatory between glycosylated and non-glycosylated asparagines.

Very few exceptions to this sequon have been found *N*-glycosylated, the most notable (but still rare) exception being the substitution of the +2 position serine/threonine with cysteine [23]. Proline at the Xaa position is a strong deterrent to a sequon acting as an acceptor site, presumably due to the conformational constraints it induces [21]. *In vivo*, position +2 threonine-containing sequences are almost three times more likely to be glycosylated than the corresponding serine-containing analogs [24, 25]. However, the efficiency of NXT glycosylation *in vitro* exceeds that of NXS sequences by as much as 40-fold [26].

9.4.1.2 Sequence Logo. Figure 9.3 illustrates the sequence context around the asparagine acceptor site for *N*-glycosylation based on 469 known sites from 192 proteins. It is hard to see any further information than the sequon (asparagine at the acceptor position, absence of proline at +1 and serine or threonine at position +2). Detailed analysis shows that proline is disfavored not only at position +1 (which was previously known) but also at position

+3. Apart from this, little new can be said about individual sequence positions in terms of *N*-glycosylation preference.

9.4.1.3 Existing Predictors. All new Swiss-Prot [27] sequences are routinely scanned for sequons indicating potential *N*-glycosylation sites and the results are noted in the Swiss-Prot entry using the following assignment in the feature list: “CARBOHYD . . . N-linked (GlcNAc . . .) (Potential)”. The method is publicly available and can be used for any protein sequence at the PROSITE web page: <http://www.expasy.org/prosite> [28]. This method identifies sequons, but does not attempt to discriminate between glycosylated and non-glycosylated sequons.

NetNGlyc is a prediction method for *N*-glycosylation sites, based on artificial neural networks that examine local sequence context around the sequon. The method was trained on a carefully curated human data set containing 469 positive sites in 192 proteins and 309 negative sites in 218 proteins. From over 1000 neural networks with different parameters that were trained, a jury of nine neural networks was selected with maximum sequence window size 21. The additional inputs of amino acid composition, protein length, and relative position of site on chain also enhanced the performance slightly. In a cross-validated performance, the networks could identify 86% of the glycosylated and 61% of the non-glycosylated sequons, with an overall accuracy of 76% (Matthews correlation coefficient 0.486). The method can be optimized for high specificity or high sensitivity by choosing a different threshold. For example, over 95% sensitivity (up to 98%) can be obtained at the cost of 50% specificity (high false positive rate). Alternatively, very high specificity (90%) can be achieved for the top 20% of scores (see Figure 9.4). Post-filtering rules, such as the fact that a signal peptide is required for *N*-glycosylation to occur, further improve performance. The

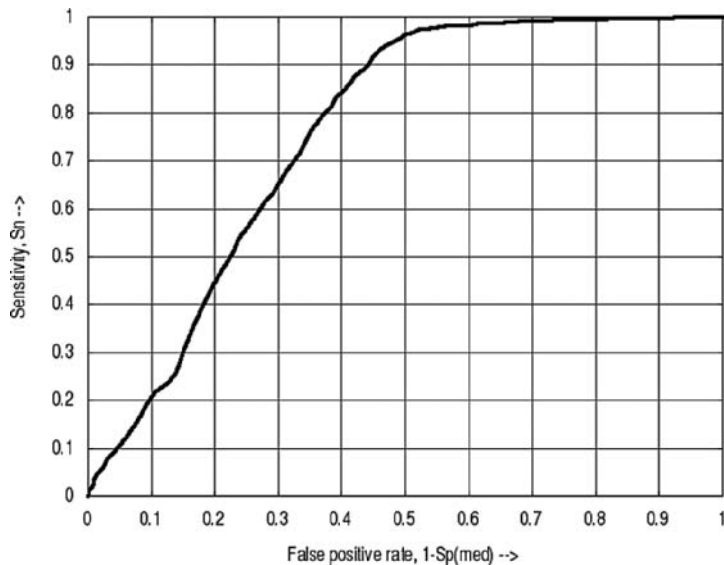


Figure 9.4 The ROC curve illustrates cross-validated performance of NetNGlyc. Over 95% sensitivity can be achieved if allowing for a 50% false positive rate on NX[ST] sequons. 90% specificity (a false positive rate of 0.1) can be obtained for the top 20% of the scores. The area below the curve is approximately 0.75.

method is publicly available for prediction at <http://www.cbs.dtu.dk/services/NetNGlyc/> and reports a score that helps in selecting sites at either high sensitivity or high specificity.

Senger and Karim [29] used artificial neural networks to predict variable site occupancy for *N*-linked glycosylation sites. Their model, based on a 10 amino acid sequence window around the asparagine acceptor site, employed a reduced amino acid alphabet grouping amino acids based on, for example, charge, size, and hydrophobicity. The model distinguished robust glycosylation (always occupied asparagine) from variable site occupancy. The data set was limited to 48 sequences where five sequences were reserved as a test set. Trained networks were used to predict site occupancy characteristics of wild-type and mutants of the rabies virus glycoprotein (rgp), and the results correlated well with earlier published experiments. The model suggested a decrease in glycosylation site occupancy robustness in the presence of positively charged residues. The authors discussed the impact of using such a model in the pharmaceutical industry to eliminate variable site occupancy glycosylation in recombinant products. As of mid-2007, their predictor model was not available publicly for use. A limitation of the predictor, which the authors acknowledge, is the limited training data set used for building the model. If available, more variable site occupancy data could be used to train a method that, if made publicly available, would be a good complement to NetNGlyc.

9.4.2 Mammalian Mucin-type (*O*-GalNAc) Glycosylation

One of the most abundant types of mammalian glycosylation is when an *N*-acetylgalactosamine (GalNAc) is α -1 linked to the hydroxyl group of a serine or threonine residue. This type of glycosylation is also called “mucin-type”. Mucin-type glycans are found on many secreted and membrane-bound mucins, but also on other glycoproteins. Mucins typically have very high carbohydrate content (>50% of the dry weight) and are the principal component of mucus, the gel that protects epithelial surfaces from dehydration, mechanical injury, proteases, and pathogens [30]. The protein backbone of a mucin contains a number of repetitive sequences, including virtually all the *O*-linked oligosaccharide attachment sites. Although these differ in terms of length and sequence from mucin to mucin, they all have a high serine, threonine, and proline content and are sometimes referred to as Ser/Thr/Pro-rich domains. Due to the steric hindrance introduced by the glycans, these domains adopt a stiff extended conformation, with an average length of 2.5 Å per amino acid residue [31, 32]. The exact locations of the *O*-glycosylation sites in these mucins are not specifically conserved between mammalian species, suggesting that the function of mucin-type glycosylation is directed not so much at specific interactions, but rather to changing the bulk properties of the protein [9].

Mucin-type glycosylation takes place in the rough endoplasmic reticulum and the Golgi complex after *N*-glycosylation, folding, and oligomerization. The process is mediated by at least 21 different UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferases [31]. From sequence similarity, it is estimated that there are up to 24 unique GalNAc-transferase genes; see [33] for a review. The different transferases have overlapping, but different, specificities and are differentially expressed [31, 33].

9.4.2.1 Proposed Sequence Motif. Although no consensus sequence has been formulated, many studies have noted the skew in amino acid composition around mucin-type *O*-glycosylation sites [5, 34, 35] with a higher frequency of prolines, serines, threonines,

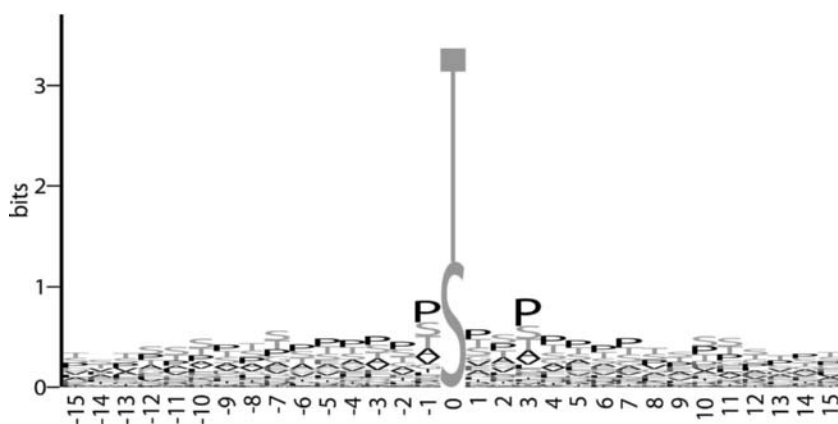


Figure 9.5 Shannon sequence logo for mammalian mucin-type *O*-glycosylation sites. Position zero denotes the location of the glycosylated serine or threonine residue. There are no strong position-specific signals for any particular amino acid at any position, but prolines, serines, and threonines are clearly over-represented on both sides of the glycosylation site. A full-color version of this figure is included in the Plate section of this book.

and alanines than expected. A number of studies have investigated the effect of flanking residues in *in vitro* experiments on synthetic peptides and especially the importance of prolines at certain positions has been confirmed [31, 33].

9.4.2.2 Sequence Logo. Figure 9.5 illustrates the sequence context around the accepting serine/threonine of mucin-type glycosylation based on 421 experimentally verified sites from 86 proteins. It can be seen that threonine is more common than serine as an acceptor. No consensus can be formulated from the logo, but it is evident that the amino acids proline, serine, and threonine are over-represented in a fairly large region around the site. Other amino acids with some over-representation are alanine, valine, glutamate, and glycine. The favored residues seem to be especially important in positions -1 and $+3$.

9.4.2.3 Existing Predictors. Early prediction methods for mucin-type *O*-glycosylation sites include a weight matrix method [34], a vector projection method [36], and two based on neural network method [5, 10]. The most widely used predictor, NetOGlyc 3 [9], was trained using a neural network-based method on over 40% more data than any of the previous methods (in total 421 positive and 2063 negative sites), due to experimental advances in the past few years. It is based on a fairly new principle, since it is trained not only on sequence data, but also on sequence-derived features with amino acid composition around the glycosylation site turning out to be the most important one. NetOGlyc 3.1 correctly predicts 76% of glycosylated sites (S_n) and 93% of non-glycosylated sites (S_p^{med}) and is publicly available at <http://www.cbs.dtu.dk/services/NetOGlyc/>. See also Figure 9.2, which shows the ROC curve of NetOGlyc 3.1. The prediction server generates warnings for sequences containing no predicted signal peptide [37], since such sequences are unlikely to be modified by mucin-type glycosylation.

More recently, a prediction method called Oglyc was developed using support vector machines [38]. The positive dataset was constructed from Swiss-Prot annotations of mammalian mucin-type glycosylation sites without any further verification. The negative dataset was constructed by random selection of serines and threonine sites in mammalian

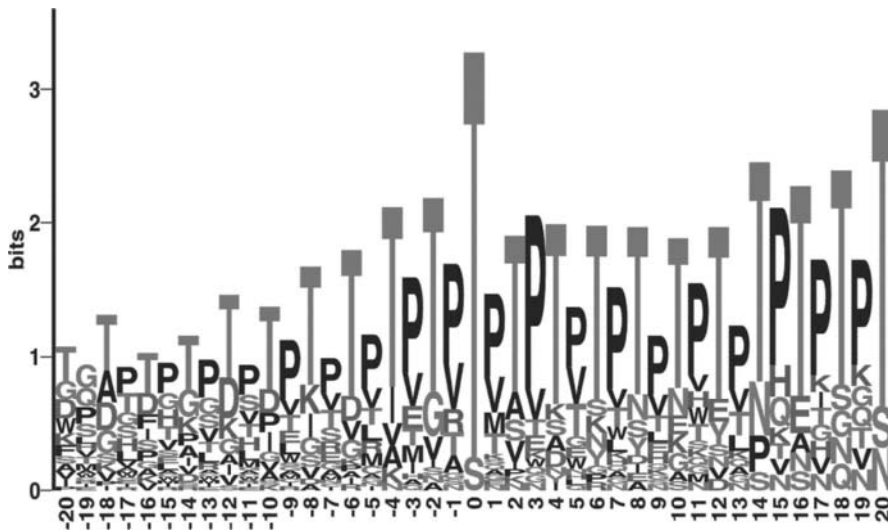


Figure 9.6 Shannon sequence logo for O - α -GlcNAc glycosylation sites in simple eukaryotes. Threonine is strongly preferred over serine as the acceptor site. Note also the peculiarly high occurrence of prolines/valines alternating with threonine. The even positioning of glycosylation sites may suggest a β sheet-like structure. A full-color version of this figure is included in the Plate section of this book. By permission of Oxford University Press [43].

Swiss-Prot sequences, regardless of glycosylation annotation. The reported performance is 78% on glycosylated sites (S_n) and 92% on non-glycosylated sites (S_p^{med}) when training on 261 positive sites and 522 negative sites. This version of the predictor is available at <http://www.biosino.org/Oglyc>. The weakness of this predictor is the data set, with a positive data set that may contain negative sites and a negative data set that may contain positive sites.

9.4.3 O - α -GlcNAc Glycosylation in Simple Eukaryotes

O -GlcNAc residues are found in two conformations attached to the polypeptide backbone: an alpha anomeric configuration and a beta anomeric configuration. This leads to the terminology O - α -GlcNAc residues (found on membrane and secreted proteins of lower eukaryotes, discussed in this section) and O - β -GlcNAc (found on cytoplasmic and nuclear proteins, discussed in the next section). Reducing terminal O - α -GlcNAc residues are found on secreted and membrane proteins of *Dictyostelium discoideum* and protozoan parasites such as *Plasmodium* [39], *Schistosoma* [40], *Trypanosoma* [41], and *Entamoeba* [42].

9.4.3.1 Sequence Logo. A sequence logo based on 39 experimentally verified O - α -GlcNAc sites from eight different *Dictyostelium discoideum* proteins is shown in Figure 9.6 [43]. It can be seen that threonine is highly preferred as an acceptor compared with serine. Although no clear consensus is observable, a high occurrence of proline and valine residues alternating with threonine residues is observed. This is largely due to a high degree of clustering of glycosylated positions in the investigated sequences and the fact that all the glycosylation sites are situated at even positions with respect to each other.

9.4.3.2 Existing Predictors. A prediction method, DictyOGlyc [43], has been developed using a neural networks-based method. Probably due to the fairly clear alternating pattern of the sequence logo, the performance is impressive with a cross-validated Matthews correlation coefficient of 0.93. DictyOGlyc is publicly available at <http://www.cbs.dtu.dk/services/DictyOGlyc/>.

9.4.4 *O*- β -GlcNAc Glycosylation on Cytoplasmic/Nuclear Proteins

In 1984, it was found that *O*-glycosylation is not only restricted to proteins which enter the endoplasmic reticulum (ER) co-translationally, but it is a modification that also occurs on nuclear and cytoplasmic proteins [44]. Nuclear and cytoplasmic glycoproteins are modified on multiple sites with a single *N*-acetylglucosamine residue (*O*- β -GlcNAc) [45]. In contrast to *O*-glycosylation on membrane/secreted proteins, the attachment to cytoplasmic/nuclear proteins is via a beta anomeric linkage to produce *O*- β -GlcNAc sites. The process is sometimes referred to as “*O*-GlcNAcylation”. A high concentration of *O*-GlcNAcylated proteins is found in the nuclear pore complex [46].

O-GlcNAcylation has been described as a dynamic process with turnover rates much higher than the protein backbones to which the carbohydrate is attached [47]. The enzymes for the addition of *O*-linked GlcNAc to the protein (UDP-*N*-acetylglucosamine: peptide *N*-acetylglucosaminyltransferase) and for the specific removal of GlcNAc from the polypeptide chain (*N*-acetyl- β -D-glucosaminidase) have been cloned [48]. As many *O*-linked GlcNAc-containing proteins examined to date are also phosphoproteins, and in some instances Ser(Thr)-*O*-GlcNAc and Ser(Thr)—phosphate appear to occupy reciprocally the same hydroxyl groups [49], it is proposed that *O*-GlcNAcylation is a regulatory protein modification [50–52].

9.4.4.1 Proposed Sequence Motifs. Although the protein sequence of the *O*-GlcNAc transferase gene has been compared with other glycosyltransferases [53], details of peptide substrate recognition are unknown. Only a small subset of serine and threonine residues are actually glycosylated and these usually occur in the vicinity of proline, valine, or other serine and threonine residues. Variations exist, however, and no clear consensus pattern emerges [13, 47, 52].

9.4.4.2 Sequence Logo. Figure 9.7 illustrates the sequence context around the accepting serine/threonine of *O*- β -GlcNAc glycosylation based on 40 experimentally verified sites from 17 proteins. The amino acid residues proline, valine, serine, and threonine are all over-represented in the vicinity of the glycosylation site, in agreement with what has been noted previously. More specifically, proline is preferred at positions -3 and -2 , valine at position -1 and serine at positions -10 , -8 , $+1$, $+2$, $+4$, $+6$, $+7$, and $+10$. Detailed analysis shows that glutamate residues generally are disfavored around the glycosylation site. A non-specific 2–3-residue pocket was observed six residues N-terminal to the acceptor site, which could correspond to a gap in the binding induced, for instance, by a cleft in the glycosyltransferase, or loop in the acceptor site.

9.4.4.3 Existing Predictors. YinOYang [54] is a prediction server based on neural networks that examines the local sequence context and surface accessibility around the acceptor Ser or Thr. The method was trained on a set of 17 protein sequences, with 40 glycosylated sites, and 1251 uncharacterized (presumed negative in the first instance) Ser/Thr positions.



Figure 9.7 Shannon sequence logo for *O*- β -GlcNAc glycosylation sites in cytoplasmic/nuclear proteins. No clear consensus emerges around the acceptor serine/threonine except the usual high occurrence of proline, valine, and other serine/threonine residues. A full-color version of this figure is included in the Plate section of this book.

A jury of nine neural networks was constructed after evaluating performance on many neural network architectures. The maximum sequence window size used was 21. In a cross validation, the method identified 72.5% of the glycosylated and 79.7% non-glycosylated sites, revealing a Matthews correlation coefficient of 0.22 on the original data, and 0.84 on an augmented data set (created by pruning possible falsely annotated negative sites).

The prediction server generates warnings for sequences containing possible signal peptides [37], since such sequences are unlikely to be modified in the cytoplasm or nucleus as required for the *O*- β -GlcNAc modification. Furthermore, the prediction server runs Net-Phos, a phosphorylation site predictor [55]. Sites that are predicted to be *O*- β -GlcNAcylated in addition to phosphorylated are identified as “Yin-yang” sites. The method is publicly available for prediction at <http://www.cbs.dtu.dk/services/YinOYang/>.

9.4.5 Proteoglycans

Proteoglycans are glycosylated proteins with highly anionic glycosaminoglycans (GAGs) attached [56]. The glycosaminoglycans are very large compared with other glycans and may consist of more than 100 sugar residues. There can be wide variations in the exact chemical composition of a GAG, but the two main types are glucosaminoglycans (heparan sulfate) and galactosaminoglycans (chondroitin/dermatan sulfate). Regardless of the nature of the sugar polymer, the assembly always starts with the attachment of xylose with an *O*- β -linkage to the hydroxyl group of a specific serine residue within the core protein by a polypeptide xylosyltransferase [57].

Proteoglycans are very important components of the cell surface, the extracellular matrix, and the connective tissue, where proteoglycan sugar chains function in a broad variety of cellular and physiological activities, such as cell differentiation, signaling, adhesion and division, blood coagulation, and wound repair [56, 58]. Proteoglycans exist in vertebrates, *Caenorhabditis elegans* and *Drosophila*, but not in prokaryotes or yeast. It appears that proteoglycans arose with the emergence of multicellularity in the metazoans [59]. Knockout experiments of enzymes that mediate GAG biosynthesis have severe consequences in mice, *Drosophila* and *C. elegans* [60–63], of which the most severe lead to early embryonic



Figure 9.8 Shannon sequence logo for 95 proteoglycan sites in proteins. The strongest preference is for glycine in position +1, and except for glycine only alanine is accepted here. An alternating pattern is discernible where acidic residues are favoured in even and glycine residues in uneven positions around the modification site. A full-color version of this figure is included in the Plate section of this book.

lethality. There are also a number of human disorders that are caused by aberrant GAG biosynthesis or a mal-functioning form of a specific proteoglycan [57, 64].

9.4.5.1 Proposed Sequence Motifs. On the basis of a limited data set, the motif Ser–Gly–Xaa–Gly (where Xaa is any amino acid residue) was originally proposed as a consensus sequence for xylosylation [65]. Further studies showed that either glycine residue could be replaced by alanine without any effect of GAG chain addition [66]. More recent studies have shown the importance of at least two acidic residues on one or both sides of the serine residue [67, 68]. A study based on some experimental sites together with a large number of predicted sites found that the immediate region around the attachment site is rich in aspartate, glutamate, glycine, alanine, serine, threonine, phenylalanine, valine, and leucine [67]. In this and other studies, a simple Ser–Gly consensus motif seems to be used for prediction of possible locations of attachment sites. No threonine sites have been found.

9.4.5.2 Sequence Logo. The sequence preference of the polypeptide xylosyltransferase has been found to be highly conserved throughout all organisms for which there is proteoglycan site information. Figure 9.8 illustrates the sequence context around the accepting serine of a proteoglycan, based on 95 experimentally verified sites from mammals, chicken, worm and fly. The strongest preference is for glycine in position +1, and except for glycine only alanine is accepted here. Glycine or alanine is also preferred in position –1, although other residues are accepted. Glycine is also the most common residue in positions –3, +3 and +5, where +3 is in agreement with the first proposed consensus pattern, Ser–Gly–Xaa–Gly [65]. There is an overall preference for acidic residues, which is particularly high in position –6, –4, –2, and +2. There might be an alternating pattern between acidic residues and glycines, where the former are preferred in even positions and the latter in uneven [69]. The occurrence of basic residues is overall very low, so there is a clear preference for a negatively charged sequence. Compared to previous findings, we can confirm high overall occurrences of aspartate, glutamate, glycine and serine but the claimed over-representation of alanine, threonine, phenylalanine, valine and leucine residues is much less clear [67].

9.4.5.3 Existing Predictors. The first proteoglycan site predictor, NetPGlyc 1.1 [69], consists of two neural network-based prediction methods: one trained on mammalian sequences only and one trained also on sequences from chicken and *C. elegans*. Both are large

improvements over the SG consensus motif, which identifies most existing sites, but severely overpredicts, giving rise to a large number of false positive predictions. As a comparison, predictions were performed on all extracellular *C. elegans* transcripts. Using the SG pattern only for prediction, 42% of the secreted proteins would be predicted to be proteoglycans, while 11% or 20% would be using the all-organism or mammalian predictor respectively and 8% if only sites predicted by both NetPGlyc predictors were trusted [69]. Both predictors are available through the same user interface at <http://www.cbs.dtu.dk/services/NetPGlyc-1.1/>.

9.4.6 GPI Anchors

Glycosylphosphatidylinositol (GPI)-anchored proteins are a group of proteins that are stably anchored at the surface of eukaryotic cells, through linkage of their C-terminal amino acid to phosphatidylinositol lipid anchors. A typical core structure of GPI is composed of ethanolamine phosphate, trimannoside, glucosamine, and inositol phospholipid, in this order [70], with various substituents introduced onto the mannose or inositol residues. GPI-anchored proteins are found in eukaryotes, vertebrates, plants, mollusks, insects, schistosomes, fungi, and protozoa, and possibly also exist among a subset of archaean species, but are probably absent among all other Archaea and all Eubacteria [70].

The biosynthesis of GPI-anchored proteins is carried out in the ER [71]. A transamidase is responsible for the cleavage of the C-terminal signal sequence of the protein and the concomitant addition of a GPI anchor [72]. After a polypeptide chain enters into the ER and the N-terminal signal peptide is cleaved off, the nascent form of GPI-anchored protein is bound to the lumen side of the ER-membrane directed by its C-terminal signal sequence, and then translocated to the site of the transamidase complex and GPI precursor [70]. Then, the C-terminal peptide beyond the cleavage site (the so called Ω -site) is removed from the protein, and the GPI precursor is linked to the resulting new C-terminus (Ω -site). Finally, the protein bearing GPI anchor is transported with secretory vesicles, and inserted into the plasma membrane as a mature GPI-anchored protein [73].

Several studies have now established that GPI-anchored proteins are involved in a number of functions, ranging from enzymatic catalysis to adhesion, and in some cases they have been shown to mediate signal transduction across the plasma membrane [74]. The GPI-anchored proteins have great medical importance. Inhibitors of GPI biosynthesis would be promising drug candidates against fungal infections [75], or could be used in chemotherapy of the diseases caused by pathogenic protozoa [70].

9.4.6.1 Proposed Sequence Motifs. The precursors of proteins to be GPI anchored have several structural characteristics in common: an N-terminal signal peptide, absence of transmembrane domains, and a GPI modification signal sequence at the C-terminus [76]. The GPI modification motif most recently proposed by Eisenhaber *et al.* consists of four sequence regions [77] (Figure 9.9): (i) an unstructured linker region of 10 residues from $\Omega - 11$ to $\Omega - 1$, which is flexible and polar; (ii) a region of small residues (almost exclusively Ser, Asp, Asn, Ala, Gly, and Cys) from $\Omega - 1$ to $\Omega + 2$ including the Ω site for propeptide cleavage and GPI attachment; (iii) a spacer sequence ($\Omega + 3$ to $\Omega + 8$) of moderately polar regions with intervening hydrophobic residues $\Omega + 4$ and $\Omega + 5$; and (iv) a hydrophobic tail from $\Omega + 9$ to the C-terminal end.

9.4.6.2 Sequence Logo. The sequence patterns for 131 GPI-anchor sites are shown as a sequence logo in Figure 9.10. The sequence logo depicts an alignment of GPI modification

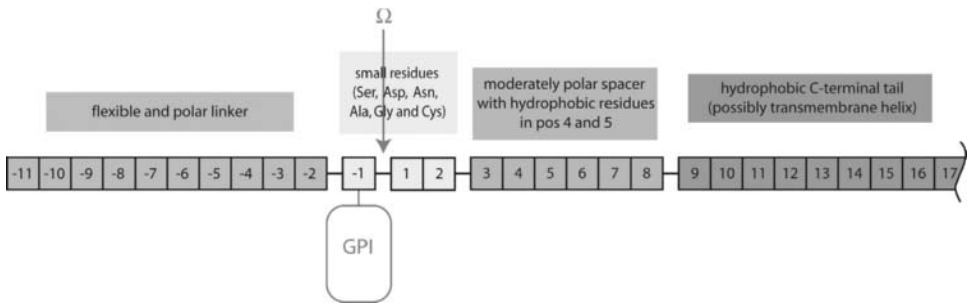


Figure 9.9 Schematic diagram of the proposed sequence motif for GPI anchor attachment [79]: (i) an unstructured linker region, which is flexible and polar; (ii) a region of small residues (almost exclusively Ser, Asp, Asn, Ala, Gly, and Cys) including the Ω site for propeptide cleavage and GPI attachment; (iii) a spacer sequence of moderately polar regions with intervening hydrophobic residues at positions +4 and +5; (iv) a hydrophobic tail from position +9 to the C-terminal end that possibly forms a transmembrane helix.

signal sequences aligned by the cleavage site. Although it does not totally agree with the sequence motif proposed by Eisenhaber *et al.*, the four sequence regions are clearly visible. The cleavage site pattern shows that the residues at position $\Omega - 1$ to $\Omega + 2$ must be small (almost always Ser, Asp, Asn, Ala, Gly, Cys, and Thr). The hydrophobic tail is a stretch of 8–20 amino acids, starting most likely from position $\Omega + 9$ or $\Omega + 10$, and dominated by Leu with some occurrence of Val, Ala, Phe, and Ile. The upstream region of the cleavage site and the spacer region do not show any significant amino acid preference at any position, but the upstream region is more negatively charged, whereas the spacer region carries more positive charge.

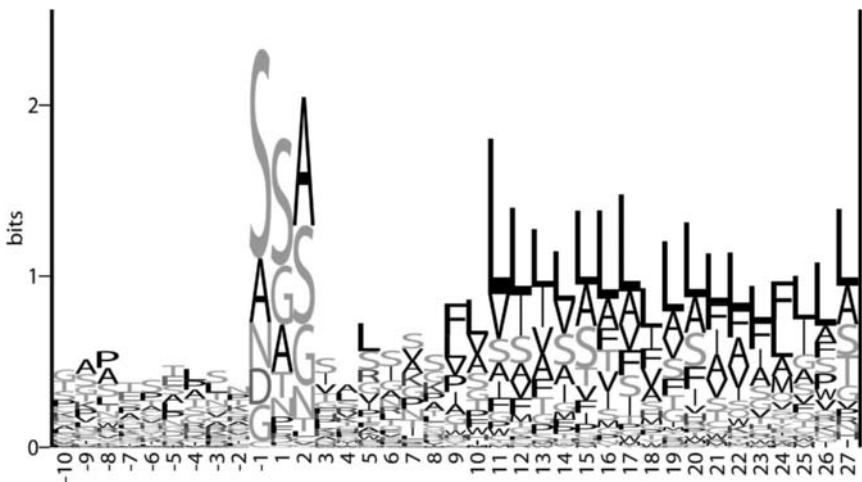


Figure 9.10 Shannon sequence logo for GPI-anchor sites. Position -1 denotes the location of the cleavage site (Ω site). The four sequence regions are clearly visible. The cleavage area is dominated by small amino acids (Ser, Asp, Asn, Ala, Gly, Cys, and Thr), followed by a spacer region and a hydrophobic tail. A full-color version of this figure is included in the Plate section of this book.

9.4.6.3 Existing Predictors. The GPI prediction problem can be seen as twofold: (i) predicting proteins that are GPI anchored and (ii) predicting the cleavage site in these proteins. Programs that only predict cleavage sites will do so even within proteins that do not have GPI anchors, and hence cannot be used reliably on unknown proteins. Furthermore, since GPI anchoring signals are only meaningful inside the endoplasmic reticulum, it is prudent to ensure that the protein in question has an N-terminal secretory signal. Such protein export signals can be predicted, for instance, by SignalP [37].

Over the last decade, numerous methods for prediction of GPI-anchored proteins have been developed. Some prediction methods use sequence motifs or consensus sequences [78–80], whereas others make use of sequence features, for example, amino acid composition [81], sequence hydrophobicity, and amino acid preference at the cleavage site [82, 83]. The most widely used prediction method to date is big-PI [84], which integrates amino acid preference with sequence physical properties extracted from a learning set. Although these methods have been used to various extents, there are some potential areas for improvement. Some of the methods were partly based on predicted data, which might result in a progression of “cyclic errors”. The sequence motif and features have not been thoroughly and systematically analyzed, therefore the sequence motifs used for prediction might be too simple to describe comprehensively the characteristics and the correlations between positions which are necessary for GPI modification. A relatively recent and promising predictor is GPI-SOM [85], which is based on self-organizing maps. It reports high accuracy and better ability to generalize on unknown proteins compared with statistical methods based on only a positive set. With its high sensitivity, GPI-SOM is a good complement to earlier predictors.

9.4.7 C-Mannosylation

C-Mannosylation is the attachment of an α -mannopyranosyl residue to the indole C2 of tryptophan via a C–C link [86]. The first example of glycosylation of a tryptophan residue (with a hexose of unknown type) was discovered in a neuropeptide from a stick insect [86]. Since then, numerous C-mannosylation sites have been found in mammalian proteins [86]. In all mammalian cases, the glycan has been found to be a single α -mannopyranose. The transfer of the mannose to the protein is catalyzed by an enzyme, C-mannosyltransferase, and probably occurs in the ER or the Golgi [86]. C-Mannosyltransferase activity towards peptides derived from human RNase has been found in *C. elegans*, amphibians, birds, and mammals, but not in *E. coli*, insects, and yeast [86]. Little is still known about the function of C-mannosylation, but two recent studies indicate that it is likely to be required for proper folding of Cys subdomains in mucins [87], and that it may have a pathological role in the development of diabetic complications under hyperglycemic conditions [88].

9.4.7.1 Proposed Sequence Motifs. A study involving site-directed mutagenesis of RNase 2 showed the sequence Trp–Xaa–Xaa–Trp in which the first Trp becomes mannosylated, to be the specificity determinant of C-mannosylation [89]. In thrombospondin repeats containing the sequence WXXWXXWXXC (in some cases with one or two tryptophan residues substituted by other amino acids), C-mannosylation was found on one, two, or all three tryptophans [90]. The shortest peptide still valid as a substrate for C-mannosyltransferase found so far is WAKW [91].

9.4.7.2 Sequence Logo. The sequence patterns for 49 C-mannosylation sites from 11 proteins are shown as a sequence logo in Figure 9.11. The strongest discrimination is

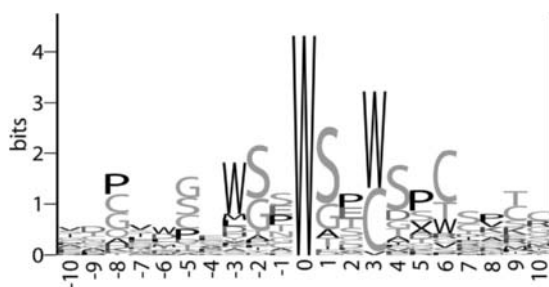


Figure 9.11 Shannon sequence logo for C-mannosylation sites. Position zero denotes the location of the glycosylated tryptophan residue. The strongest discrimination is for tryptophan and cysteine in position +3, but also seen is a strong preference for small and/or polar residues like serine, glycine, alanine, and threonine in position +1 as well as a repetition pattern every three residues. A full-color version of this figure is included in the Plate section of this book.

clearly in position +3, where mostly tryptophan and cysteine are accepted, but a strong preference for small and/or polar residues such as serine, glycine, alanine, and threonine in position +1, not previously reported, can also be observed. A repetitive pattern where a tryptophan adjacent to a serine/glycine is repeated every three residues on either side of the glycosylation site is also evident. This is probably arising from C-mannosylation sites located in thrombospondin repeats.

9.4.7.3 Existing Predictors. The first C-mannosylation predictor, NetCGlyc 1.0 [92], is based on 69 known mammalian sites using a neural network-based method. NetCGlyc 1.0 correctly predicts 93% of both positive and negative C-mannosylation sites (S_n and S_p^{med}). This is a significant improvement over the WXXW consensus motif itself, which identifies only 67% of positive sites. NetCGlyc 1.0 is available at <http://www.cbs.dtu.dk/services/NetCGlyc/>.

9.5 Glycation

Glycation is a non-enzymatic process in which proteins react with reducing sugar molecules to form Advanced Glycation End-products (AGEs). The process proceeds through several steps. Glycation is the result of a breakdown process, rather than a true post-translational modification. The initial reaction occurs between the α amino group of the N-terminal amino acid or the side-chains of lysines and the aldehyde or keto group of a reducing sugar [93, 94]. Furthermore, AGE formation has been found on the side-chains of arginines and histidines [95]. Extracellularly, the most important glycating agent is glucose. Intracellularly, the process is more complex, however, since glucose metabolites such as fructose, especially in diabetic subjects, also react to form AGEs [93, 96]. AGEs accumulate and cause deleterious effects. This accumulation is likely to be more important for proteins with a long biological half-life such as collagen and is, for example, associated with the pathogenesis of aging and complications of diabetes [97]. Other human diseases in which AGEs are implicated are atherosclerosis, amyloidosis, Alzheimer's disease, Pick's disease, Parkinson's disease, Lewy body disease, and actinic elastosis [98].

9.5.1 Proposed Sequence Motif

Most knowledge has been obtained for glycation of lysine side-chains and we will therefore concentrate on this topic in the following discussion. In general, amino groups with lower pK_a values should be expected to be more reactive towards glycation because of their greater nucleophilicity [99]. However, other factors appear to be more important [100]. For lysines, it has been suggested that the properties of nearby amino acids play a role in determining whether a given lysine is glycated or not. Positively charged amino acids placed close to a lysine either in primary or tertiary structure have been proposed to catalyze glycation of that lysine [100, 101]. The catalytic power of histidines has, furthermore, been suggested to be mediated via the hydroxyl group of threonines [102].

9.5.2 Sequence Logo

A sequence logo based on 100 glycation sites from 22 proteins displays which amino acids are characteristic in the vicinity of the glycated lysines in the sequence (Figure 9.12). The first thing that is obvious is that it does not contain strong position-specific signals for any particular amino acid at any position over the 50-residue window. A reason for the lack of a position-specific signal in the logo could be that 3D space interactions between lysines and other amino acids situated far away in the primary structure also play a role in determining whether the given lysine is glycated or not. The amino acids that catalyze glycation of lysine may not need to be placed at specific positions, thus smearing the information content over the window.

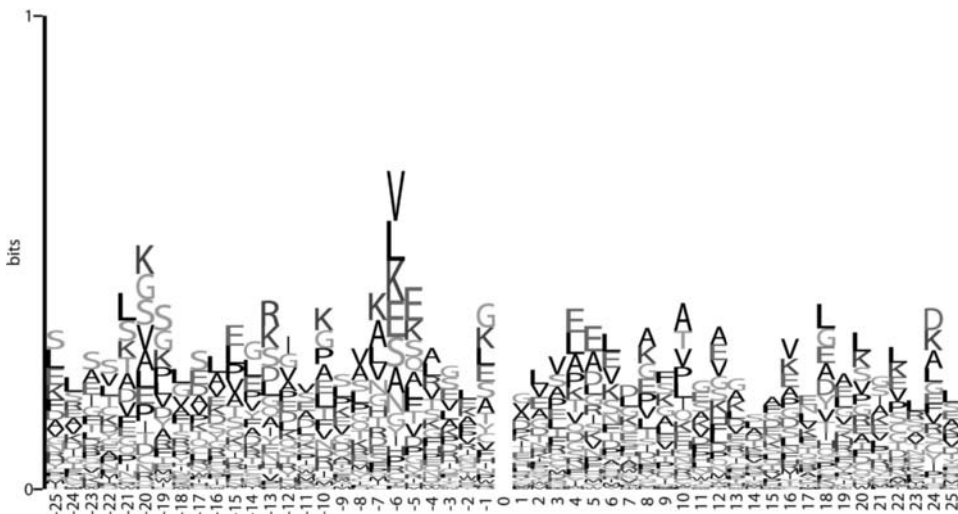


Figure 9.12 Shannon sequence logo for lysine glycation sites. Position zero denotes the location of the glycated lysine residue. The central K symbol (the glycation site) is removed here for scaling purposes (it would be 4.3 bits high). The logo does not contain strong position-specific signals for any particular amino acid at any position, but there is some over-representation of lysines and glutamates next to the glycation site. A full-color version of this figure is included in the Plate section of this book.

The logo shows that lysine and glutamate residues are over-represented in the vicinity of the glycation site, suggesting that they may have a role in catalyzing the glycation reaction. The signal is weaker for arginines and aspartates, and for histidine and threonine there is absolutely no over-representation. This suggests that the role of histidine and threonine is more likely through three-dimensional spatial interactions. Although it is difficult to judge if it is significant, the logo suggests a correlation between glycation of lysines, and valine or leucine at position -6 .

9.5.3 Existing Predictors

The analysis of the logo shows that prediction of glycation sites is a complex problem, since there is so little position-specific information in the logo. The difficult nature of this post-translational modification compared with others may be attributed to its non-enzymatic character: enzyme involvement presumably offers a more specific reaction and often has a biased sequence motif. However, a neural network-based predictor that predicts lysine glycation sites with a cross-validated Matthews correlation coefficient of 0.583 is available at www.cbs.dtu.dk/services/NetGlycate [103].

9.6 Conclusion

This chapter has taken the reader through many glycosylation linkages, and strategies to predict amino acid positions that are likely to be glycosylated. Most glycosylation linkages rely on a suitable local spatial orientation in the protein, and this suitability is predictable from sequence patterns in the neighboring environment. No clear consensus pattern (that is both necessary and sufficient) is evident for most glycosylation linkages. However, with the aid of machine learning, correlations in sequence positions flanking potential acceptor sites can be utilized to predict glycosylation sites with varying degrees of sensitivity and specificity.

A somewhat different utility of glycosylation site predictions *en masse* is due to the contribution of glycosylation to protein structure and function. The availability of many complete genome sequences has led to the pursuit of whole genome functional mapping. A challenge has been orphan proteins (comprising up to 40% of the proteome) that do not exhibit sequence similarity to other functionally known proteins, and hence where functional annotation cannot be transferred from similar, already experimentally characterized proteins. Post-translational modifications such as glycosylation and phosphorylation add information about protein function by their presence, their location on the protein, and their frequencies of occurrence. Using such predicted information to predict further the functional category of a protein has been attempted with some success [4, 54, 104]. Glycosylation is an important feature in the determination of protein function, and has been implicated in protein structure and stability, molecular recognition, and signaling. Not surprisingly, glycosylation site occurrence and frequency were shown to be useful in identifying transport and binding proteins and also replication and transcription proteins, particularly due to the *O*- β -GlcNAc regulatory modification.

It is estimated that as many as 50% of all proteins could be glycosylated [105]; however only a small proportion ($\sim 5\%$) of proteins in public databases have site-mapped glycosylation information available. To bridge this gap, computational prediction methods are

useful tools. Predictions are not a substitute for experimental validation, but having such predictor services made available publicly helps the scientific community in prioritizing site-mapping experiments. In return, newly available experimental data can be used to retrain and improve the accuracy of machine learning methods.

Abbreviations

AGE	Advanced Glycation End-product
ANN	artificial neural network
CC	Matthews correlation coefficient
ER	endoplasmic reticulum
GAG	glycosaminoglycan
GalNAc	<i>N</i> -acetylgalactosamine
GlcNAc	<i>N</i> -acetylglucosamine
GPI	glycosylphosphatidylinositol
HHM	hidden Markov model
PPV	positive predictive value
SRS	the Sequence Retrieval System from the European Bioinformatics Institute

Glossary

Artificial neural network	Joint name for a number of machine learning techniques somewhat inspired by the architecture of the human brain
Bits of information	In information theory, every bit of information represents the amount of information that can be obtained by answering one yes/no question
C-Mannosylation	The attachment of an α -mannopyranosyl residue to the indole C2 of tryptophan via a C–C link
Cross-validation	In machine learning, this is when the data are divided into a number of subsets. One subset is used as test data and the rest as training data. The predictive performance is recorded and the procedure is repeated with another set as the test set until all subsets have been test set exactly once. The performances, recorded for each test set, are compiled and presented as the cross-validated performance of the method
Generalization ability	In machine learning, this is the ability of a trained method to recognize similar but not identical data compared with the training data
Glycation	A non-enzymatic reaction between reducing sugar molecules and proteins that through several steps leads to formation of Advanced Glycation End-products (AGEs)

Glycosaminoglycans (GAGs)	Very large anionic sugar polymers, consisting of as many as 100 residues. GAGs are the glycan part of proteoglycan molecules
Glycosylphosphatidylinositol (GPI)-anchored protein	A type of protein associated with the membrane through linkage of its C-terminal amino acid to a phosphatidylinositol lipid anchor
Glycosyltransferase	An enzyme that catalyzes the transfer of sugar from the activated sugar donor to the acceptor and is thus responsible for the formation of the glycosidic linkage
Hidden Markov model	A machine learning technique commonly used to recognize protein or DNA sequence patterns
Matthews correlation coefficient	In machine learning, this is an overall performance estimator that takes both sensitivity and specificity into account
Mucin-type glycosylation	A GalNAc is α -1 linked to the hydroxyl group of a serine or threonine residue of an exported or membrane-bound mammalian protein
<i>O</i> - α -GlcNAc glycosylation	<i>O</i> -Glycosylation in lower eukaryotes corresponding to mucin-type glycosylation in mammals
<i>O</i> - β -GlcNAc glycosylation	<i>O</i> -Glycosylation on cytoplasmic/nuclear proteins
Over-training	In machine learning, this is when a method has learned the training data perfectly, but lost its generalization ability
ROC curve	In machine learning and medical diagnostics, this is a commonly used illustration of the quality of a classification method. It shows the trade-off between making many positive predictions, of which an increasing proportion are false, and making few predictions with higher quality and thereby missing some
Sensitivity	In machine learning, this is the proportion of positive sites that the method can correctly identify
Shannon sequence logo	A logo that shows the frequencies of amino acid residues at each position, as the relative heights of letters, along with the degree of sequence conservation as the total height of a stack of letters, measured in bits of information
Specificity	In machine learning, this is either (i) the proportion of negative sites correctly identified (S_p^{med}) or (ii) the proportion of positive predictions that are in fact true (S_p^{PPV})
Test data	In machine learning, these are the data used to test the method
Training data	In machine learning, these are the data used to train the method

References

1. Hart GW: Glycosylation. *Curr Opin Cell Biol* 1992, **4**:1017–1023.
2. Spiro RG: Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* 2002, **12**:43R–56R.
3. Varki A: Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* 1993, **3**:97–130.
4. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S: Prediction of human protein function according to gene ontology categories. *Bioinformatics* 2003, **19**:635–642.
5. Hansen JE, Lund O, Tolstrup N, *et al.*: NetOglyc: prediction of mucin type *O*-glycosylation sites based on sequence context and surface accessibility. *Glycoconj J* 1998, **15**:115–130.
6. Presnell SR, Cohen FE: Artificial neural networks for pattern recognition in biochemical sequences. *Annu Rev Biophys Biomol Struct* 1993, **22**:283–298.
7. Baldi P, Brunak S: *Bioinformatics: the Machine Learning Approach*, 2nd edn. Cambridge, MA: MIT Press; 2002.
8. Matthews BW: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975, **405**:442–451.
9. Julenius K, Molgaard A, Gupta R, Brunak S: Prediction, conservation analysis, and structural characterization of mammalian mucin-type *O*-glycosylation sites. *Glycobiology* 2005, **15**:153–164.
10. Hansen JE, Lund O, Engelbrecht J, *et al.*: Prediction of *O*-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase. *Biochem J* 1995, **308**:801–813.
11. Schneider TD, Stephens RM: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990, **18**:6097–6100.
12. Taylor ME, Drickamer K: *Introduction to Glycobiology*. New York: Oxford University Press; 2003.
13. Snow DM, Hart GW: Nuclear and cytoplasmic glycosylation. *Int Rev Cytol* 1998, **181**:43–74.
14. Imperiali B, Hendrickson TL: Asparagine-linked glycosylation: specificity and function of oligosaccharyl transferase. *Bioorg Med Chem* 1995, **3**:1565–1578.
15. Knauer R, Lehle L: The *N*-oligosaccharyltransferase complex from yeast. *FEBS Lett* 1994, **344**:83–86.
16. Marquardt T, Denecke J: Congenital disorders of glycosylation: review of their molecular bases, clinical presentations and specific therapies. *Eur J Pediatr* 2003, **162**:359–379.
17. Cai G, Salonikidis PS, Fei J, *et al.*: The role of *N*-glycosylation in the stability, trafficking and GABA-uptake of GABA-transporter 1. Terminal *N*-glycans facilitate efficient GABA-uptake activity of the GABA transporter. *FEBS J* 2005, **272**:1625–1638.
18. Clark SE, Muslin EH, Henson CA: Effect of adding and removing *N*-glycosylation recognition sites on the thermostability of barley alpha-glucosidase. *Protein Eng Des Sel* 2004, **17**:245–249.
19. Ha SJ, Chang J, Song MK, *et al.*: Engineering *N*-glycosylation mutations in IL-12 enhances sustained cytotoxic T lymphocyte responses for DNA immunization. *Nat Biotechnol* 2002, **20**:381–386.
20. Heatherington AC, Schuller J, Mercer AJ: Pharmacokinetics of novel erythropoiesis stimulating protein (NESP) in cancer patients: preliminary report. *Br J Cancer* 2001, **84**:11–16.
21. Bause E: Structural requirements of *N*-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem J* 1983, **209**:331–336.
22. Roitsch T, Lehle L: Structural requirements for protein *N*-glycosylation. Influence of acceptor peptides on cotranslational glycosylation of yeast invertase and site-directed mutagenesis around a sequon sequence. *Eur J Biochem* 1989, **181**:525–529.

23. Miletich JP, Broze GJ Jr: Beta protein C is not glycosylated at asparagine 329. The rate of translation may influence the frequency of usage at asparagine-X-cysteine sites. *J Biol Chem* 1990, **265**:11397–11404.
24. Gavel Y, von Heijne G: Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng* 1990, **3**:433–442.
25. Kasturi L, Eshleman JR, Wunner WH, Shakin-Eshleman SH: The hydroxy amino acid in an Asn-X-Ser/Thr sequon can influence *N*-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein. *J Biol Chem* 1995, **270**:14756–14761.
26. Bause E: Model studies on *N*-glycosylation of proteins. *Biochem Soc Trans* 1984, **12**:514–517.
27. Boeckmann B, Bairoch A, Apweiler R, *et al.*: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003, **31**:365–370.
28. Sigrist CJ, Cerutti L, Hulo N, *et al.*: PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 2002, **3**:265–274.
29. Senger RS, Karim MN: Variable site-occupancy classification of *N*-linked glycosylation using artificial neural networks. *Biotechnol Prog* 2005, **21**:1653–1662.
30. Strous GJ, Dekker J: Mucin-type glycoproteins. *Crit Rev Biochem Mol Biol* 1992, **27**:57–92.
31. Hang HC, Bertozzi CR: The chemistry and biology of mucin-type *O*-linked glycosylation. *Bioorg Med Chem* 2005, **13**:5021–5034.
32. Jentoft N: Why are proteins *O*-glycosylated? *Trends Biochem Sci* 1990, **15**:291–294.
33. Ten Hagen KG, Fritz TA, Tabak LA: All in the family: the UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferases. *Glycobiology* 2003, **13**:1R–16R.
34. Elhammer AP, Poorman RA, Brown E, *et al.*: The specificity of UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase as inferred from a database of *in vivo* substrates and from the *in vitro* glycosylation of proteins and peptides. *J Biol Chem* 1993, **268**:10029–10038.
35. Wilson IB, Gavel Y, von Heijne G: Amino acid distributions around *O*-linked glycosylation sites. *Biochem J* 1991, **275**:529–534.
36. Chou KC, Zhang CT, Kezdy FJ, Poorman RA: A vector projection method for predicting the specificity of GalNAc-transferase. *Proteins* 1995, **21**:118–126.
37. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 2004, **340**:783–795.
38. Li S, Liu B, Zeng R, *et al.*: Predicting *O*-glycosylation sites in mammalian proteins by using SVMs. *Comput Biol Chem* 2006, **30**:203–208.
39. Nasir-ud-Din, Drager-Dayal R, Decrind C, *et al.*: *Plasmodium falciparum* synthesizes *O*-glycosylated glycoproteins containing *O*-linked *N*-acetylglucosamine. *Biochem Int* 1992, **27**:55–64.
40. Nyame K, Cummings RD, Damian RT: *Schistosoma mansoni* synthesizes glycoproteins containing terminal *O*-linked *N*-acetylglucosamine residues. *J Biol Chem* 1987, **262**:7990–7995.
41. Previato JO, Jones C, Goncalves LP, *et al.*: *O*-Glycosidically linked *N*-acetylglucosamine-bound oligosaccharides from glycoproteins of *Trypanosoma cruzi*. *Biochem J* 1994, **301**:151–159.
42. Stanley SL Jr, Tian K, Koester JP, Li E: The serine-rich *Entamoeba histolytica* protein is a phosphorylated membrane protein containing *O*-linked terminal *N*-acetylglucosamine residues. *J Biol Chem* 1995, **270**:4121–4126.
43. Gupta R, Jung E, Gooley AA, *et al.*: Scanning the available *Dictyostelium discoideum* proteome for *O*-linked GlcNAc glycosylation sites using neural networks. *Glycobiology* 1999, **9**:1009–1022.
44. Torres CR, Hart GW: Topography and polypeptide distribution of terminal *N*-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for *O*-linked GlcNAc. *J Biol Chem* 1984, **259**:3308–3317.
45. Hart GW, Haltiwanger RS, Holt GD, Kelly WG: Glycosylation in the nucleus and cytoplasm. *Annu Rev Biochem* 1989, **58**:841–874.

46. Holt GD, Snow CM, Senior A, *et al.*: Nuclear pore complex glycoproteins contain cytoplasmically disposed *O*-linked *N*-acetylglucosamine. *J Cell Biol* 1987, **104**:1157–1164.
47. Haltiwanger RS, Kelly WG, Roquemore EP, *et al.*: Glycosylation of nuclear and cytoplasmic proteins is ubiquitous and dynamic. *Biochem Soc Trans* 1992, **20**:264–269.
48. Kreppel LK, Blomberg MA, Hart GW: Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique *O*-GlcNAc transferase with multiple tetratricopeptide repeats. *J Biol Chem* 1997, **272**:9308–9315.
49. Chou TY, Hart GW, Dang CV: c-Myc is glycosylated at threonine 58, a known phosphorylation site and a mutational hot spot in lymphomas. *J Biol Chem* 1995, **270**:18961–18965.
50. Hart GW: Dynamic *O*-linked glycosylation of nuclear and cytoskeletal proteins. *Annu Rev Biochem* 1997, **66**:315–335.
51. Hart GW, Greis KD, Dong LY, *et al.*: *O*-Linked *N*-acetylglucosamine: the “yin-yang” of Ser/Thr phosphorylation? Nuclear and cytoplasmic glycosylation. *Adv Exp Med Biol* 1995, **376**:115–123.
52. Haltiwanger RS, Busby S, Grove K, *et al.*: *O*-Glycosylation of nuclear and cytoplasmic proteins: regulation analogous to phosphorylation? *Biophys Res Commun* 1997, **231**:237–242.
53. Roos MD, Hanover JA: Structure of *O*-linked GlcNAc transferase: mediator of glycan-dependent signaling. *Biochem Biophys Res Commun* 2000, **271**:275–280.
54. Gupta R, Brunak S: Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput* 2002, **7**:310–322.
55. Blom N, Gammeltoft S, Brunak S: Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 1999, **294**:1351–1362.
56. Kjellen L, Lindahl U: Proteoglycans: structures and interactions. *Annu Rev Biochem* 1991, **60**:443–475.
57. Sugahara K, Kitagawa H: Recent advances in the study of the biosynthesis and functions of sulfated glycosaminoglycans. *Curr Opin Struct Biol* 2000, **10**:518–527.
58. Jackson RL, Busch SJ, Cardin AD: Glycosaminoglycans: molecular properties, protein interactions, and role in physiological processes. *Physiol Rev* 1991, **71**:481–539.
59. Perrimon N, Bernfield M: Cellular functions of proteoglycans – an overview. *Semin Cell Dev Biol* 2001, **12**:65–67.
60. Hwang HY, Olson SK, Esko JD, Horvitz HR: *Caenorhabditis elegans* early embryogenesis and vulval morphogenesis require chondroitin biosynthesis. *Nature* 2003, **423**:439–443.
61. Mizuguchi S, Uyama T, Kitagawa H, *et al.*: Chondroitin proteoglycans are involved in cell division of *Caenorhabditis elegans*. *Nature* 2003, **423**:443–448.
62. Lander AD, Selleck SB: The elusive functions of proteoglycans: in vivo veritas. *J Cell Biol* 2000, **148**:227–232.
63. Forsberg E, Kjellen L: Heparan sulfate: lessons from knockout mice. *J Clin Invest* 2001, **108**:175–180.
64. Iozzo RV: Heparan sulfate proteoglycans: intricate molecules with intriguing functions. *J Clin Invest* 2001, **108**:165–167.
65. Bourdon MA, Oldberg A, Pierschbacher M, Ruoslahti E: Molecular cloning and sequence analysis of a chondroitin sulfate proteoglycan cDNA. *Proc Natl Acad Sci USA* 1985, **82**:1321–1325.
66. Mann DM, Yamaguchi Y, Bourdon MA, Ruoslahti E: Analysis of glycosaminoglycan substitution in decorin by site-directed mutagenesis. *J Biol Chem* 1990, **265**:5317–5323.
67. Esko JD, Zhang L: Influence of core protein sequence on glycosaminoglycan assembly. *Curr Opin Struct Biol* 1996, **6**:663–670.
68. Thinakaran G, Slunt HH, Sisodia SS: Novel regulation of chondroitin sulfate glycosaminoglycan modification of amyloid precursor protein and its homologue, APLP2. *J Biol Chem* 1995, **270**:16522–16525.

69. Hagen FK, Wang H, Sievert M, Hryhorenko J, Julenius K: Proteoglycan site mapping and mutagenesis in worms integrates well with data from other metazoans for neural network training of an all-organism predictor. manuscript in preparation.
70. Ikezawa H: Glycosylphosphatidylinositol (GPI)-anchored proteins. *Biol Pharm Bull* 2002, **25**:409–417.
71. Gerber LD, Kodukula K, Udenfriend S: Phosphatidylinositol glycan (PI-G) anchored membrane proteins. Amino acid requirements adjacent to the site of cleavage and PI-G attachment in the COOH-terminal signal peptide. *J Biol Chem* 1992, **267**:12168–12173.
72. Maxwell SE, Ramalingam S, Gerber LD, *et al.*: An active carbonyl formed during glycosylphosphatidylinositol addition to a protein is evidence of catalysis by a transamidase. *J Biol Chem* 1995, **270**:19576–19582.
73. Nosjean O, Briolay A, Roux B: Mammalian GPI proteins: sorting, membrane residence and functions. *Biochim Biophys Acta* 1997, **1331**:153–186.
74. Sharom FJ, Lehto MT: Glycosylphosphatidylinositol-anchored proteins: structure, function, and cleavage by phosphatidylinositol-specific phospholipase C. *Biochem Cell Biol* 2002, **80**:535–549.
75. Tsukahara K, Hata K, Nakamoto K, *et al.*: Medicinal genetics approach towards identifying the molecular target of a novel inhibitor of fungal cell wall assembly. *Mol Microbiol* 2003, **48**:1029–1042.
76. Udenfriend S, Micanovic R, Kodukula K: Structural requirements of a nascent protein for processing to a PI-G anchored form: studies in intact cells and cell-free systems. *Cell Biol Int Rep* 1991, **15**:739–759.
77. Eisenhaber B, Bork P, Eisenhaber F: Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng* 1998, **11**:1155–1161.
78. Caro LH, Tettelin H, Vossen JH, *et al.*: In silico identification of glycosyl-phosphatidylinositol-anchored plasma-membrane and cell wall proteins of *Saccharomyces cerevisiae*. *Yeast* 1997, **13**:1477–1489.
79. De Groot PW, Hellingwerf KJ, Klis FM: Genome-wide identification of fungal GPI proteins. *Yeast* 2003, **20**:781–796.
80. Hamada K, Fukuchi S, Arisawa M, *et al.*: Screening for glycosylphosphatidylinositol (GPI)-dependent cell wall proteins in *Saccharomyces cerevisiae*. *Mol Gen Genet* 1998, **258**:53–59.
81. Chou KC, Elrod DW: Prediction of membrane protein types and subcellular locations. *Proteins* 1999, **34**:137–153.
82. Borner GH, Sherrier DJ, Stevens TJ, *et al.*: Prediction of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A genomic analysis. *Plant Physiol* 2002, **129**:486–499.
83. Krongeg J, Buloz D: Detection/prediction of GPI cleavage site (GPI-anchor) in a protein (DGPI). 1999. Retrieved from <http://129.194.185.165/dgpi/>.
84. Eisenhaber B, Bork P, Eisenhaber F: Prediction of potential GPI-modification sites in proprotein sequences. *J Mol Biol* 1999, **292**:741–758.
85. Fankhauser N, Maser P: Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* 2005, **21**:1846–1852.
86. Furmanek A, Hofsteenge J: Protein C-mannosylation: facts and questions. *Acta Biochim Pol* 2000, **47**:781–789.
87. Perez-Vilar J, Randell SH, Boucher RC: C-Mannosylation of MUC5AC and MUC5B Cys subdomains. *Glycobiology* 2004, **14**:325–337.
88. Ihara Y, Manabe S, Kanda M, *et al.*: Increased expression of protein C-mannosylation in the aortic vessels of diabetic Zucker rats. *Glycobiology* 2005, **15**:383–392.
89. Krieg J, Hartmann S, Vicentini A, *et al.*: Recognition signal for C-mannosylation of Trp-7 in RNase 2 consists of sequence Trp-x-x-Trp. *Mol Biol Cell* 1998, **9**:301–309.

90. Hofsteenge J, Blommers M, Hess D, *et al.*: The four terminal components of the complement system are C-mannosylated on multiple tryptophan residues. *J Biol Chem* 1999, **274**:32786–32794.
91. Hartmann S, Hofsteenge J: Properdin, the positive regulator of complement, is highly C-mannosylated. *J Biol Chem* 2000, **275**:28569–28574.
92. Julenius K: NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. *Glycobiology* 2007, **17**:868–876.
93. Schalkwijk CG, Stehouwer CD, van Hinsbergh VW: Fructose-mediated non-enzymatic glycation: sweet coupling or bad modification. *Diabetes Metab Res Rev* 2004, **20**:369–382.
94. Suji G, Sivakami S: Glucose, glycation and aging. *Biogerontology* 2004, **5**:365–373.
95. Bidasee KR, Zhang Y, Shao CH, *et al.*: Diabetes increases formation of advanced glycation end products on sarco(endo)plasmic reticulum Ca²⁺-ATPase. *Diabetes* 2004, **53**:463–473.
96. Verbeke P, Clark BF, Rattan SI: Modulating cellular aging *in vitro*: hormetic effects of repeated mild heat stress on protein oxidation and glycation. *Exp Gerontol* 2000, **35**:787–794.
97. Paul RG, Bailey AJ: Glycation of collagen: the basis of its central role in the late complications of ageing and diabetes. *Int J Biochem Cell Biol* 1996, **28**:1297–1310.
98. Ling X, Sakashita N, Takeya M, *et al.*: Immunohistochemical distribution and subcellular localization of three distinct specific molecular structures of advanced glycation end products in human tissues. *Lab Invest* 1998, **78**:1591–1606.
99. Bunn HF, Shapiro R, McManus M, *et al.*: Structural heterogeneity of human hemoglobin A due to nonenzymatic glycosylation. *J Biol Chem* 1979, **254**:3892–3898.
100. Baynes JW, Watkins NG, Fisher CI, *et al.*: The Amadori product on protein: structure and reactions. *Prog Clin Biol Res* 1989, **304**:43–67.
101. Venkatraman J, Aggarwal K, Balaram P: Helical peptide models for protein glycation: proximity effects in catalysis of the Amadori rearrangement. *Chem Biol* 2001, **8**:611–625.
102. Shilton BH, Campbell RL, Walton DJ: Site specificity of glycation of horse liver alcohol dehydrogenase *in vitro*. *Eur J Biochem* 1993, **215**:567–572.
103. Johansen MB, Kiemer L, Brunak S: Analysis and prediction of mammalian protein glycation. *Glycobiology* 2006, **16**:844–853.
104. Jensen LJ, Gupta R, Blom N, *et al.*: Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol* 2002, **319**:1257–1265.
105. Apweiler R, Hermjakob H, Sharon N: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999, **1473**:4–8.

**Section 4:
Experimental
Methods – Bioinformatic
Requirements**

Experimental Methods for the Analysis of Glycans and Their Bioinformatics Requirements

Claus-Wilhelm von der Lieth

*Formerly at Deutsches Krebsforschungszentrum (German Cancer Research Center),
Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany*

10.1 Introduction

A comprehensive understanding of the biological functions of complex carbohydrates requires the detailed knowledge of all their structural features (see Chapter 2). In this chapter, often used analytical techniques for primary structure determination will be discussed with respect to their implications for informatics. The aspects of 3D structures and the conformational space that oligosaccharides can access are discussed in Section 5. The inherent complexity of protein glycosylation, the sensitivity with which it can be analyzed, and also the relatively extensive and time-consuming analytical procedures needed, have long been limiting factors for applications in many biomedical and glycobiological studies.

The primary analysis of glycan structures is hampered by various factors:

1. Oligosaccharides exhibit an intrinsic structural diversity (see Chapter 2), which makes their detection more difficult than DNA and protein sequences. The structural space of the monosaccharide building blocks found in Nature is much larger than that for the other two classes of biological information encoding macromolecules, which are composed of a limited number of building blocks (residues) – four nucleotides each for DNA and RNA and 20 amino acids for proteins. However, in specific areas such as *N*-glycosylation of mammalian proteins, only a restricted number of residues – mannose, *N*-acetylglucosamine, fucose, galactose and sialic acid – are found. Nevertheless, a recent study estimated [1] that more than 7500 different *N*-glycans can be generated on applying 20 rules for enzyme activity that were derived when formalizing the fairly well understood *N*-glycosylation pathway. This large number of possible *N*-glycan structures, which consists of only five different residues and which have additionally the same structural core in common, is impressive evidence that the potential to connect monosaccharides in different ways and to form branches is probably the major reason for the diversity of carbohydrate structures found in Nature.
2. The biosynthesis of glycans is a non-template-driven process so that no biological amplification methods exists. Consequently, carbohydrates have either to be analyzed

at their physiological concentration or time-consuming enrichment procedures have to be applied.

3. Although the same glycosylation machinery is available to all proteins in a given cell, most glycoproteins emerge with characteristic glycosylation patterns and heterogeneous populations of glycans at each glycosylation site (glycoforms). More than 100 glycoforms have been reported for a single glycosylation site of a protein, and more than 10 glycoforms are often reported [2].
4. The building blocks of complex carbohydrates consist of monosaccharides, which often have the same molecular weight and chemical constitution and differ only in the stereochemistry of the attached hydroxyl groups.

Therefore, a detailed analysis of glycans, including all structural aspects is difficult and time consuming and often needs relatively large amounts (in the microgram range) of isolated oligosaccharides. Often methods are applied, which cannot detect all structural details of glycans. However, modern analytical methods have the ability to elucidate most structural details at the concentration levels required for proteomics/glycomics projects.

10.2 Varying Explanatory Power of Analytical Procedures

Because of the above-outlined inherent problems, a universal protocol for oligosaccharide analysis does not exist and often several separations and/or enrichment methods are used in combination with various detection modes to obtain the desired structural detail. In many papers, often only a gross structural characterization such as the composition are reported. However, since compositions are often the only available data, they should not be neglected. The varying explanatory power of analytical procedures to determine glycan structures at different levels of structural completeness constitutes a major challenge for the design of glyco-related databases. The scope and limitations of each experimental technique and its reliability have to be reflected when storing the data. Users should be able to evaluate only those structures which have been assigned using analytical procedures that are indeed adequate to answer the question asked.

Mass spectrometry (MS)-based *N*-glycan profiling papers often provide completely assigned *N*-glycan structures [3, 4], although their composition is the only experimentally determined and reliable value. The justification for assigning complete structures instead of just giving compositions is that the *N*-glycosylation pathways in mammals are well known and that it is justified to assign explicit *N*-glycan structures when merging these two independent sources of information. However, an unambiguous assignment of *N*-glycan structures based on MS approaches requires higher order spectra (MS^n). A database should be able to handle various types of reliability of the data on the basis of the underlying analytical methods.

Another example where current databases pretend to have a higher exactness than the applied experimental technique really provides, is the use of exoglycosidases for glycan structure determination. Exoglycosidases are used to cleave terminal monosaccharides from glycans. This technique is often used in combination with high-performance liquid chromatographic (HPLC) methods to ascertain the presence of specific terminal residues. Although it is often assumed that specific linkages and/or specific terminal residues are

recognized by specific exoglycosidases, many enzymes indeed cleave unspecifically several types of linkages. The sialidase Au Alpha-(2–3,6,8,9), for example, cleaves all non-reducing terminal branched and unbranched sialic acids. Moreover, Schauer [5] and Angata and Varki [6] pointed out that sialic acids comprise a family of over 40 neuraminic acid derivatives. A cleavage with sialidase provides only the information that one member of this family was attached to the glycan core. Nevertheless, many databases such as CarbBank indicate that Neu5Ac – the most frequently occurring member of the sialic acid family in humans – is present. However, the applied experimental technique is not able to prove this fact. For future use in databases, a name describing all types of sialic acids, like Sia, should be used in these cases.

10.3 Analytical Methods

The analytical methods currently applied to determine glycan structures in biological samples (see Table 10.1) can be divided into those which directly detect intrinsic physicochemical properties of carbohydrates – such as the mass of fragments and nuclear spin systems of atoms in a certain chemical environment – and those which have to modify the sugars to be able to measure their presence indirectly through detectable properties of the attached substitution. Here only those methods will be described which require informatics for the rapid and automatic interpretation of the data produced. A general overview of analytical procedures for carbohydrates can be found in general text books on glycobiology (e.g. [7]).

10.3.1 NMR

NMR methods permit a complete and unambiguous assignment of all structural features of glycans – stereochemistry of monosaccharide units, the type of linkage between connected units, and even their conformational preferences – using the same experimental setup and in a non-destructive way. However the main drawback is the large amount of purified material – in the range 10–100 µg of purified glycan are required – to obtain high-resolution spectra. This prerequisite normally excludes NMR from being used in glycomics projects, where only a physiological amount of material is available. Another obstacle is that the equipment is expensive and well-trained scientists are needed to adjust the analytical procedures to extract the maximum quality of data and precise structural information. Nevertheless, NMR methods are intensively used to determine glycan structures whenever a sufficient quantity of pure compound is available. The main areas of NMR-based structural determination of carbohydrate structures are bacterial carbohydrates [8] and polysaccharides [9] including maritime plants and also synthetically produced sugars.

The contributions in the NMR section of this book summarize the concepts of how NMR data are stored and assigned to atoms in glycan structures in databases. Several approaches are described to how the stored data can be used for the automatic assignment of NMR spectra, the estimation of chemical shifts for given structures, and automatic identification of glycan structures using library searches and artificial neural networks. One intriguing property of NMR resonances is that each single chemical shift can be assigned unambiguously to exactly one atom in a given structure. Additionally, the exact value of the chemical shift depends on the atom's chemical environment and is essentially influenced

Table 10.1 Overview of experimental methods used for glycan analysis.

Method	Structural data	Material needed (μg)	Saccharide detection level	Reliability of results	Potential for automation	Advantages	Disadvantages
NMR	Complete structure including stereochemistry	100	nmol	High	Low	Non-destructive technique, quantification applicable in high-throughput projects	Pure probes, expensive equipment
MS Profiling (MALDI-TOF, ESI)	Composition	10	pmol–fmol	High	High	Applicable in high-throughput projects	Quantification difficult
MS ⁿ (MALDI-TOF, ESI)	Sequence – no stereochemistry	10	pmol–fmol	High	High	Applicable in high-throughput projects	Quantification difficult, derivatization
MS ion trap	Sequence – no stereochemistry	10	pmol–fmol	High	Medium	Detailed structural information	Quantification difficult, derivatization
HPLC fluorescent label	Sequence	1	fmol	Medium	Low	Quantification of relative amounts	No direct proof, internal calibration
Exoglycosidases + HPLC or MALDI-TOF	Terminal residues	1	fmol	HPLC: medium MALDI: high	Low	Fast method, no derivatization required	Only reliable in combination with other methods
Lectin binding	Terminal residues	100	nmol	Low	Low	Good for purification + enrichment	Only good for preliminary identification

by the type of bonds formed with the directly adjacent atoms. Several computational approaches which make use of these fundamental properties of NMR will be discussed.

10.3.2 MS

The development of MS methods and equipment capable of detecting oligosaccharides in the low pico- to femtomole range with great accuracy in mass detection has made MS a key technology for glycomics analysis, as it is also for the detection of peptides in proteomics projects. The emerging high-throughput glycomics and glycoproteomics projects aim to characterize all forms of glycoproteins in different tissues and organisms. The advantages of MS techniques are

1. High sensitivity – for example, *N*-glycan profiles from the brain of a single mouse embryo can be detected with high resolution.
2. High potential to integrate MS detection methods into high-throughput glycan screening procedures and direct coupling with separation methods.
3. Potential to develop automatic assignment procedures. These algorithms, which are currently being developed, will enable a rapid automatic assignment of detected glycans as is nowadays accomplished for the identification of proteins.

MS techniques are not able to distinguish between different stereoisomers such as galactose, glucose, and mannose, which have the same mass. Several techniques are available to determine the linkage type in complex glycans. In contrast to the proteomics area, where fairly standardized MS-based procedures for the identification of proteins are now applied worldwide, a broad variety of MS methods in combination with separation/enrichment techniques and chemical derivatization are used to analyze glycans. The contribution of Niclas Karlsson and Nicolle Packer (Chapter 12) summarizes the information that one can obtain with the currently used glycomic/glycoproteomic mass spectrometric methodologies.

10.3.3 HPLC and Enzymatic Digestion

HPLC techniques for separation and detection of mixtures of *N*- and *O*-glycans have been used since the 1980s and are now well established [10]. A wide range of techniques have been developed for the detection of glycans.

Chemical derivatization is now the most common method used for labeling glycans at their reducing ends, for example by reductive amination. A single fluorescently labeled molecule can be incorporated to each mono- or oligosaccharide. The sensitivity of detection by this technique is in the low-femtomole range.

HPLC techniques are often used for the mapping of oligosaccharides released either chemically or enzymatically from glycoproteins. As internal standard, a source of glucose oligomers – a so-called dextran ladder – is used and the elution position of each peak is expressed in glucose units (gu). The elution positions of peaks in an unknown glycan pool are assigned an overall gu value by comparison with the standard dextran ladder.

The assignments of glycans may be confirmed by sequential exoglycosidase digestion using specific enzymes. The products of such digestions are oligosaccharide fragments from which the enzyme(s) have removed specific terminal residues. A re-run of the HPLC should result in peaks with gu values characteristic of the cleaved products.

For an unambiguous assignment of glycan structures, HPLC is most often used in combination with various MS techniques [11]. A combination of HPLC gu values of fluorescently labeled sugars, MS composition and mass fragmentation data and also exoglycosidase digestions provides highly trustworthy assignments of glycans at physiological concentrations [10].

10.4 Ensuring the Quality of the Data Stored in Databases

Another often important aspect for practical use in bioinformatics applications is the varying quality of analytical data produced using the same experimental procedures. The reasons for this variability may be manifold: insufficient biological material, varying purification and enrichment procedures and a slightly changed experimental setup of the detection method used. Therefore, it is important for the database design to develop an assessment schema for each experimental method. Additionally, each schema should be able to evaluate and compare the results obtained by various experimental approaches.

Whenever possible, a plausibility check of newly entered data should be performed and an automatically generated quality score should be assigned. For NMR, where well-established and fast estimation tools are available, for each shift value the environment shift code (see Chapter 15) should be looked up in appropriate databases. Alternatively, an increment-based approach, as implemented, for example, in CASPER, can be used to predict shift values. Obviously wrong assignments should be rejected. Missing and obviously wrong assignments should be reported to the database curator.

For mass spectra (e.g. *N*-glycan profiling), a quality score based on the assigned peaks and their theoretical values can be easily calculated. In the case of MS^n spectra, all possible fragment ions can be calculated if the structure is assigned. Combined with a probabilistic score that certain fragments occur more often than others, a quality score for MS^n spectra can be assigned. The evaluation of correlations between (sub)structural features and marker fragment ions will be an important scientific task, which needs reliably assigned experimental data. The derived probabilistic score will not only be useful as part of a quality score, but will also be very useful for evaluating the plausibility of automatically assigned structures. However, the fragmentation scheme found may depend heavily on the activation method. It has been found, for example, that when analyzing glycopeptides by collision-induced dissociation, the mass spectra obtained are usually dominated by cleavages of the glycosidic linkages.

References

1. Krambeck F, Betenbaugh M: A mathematical model of *N*-linked glycosylation. *Biotechnol Bioeng* 2005, **92**:711–728.
2. Rudd P, Dwek R: Glycosylation: heterogeneity and the 3D structure of proteins. *Crit Rev Biochem Mol Biol* 1997, **32**:1–100.
3. Sutton-Smith M, Morris H, Grewal P, Hewitt J, Bittner R, Goldin E, Schiffmann R, Dell A: MS screening strategies: investigating the glycomes of knockout and myodystrophic mice and leukodystrophic human brains. *Biochem Soc Symp* 2002, **69**:105–115.
4. Goldberg D, Sutton-Smith M, Paulson J, Dell A: Automatic annotation of matrix-assisted laser desorption/ionization *N*-glycan spectra. *Proteomics* 2005, **5**:865–875.

5. Schauer R: Achievements and challenges of sialic acid research. *Glycoconj J* 2000, **17**:485–499.
6. Angata T, Varki A: Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective. *Chem Rev* 2002, **102**:439–469.
7. Brooks S, Dwek M, Schumacher U: *Functional and Molecular Glycobiology*. Oxford: BIOS Scientific; 2002.
8. Feng L, Senchenkova S, Wang W, Shashkov A, Liu B, Shevelev S, Liu D, Knirel Y, Wang L: Structural and genetic characterization of the *Shigella boydii* type 18 O antigen. *Gene* 2005, **355**:79–86.
9. Ahrazem O, Prieto A, Giménez-Abián M, Leal J, Jiménez-Barbero J, Bernabé M: Structural elucidation of fungal polysaccharides isolated from the cell wall of *Plectosphaerella cucumerina* and *Verticillium* spp. *Carbohydr Res* 2005, **341**:246–252.
10. Royle L, Radcliffe C, Dwek R, Rudd P: Detailed structural analysis of *N*-glycans released from glycoproteins in SDS-PAGE gel bands using HPLC combined with exoglycosidase array digestions. *Methods Mol Biol* 2006, **347**:125–143.
11. Wuhrer M, Deelder A, Hokke C: Protein glycosylation analysis by liquid chromatography–mass spectrometry. *J Chromatogr B Anal Technol Biomed Life Sci* 2005, **825**:124–133.

11

Analysis of *N*- and *O*-Glycans of Glycoproteins by HPLC Technology

Anthony H. Merry¹ and Sviatlana A. Astrautsova²

¹*Glycosciences Consultancy, Charlbury OX7 3HB, UK*

²*Department of Microbiology, Medical University of Grodno, Grodno, Belarus*

11.1 Introduction

The majority of proteins are glycosylated [1] and the characterization of the associated glycans therefore forms an important part of their structural analysis. The initial step in the analysis of protein glycosylation is to decide how much detail and information is required. Unlike the sequence analysis of nucleic acids or proteins, both of which have single linear sequences, many of the glycans attached to glycoproteins will have branched structures, there may be several attachment sites, and also many different glycans may be attached at any given site.

The complete analysis of a glycoprotein would therefore require all of these glycans to be identified, the distribution of glycans at each site measured, and the degree of glycosylation (site occupancy) determined. This is an immense task and few glycoproteins have been completely characterized at this level of detail, although there are notable exceptions.

In practice, it is generally possible to use the techniques described here to obtain a profile of the major glycans present in most glycoproteins and to identify those which make up the majority of the glycans present. Such information can now be obtained fairly readily and analysis or large-scale screening of samples is a realistic possibility. Monitoring the glycosylation of proteins and changes in glycosylation found in disease states, or seeing how a deficiency in the activity of given enzymes required for glycan synthesis affects the glycosylation profiles, is realistic. This has been demonstrated in such applications as monitoring changes in glycosylation related to cancer diseases [2–4], where individuals have congenital enzyme deficiencies [5] or where enzymes have been deleted in knockout mice [6, 7]. Characterization of such changes, and looking at them in relation to the disease process, will help in our understanding of the biological roles of glycosylation. A typical scheme for protein glycans analysis is shown in Figure 11.1.

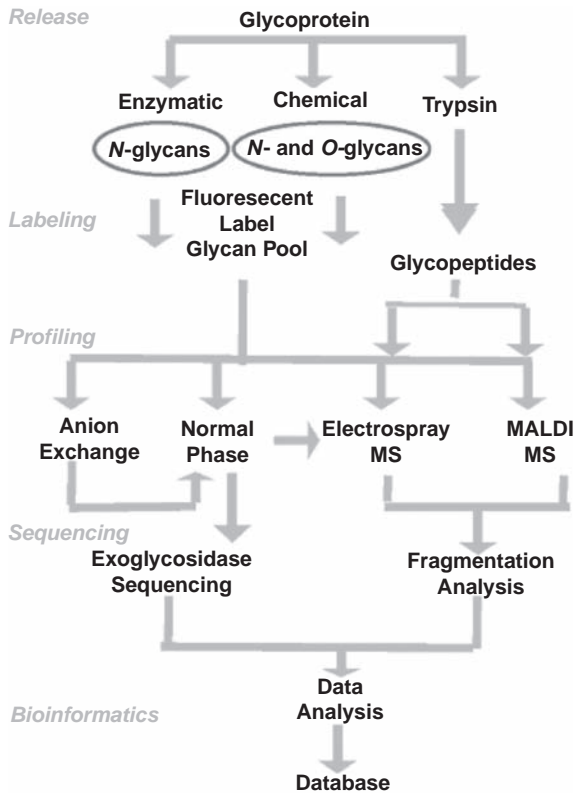


Figure 11.1 Glycan analysis scheme. Diagram of the stages of release, labeling, profiling, and sequencing in glycan analysis using HPLC and mass spectrometry.

11.2 Techniques

11.2.1 Glycan Release

Generally, the analysis of glycans requires their release from the peptide backbone. This may be achieved by either chemical means such as base-catalyzed β -elimination [8, 9] or hydrazinolysis [10–14], or by enzymatic means such as the use of the glycosaminidase peptide-*N*-glycosidase F [15], or endoglycosidases such as endo-H [16] and endo-F [17].

The most widely used technique for the analysis of *N*-glycans involves the release of the glycans with peptide-*N*-glycanase F (PNGaseF). This may present some difficulties connected with selective or incomplete release of glycans if the incubation conditions are not optimized for the glycoprotein in question [14, 18]. Hence problems may arise if working with complex mixtures such as whole serum, where in this case chemical release is preferable. In practice, the method proves very reliable especially on denatured proteins and may be used very conveniently to release *N*-glycans from glycoproteins separated by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) [19, 20]. One limitation is that the enzyme will not act on glycans having a fucose substitution at the 3-position of the GlcNAc connected to the asparagine residue on the peptide [21]. Such

substitution occurs frequently in plant glycans but in this case another enzyme, peptide-*N*-glycanase A, may be used in combination with trypsin, and this is also effective on SDS-PAGE proteins bands [22, 23].

When the glycans are to be released from a mixture of proteins or where *O*-linked glycans are present, then a chemical procedure must be used. Beta elimination under base conditions is effective, but in order to reduce the modification of the glycan it should be performed under strong reducing conditions [8]. This procedure, however, will convert the released glycan to an alditol form which is difficult to derivatize. That may not be a problem for analysis by mass spectrometry, but does preclude the use of high-sensitivity high-performance liquid chromatographic (HPLC) techniques. In order to preserve a non-reducing terminal monosaccharide, chemical release by hydrazinolysis may be employed and optimized conditions for such release have been evaluated [14, 24].

11.2.2 Glycan Labeling

Detection of underivatized glycans is difficult and lacks sensitivity, as glycans have no good absorbance or fluorescent properties, and therefore the sensitive detection of these molecules generally requires the introduction of some label. The only practical alternative is the use of pulsed amperometric detection (PAD) [25, 26a], as discussed later

In mass spectrometry, although the underivatized glycans can be detected, derivatization may aid their ionization [27–29]. Some form of labeling is therefore generally used following glycan release. The problems in introducing suitable labels was first overcome by the use of sodium borotritide to introduce tritium into the non-reducing form of the glycan in a reductive reaction [12, 30, 31]. This process may be highly efficient and stoichiometric incorporation of tritium can be achieved regardless of the nature of the glycan [32]. The problems with such an approach are the detection of a weak beta emitter such as tritium and the requirement for the use of high specific activity sodium borotritide, which is increasingly difficult in the current regulatory environment in most countries.

Fluorescent detection offers a good alternative [33]. The most commonly used derivatization is now the introduction of a group with fluorescent properties at the reducing terminal of the glycan [33–37]. The label 2-aminopyridine has been used very successfully in a large number of studies [38] but can present some difficulties in removal of free label from the labeled glycan. Alternatives are the fluorophores 2-aminobenzoic acid [39] and 2-aminobenzamide [37], which are more readily separated and can be detected at high sensitivity by modern fluorescent detectors.

Such derivatization improves sensitivity and allows quantitation in the low pictogram range under certain circumstances. Commonly used fluorophores are 2-aminopyridine, 2-aminobenzamide (2-AB) [37], 2-aminobenzoic acid, 2-aminoanthranilic acid [40] and ANTS [41]. The choice of label depends on the separation and detection method, with 2-AB and 2-AP [33] being used for HPLC whereas ANTS [41, 42] is used in capillary electrophoresis or SDS-PAGE. The label 2-aminoanthranilic acid may be used in both HPLC [43, 44] and electrophoresis techniques [37].

11.2.2.1 Removal of Free Label from Labeled Glycans. The removal of free label from the glycans is not a trivial process. In order to obtain maximum efficiency and stoichiometry of labeling the concentration of the free labeling agent is often very high. For example, for 2-AB labeling a final concentration of 0.5 M is generally used. When the concentration of

the glycans will rarely exceed 500 pM and is frequently lower, there is obviously a large amount of the free label to remove. In addition, when using a relatively non-hydrophobic label it may be particularly difficult to separate this from small glycans. The classical method has been to use paper chromatography, but with the high concentration of free dye present trace contaminants in the chromatography paper will become readily labeled. Hence the procedure requires an optimized washing and separation procedure in order to minimize background contamination. This is not a great problem with the relatively larger *N*-glycans, but with smaller *O*-glycans large background peaks may be found. [24].

11.2.3 Glycan Profiling

Once labeled, the glycans may be separated on the basis of certain properties to obtain a profile of all the glycans present. Separations are generally based on either size (or hydrophilicity, which may be equivalent to mass under certain conditions) or by charge (in *N*- and *O*-glycans from glycoproteins this is most commonly derived from the presence of sialic acids). Analysis of complex mixtures will generally require the use of at least two types of separation. For example, different types of matrix can be used in HPLC separations [45–47] or both HPLC and mass spectrometry may be performed [48–51].

The types of profiling now commonly used are HPLC, gel or capillary electrophoresis and mass spectrometry. Various matrices have been applied for glycan separations and the multi-dimensional techniques described by Takahashi and colleagues will provide identification of glycans by reference to standards. This does require extreme precision on the HPLC system. however, and the separation on amide columns [52, 53] using an internal standard of hydrolyzed dextran described by Guile *et al.* [54] provides a simpler and more robust approach. The choice of HPLC system and detector is still important, however, as the gradient of acetonitrile–water used must be precise and reproducible. When there are a large number of similar glycans, however, separations are incomplete and initial fractionation on the basis of charge using weak anion-exchange chromatography should be performed [46, 55].

Capillary electrophoresis requires careful standardization and optimization to obtain reliable results [40]. Once this has been done, however, the technique can be used routinely with high throughput for screening [56]. Gel electrophoresis techniques are also very popular for screening and can provide rapid visual information on the types of glycan present [57]. Recently, DNA sequencing instrumentation has also been applied to glycan profiling and this adds full automation to the analysis [58].

11.2.4 Glycan Quantification

Quantification generally requires derivatization with a chromo- or fluorophore. Semiquantitative data may be obtained from electrochemical detection [59a, 60] or mass spectrometry [61], but unless the system has been fully calibrated with all appropriate standards the results may not be reliable. The introduction of a specific fluorophore at the reducing terminus of the glycan can allow the precise and sensitive quantitation of glycans. The label 2-AB has been used extensively for this purpose, as has the related label 2-aminoanthanilic acid. The use of a fluorophore introduces an additional step in the sample preparation and also requires cleanup of the large excess of label used. Although the labeling is quantitative and

essentially complete for a wide range of glycan structures, it should be borne in mind that larger structures may be labeled at a lower efficiency (D. Neville, personal communication).

Quantification by mass spectrometry can provide fairly good quantitative data under certain conditions and particularly where a series of similar glycans are studied. However, where different types of glycans, such as charged and non-charged, are present the differences in ionization can lead to differences in the efficiency of detection of the separate glycans. This was fully discussed by Harvey [62, 63].

11.2.4.1 Effect of Label on Separation. The choice of label may substantially affect the properties of the glycans during chromatography or electrophoresis, hence the introduction of a relatively hydrophobic label such as 2-aminopyridine can affect the separation when based on the hydrophilicity of the glycan. [33]. The use of the relatively highly charged and bulky ANTS label will have substantial effects, but is very beneficial for electrophoretic separation [41]. The label 2-AB, which is comparatively small, non-charged, and fairly hydrophilic, is very useful in separations under normal-phase conditions based on hydrophilic interactions [37], and the free acid form 2-aminobenzoic acid (anthranilic acid, 2AA) can similarly be used [43] and may give improved separation and detection sensitivity [64].

11.2.5 Glycan Sequencing

The complete sequence analysis of all glycans is still not a trivial matter for most glycoproteins. Many glycans are very similar, or indeed are isometric forms, and therefore difficult to separate and may be present in only low amounts. In many cases it is possible, however, to characterize all the major glycans, and in the case when the glycoprotein comes from a source where the activities of biosynthetic enzymes are known, and where characteristic glycoproteins have already been sequenced, a lot of information can be obtained, and sequences predicted, without the need for rigorous sequence analysis. This applies to many glycoproteins of biological interest.

The peaks separated in the profiles cannot generally be identified directly as several glycans may have the same properties, and would not be separated. For example, they might have the same mass or charge. Hence the sequence analysis requires further information about the component and this is generally achieved by breaking it down in some way. One method which has proved very successful is the use of specific exoglycosidases [65–70], which only cleave given monosaccharides in a certain linkage. By application of a number of different combinations of such exoglycosidases, it is often possible to sequence the glycan completely. Another technique that is useful when exoglycosidases of the required specificity are not available is to induce fragmentation in the molecule while analyzing it by mass spectrometry [20, 71, 72].

An example of exoglycosidase sequencing is shown in Figure 11.2 [46]. It should be borne in mind, however, that the monosaccharides removed by the glycosidases may be more readily removed from some positions in the glycan than others and conditions may need to be optimized for complete removal. In such cases, it may be necessary to check the specificity *under the incubation conditions used* rather than those used by the supplier as low activities of contaminating enzymes may be present which could prevent interpretation of the data. It is advisable to test all batches of enzyme on known standards under the same methods as used for unknown samples. One advantage of this technique in combination with the fluorescent label described above is that the label is retained after the removal of the monosaccharide

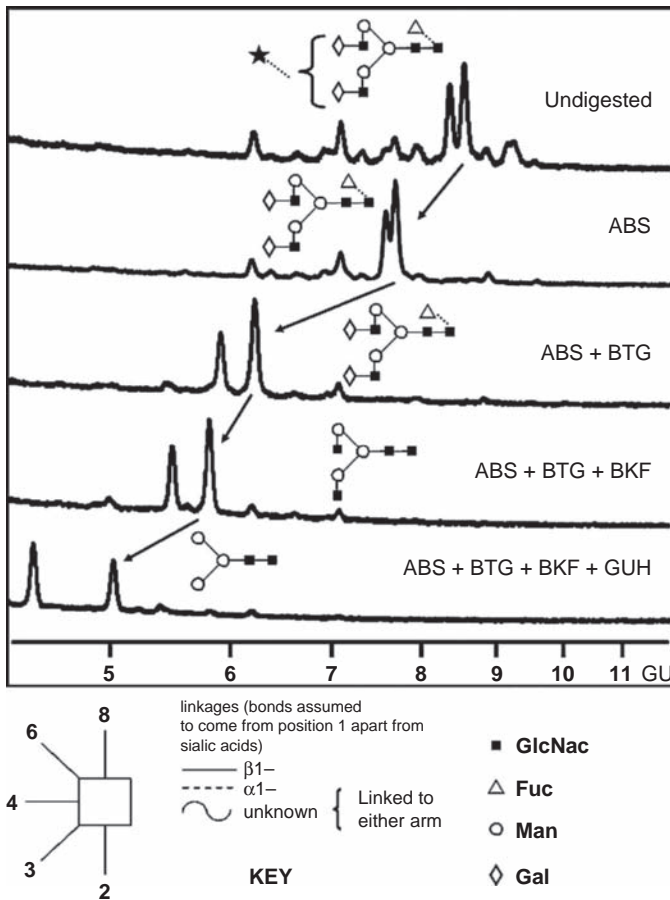


Figure 11.2 An example of exoglycosidase sequencing. Exoglycosidase sequencing of glycan pool from from IgD. The sequential degradation of peak A is indicated showing the intermediate structures resulting from each glycosidase step. Enzyme abbreviations: ABS, *Arthrobacter ureafaciens* sialidase; BTG, bovine testes β -galactosidase; BKF, bovine kidney α -fucosidase; GUH, recombinant β -*N*-acetylglucosaminidase.

by the enzyme. It is also possible to use combinations of many of the exoglycosidases as they have sufficiently similar pH optima and buffer requirements for a single buffer to be used [66, 69]. It is also possible to digest a pool of glycans rather than to digest them separately. A list of common enzymes and incubation conditions is given in Table 11.1.

A very useful feature is that each monosaccharide in the glycan was found to contribute a discrete glucose unit (gu) value to that of the glycan. Following digestion, there is information about the nature of the residue removed, its linkage, the number of residues removed, and in some cases the location in the glycan structure (it was found that the same residue in different positions on the glycan may have different gu values; for example, when GlcNac is in a β 1,3 “bisected” linkage to the trimannosyl core the structure has a gu value 4.95, giving a value of 0.55 for the GlcNac residue, but when it is in a β 1,2 linkage on the upper mannose arm the value for the structure is 4.92 gu giving a value of 0.52 gu for

Table 11.1 Commonly used exoglycosidases and incubation conditions.

Enzyme	Specificity	Buffer	Concentration and incubation time ^a
Sialidase <i>Arthrobacter ureafaciens</i> EC 3.2.1.18	α 2–3, 6, 8 sialic acid	50 mM sodium acetate, pH 5.5	1 μ U μ l ⁻¹ 3–16 h
Sialidase From Newcastle disease virus (recombinant) EC 3.2.1.26	α 2,3 sialic acid only	50 mM sodium phosphate, pH 6.0	1 μ U μ l ⁻¹ 16 h
β -Galactosidase Bovine testes EC 3.5.1.52	β 1, 3 or 4 Gal	100 mM citrate–phosphate, pH 5.0	1 mU μ l ⁻¹ 3–16 h
β -Galactosidase <i>Streptomyces pneumoniae</i> EC 3.2.1.23	β 1,3 Gal	100 mM sodium acetate, pH 6.0	80 μ U μ l ⁻¹ 16 h
β -N-Acetylhexoseaminidase Recombinant EC 3.2.1.23	β 1, 3 or 4 GlcNAc or GalNAc	100 mM sodium citrate–phosphate, pH 6.0	10 μ U μ l ⁻¹ 16 h
β -N-Acetylglucosaminidase <i>Streptomyces pneumoniae</i> EC 3.2.1.30	β 1, 3 or 4 GlcNAc	100 mM sodium citrate–phosphate, pH 6.0	120 μ U μ l ⁻¹ 16 h
α -Fucosidase Almond meal EC 3.2.1.111	α 1-3 Fuc	50 mM sodium acetate, pH 5.0	1 mU μ l ⁻¹ 16 h
α -Fucosidase Bovine kidney EC 3.2.1.51	α 1–4, 6 Fuc	100 mM sodium citrate, pH 6.0	1 mU μ l ⁻¹ 16 h
α -Mannosidase Jack bean EC 3.2.1.23	α 1–2, 4 Man	100 mM sodium acetate, 2 mM Zn, pH 5.0	67 mU μ l ⁻¹ 2x 16 h
Endo- β -galactosidase <i>Bacteroides fragilis</i> EC 3.2.1.102	β 1–3,4 Gal in poly N-acetyllactosamine	50 mM sodium acetate, pH 5.8	100 μ U μ l ⁻¹ 3 h

^aIt should be noted that for most of these exoglycosidases the enzyme concentration and the incubation times have not been optimized for the type of glycan or the amount used in studies with fluorescent glycans. The conditions are generally those which can be expected to give complete digestion of the particular linkages. In practice, lower enzyme concentrations and shorter incubation times could often be used. It is advisable to check any enzymes against known standard glycans of different types before use in sequencing.

the GlcNAc, whereas on the lower mannose arm also in a β 1,2 linkage the gu value is 5.06, giving a value of 0.66 gu for the GlcNAc). Sequencing by use of exoglycosidase digests was initially performed using analysis by P4 size-exclusion chromatography, which was a tedious and lengthy task. The digestion with each enzyme was performed individually and the fractions collected were then subject to the next exoglycosidase in the sequence. Considering the long run times (48 h), the large volumes collected, and the fact that the digests had to be rendered salt free before the chromatography, along with the relatively low sensitivity of detection of tritium, this was a monumental effort. Despite this, a very large number of glycans were identified and sequenced mainly by the groups of Kobata and Dwek.

With the advent of suitable HPLC technology and fluorimetric detection, it was soon realized that the process could be made quicker and more sensitive. A joint development

by the Oxford *Glycobiology* Institute and Oxford GlycoSciences led to a further advance in that digests could be performed with several enzymes simultaneously. It was found that by selection of suitable sources of the exoglycosidases, suitable buffer conditions could be found where all the enzymes were active (although not necessarily at their pH optimum). This system was termed the random array analysis method (RAAM) and was incorporated into a commercial product by OGS.

The technique did, however, still require the purification of individual glycans before sequencing. Not only did this greatly increase the work load, but also it was sometimes extremely difficult to separate glycans completely on a single column system. Work by Rudd and co-workers showed that it was possible to deconvolute the information on the digest of the entire glycan pool without separation. This was exemplified by the complete sequencing of the glycan of GP59, a total of 109 structures being identified.

This technique, along with the increased sensitivity of fluorimetric detection, considerably reduced the time and sensitivity of glycan sequencing by HPLC. However, the complexity of the data interpretation increased considerably. The resolution of the system is such that several glycans may be eluting in the same peak. Hence the shift in profile during the exoglycosidases digest of a pool may result in changes in the relative areas of peaks rather than their complete removal. Considering that the peaks resulting from the digest may also consist of several glycan species, the interpretation of the digests is a highly skilled task. Attempts have been made to computerize the data interpretation, using for example the PeakTime program developed by the Oxford *Glycobiology* Institute. This system was never made generally available, however, and a revised version of the software known as GlycoBase has now been developed within the EuroCarbDB program.

Sequencing by exoglycosidases is generally performed along with HPLC analysis, particularly using normal-phase HPLC, where changes in the positions of the peaks relative to the dextran standard (gu values) may be related to both the type and the number of monosaccharides removed. It is also possible to perform enzymatic sequencing with mass spectrometric analysis [73], but increasingly fragmentation analysis is being employed with various techniques and this has the advantage that not all of the desired specificities of exoglycosidase exist. A drawback is that the fragmentation is very dependent on the technique and instrumentation and is difficult to predict. The relative merits of different techniques and their application have been discussed [63]. In practice, a combination of enzymatic and fragmentation analysis may be used to give as much glycan sequence information as possible.

11.3 Separation by Capillary Electrophoresis (CE)

Capillary electrophoresis can allow rapid and precise profiling of glycans. There are, however, a number of points to consider when using this technique. Unlike HPLC techniques, where the number of columns with suitable media is strictly limited, there are a number of choices for separation by CE. The various types of technique were discussed fully by Taverna *et al.* [56] and by Cammelleri and co-workers [74] and will not be considered further here.

In general, CE is a technique used for profiling where a comparison of glycosylation between samples is required rather than a detailed analysis. Frequently an initial analysis has been performed by other techniques, so reference to this will allow identification of the glycans from a profile. The variations in techniques and labels and the large effect of

minor changes in running conditions may make it difficult to produce standard values, but a library of profiles obtained by various techniques would be of some value.

11.4 Separation by HPLC

Amide and reversed-phase HPLC have been in use for several years. The use of an amine column (AX-5) was initially described in 1985 [75] and later refined by Takeuchi *et al.* [76]. The use of high-pH anion-exchange columns was described by Lee and co-workers [59b, 77, 78] and subsequently developed by Dionex Corporation as high-performance anion-exchange chromatography in combination with pulsed amperometric electrochemical detection (HPAEC-PAD) [26b, 79, 80]. These techniques are now the most widely used for the separation of complex glycans.

11.4.1 Anion Exchange

Many of the original separations of glycans were carried out on strong anion-exchange media (SAX). The procedures used conditions under which either certain groups on the glycan become charged or where a native charged group was present, such as sialic acids or sulfate. Indeed, this is still the method of choice for the separation of the highly charged glycosamine glycan from proteoglycans [81]. For general separation of glycans, weak anion-exchange chromatography (WAX) is now generally performed. This will readily separate the various sialylated forms of N-linked glycans, for example [46]. The separation of charged and non-charged glycans from a recombinant form of the HIV surface coat protein gp160 on a weak anion-exchange medium is shown in Figure 11.3.

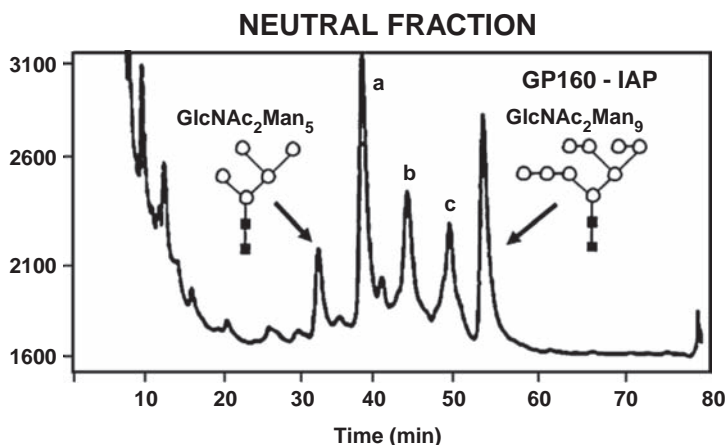


Figure 11.3 Glycan separation by capillary electrophoresis. The separation of the neutral glycans from GP160 by CE is shown. The separation of the oligomannose glycans ranging from oligomannose 5 to oligomannose 9 is shown. Structures a–c represent oligomannose glycans with six, seven and eight glycans respectively. Reprinted from Taverna M, Nguyet TT, *et al.*: A multi-mode chromatographic method for the comparison of the N-glycosylation of a recombinant HIV envelope glycoprotein (gp160s-MN/LAI) purified by two different processes. *Journal of Biotechnology* **68**, 37–48, Copyright 1999, with permission from Elsevier.

11.4.2 High-pH Anion Exchange with Pulsed Amperometric Detection (HPAEC-PAD)

In this variant of anion-exchange chromatography, a charge is induced in the oxyanion at high pH and this is then used to allow separation on a pellicular ion-exchange medium. The very high surface area and size of pores present allow the separation of either monosaccharides or larger glycan structures. Problems with this system are the lack of predictability of the separation and also variations in elution times on different systems, hence standardization for purposes of identification of glycans can prove difficult [82].

The system is very popular, however, as it requires no derivatization for high-sensitivity analysis and is widely used for the routine profiling of glycans where these have previously been identified by other techniques [83]. An example of a glycan profile on HPAEC is shown in Figure 11.4

11.4.3 Normal Phase

The use of normal-phase chromatography on amide-based columns has proved very useful in the characterization and sequencing of both *N*- and *O*-linked glycans. This was originally described for the separation of 2-AP-labeled glycans by normal-phase HPLC [38] and was combined with other separation methods to give the identity of glycans by 2D mapping [38]. The separation of glycans on normal-phase amide columns has permitted profiling of glycans from many glycoproteins of biological interest. As an example, the glycans for gp 160 are shown in Figure 11.5.

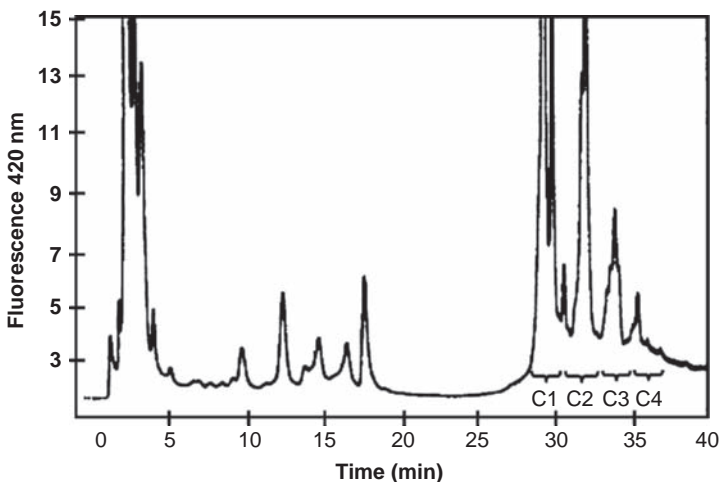


Figure 11.4 Glycan profiling on a weak anion-exchange column. Data on glycans from a preparation of recombinant gp160 showing separation of neutral and charged glycans. The separation of neutral glycans and of peaks containing glycans in different charge states from one (C1) to four (C4) is shown. Reprinted from Taverna M, Nguyet TT, *et al.*: A multi-mode chromatographic method for the comparison of the *N*-glycosylation of a recombinant HIV envelope glycoprotein (gp160s-MN/LAI) purified by two different processes. *Journal of Biotechnology* **68**, 37–48, Copyright 1999, with permission from Elsevier.

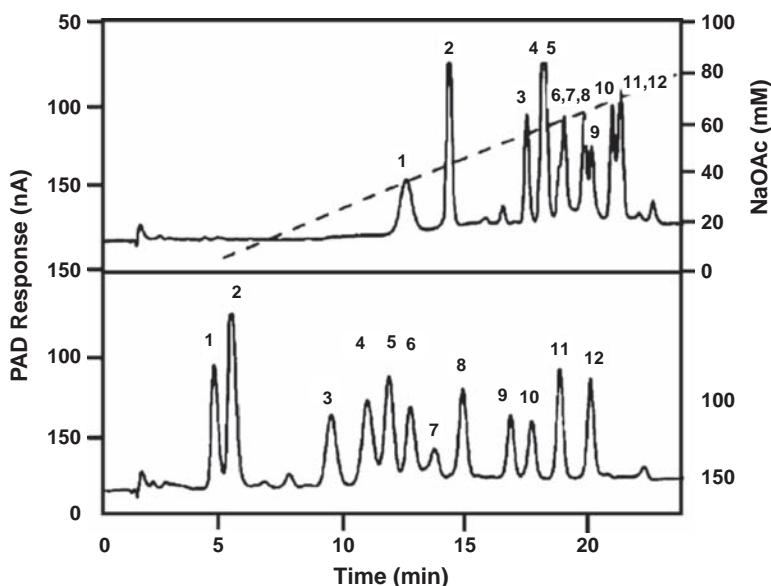


Figure 11.5 Glycan profiling by high-pH anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD): separation of a range of neutral *N*-linked glycans on a Carbopac PA1000 anion-exchange column. The separation of 12 glycans using isocratic conditions and a sodium acetate gradient for elution is shown. Reprinted from Cooper G, Rohrer I: Separation of neutral asparagine-linked oligosaccharides by high-pH anion-exchange chromatography with pulsed amperometric detection. *Analytical Biochemistry* **226**, 182–184, Copyright 1995, with permission from Elsevier.

11.4.4 Reversed Phase

Reversed-phase chromatography on an octadecylsilica column (ODS C_{18}) has also been used to separate fluorescently labeled glycans [70] and may allow the separation of glycans that cannot be resolved by other techniques. An example is shown in Figure 11.6, where the differential separation of glycans on both normal- and reversed-phase media is shown.

11.5 Data Analysis and Databases

11.5.1 Capillary Electrophoresis

The data from capillary electrophoresis are generally recorded as a profile of the detector output against time. This may be compiled into a table with a list of peaks which are sometimes identified if a complete analysis of the glycans has been performed. Although there are numerous reports and examples of separations, there is no database of such values at present.

11.5.2 Data Obtained from HPLC Analysis

Generally, a single detection system is used with HPLC for glycan detection. This is either an electrochemical detector as in HPAEC-PAD, or a fluorescence detector. In HPAEC-PAD,

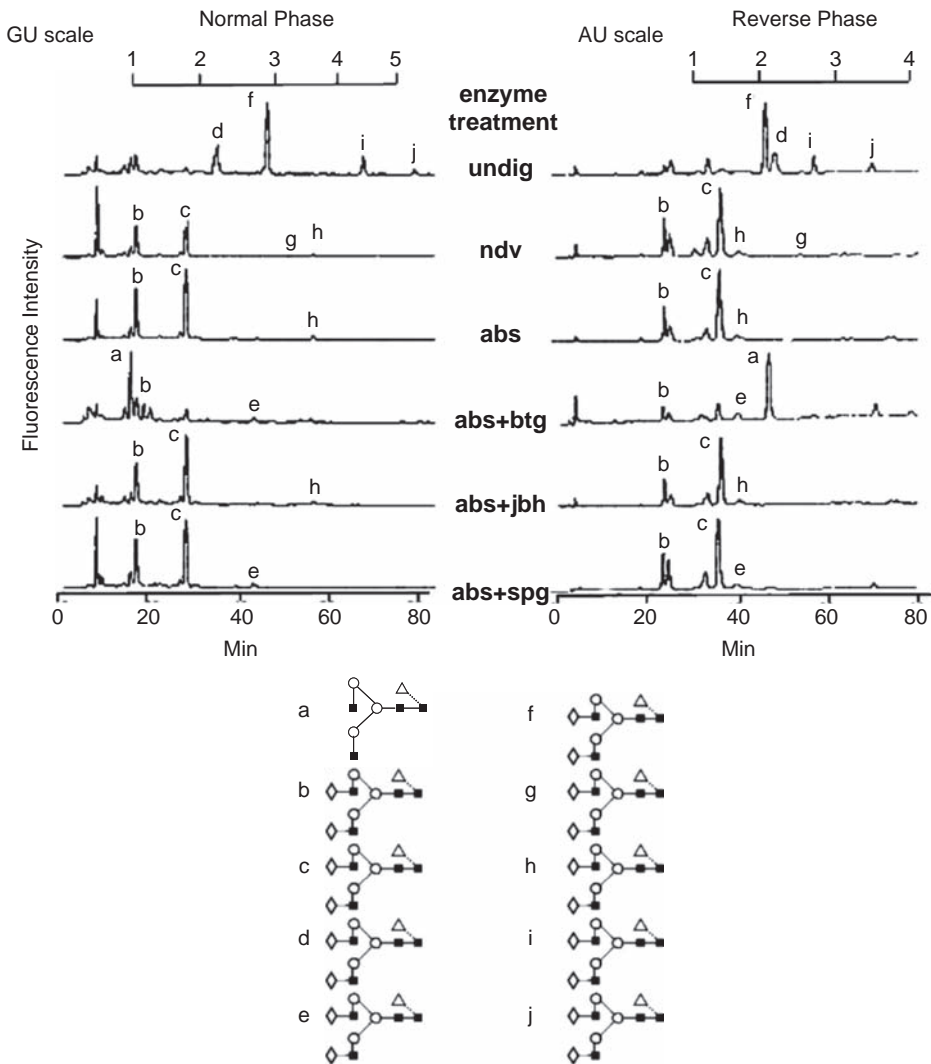


Figure 11.6 Glycan separation by normal-phase chromatography (amide column) and by reversed-phase chromatography (ODS C₁₈). GU, glucose units; AU, arabinose units; SPG, *Streptomyces pneumoniae* β -galactosidase; enzyme abbreviations as in Figure 11.2.

standards are used in monosaccharide analysis but are not generally employed in glycan analysis. In contrast, in HPLC, a standard is usually used to reference the elution of different glycans. The data are recorded as a profile of detector response against time, but are then generally calibrated against the dextran ladder and so each peak is given a gu value. If the sample is run using different types of chromatography, then a number of gu values on each type of chromatographic system are recorded. In some cases a value from another polymer may be used, such as arabinose units. Tables may then be compiled with gu values for known glycans on the same type of separation. It is also possible to construct a table

which shows the incremental values for each type of monosaccharide, often in terms of the linkage type.

Further to this, exoglycosidase digestions are often performed to aid glycan identification, so it is possible to predict the shift in a peak's *gu* value following treatment with any given glycosidase. After using a series of such exoglycosidases, it is frequently possible to identify many of the glycans, although it is advisable to check these by mass spectrometry.

This is generally in the form of an external standard of a series of glucose oligomers obtained by the partial acid hydrolysis of dextran, the so-called glucose units (*gu*). In some cases, an alternative series of oligomers may be used, for example arabinose for reversed-phase HPLC [84].

11.5.3 Use of Glucose Units (*gu*)

The analysis of complex glycans was originally carried out on polysaccharides, and for purposes of standardization a ladder consisting of oligomers of glucose was used. With the advent of the analysis of glycans from glycoproteins, this method of standardization was still employed especially when the initial analysis was carried out by gel permeation chromatography such as on the medium BioGel P4. The use of glucose units as a means of standardization was advocated by Kobata *et al.* [85].

As the analysis of glycan progressed to HPLC techniques, the use of glucose units was retained and further developed so that glucose units determined on two [86] and later three different types of separation could be used to identify a glycan uniquely from comparison of *gu* values on each type of column against a reference list of values for standard glycans [86, 87].

At present, no commercial software is available for the interpretation of data from HPLC or mass spectrometric analysis. The interpretation of data therefore requires detailed knowledge of the properties of the glycans and of exoglycosidases or fragmentation patterns. This will limit the throughput in the sequence analysis and attempts are being made to enable software to be used for interpretation. An example of this is the PeakTime software described by Rudd *et al.* [70]; however, this is not available for distribution and is specifically designed for use with one type of HPLC system. A database of *gu* values for 2-AP-labeled glycans separated on three different media is now available that allows the identification of glycans by 3D mapping as described by Takasaki and co-workers, and this database is freely available and has more general applicability. A comprehensive database of glucose unit values for 3D mapping has recently been established by Takasaki's group at http://www.gak.co.jp/ECD/Hpg_eng/hpg_eng.htm. This database is very comprehensive and has extensive search capabilities.

At present, there is no comprehensive database which is freely available for glycan structures that have been identified experimentally. The original definitive database CarbBank [88] is now out of date and some of the other databases (GlycosuiteDB and Glycomics Database) are only available commercially. A number of specialist databases for HPLC, mass spectrometry, exoglycosidases and NMR do exist and these are listed in Table 11.2.

11.5.4 Display of Glycan Structures

The presentation of a large number of structures using the standard IUPAC nomenclature is difficult and somewhat cumbersome. It is also difficult to assimilate information by eye. The established IUPAC nomenclature [89] gives a complete and unambiguous description of the

Table 11.2 Typical arrays for *N*-glycan sequencing.

Array	Vol. enzyme (μ l)	Vol. buffer (μ l)	Vol. water (μ l)
ABS	1	2	7
ABS BTG	1, 2	2	5
ABS BTG BKF	1, 2, 1	2	4
ABS BTG BKF AMF	1, 2, 1, 1	2	3
ABS BTG BKF AMF JBH	1, 2, 1, 1, 2	2	1
JBM (high)	10		
JBM (low)	0.8		9.2

glycan, but it is difficult to interpret when many different glycans are shown. It is common practice to use some kind of pictorial scheme to show the structure in a table or profile along with its abundance. Unfortunately, there is no common scheme in place and different types of representation are used by different groups. A unified scheme incorporating features from several existing schemes has been proposed and is explained in a separate chapter (see Chapter 3). The interlinking of separate databases and provision of crosslinking with other databases such as protein structure databases would help with the analysis and cross-referencing of glycan structures [90] in addition to helping to correlate glycan structure with glycoprotein function.

11.6 Future Developments

Many techniques are now being used for glycan profiling with a view to rapid analysis. Capillary electrophoresis is one such technique and may be optimized for glycan profiling [91, 92]. The use of automated equipment adapted from DNA sequencers [58, 93] could allow high-throughput screening for comparative purposes. Further developments in these fields and in others such as microchips [92] may increase sensitivity and increase the speed of analysis. It is envisaged that the further development of mass spectrometry will lead to both an increase in the capability and a reduction in the cost of such systems. Developments using HPLC with on-line mass spectrometry would be an advantage for comprehensive analysis. Systems based on the identification of glycan structures through lectin-based recognition have now also been developed but remain to be fully evaluated. The refinements in mass spectrometry could also make glycopeptide analysis feasible in more cases. The greatest advances, however, are likely to be those in the automated data analysis and provision of more comprehensive and accessible databases such as those being developed in the EuroCarbDB project (see Chapter 1).

References

1. Apweiler R, Hermjakob H, *et al.*: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999, **1473** (1):4–8.
2. Elliott HG, Elliott MA, *et al.*: Chromatographic investigation of the glycosylation pattern of alpha-1-acid glycoprotein secreted by the HepG2 cell line; a putative model for inflammation? *Biomed Chromatogr* 1995, **9** (5):199–204.

3. Litynska A, Przybylo M, *et al.*: Differences of alpha3beta1 integrin glycans from different human bladder cell lines. *Acta Biochim Pol* 2000, **47** (2):427–434.
4. Hakomori S: Tumor-associated carbohydrate antigens defining tumor malignancy: basis for development of anti-cancer vaccines. *Adv Exp Med Biol* 2001, **491**:369–402.
5. Butler M, Quelhas D, *et al.*: Detailed glycan analysis of serum glycoproteins of patients with congenital disorders of glycosylation indicates the specific defective glycan processing step and provides an insight into pathogenesis. *Glycobiology* 2003, **13** (9):601–622.
6. Sutton-Smith M, Morris HR, *et al.*: MS screening strategies: investigating the glycomes of knockout and myodystrophic mice and leukodystrophic human brains. *Biochem Soc Symp* 2002, (69):105–115.
7. Kui Wong N, Easton RL, *et al.*: Characterization of the oligosaccharides associated with the human ovarian tumor marker CA125. *J Biol Chem* 2003, **278** (31):28619–28634.
8. Aminoff D, Gathmann WD, *et al.*: Quantitation of oligosaccharides released by the beta-elimination reaction. *Anal Biochem* 1980, **101** (1):44–53.
9. Strecker G, Wieruszkeski JM, *et al.*: Primary structure of 12 neutral oligosaccharide-alditols released from the jelly coats of the anuran *Xenopus laevis* by reductive beta-elimination. *Glycobiology* 1995, **5** (1):137–146.
10. Fukuda M, Kondo T, *et al.*: Studies on the hydrazinolysis of glycoproteins. Core structures of oligosaccharides obtained from porcine thyroglobulin and pineapple stem bromelain. *J Biochem (Tokyo)* 1976, **80** (6):1223–1232.
11. Strecker G, Pierce-Cretel A, *et al.*: Characterization by gas–liquid chromatography–mass spectrometry of oligosaccharides resulting from the hydrazinolysis–nitrous acid deamination reaction of glycopeptides. *Anal Biochem* 1981, **111** (1):17–26.
12. Yamashita K, Ohkura T, *et al.*: Comparative study of the oligosaccharides released from baby hamster kidney cells and their polyoma transformant by hydrazinolysis. *J Biol Chem* 1984, **259** (17):10834–10840.
13. Bendiak B, Cumming DA: Hydrazinolysis-*N*-reacetylation of glycopeptides and glycoproteins. Model studies using 2-acetamido-1-*N*-(L-aspart-4-oyl)-2-deoxy-beta-D-glucopyranosylamine. *Carbohydr Res* 1985, **144** (1):1–12.
14. Patel T, Bruce J, *et al.*: Use of hydrazine to release in intact and unreduced form both *N*- and *O*-linked oligosaccharides from glycoproteins. *Biochemistry* 1993, **32** (2):679–693.
15. Waddling CA, Plummer TH Jr, *et al.*: Structural basis for the substrate specificity of endo-beta-*N*-acetylglucosaminidase F (3). *Biochemistry* 2000, **39** (27):7878–7885.
16. Robbins PW, Wirth DF, *et al.*: Expression of the *Streptomyces* enzyme endoglycosidase H in *Escherichia coli*. *J Biol Chem* 1981, **256** (20):10640–10644.
17. Plummer TH Jr, Tarentino AL: Purification of the oligosaccharide-cleaving enzymes of *Flavobacterium meningosepticum*. *Glycobiology* 1991, **1** (3):257–263.
18. Takasaki S, Mizuochi T, *et al.*: Hydrazinolysis of asparagine-linked sugar chains to produce free oligosaccharides. *Methods Enzymol* 1982, **83**:263–268.
19. Kuster B, Wheeler SF, *et al.*: Sequencing of *N*-linked oligosaccharides directly from protein gels: in-gel deglycosylation followed by matrix-assisted laser desorption/ionization mass spectrometry and normal-phase high-performance liquid chromatography. *Anal Biochem* 1997, **250** (1):82–101.
20. Wheeler SF, Harvey DJ: Extension of the in-gel release method for structural analysis of neutral and sialylated *N*-linked glycans to the analysis of sulfated glycans: application to the glycans from bovine thyroid-stimulating hormone. *Anal Biochem* 2001, **296** (1):92–100.
21. Altmann F, Schweiszer S, *et al.*: Kinetic comparison of peptide: *N*-glycosidases F and A reveals several differences in substrate specificity. *Glycoconj J* 1995, **12** (1):84–93.
22. Ko K, Tekoah Y, *et al.*: Function and glycosylation of plant-derived antiviral monoclonal antibody. *Proc Natl Acad Sci USA* 2003, **100** (13):8013–8018.

23. Tekoah Y, Ko K, *et al.*: Controlled glycosylation of therapeutic antibodies in plants. *Arch Biochem Biophys* 2004, **426** (2):266–278.
24. Merry AH, Neville DC, *et al.*: Recovery of intact 2-aminobenzamide-labeled *O*-glycans released from glycoproteins by hydrazinolysis. *Anal Biochem* 2002, **304** (1):91–99.
25. Martens DA, Frankenberger WT Jr: Determination of saccharides in biological materials by high-performance anion-exchange chromatography with pulsed amperometric detection. *J Chromatogr* 1991, **546** (1–2):297–309.
26. (a) Weitzhandler M, Hardy M, *et al.*: Analysis of carbohydrates on IgG preparations. *J Pharm Sci* 1994, **83** (12):1670–5; (b) Weitzhandler M, Hardy M: Sensitive blotting assay for the detection of glycopeptides in peptide maps. *J Chromatogr* 1990, **510**:225–232.
27. Harvey DJ: Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrom Rev* 1999, **18** (6):349–450.
28. Harvey DJ: *N*-(2-Diethylamino)ethyl-4-aminobenzamide derivative for high sensitivity mass spectrometric detection and structure determination of *N*-linked carbohydrates. *Rapid Commun Mass Spectrom* 2000, **14** (10):862–871.
29. Harvey DJ: Identification of protein-bound carbohydrates by mass spectrometry. *Proteomics* 2001, **1** (2):311–328.
30. Endo Y, Yamashita K, *et al.*: Structures of the asparagine-linked sugar chains of human chorionic gonadotropin. *J Biochem (Tokyo)* 1979, **85** (3):669–679.
31. Mizuochi T, Taniguchi T, *et al.*: Comparative studies on the structures of the carbohydrate moieties of human fibrinogen and abnormal fibrinogen Nagoya. *J Biochem (Tokyo)* 1982, **92** (1):283–293.
32. Kobata A: A journey to the world of glycobiology. *Glycoconj J* 2000, **17** (7–9): 443–464.
33. Hase S, Hara S, *et al.*: Tagging of sugars with a fluorescent compound, 2-aminopyridine. *J Biochem (Tokyo)* 1979, **85** (1):217–220.
34. Hase S, Ikenaka T, *et al.*: Structure analyses of oligosaccharides by tagging of the reducing end sugars with a fluorescent compound. *Biochem Biophys Res Commun* 1978, **85** (1):257–263.
35. Camilleri JP, Nlom MO, *et al.*: Capillary perfusion patterns in reperfused ischemic subendocardial myocardium: experimental study using fluorescent dextran. *Exp Mol Pathol* 1983, **39** (1):89–99.
36. Prakash C, Vijay IK: A new fluorescent tag for labeling of saccharides. *Anal Biochem* 1983, **128** (1):41–46.
37. Bigge JC, Patel TP, *et al.*: Nonselective and efficient fluorescent labeling of glycans using 2-aminobenzamide and anthranilic acid. *Anal Biochem* 1995, **230** (2):229–238.
38. Tomiya N, Awaya J, *et al.*: Analyses of *N*-linked oligosaccharides using a two-dimensional mapping technique. *Anal Biochem* 1988, **171** (1):73–90.
39. Caesar JP Jr, Sheeley DM, *et al.*: Femtomole oligosaccharide detection using a reducing-end derivative and chemical ionization mass spectrometry. *Anal Biochem* 1990, **191** (2):247–252.
40. Camilleri P, Harland GB, *et al.*: High resolution and rapid analysis of branched oligosaccharides by capillary electrophoresis. *Anal Biochem* 1995, **230** (1):115–122.
41. Jackson P: Fluorophore-assisted carbohydrate electrophoresis: a new technology for the analysis of glycans. *Biochem Soc Trans* 1993, **21** (1):121–125.
42. Jackson P: Polyacrylamide gel electrophoresis of reducing saccharides labeled with the fluorophore 2-aminoacridone: subpicomolar detection using an imaging system based on a cooled charge-coupled device. *Anal Biochem* 1991, **196** (2):238–244.
43. Anumula KR, Dhume ST: High resolution and high sensitivity methods for oligosaccharide mapping and characterization by normal phase high performance liquid chromatography following derivatization with highly fluorescent anthranilic acid. *Glycobiology* 1998, **8** (7):685–694.
44. Anumula KR, Du P: Characterization of carbohydrates using highly fluorescent 2-aminobenzoic acid tag following gel electrophoresis of glycoproteins. *Anal Biochem* 1999, **275** (2):236–242.
45. Takahashi N, Matsuda T, *et al.*: A structural study of the asparagine-linked oligosaccharide moiety of duck ovomucoid. *Glycoconj J* 1993, **10** (6):425–434.

46. Taverna M, Nguyet TT, *et al.*: A multi-mode chromatographic method for the comparison of the N-glycosylation of a recombinant HIV envelope glycoprotein (gp160s-MN/LAI) purified by two different processes. *J Biotechnol* 1999, **68** (1):37–48.
47. Rudd PM, Opdenakker G, *et al.*: Holistic approaches to glycobiology. *Nat Biotechnol* 2001, **19** (6):531–532.
48. Suzuki-Sawada J, Umeda Y, *et al.*: Analysis of oligosaccharides by on-line high-performance liquid chromatography and ion-spray mass spectrometry. *Anal Biochem* 1992, **207** (2):203–207.
49. Medzihradsky KF, Besman MJ, *et al.*: Structural characterization of site-specific N-glycosylation of recombinant human factor VIII by reversed-phase high-performance liquid chromatography–electrospray ionization mass spectrometry. *Anal Chem* 1997, **69** (19):3986–3994.
50. Charlwood J, Birrell H, *et al.*: Carbohydrate release from picomole quantities of glycoprotein and characterization of glycans by high-performance liquid chromatography and mass spectrometry. *J Chromatogr B Biomed Sci Appl* 1999, **734** (1):169–174.
51. Kawasaki N, Ohta M, *et al.*: Usefulness of sugar mapping by liquid chromatography/mass spectrometry in comparability assessments of glycoprotein products. *Biologicals* 2002, **30** (2):113–123.
52. Lee YC, Lee BI, *et al.*: Parameterization of contribution of sugar units to elution volumes in reverse-phase HPLC of 2-pyridylaminated oligosaccharides. *Anal Biochem* 1990, **188** (2):259–266.
53. Takahashi N, Wada Y, *et al.*: Two-dimensional elution map of GalNAc-containing N-linked oligosaccharides. *Anal Biochem* 1993, **208** (1):96–109.
54. Guile GR, Rudd PM, *et al.*: A rapid high-resolution high-performance liquid chromatographic method for separating glycan mixtures and analyzing oligosaccharide profiles. *Anal Biochem* 1996, **240** (2):210–226.
55. Tran NT, Taverna M, *et al.*: A sensitive mapping strategy for monitoring the reproducibility of glycan processing in an HIV vaccine, RGP-160, expressed in a mammalian cell line. *Glycoconj J* 2000, **17** (6):401–406.
56. Taverna M, Tran NT, *et al.*: Electrophoretic methods for process monitoring and the quality assessment of recombinant glycoproteins. *Electrophoresis* 1998, **19** (15):2572–2594.
57. Frears ER, Merry AH, *et al.*: Screening neutral and acidic IgG N-glycans by high density electrophoresis. *Glycoconj J* 1999, **16** (6):283–290.
58. Callewaert N, Geysens S, *et al.*: Ultrasensitive profiling and sequencing of N-linked oligosaccharides using standard DNA-sequencing equipment. *Glycobiology* 2001, **11** (4):275–281.
59. (a) Townsend RR, Hardy MR, *et al.*: High-performance anion-exchange chromatography of oligosaccharides using pellicular resins and pulsed amperometric detection. *Anal Biochem* 1988, **174** (2):459–470; (b) Townsend RR, Hardy MR, *et al.*: Separation of oligosaccharides using high-performance anion-exchange chromatography with pulsed amperometric detection. *Methods Enzymol* 1989, **179**:65–76.
60. Honda S, Suzuki S, *et al.*: Analysis of N- and O-glycosidically bound sialooligosaccharides in glycoproteins by high-performance liquid chromatography with pulsed amperometric detection. *J Chromatogr* 1990, **523**:189–200.
61. Gennaro LA, Harvey DJ, *et al.*: Reversed-phase ion-pairing liquid chromatography/ion trap mass spectrometry for the analysis of negatively charged, derivatized glycans. *Rapid Commun Mass Spectrom* 2003, **17** (14):1528–1534.
62. Harvey DJ: Quantitative aspects of the matrix-assisted laser desorption mass spectrometry of complex oligosaccharides. *Rapid Commun Mass Spectrom* 1993, **7** (7):614–619.
63. Harvey DJ: Ionization and collision-induced fragmentation of N-linked and related carbohydrates using divalent cations. *J Am Soc Mass Spectrom* 2001, **12** (8):926–37.
64. Dhume ST, Ebert MB, *et al.*: Monitoring glycosylation of therapeutic glycoproteins for consistency using highly fluorescent anthranilic acid. *Methods Mol Biol* 2002, **194**:127–142.

65. Mizuochi T, Yonemasu K, *et al.*: The asparagine-linked sugar chains of subcomponent C1q of the first component of human complement. *J Biol Chem* 1978, **253** (20):7404–7409.
66. Edge CJ, Rademacher TW, *et al.*: Fast sequencing of oligosaccharides: the reagent-array analysis method. *Proc Natl Acad Sci USA* 1992, **89** (14):6338–6342.
67. Kuster B, Naven TJ, *et al.*: Effect of the reducing-terminal substituents on the high energy collision-induced dissociation matrix-assisted laser desorption/ionization mass spectra of oligosaccharides. *Rapid Commun Mass Spectrom* 1996, **10** (13):1645–1651.
68. Prime S, Dearnley J, *et al.*: Oligosaccharide sequencing based on exo- and endoglycosidase digestion and liquid chromatographic analysis of the products. *J Chromatogr A* 1996, **720** (1–2):263–274.
69. Prime S, Merry T: Exoglycosidase sequencing of *N*-linked glycans by the reagent array analysis method (RAAM). *Methods Mol Biol* 1998, **76**:53–69.
70. Rudd PM, Colominas C, *et al.*: A high-performance liquid chromatography based strategy for rapid, sensitive sequencing of *N*-linked oligosaccharide modifications to proteins in sodium dodecyl sulfate polyacrylamide electrophoresis gel bands. *Proteomics* 2001, **1** (2):285–294.
71. Dell A, Tiller PR: A novel mass spectrometric procedure to rapidly determine the position of *O*-acylated residues in the sequence of naturally occurring oligosaccharides. *Biochem Biophys Res Commun* 1986, **135** (3):1126–1134.
72. Harvey DJ, Naven TJ, *et al.*: Comparison of fragmentation modes for the structural determination of complex oligosaccharides ionized by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* 1995, **9** (15):1556–1561.
73. Settineri CA, Burlingame AL: Structural characterization of protein glycosylation using HPLC/electrospray ionization mass spectrometry and glycosidase digestion. *Methods Mol Biol* 1996, **61**:255–278.
74. Hadley M, Gilges M, *et al.*: Capillary electrophoresis in the pharmaceutical industry: applications in discovery and chemical development. *J Chromatogr B Biomed Sci Appl* 2000, **745** (1):177–188.
75. Dua VK, Goso K, *et al.*: Characterization of lacto-*N*-hexaose and two fucosylated derivatives from human milk by high-performance liquid chromatography and proton NMR spectroscopy. *J Chromatogr* 1985, **328**:259–269.
76. Takeuchi M, Takasaki S, *et al.*: Sensitive method for carbohydrate composition analysis of glycoproteins by high-performance liquid chromatography. *J Chromatogr* 1987, **400**:207–213.
77. Hardy MR, Townsend RR, *et al.*: Monosaccharide analysis of glycoconjugates by anion exchange chromatography with pulsed amperometric detection. *Anal Biochem* 1988, **170** (1):54–62.
78. Hayase T, Sheykhazari M, *et al.*: Separation and identification of *O*-linked oligosaccharides derived from glycoproteins by high-pH anion-exchange chromatography. *Anal Biochem* 1993, **211** (1):72–80.
79. Wang WT, Erlansson K, *et al.*: High-performance liquid chromatography of sialic acid-containing oligosaccharides and acidic monosaccharides. *Anal Biochem* 1990, **190** (2):182–187.
80. Anumula KR, Taylor PB: Rapid characterization of asparagine-linked oligosaccharides isolated from glycoproteins using a carbohydrate analyzer. *Eur J Biochem* 1991, **195** (1):269–280.
81. Merry CL, Lyon M, *et al.*: Highly sensitive sequencing of the sulfated domains of heparan sulfate. *J Biol Chem* 1999, **274** (26):18455–18462.
82. Davies MJ, Hounsell EF: HPLC and HPAEC of oligosaccharides and glycopeptides. *Methods Mol Biol* 1998, **76**:79–100.
83. Goodarzi MT, Turner GA: Reproducible and sensitive determination of charged oligosaccharides from haptoglobin by PNGase F digestion and HPAEC/PAD analysis: glycan composition varies with disease. *Glycoconj J* 1998, **15** (5):469–475.
84. Guile GR, Wong SY, *et al.*: Analytical and preparative separation of anionic oligosaccharides by weak anion-exchange high-performance liquid chromatography on an inert polymer column. *Anal Biochem* 1994, **222** (1):231–235.

85. Kobata A, Yamashita K, *et al.*: BioGel P-4 column chromatography of oligosaccharides: effective size of oligosaccharides expressed in glucose units. *Methods Enzymol* 1987, **138**:84–94.
86. Tomiya N, Lee YC, *et al.*: Calculated two-dimensional sugar map of pyridylaminated oligosaccharides: elucidation of the jack bean alpha-mannosidase digestion pathway of Man9GlcNAc2. *Anal Biochem* 1991, **193** (1):90–100.
87. Tomiya N, Takahashi N: Parameterization of contribution of sugar units to elution volumes in HPLC of pyridyl-2-aminated oligosaccharides. *Tanpakushitsu Kakusan Koso* 1991, **36** (1):63–68 (in Japanese).
88. Doubet S, Albersheim P: CarbBank. *Glycobiology* 1992, **2** (6):505.
89. IUPAC–IUB Joint Commission on Biochemical Nomenclature (JCBN): Abbreviated terminology of oligosaccharide chains. Recommendations 1980. *Eur J Biochem* 1982, **126** (3):433–437.
90. Loss A, Bunsmann P, *et al.*: SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res* 2002, **30** (1):405–408.
91. Charlwood J, Langridge J, *et al.*: Profiling of 2-aminoacridone derivatized glycans by electrospray ionization mass spectrometry. *Rapid Commun Mass Spectrom* 1999, **13** (2):107–112.
92. Suzuki S, Honda S: Miniaturization in carbohydrate analysis. *Electrophoresis* 2003, **24** (21):3577–3582.
93. Callewaert N, Schollen E, *et al.*: Increased fucosylation and reduced branching of serum glycoprotein N-glycans in all known subtypes of congenital disorder of glycosylation I. *Glycobiology* 2003, **13** (5):367–375.

Glycomic Mass Spectrometric Analysis and Data Interpretation Tools

Niclas G. Karlsson¹ and Nicolle H. Packer²

¹*Centre for BioAnalytical Sciences, Chemistry Department, NUI Galway, Galway Ireland*

²*Biomolecular Frontiers Research Centre, Department of Chemistry and Biomolecular Sciences, Macquarie University, North Ryde, Sydney, NSW 2109, Australia*

12.1 Introduction

This introduction to glycomic mass spectrometry is intended to provide bioinformaticians with a basic understanding of the information they are likely to obtain with the glycomic/glycoproteomic mass spectrometric methodologies currently used. More detailed reviews of mass spectrometric applications to glycobiology are available [1, 2], but this chapter is intended to help scientists know how this MS information can be analyzed and interpreted using glycomic software applications. Glycomic analysis has in many senses trailed proteomic analysis, and the development of software for glycomic analysis, although similar to the proteomic process, is different since the biological questions involved in functional glycomics are not the same.

There is a need to differentiate the type of research involved in glycomic science as opposed to that of glycoprotein research. In our view, glycomics involves looking at global glycosylation and mechanisms of glycosylation. Glycomics by definition involves all types of glycoconjugates such as free oligosaccharides, glycolipids, including lipopolysaccharides, glycosylinositolphospholipids, and various types of protein-linked glycosylation (glycoproteins, peptidoglycans, proteoglycans), and also includes investigation into the glycosylation machinery itself. Analogous to the proteomic definition, a glycome can be defined as all glycoconjugates synthesized and expressed by a cell or tissue at any defined time. The general definition of a glycoconjugate also needs to be defined; where molecules such as DNA and RNA (containing the sugars deoxyribose and ribose, respectively) are excluded, as are glycosides.

The developments in mass spectrometry together with the parallel development of isolation techniques using gel electrophoresis and chromatography have allowed the analysis of glycosylation on a global level. The use of lectin affinity for the isolation of glycoproteins provides an additional bonus, allowing subfractionation of specific glycoprotein

subclasses. However, up until the late 1990s, it was believed that the sensitivity of glycoconjugate analysis with mass spectrometry was substantially lower than that achieved by the sophistication and sensitivity of peptide analysis work. Developments at the start of the new millennium now allow subpicomole detection of oligosaccharides to become routine, and this has opened up the field for glyco-screening of biological samples at a new level. Not surprisingly, this has led to the need for the more efficient interpretation of this increased amount of data, and scientists are now developing bioinformatic tools to interpret their mass spectrometric glyco-data and make sense of them in a systems biology context.

Mass spectrometry has for a long time played a vital role in the characterization of oligosaccharides. One of the earliest basic tools determining the linkage between monosaccharides was gas chromatography–mass spectrometry (GC–MS) of hydrolyzed and derivatized components of an isolated glycoconjugate [3]. The breakthrough for biological glycoconjugate characterization by mass spectrometry came with the ability to analyze intact glycoconjugates by fast atom bombardment (FAB)-MS and liquid secondary ion (LSI)-MS (reviewed by Peter-Katalinic [4]). Since then, mass spectrometry has remained the fundamental tool for glycoconjugate analysis. This has not completely supplanted the need for methods such as NMR, monosaccharide composition analysis, exoglycosidase digestion, and other techniques, but the sensitivity and quality of the data achieved by mass spectrometry even in its simplest form (providing a molecular mass) give substantial information. Which other techniques are then needed will depend on the question being asked.

In the early phase of mass spectrometry using FAB-MS/LSIMS or the traditional electron impact ionization techniques, it was found that glycoconjugates such as free oligosaccharides and glycosphingolipids showed substantial fragmentation of the oligosaccharide backbone along with the pseudomolecular ion. From this, the nomenclature for oligosaccharide fragmentation was proposed [5] (Figure 12.1), which is still the basis for oligosaccharide mass spectrometric annotation with only minor modifications [6]. For newcomers to the field of oligosaccharide fragmentation using mass spectrometry, it must be emphasized that the nomenclature is there to allow the description of fragmentation rather than as a prediction of the fragmentation itself. To some extent the fragmentation can be directed using different fragmentation techniques, ionization modes (negative or positive), charge states of parent ion, types of mass spectrometer, and energy transmission in the collision event (Table 12.1). Some of this is exemplified in Figure 12.2, where the hexasaccharide alditol Fuc α 1–2Gal β 1–4GlcNAc β 1–6(Fuc α 1–2Gal β 1–3)GalNAcol is fragmented in the negative mode $\{([M - 2H]^{2-}$ ions and $[M - H]^-$ ions $\}$ and in the positive mode $\{([M + Na]^+$ ions $\}$ in a low-energy collision ion trap mass spectrometer. From the spectra, it is obvious that whereas the negative mode in these circumstances produce mostly Z-type fragmentation, the positive mode of the sodiated adduct gives mostly Y-type fragmentation. The differences between the fragmentations of the two different charge states in the negative mode is more subtle, where collision of the higher charge state $[M - 2H]^{2-}$ seems to produce more cross-ring cleavages (A-type) than the lower charge state $[M - H]^-$.

12.2 Sample Preparation for Mass Spectrometric Analysis of the Glycoproteome

This chapter will focus on a subset of the glycome, namely the analysis of the glycosylation of glycoproteins, and more specifically on the mammalian glycosylation found attached

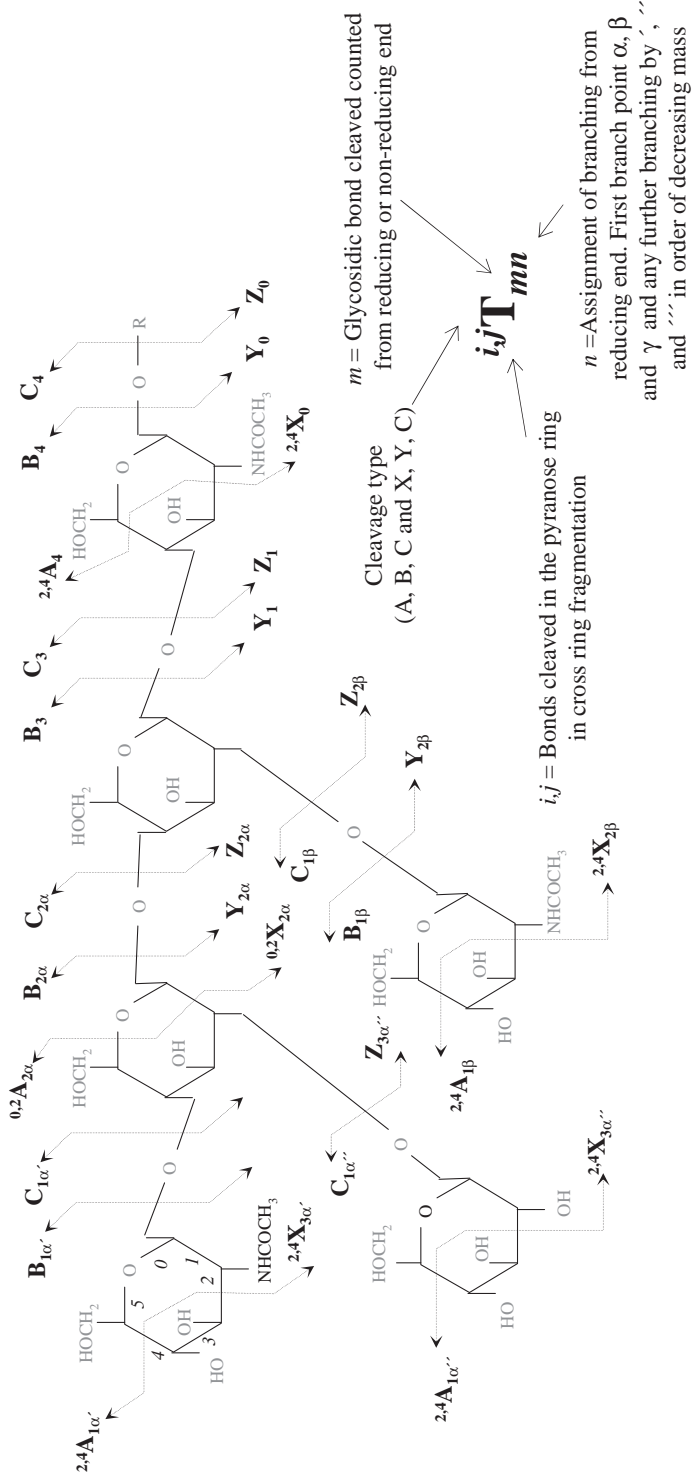
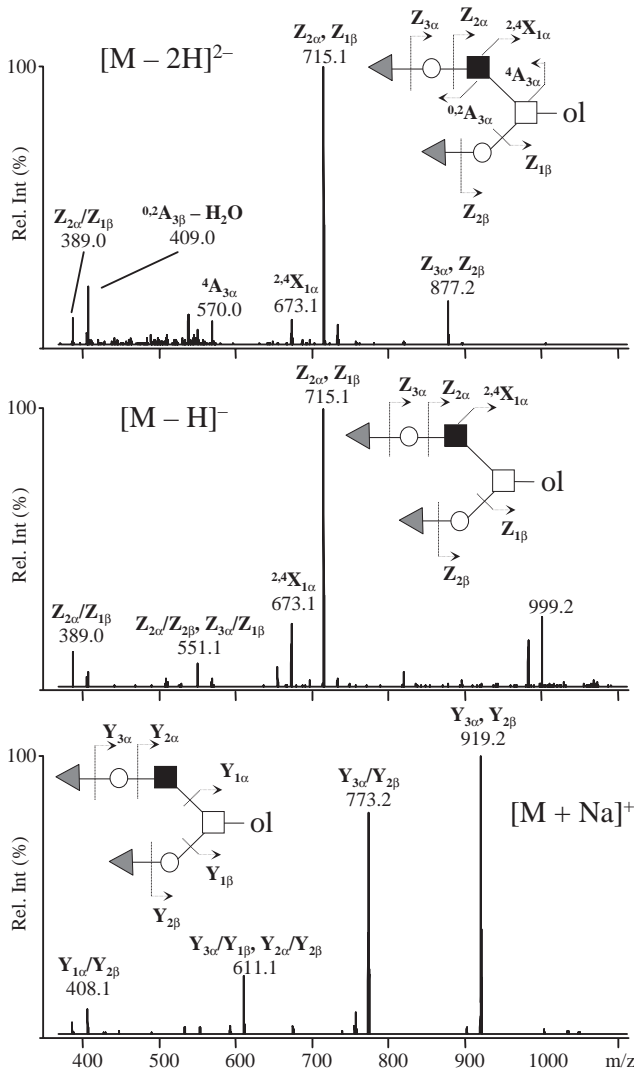


Figure 12.1 Glycan fragmentation nomenclature as proposed by Domon and Costello [5].

Table 12.1 Parameters that influence oligosaccharide fragmentation pattern.

Parameter	Example
Mode	Positive (+) and negative (-) modes
Charge state	$[M + H]^+$, $[M + 2H]^{2+}$, $[M + 3H]^{3+}$
Adducts	$[M + H]^+$, $[M + 2Na]^{2+}$, $[M - 2H]^-$, $[M + Cl]^-$
Derivatization	Permethylation, reduction, reductive amination of reducing end
Energy transfer	Collision-induced/associated dissociation (CID/CAD), electron capture dissociation (ECD), post-source decay fragmentation
Collision Energy	High-energy collisions (electronic excitation >1 keV), low-energy collisions (vibrational dissociation <0.5 keV)

**Figure 12.2** MS/MS fragmentation of the Fuc α 1-2Gal β 1-4GlcNAc β 1-6(Fuc α 1-2Gal β 1-3)GalNAcol oligosaccharide in different ion modes, charge states, and adducts.

to asparagine residues (*N*-linked) or to serine/threonine (*O*-linked). This kind of research should be considered as glycoproteomics, and is a subdivision of glycomic research which includes all the glycoconjugates. Glycoproteomic analysis of released oligosaccharides and glycopeptides provides two different types of information. One is the global identification of oligosaccharides that are differentially synthesized between samples (i.e. differential display), and the other is the characterization of the site glycosylation of a protein. This division of analysis can be justified to some extent as the two types reflect different biological cellular processes: the site determination measures the initiation of glycosylation (often co-translational) whereas the oligosaccharide structure results from the post-translational elongation mechanisms of glycosylation in the endoplasmic reticulum and Golgi apparatus.

Mass spectrometric glycoproteomic research so far has mainly been performed on either released oligosaccharides or glycopeptides. The more traditional glycoproteomic approach mimics the approach of proteomic research using gel electrophoresis to isolate glycoproteins followed by release of oligosaccharides or generation of glycopeptides from individual protein spots. This approach can be applied both to the total mixture of glycoproteins and proteins in a proteome, and also after enrichment of glycoproteins using, for instance, lectin affinity isolation [7]. Another approach to glycomic analysis is the analysis of all the released oligosaccharides from a pool of proteins. This global analysis can be used as a first screening method, as it reflects the synthetic pathway of glycosylation in a cell. Hence, by comparing the global glyco-profiles, disease-related glycosylation differences can be identified as the products of altered glycosylation machinery. Recently, alternative global glycomic strategies have been developed for the identification of glycopeptides, where the site of glycosylation is determined by the isolation of the glycopeptide complement of the cell [8, 9]. Both of those approaches generate complementary data on glycoproteome research. Sophisticated techniques such as stable isotope labeling of glycoconjugates for quantitation [10, 11] and 2D graphics to display and quantify oligosaccharide mass spectrometric data from MS and LC-MS are starting to emerge in glycomic applications [12].

We will deal separately with the two main approaches to the analysis of the glycosylation of proteins by mass spectrometry: first, the analysis of oligosaccharides released from glycoproteins, and second, the analytical procedures used to determine the site glycosylation of peptides.

12.3 Analysis of Released Oligosaccharides

Oligosaccharides can be released from glycoproteins either by using chemical methods such as hydrazinolysis (*N*-linked) and β -elimination (*O*-linked), or enzymatic methods using PNGase F or A (*N*-linked). There are currently no generic enzymatic methods to release all *O*-linked oligosaccharides, but there is an enzyme, *O*-glycanase, that is frequently used in combination with sialidase to release simple *O*-linked oligosaccharides such as NeuAc α 2-3Gal β 1-3(NeuAc α 2-6)GalNAc, NeuAc α 2-3Gal β 1-3GalNAc, and Gal β 1-3GalNAc commonly found on mammalian hemic (in blood) glycoproteins. By releasing oligosaccharides from the protein backbone, the downstream separation and analysis are dependent only on the oligosaccharide properties. For global analysis of glycosylation this is advantageous, since the analysis is confined to a homogeneous type of molecule spread within a limited mass range. As described in Chapter 11, the release of glycans from the peptide also allows analysis by HPLC techniques. There is also a biological argument

for confining the analysis to the sugar molecules which are secreted and displayed on the surface of cells and which present as the first point of contact with mammalian cells.

Dedicated high-resolution HPLC for oligosaccharide separation such as graphitized carbon [12–14] or normal-phase amine or amide chromatography [15, 16] can be used on-line or off-line before mass spectrometric analysis in order to resolve oligosaccharides isomerically and/or by molecular weight. A more detailed description of HPLC techniques is given in Chapter 11. Compositional interpretation of full-scan spectra is straightforward from the molecular ion masses and the data obtained by fragmentation of oligosaccharides are not complicated by the fragmentation of protein or peptide. Although losing the information on the attachment site of the oligosaccharide, there is more to gain in the identification of monosaccharide sequences, linkage position and configuration by analyzing the oligosaccharides separately from the protein/peptide. Bioinformatically, the structural information on the oligosaccharide can then be directly correlated with the action of specific glycosyltransferases, and thus information can be obtained on the glycosylation machinery in the endoplasmatic reticulum and the Golgi apparatus. It is essential to know this fine detail of oligosaccharide structure for studies of glycomic interactions with ligand molecules such as lectins, proteins, and lipids.

The glycosylation process should be considered as less exact than the translational (protein) and transcriptional (RNA) processes. Rather than producing a single oligosaccharide structure on a defined glycosylation site, there is a tendency to have a spectrum of closely related structures on a single site. These oligosaccharides are defined by the glycosylation machinery or, more precisely, the repertoire and distribution of glycosyltransferases, glycosidases, and monosaccharide donors. Hence oligosaccharides from a specific tissue are often made up of common core structures with various extensions and terminal epitopes. This is a feature that distinguishes glycomic mass spectrometry from proteomic proteolytic digest mass spectrometry. In peptide mass fingerprinting, there is a seemingly random distribution of peptides (Figure 12.3a), and it is not until the protein has been identified that the relation between the peptides is revealed. Mass spectrometry of released oligosaccharides, however, will usually produce spectra where the low-mass structures are building blocks

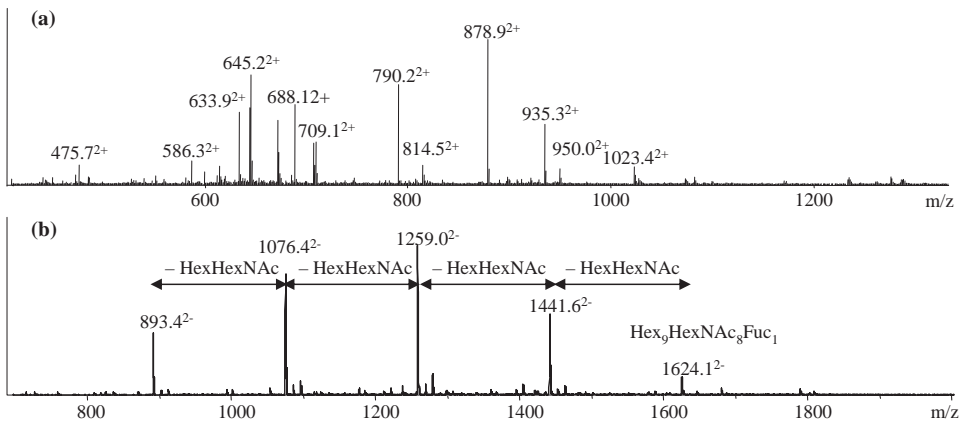


Figure 12.3 ESI-MS of peptides generated by tryptic digestion of a protein (a) and *N*-linked oligosaccharides released by PNGase F (b) illustrating the relatedness of oligosaccharides on a glycoprotein, whereas peptide masses appear to be random.

for the higher-mass, more complex structures (Figure 12.3b), creating envelopes of related structures and glycosylation “themes” reflected in the mass spectra. This feature of relatedness is something that can be applied bioinformatically to the interpretation of the data.

12.3.1 Released Underivatized Oligosaccharides

The general rule for maximum sensitivity of analysis is the less manipulation of the sample the better. The aim is to ensure minimal sample loss, without artifacts due to sample preparation. One of the key developments for glycoproteomics is the use of microscale clean-up methods such as cation exchange, C₁₈, and graphitized carbon [12, 17–19]. This has been used for glycoproteomic work for both matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) mass spectrometry. The most widely used method for releasing free *N*-linked oligosaccharides from proteins is enzymatic PNGase F release. *O*-linked oligosaccharides are usually released by reductive β -elimination, which results in the reduction of the terminal sugar, (Figure 12.4a) while at the same time preventing “peeling” degradation [20]. The alkaline conditions used do not preserve labile groups such as *O*-acetyls that may have been attached to the *O*-linked oligosaccharides. Strictly this process should be regarded as a derivatization which simplifies the separation of structural isomers of oligosaccharides by reducing the complexity of having α - and β -configurations possible at the reducing terminus. As such, reduction of released *N*-linked oligosaccharides is also recommended prior to chromatographic separations [12].

There is a significant amount of data on the mass spectrometry of underivatized and reduced oligosaccharides. Reports describing high-throughput analysis of oligosaccharides from recombinant glycoprotein production and extraction of released oligosaccharides from sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)-separated proteins [18, 21, 22] showed that miniaturized sample preparation methods could be applied to release oligosaccharides for identification with MALDI-MS and ESI-MS.

12.3.2 Permethylation/Peracetylation

Permethylation and peracetylation (Figure 12.4c and d) are methods historically used to render sugar derivatives more volatile and allow small oligosaccharides to vaporize in vacuum, a property that is necessary for electron impact MS. With the introduction of ESI and MALDI, volatility of samples/ions is no longer necessary for transfer into the gas phase. However, permethylation remains popular since the derivatives exhibit some useful characteristics in MS and MS/MS modes, where capping of the hydroxyl groups on carbohydrates by methyl groups allows internal fragments to be distinguished from fragments of a single cleavage event [1]. Without derivatization, misinterpretation of the fragmentation spectrum can occur since an internal fragmentation can have the same mass as a single cleavage sequence ion. This particular feature of permethylation, and to some extent peracetylation, make it attractive for non-ambiguous automated data interpretation (described later). Permethylation also allows the detection of sialylated and neutral oligosaccharides with the same mass spectrometric approach (e.g. positive ion MALDI-MS), since the negatively charged sialic acid residues are converted into permethylated methyl esters (see, for example, Kui *et al.* [23] and Manna *et al.* [24]). The major drawback of this type of derivatization, apart from the losses from the additional chemistry and clean-up, is that the

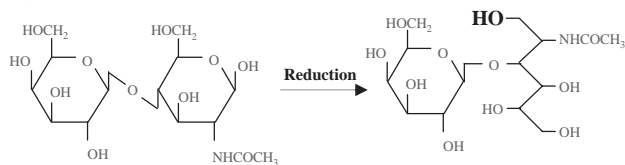
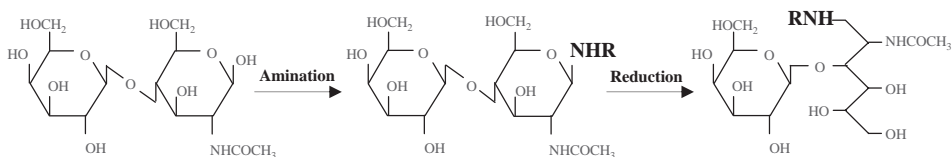
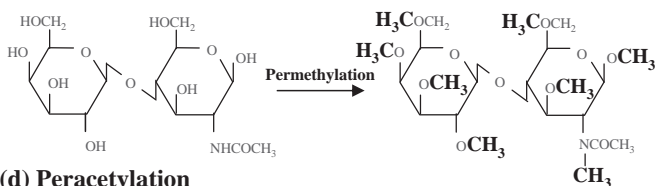
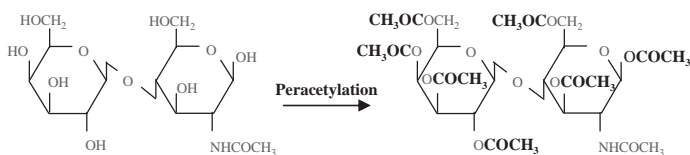
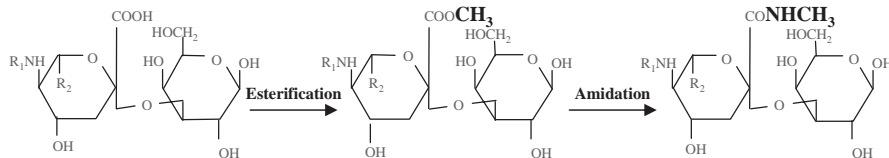
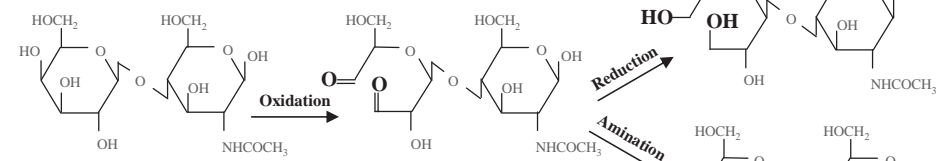
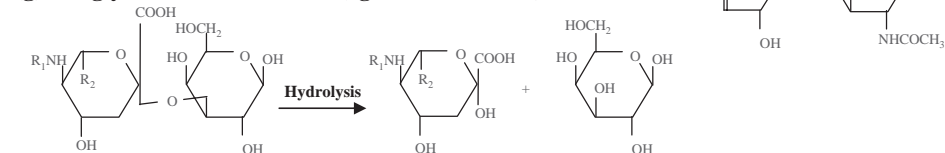
(a) Reduction**(b) Reducing end derivatization (reductive amination)****(c) Permethylation****(d) Peracetylation****(e) Neutralization of sialic acids (carboxylic acids)****(f) Periodate oxidation/reduction****(g) Exoglycosidase treatment (eg neuraminidase)**

Figure 12.4 Common derivatization and chemical reactions of oligosaccharides applied in mass spectrometric glycomics.

standard conditions for methylation are not compatible with the analysis of labile groups such as sulfate and *O*-acetyl groups.

12.3.3 Reducing End Derivatization

Reducing end derivatization provides an efficient way of introducing functional groups into oligosaccharides. The most commonly used derivatization technique is reductive amination (Figure 12.4b). A wide variety of reducing end derivatives have been used for the analysis of oligosaccharides (Table 12.2), but the most frequently used are low molecular weight amines that do not significantly change the mass range in which the oligosaccharides can be detected by mass spectrometry. The added functional group can include features such as a chromophore or fluorophore, making the oligosaccharide detectable using UV or fluorescence detection [25]. This also has the advantage of permitting sensitive detection on HPLC. Positively or negatively charged groups can be introduced into the oligosaccharide in order to increase the ionization ability and manipulate the fragmentation pattern in the positive or negative ion mode, respectively or to permit electrophoretic separation. Stable isotope labeling of derivatives can also be performed, in order to quantitate the up- or down-regulation of oligosaccharides in different samples [10]. Attaching a hydrophobic chromophore or fluorophore to the oligosaccharide allows the use of reversed-phase chromatography prior to MS detection. The chemistry of reductive amination can be difficult to drive to completion so that incomplete derivatization of the oligosaccharides can occur.

12.3.4 Periodate Oxidation

Oxidation of oligosaccharides with periodic acid destroys the oligosaccharide structure but generates reactive aldehydes which can be derivatized to gain additional structural information. Periodate treatment (Figure 12.4f) relies on the presence of *cis*-diols, such as those present on, for example, terminal galactose, fucose, and sialic acid. This reaction is commonly used for general color or fluorescent staining of glycoproteins on gels and tissues, after reaction of the aldehyde with a Schiff base (e.g. periodic acid–Schiff base (PAS), Emerald Green). Although periodic acid oxidation together with reduction and acidic hydrolysis (Smith degradation) used to be popular for the sequence analysis of oligosaccharides, it has not been widely used for detailed glycoprotein glycan structural determination, probably because of the amount of material required. However, the method of periodate oxidation and reaction of the resultant aldehydes with biotinylated hydrazide has recently been described for the global identification of glycosylation sites in glycopeptides (described later).

There are other techniques that can also be applied to both released oligosaccharides and intact glycopeptides to assist in the MS analysis of the oligosaccharide structure.

12.3.5 Terminal Sialic Acid Esterification

One of the major problems with positive ion mode mass spectrometry of oligosaccharides is the presence of negatively charged residues such as sialic acid and sulfate. Sialic acid has been shown to be easily lost in MALDI time-of-flight (TOF) analysis and the desialylated oligosaccharides are detected as metastable ions in the reflectron mode because of post-source decay (PSD) fragmentation. In order to stabilize this and to neutralize the negative charge for positive MALDI, sialic acids can be converted into their methyl esters

Table 12.2 Reducing end derivatives.

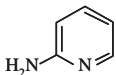
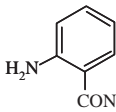
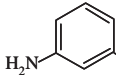
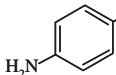
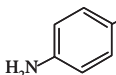
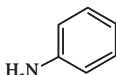
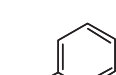
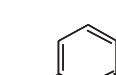
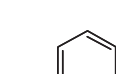
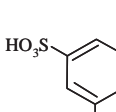
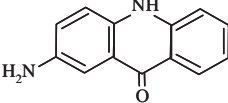
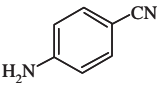
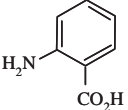
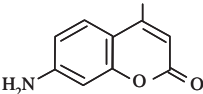
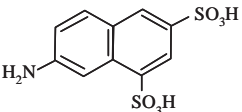
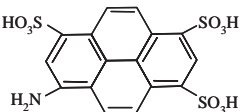
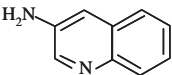
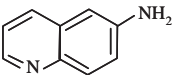
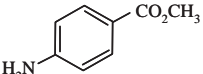
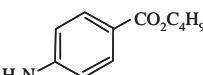
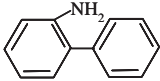
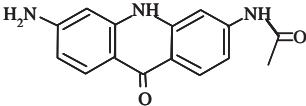
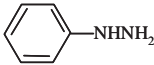
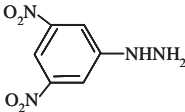
Reagent	Abbreviation	Formula	Monoisotopic mass increment
None			18.0105646
Sodium borohydride			20.0262146
2-Aminopyridine	2-AP		96.068748
2-Aminobenzamide	2-ABAD		138.0793126
3-Aminobenzamide	3-ABAD		138.0793126
Ethyl 4-aminobenzoate	ABEE		167.0946282
Hexyl 4-aminobenzoate	ABHE		223.1572282
4-Amino-N-(2-diethylaminoethyl) benzamide	DEAEAB		237.1841116
Trifluoroacetamidoaniline	TFAN		206.0666973
2-Amino-6-cyanoethylpyridine	ACP		149.095297
2-Amino-6-amidobiotinylpyridine	BAP		337.157246
8-Aminonaphthalene-1,3,6-trisulfonic acid	ANTS		384.9595958

Table 12.2 (Continued)

Reagent	Abbreviation	Formula	Monoisotopic mass increment
2-Aminoacridone	AMAC		212.0949626
4-Aminobenzonitrile	4-ABN		120.068748
2-Aminobenzoic acid	ABA		139.0633282
7-Aminoethyl-4-methylcoumarin	AMC		177.0789782
7-Aminonaphthalene-1,3-disulfonic acid	ANDS		305.0027802
8-Aminopyrene-1,3,6-trisulfonic acid	APTS		458.9752458
3-Aminoquinoline	3-AQ		146.084398
6-Aminoquinoline	6-AQ		146.084398
Methyl 4-aminobenzoate	ABME		153.0789782
n-Butyl 4-aminobenzoate	ABBE		195.1259282

(Continued)

Table 12.2 (Continued)

Reagent	Abbreviation	Formula	Monoisotopic mass increment
2-Aminobiphenyl	2-ABP		171.104799
3-Acetylamino-6-aminoacridone	AAMC		269.1164262
Phenylhydrazine	PHN		108.068748
2,4-Dinitrophenylhydrazine	DNPH		198.0389044

[26, 27] (Figure 12.4e). This method has been widely used for the MALDI-MS analysis of gel-separated glycoproteins, and recently has been shown to stabilize sialic acid in the fragmentation of negatively charged *N*-linked oligosaccharides to provide more informative MS/MS spectra [28].

12.3.6 Enzymatic Sequencing

Exoglycosidase treatment (Figure 12.4g) is another approach to determine oligosaccharide sequence and linkage, when fragmentation spectra are not sufficient for confirmation of a particular sequence or linkage. Most of these exoglycosidases have been isolated from bacteria or legumes. There are a variety of sialidases, galactosidases, β -acetylhexosaminidases, and mannosidases with different linkage substrate specificity that are used for the structural elucidation of oligosaccharides [29]. A list of suitable sequencing enzymes is given in Chapter 11. For glycoproteomic mass spectrometric work, the enzymes that have attracted the most attention are the sialidases, since they can be used specifically to remove sialic acid, thus simplifying the detection and sequencing of the resultant neutral glycoconjugates by mass spectrometry.

12.4 Mass Spectrometry of Released Oligosaccharides

12.4.1 MALDI-MS

One of the key technical developments for glycoproteomics was the use of PNGase F to release oligosaccharides from gel-separated glycoproteins [19, 30]. This combined a

high-resolution multiprotein separation technique with the high-quality data for oligosaccharides generated by mass spectrometry. This approach has been used for both MALDI-MS and ESI-MS analysis of 2D gel-separated glycoproteins [22, 31, 32]. The sensitivity of MALDI-MS is limited to the high-picomole analysis of oligosaccharides, and fragmentation information has been difficult to obtain using PSD in the traditional MALDI-TOF instruments. The later introduction of hybrid instruments such as Q-TOF, TOF-TOF and QIT-TOF provided more efficient fragmentation of oligosaccharide precursor ions in MALDI-MS [33–36]. Also, novel fragmentation techniques such as infrared multiphoton dissociation (IRMPD), developed for fragmentation in Fourier transform (FT) MS instrumentation shows that high-resolution informative spectra for the structural characterization of sugars can be generated using FT-MS [37]. Positive ion MALDI is usually used for the analysis of oligosaccharides, and it works well for neutral and low or desialylated oligosaccharides both for molecular mass determination and for fragmentation analysis. However, MALDI produces mainly singly charged ions, and the low-energy fragmentation results in the loss of labile sialic acid and sulfate as the major fragment ions, with limited additional sequence information. With correct derivatization, highly sialylated released oligosaccharides can be detected using positive MALDI-MS after permethylation, methyl esterification, or reducing end derivatization.

Negative ion mode MALDI-MS has for a long time provided a challenge for glycomic analysis. The theory that sialylated oligosaccharides will ionize better in the negative mode is attractive but is counteracted by a lack of suitable matrices. The use of trihydroxyacetophenone and other matrices has recently been shown to be useful for this type of application [32]. Labile groups such as sulfate and sialic acid are easily lost in the ionization or in the analyzer, rendering spectra containing both pseudomolecular ions, fragment ions and/or metastable ions, but novel approaches for ionization in MALDI have showed that this problem could be overcome. MS/MS fragmentation of negative precursor ions by MALDI has rarely been reported, but negative ion MALDI has been performed on neutral oligosaccharides in complexes with negative adducts such as Cl^- , where fragmentation was shown to be informative [38]. Oligosaccharide mass distribution has also been obtained on the negatively charged fluorescent derivatives of oligosaccharides separated by gel electrophoresis, namely the FACE-technique [39]. However, negative ion mode MALDI-MS structural characterization using MS/MS is still to prove versatile enough for true glycoproteomic/glycomic applications.

12.4.2 *ESI-MS*

ESI-MS has shown to be applicable in both negative and positive ion modes for the characterization of oligosaccharides. In combination with LC (e.g. graphitized carbon), it has been used for the analysis of underivatized oligosaccharides, where both modes provide excellent analysis of mixtures containing both neutral and acidic oligosaccharides [12, 14, 15, 40]. This LC-ESI-MS/MS approach allows the separation of the many structural isomers and makes it attractive for the high-throughput structural analysis of oligosaccharides needed for glycomics. In particular, ion-trap ESI-MS allows for the further collision of the fragment ions to gain fine structural information. The benefit of ESI compared with MALDI is the generation of multiply charged parent ions, which usually provide more informative fragmentation spectra. Nanospray also provides a sensitive, more stable, parent ion generation, while MALDI requires that the laser beam is constantly finding “hot spots”

in the crystallized matrix to generate enough of the parent ion. MALDI, however, is easier to automate and accumulation of data is faster, although less information rich. However, the recent development of microfluidic devices permits automated high-throughput “chip” ESI-MS of oligosaccharides and other glycoconjugates [41]. A summary of the properties that are conferred by the different methods of glycosylation analysis with MALDI-MS and ESI-MS is given in Table 12.3.

Table 12.3 Characteristics of oligosaccharide MS analysis.

Mode	Derivative	MS	MS ⁿ	Other
+	Free or alditols	Pseudomolecular ions with H ⁺ , Na ⁺ , or other metal ions. Difficult to detect highly acidic sialylated and sulfated oligosaccharides	Moderate	Metastable ions detected in MALDI reflectron mode for sialylated oligosaccharides. Internal fragment masses isobaric with glycosidic sequence ions
+	Permethylated/peracetylated	Pseudomolecular ions with Na ⁺ or other metal ions. Permethylation of sulfated and oligosaccharides is difficult and information of <i>O</i> -acetylation is lost. Mixtures containing both neutral and sialylated oligosaccharides could be analyzed together. Negative charged sialic acid is converted into methyl esters or lactones	Excellent	Internal fragment masses not isobaric with glycosidic sequence ions
+	Reducing end derivatives	Derivatives of quaternary amines and amines with high pK _a gives H ⁺ adducts, other gives preferentially metal adducts	Moderate to excellent depending on derivative	Positively charged derivatives giving preferentially X-, Y-, and Z-type ions. Rearrangements reported in ESI. Internal fragment masses isobaric with glycosidic sequence ions
-	Free or alditols	[M - xH] ^{x-} ions of both neutral and acidic oligosaccharides, neutral also Cl ⁻ adducts. Mixtures containing both neutral and acidic oligosaccharides could be analyzed together by -ESI.	Moderate	Rearrangements reported. Informative fragmentation difficult for highly sulfated and sialylated structures in MALDI. Internal fragment masses isobaric with glycosidic sequence ions
-	Permethylated/peracetylated	Rarely reported	Rarely reported	
-	Reducing end derivatives	Negatively charged derivatives give [M - xH] ^{x-} ions	Rarely reported	

12.5 Analysis of Glycopeptides

There are several time-consuming laboratory procedures available to provide information on the oligosaccharide(s) at a site of glycosylation on a protein, although there is currently little bioinformatics available to assist the interpretation. To date, analysis of intact glycopeptides has rarely been performed on the global screening scale required by glycomics applications.

12.5.1 *Isolation of Glycopeptides*

In order to solve the problem of suppression of glycopeptide mass spectrometric signals in the presence of non-glycosylated peptides, there have been improvements in the methodology of enrichment of glycopeptides using different affinity methods. The use of specific proteases such as trypsin to derive the glycopeptides has both benefits and problems. On the positive side, the peptide mass can easily be predicted if the protein is known, and the same sample can be used both to identify the protein and to identify the glycosylation. The negative side is that there is limited control of the size of the glycopeptide, as masses >2000 Da (sum of tryptic peptide and oligosaccharide) is common in a tryptic digest. The separation and enrichment of tryptic glycopeptides by chromatography are also complicated by the different properties of peptide and size of both the peptide and the oligosaccharide. More generic methods for the isolation of glycopeptides using less specific proteases such as proteinase K and pronase have been suggested [42, 43]. These enzymes will digest the peptide into just a few amino acids around the glycosylation site. With the combined knowledge of peptide sequence and the characterized released oligosaccharides present on the glycoprotein, the attachment site and identity of the oligosaccharide on the site can be deduced. The advantage of this approach is that these small glycopeptides can be isolated and chromatographed with the standard methods used for oligosaccharide analysis, since the peptide part will play only a minor role in the chromatographic properties of the glycopeptide.

Alternatively, a method for generic glycopeptide isolation using the hydrophilic interaction between oligosaccharides and a solid support such as cellulose and Sepharose has been demonstrated to enrich most of the larger glycopeptides of fibronectin [44], which were then separated by reversed-phase chromatography.

12.5.2 *Analysis of Intact Glycopeptides*

The ultimate goal when analyzing glycopeptides is to obtain information about both the oligosaccharide structural heterogeneity and the peptide sequence in a single analysis step. In a glycoproteomic context, this has to be performed on a global level on multiple glycoproteins from a single tissue. The first practical aspect of glycopeptide analysis is that they are generally difficult to detect using the standard techniques for peptide analysis. If the problems with suppression of the ionization of glycopeptides in favor of other peptides in a tryptic mixture (e.g. as generally seen in MALDI-MS) could be overcome, there is still the problem with the determination of the glycosylation heterogeneity on a single glycosylation site. Hence individual glycopeptides are usually seen in smaller quantities than non-glycosylated peptides in a tryptic mixture of even a single glycoprotein. Using 2D SDS-PAGE, individual glycoprotein isoforms can be separated into multiple spots due to charge (e.g. sialic acid and sulfate) and size. They will often be distinguished as trains

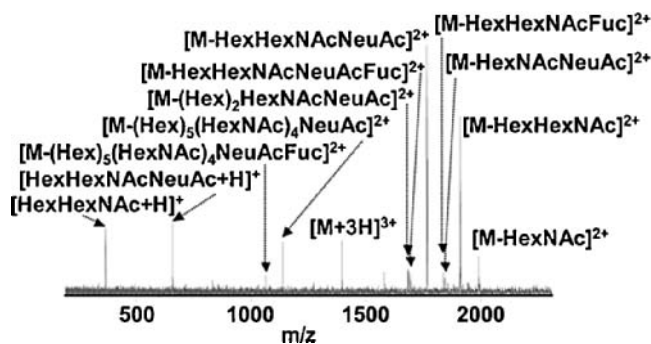


Figure 12.5 Infrared multiphoton dissociation FT-MS product ion spectrum from a triply protonated β -trace glycoprotein glycopeptide from cerebrospinal fluid. Reprinted with permission from Håkansson *et al.* [46]. Copyright 2003 American Chemical Society.

of spots with multiple pI and increasing acidity (see, for example, Wilson *et al.* [22]) This separation of glycoforms provides a more homogeneous tryptic glycopeptide digest. There are only few reports of glycoproteomic analysis of glycopeptides using 2D SDS-PAGE separation. Using IRMPD, it was shown that FT-MS could be used for oligosaccharide sequencing of sialylated glycopeptides from cerebrospinal fluid [45] (Figure 12.5). LC-MS in combination with different fragmentation techniques such as IRMPD and electron capture dissociation (ECD) together with the high-resolution of FT-MS are currently the most accurate approach for combining both glyco and protein sequencing on one sample [46]. ECD has even been shown to be used for analysis of highly glycosylated *O*-linked peptides, containing multiple sites of glycosylation [47]. The strength of ECD is that it preferentially cleaves the protein backbone without affecting the oligosaccharide, whereas IRMPD fragments the oligosaccharide before the peptide backbone. Other methods of mass spectrometry could be used for global glycoproteomic analysis of glycopeptides, but reporting of the successful performance of this is scarce.

MALDI-QIT has been used to analyze an enriched glycopeptide fraction from serum using the sequential fragmentation of the ion trap to sequence both the sugar (MS^2) and the peptide (MS^3) [48]. There are several other reports describing glycopeptide analysis from purified proteins displaying an enormous amount of information obtained usually by manual interpretation of the mass spectrometric data. To do this kind of analysis on a global scale is not only time consuming but also the data sets are difficult to collate and present in order to interpret, and require that the technical laboratory procedures for glycopeptide enrichment be improved.

12.5.3 Analysis of Deglycosylated Glycopeptides

The approach for the isolation of glycoproteins using high-resolution 2D gel electrophoresis followed by detailed characterization of individual glycoproteins is labor intensive, and may not be the first choice as a quick screening method. Hence there have been attempts to decouple the two important questions in glycoproteomics: what kind of glycosylation does this tissue have, and where and how is it attached? The first question is the one that is able to be answered by a global automated analysis of the released oligosaccharides of the glycoproteome. The second question does not necessarily have to be answered by analyzing intact glycopeptides, but in its simplest form can be determined by analyzing peptides before and

after deglycosylation. Since glycopeptides are rarely detected in MALDI-MS in tryptic mixtures, using the standard technique for PMF, the differential display of peptides before and after PNGase F treatment can provide the identity of many of the sites of glycosylation [22]. The previously detected glycosylated peptides have the attaching asparagine converted into aspartic acid after PNGase F digestion, which results in a “new” peptide appearing of +1 Da mass. A similar approach using a chemical release instead has been proposed for *O*-linked sites, where clustered glycosylation sites hinder enzymatic protease digestion. Converting glycosylated serines and threonines into amines during the progression of the chemical β -elimination release has shown to be an efficient way to deglycosylate and at the same time label the sites of glycosylation for subsequent mass spectrometric identification [49].

More sophisticated methods have recently been developed in order to identify glycosylation sites in a mixture of glycoproteins. They rely on the ability to pull out specifically glycopeptides from a total mixture of tryptically digested proteins and glycoproteins. This is done using a series of carbohydrate epitope-specific lectins (non-covalent) [9] or a covalent attachment of the oligosaccharide portion of the glycopeptides after periodate oxidation (generating aldehydes and solid-phase amination described in Figure 12.4f) [8], followed by PNGase F treatment to release the deglycosylated glycopeptides. The isolated mixture of deglycosylated peptides can then be sequenced using LC-MS². The process of quantifying differences between the glycosylation samples can also be automated in both approaches using stable isotopic labeling of the peptide. These approaches do not give information on the structure of the attached oligosaccharide(s) but, in combination with the global analysis of released oligosaccharides, are a powerful approach to answering the two questions, what and where are the oligosaccharides in the glycoproteome?

12.6 Parallel Approach to Glycomic Mass Spectrometric Analysis

We have covered so far the advantages and disadvantages of the different ways in which the analysis of glycoproteins can be carried out by mass spectrometry.

In summary, we recommend that for the large-scale analysis needed for glycomic research, the glycoprotein sample be separated into two aliquots, one of which is analyzed for its oligosaccharide structural heterogeneity by the release, separation, and mass spectrometric fragmentation of its glycans, whereas the other is digested proteolytically and analyzed in respect of the glycosylation site. This approach can be used both on isolated glycoproteins and at a global level on an unseparated mixture of proteins. The first part of this chapter described the various ways in which this can be achieved. We believe that this approach is the most readily accessible for the majority of research laboratories coming into glycomics research from proteomics or genomics.

A multinational initiative has been launched by Professor Naoyuki Taniguchi for the investigation of human disease-related glycomics [Human Disease Glycomics/Proteome Initiative (HGPI); www.hgpi.jp], where different ways of screening glycoprotein glycosylation using mass spectrometry (MALDI and ESI, and free oligosaccharides and glycopeptides) is assessed for finding disease-related glycomes [50].

12.7 Interpretation of the Data

There is much scope for bioinformatics in this parallel approach, and even more scope to make the software tools robust and user friendly for scientists coming into the field of

glycomics for the first time. There has been a lot of recent activity in the glycoinformatics space to try to facilitate the interpretation of mass spectrometric data on glycans. The ideal would be to be able to obtain the complete glycan structural information solely by mass spectrometric techniques – but this requires that the composition, sequence, branching linkage, anomericity, and number of isomeric oligosaccharide structures be determined all on one sample, in addition to the attachment site and occupancy of the peptide. The question that needs to be asked here is whether all this information is always needed to answer the biological problem. It could be a “catch 22” question as “we don’t know what we need until we know what we don’t need,” but our experience is that if we can start by obtaining at least some of this information quickly and easily the extent of what we need to know will become easier to answer.

12.7.1 Compositional Analysis Tools

The first requirement, and often unfortunately the only one commonly used to suggest sugar structures, is to assign possible saccharide composition to a specific ion mass. This experimentally obtained mass can be either a glycopeptide (tryptic or otherwise) or a released and/or derivatized (e.g. Table 12.2 and Figure 12.4) oligosaccharide. In addition to the possible monosaccharide constituents (HexNAc, Hex, dHex, HexA, KDN, Pent, NeuAc, NeuGc), sulfate, phosphate, acetyl, and methyl modifications of the sugars need to be considered as do possible mass adducts (e.g. Na^+ , K^+ , TFA^-). A further consideration is that many of the monosaccharides are isobaric, so parent ion mass spectrometry cannot differentiate glucose from galactose or mannose, GlcNAc from GalNAc, xylose from arabinose, and so on.

Many laboratories cope with this calculation by setting up spreadsheets to calculate the possible compositions corresponding to the masses obtained. A more sophisticated calculator is freely available on the ExpASy web site (GlycoMod: <http://ca.expasy.org/tools/glycomod/>). As the oligosaccharide parent ion mass increases, however, the number of possible monosaccharide compositions increases exponentially, especially if other constituents such as phosphorylation, sulfation, acetylation, and methylation are considered (Figure 12.6). Fortunately for the analysts, nature has limited these possibilities by the number of specificities allowed for by the available glycosyltransferases. It is possible to refine the number of corresponding compositions matching a mass by using biological rules. For example, GlycoMod [51] ranks the possible compositions based on empirical monosaccharide rules of *N*- and *O*-linked structures, and the commercially available GlycoComp uses a biological ranking index calculated from the monosaccharide compositions previously found and collected in GlycoSuiteDB [52] (<http://glycosuitedb.expasy.org/glycosuite/glycodb>). The need for these constraints is evidenced by the actual number of “real” compositions which have been reported to correspond in these mass ranges compared to the possible permutations (Figure 12.6). Furthermore, the number of compositions which have been reported (Figure 12.7) attached to glycoproteins across species and tissue sources falls predominantly in the size range 3–13 monosaccharide residues.

The danger in simply assigning even a biologically relevant composition to an experimentally derived mass of either an oligosaccharide or a glycopeptide is the number of possible structures, comprising different sequences, monosaccharide linkages and anomericity, which can theoretically correspond to this composition, particularly as the mass increases. Table 12.4 shows that an oligosaccharide with a mass of 2368 ± 0.5 Da could correspond to

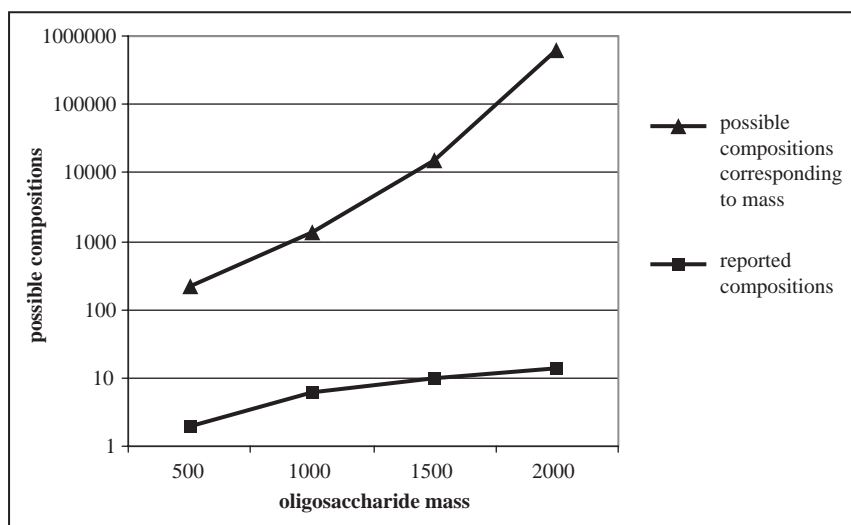


Figure 12.6 Number of theoretical and reported oligosaccharide(s) (from GlycoSuiteDB) compositions corresponding to mass ± 100 Da.

10 possible monosaccharide compositions (considering only HexNAc, Hex, dHex, NeuAc, and NeuGc as constituents). Increasing the mass accuracy will not substantially decrease these possibilities, but the fact that only two of these compositions in different species have been reported to occur in Nature to date does increase the probability of assigning the correct composition. The same monosaccharide composition has been reported to occur as several different linkage and sequence isomers depending on the biological source in which they have been found (Table 12.4).

Informatic tools querying databases of reported structures and their origins, such as GlycoSuite, Kegg GLYCAN (www.genome.jp/kegg/glycan/), and the database of the

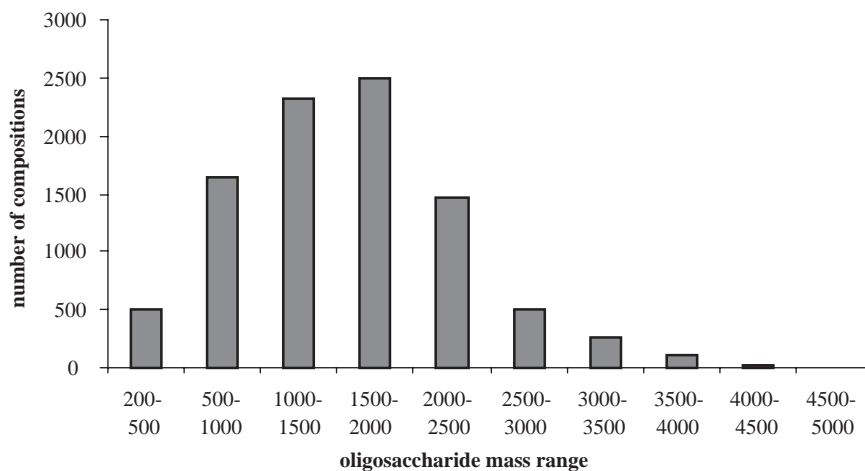







Figure 12.7 Number of reported compositions (GlycoSuiteDB) based on mass.

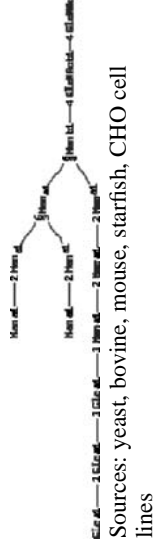
Table 12.4 Same mass, same composition, different structure.

Mass	Δ mass	Predicted composition	Reported structures
2368.841	-0.167	HexNAc ₄ Hex ₅ dHex ₁ NeuAc ₂ Complex	 <p>Sources: human, pig, virus, mouse</p>
			 <p>Sources: human, mouse, bat, CHO/BHK cells, virus, snake</p>
			 <p>Sources: human, pig, rat, chicken</p>
			 <p>Source: human multiple myeloma</p>
			 <p>Source: human</p>

2368.803

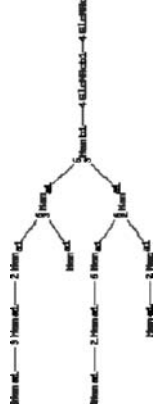
-0.205

HexNAc₂Hex₁₂
Untrimmed *N*-linked
precursors



Sources: yeast, bovine, mouse, starfish, CHO cell lines

Many high-mannose
variants



Sources: yeast, fungi

2368.841

-0.167

HexNAc₄Hex₄dHex₂NeuAc₁NeuGc₁

2368.841

-0.167

HexNAc₄Hex₃dHex₃NeuGc₂

2368.828

-0.180

Hex₁₀dHex₅

2368.866

-0.142

HexNAc₂Hex₃dHex₆NeuAc₂

2368.866

-0.142

HexNAc₂Hex₂dHex₇NeuAc₁NeuGc₁

2368.866

-0.142

HexNAc₂Hex₁dHex₈NeuGc₂

2368.891

-0.117

Hex₁dHex₁₁NeuAc₂

2368.891

-0.117

dHex₁₂NeuAc₁NeuGc₁

Consortium of Functional Glycomics (www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp), allow this type of information to be obtained and highlight the need to acquire more than just the mass of the parent ion of a glycan and its corresponding possible composition, in order to assign its structure.

12.7.2 Sequence Analysis Software Tools

To move from monosaccharide composition to two-dimensional oligosaccharide structure, it is essential to have bioinformatics tools which facilitate the interpretation of the MS/MS spectra of the *N*- and *O*-linked oligosaccharides. This is not a trivial task since the structure of an oligosaccharide, in contrast to that of a protein, involves not only the assignment of sequence of the monosaccharide residues but also the determination of branch points, linkages, and anomericity. Hence it is not sufficient to obtain a partial sequence (as it is in peptide analysis) to assign a structure reliably. Fortunately, Nature has simplified what would otherwise be a Herculean task, with almost infinite possibilities, by restricting the number of residues used, conserving the pentasaccharide core of the *N*-glycans, and limiting the specificity of enzymes available to modify the structures. Having said that, the number of enzymes involved in the biosynthesis, localization, and modification of oligosaccharide structures comprises almost 1% of the human genome [53], a remarkably large figure that would provide an enormous variety of possible oligosaccharide structures.

Since there is no template from which to read the sequence of an expressed oligosaccharide, different approaches are needed to interpret MS/MS fragmentation data for the purposes of structure assignment. In peptide sequencing, the fragmentation is across the peptide bond and a linear sequence is seen which can be matched to the sequences deduced from the relevant genome sequence. Automation of this analysis becomes possible when the genome of interest is available and the predicted peptide sequences can be deduced for all expressed proteins. The experimentally obtained fragmentation data are then automatically matched against all theoretical fragments generated from databases of protein sequences. The algorithms in Sequest, Mascot, and so on have enabled this to be implemented routinely in proteomic mass spectrometry laboratories.

This concept of matching MS/MS data from separated released glycans with a database of reported oligosaccharides structures has now been used by several groups (Table 12.5B). In summary (Figure 12.8), identification/characterization of glycans can be achieved by comparison of experimentally observed fragmentation mass fingerprints of a glycan to a constructed database of all theoretically derived glycofragment masses, after which the resulting matched list is ranked according to different criteria. Recently, this approach was validated [34] by showing that the increased cross-ring cleavages obtained by high-energy collision MS fragmentation increased the accuracy of the ranking of the matched structures by these tools.

Compared with peptide fragment mass sequencing, however, this approach to oligosaccharide structure determination is limited by the number of structures contained in these databases. Since there is no blueprint from which to read the possible glycan structures, the glycofragment mass sequencing will necessarily only predict sequences that have been previously seen (as in the case of GlycoSidIQ and GLYCOSCIENCE.de [54, 55]) or which have been synthesized (MSⁿ Spectral Library [70]). In order to match against/identify novel glycan structures, it would be necessary to construct a database of all fragments of a larger set of all theoretically possible glycans. This may be computationally feasible,

Table 12.5 Computational approaches to assignment of structure from MS data.

Name	Access	Principle	Scoring	Limitations
A. Structures deduced from compositional data based on biological constraint knowledge				
Cartoonist	Goldberg <i>et al.</i> [60]	MS data predict a composition which matches a library of biosynthetically plausible structures generated from a base set of previously assigned known structures MS data used to predict possible composition of oligosaccharides and known glycoproteins	Based on probability that the sequence is correct using synthetic pathway constraints	Infers structures of mammalian <i>N</i> -linked glycans generated by known biosynthetic pathways
GlycoMod	www.expasy.org/tools/glycomod/ Cooper <i>et al.</i> [52]		Ranking based on rules of biologically occurring monosaccharide ratios	Compositional calculation only, structures inferred
B. Mass fingerprinting approaches – MS² fragmentation spectra matched against a theoretical fragmentation of a database of glycan structures				
GlycoSidIQ	Joshi <i>et al.</i> [69]	Glycofragment mass fingerprinting by matching MS ² data to a theoretical fragmentation of a database of reported glycan structures (GlycoSuiteDB)	Uses glycosidic cleavage and fragment intensity matching criteria for ranking	Database created from literature references is limited by structures which have been published and curated
GlycoSearch-MS/GLYCOSCIENCES.de [54]	www.glycosciences.de/[54]	Mass fingerprinting by matching MS ² data to a theoretical fragmentation of a database of previously collected glycan structures (CarbBank)	Output does not rank the possible structure matches	Database not curated and content limited (CarbBank user-entry collection discontinued in 1998). Single MS/MS data entry
MS ⁿ spectral Library	Kameyama <i>et al.</i> [70]	Matching to MS ² fragmentation library of structures, commercially available and synthesized with cloned glycosyltransferases with known linkage specificity	Difference between library and experimental spectra by ion signal intensity profiles determines which mass to subject to MS ³ for differentiation of isomeric structures	Limited by availability of MS data from known and synthetic structures in library

(Continued)

Table 12.5 (Continued)

Name	Access	Principle	Scoring	Limitations
C. Catalogue Library – MS motif library derived from standard substructures and computationally reassembled (bottom-up approach) Catalogue Library	Tseng <i>et al.</i> [63]	Proposed use of combining MS ² structural motif library derived from NMR characterized oligosaccharides	Manual interpretation	Limited by number of fragment motifs in library (5 off)
GLYCH	Tang <i>et al.</i> [64]	Early-stage algorithm analyzing MS ² data for structure based on cross-ring cleavage ions	Developing an ion intensity based scoring function.	Absence of double cleavage ions in the algorithm results in similar scoring of related structures. Small number of low MW standards compared.
MS ⁿ FragLib	Zhang <i>et al.</i> [37]	Bottom–up approach of matching to fragmentation library derived from MS ⁿ of commercially available small methylated oligomers from any stage of disassembly and coupling these substructure spectra to compare with experimentally derived spectra	Spectral comparisons uses the NIST scoring approach based on the differences in intensities of the fragment ions.	Relies on the fragmentation of the component oligomers being the same when they are part of a larger oligosaccharide structure. Proof of concept on permethylated neutral structures
D. Saccharide topology – the parent ion mass constrains the compositions and the masses of possible connected branched residues are generated for matching to experimental fragments (top-down approach) STAT: Saccharide Topology Analysis Tool	Gaucher <i>et al.</i> [67]	Infers the glycan's possible compositions from the parent ion mass, generates a candidate set of all possible branching topologies and compares the ion masses of a MS ⁿ data tree	A ranked list of branching topologies is generated based on connecting substructure masses	Supports native glycans of less than 10 monosaccharides, requires manual intervention to resolve ambiguous ion compositions, and cannot sequence bisecting N-linked structures

StrOligo	Ethier <i>et al.</i> [58]	Compositional mass predicts possible structures based on known <i>N</i> -linked mammalian biosynthetic pathways, which are theoretically fragmented and matched to experimental data	Depends on the number of nodes in a relationship tree matching the structure proposed by the biosynthetic rules	Cannot differentiate isomeric structures. Dependent on biosynthetic rules of mammalian <i>N</i> -linked structures
OSCAR: Oligosaccharide Subtree Constraint Algorithm	Lapadula <i>et al.</i> [68] GlySpy http://glycome.unh.edu/tools/GlySpy/	Computation of glycan top-down topology. Given a glycan MS mass, the algorithm retrieves the matching compositions from a fragment composition database. An optimized representation of all possible glycan topologies fitting that composition is constructed based on theoretical MS ⁿ fragmentation pathways	<i>De novo</i> coupling of subtree disassembly of structures without using biosynthetic constraints or comparison against previously characterized oligosaccharides	Requires experimental data from sequential multi MS ⁿ fragmentations of sufficient material. Structural inference is derived from permethylated glycan fragmentation rules

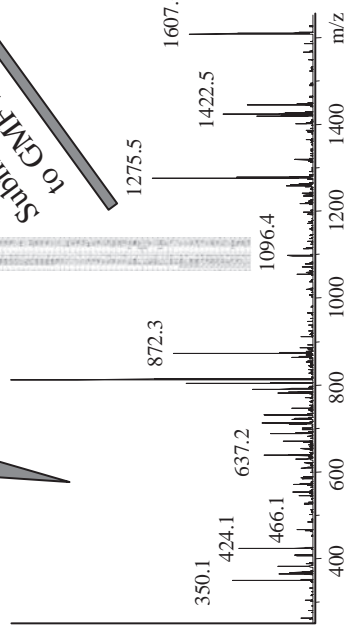
LC-MS

Single ion chromatogram of m/z 893.3



LC-MS/MS

Submission of peak list
to GMF search engine



GMF-output



Reporting

STRUCTURE	Sample A	ad
<chem>C1=CC=C(C=C1)O</chem>	%	%
<chem>C1=CC=C(C=C1)O</chem>	7.86%	2.25%
<chem>C1=CC=C(C=C1)O</chem>	1.00%	0.06%
<chem>C1=CC=C(C=C1)O</chem>	0.00%	0.00%
<chem>C1=CC=C(C=C1)O</chem>	19.31%	1.04%
<chem>C1=CC=C(C=C1)O</chem>	0.00%	0.00%
<chem>C1=CC=C(C=C1)O</chem>	1.57%	0.79%
<chem>C1=CC=C(C=C1)O</chem>	0.00%	0.00%
<chem>C1=CC=C(C=C1)O</chem>	16.65%	0.58%
<chem>C1=CC=C(C=C1)O</chem>	2.52%	0.21%
<chem>C1=CC=C(C=C1)O</chem>	0.00%	0.00%
<chem>C1=CC=C(C=C1)O</chem>	2.98%	0.31%
<chem>C1=CC=C(C=C1)O</chem>	2.47%	0.00%
<chem>C1=CC=C(C=C1)O</chem>	0.00%	0.00%
<chem>C1=CC=C(C=C1)O</chem>	2.41%	0.39%
<chem>C1=CC=C(C=C1)O</chem>	23.18%	0.59%
<chem>C1=CC=C(C=C1)O</chem>	0.49%	0.69%
<chem>C1=CC=C(C=C1)O</chem>	0.00%	0.00%
<chem>C1=CC=C(C=C1)O</chem>	2.78%	0.44%
<chem>C1=CC=C(C=C1)O</chem>	4.71%	0.98%
<chem>C1=CC=C(C=C1)O</chem>	10.42%	1.71%
<chem>C1=CC=C(C=C1)O</chem>	0.00%	0.00%

Figure 12.8 The principle of glycofragment mass fingerprinting (GMF) [69].

although practicably it would have massive hardware requirements. Even more demanding is the fact that the construction of these databases has focused mainly on the collection of glycoprotein-associated oligosaccharides; it remains to collate accurately the structures contained on other glycoconjugates such as proteoglycans, glycolipids, peptidoglycans, lipopolysaccharides, and polysaccharides.

The key to this methodology of glycofragment mass fingerprinting (GMF) [69], as in peptide mass fingerprinting, is the criteria established for ranking the possible glycan structures which correspond to the experimentally obtained MS/MS spectra. This is crucial because of the high similarity of fragment masses derived from the fragmentation of the limited number of different monosaccharide residues. These criteria need to relate to such things as the intensity of the fragment signals, the propensity for specific bond cleavages, specific diagnostic fragment ions, difference in linkage isomers, ion mode, and the identity of the equal mass monosaccharides.

Less automated MS/MS matching algorithms are available in which the user can enter a presumed structure and fragment it *in silico* to compare with their experimental MS/MS fragmentation data. With different user friendliness, there are web-based programs available: GlycoFrag [69], GlycoFragment (www.glycosciences.de) [56], and Sweet Substitute [57]. This approach can also be automated from the level of composition–structure, where the amount of fragments matched to a predetermined structure can be obtained [58]. This approach could be useful for glycomic analysis, when type of structures is known and the variety is limited.

Using the highly conserved nature of mammalian *N*-linked glycans, an automated interpretation of MS/MS data on these types of glycosylation is provided by StrOligo [58, 59] and Cartoonist [60]. Both of these algorithms result in a prediction of the most probable *N*-linked glycan structure corresponding to a composition and an MS/MS spectrum and are based on a knowledge of the biosynthesis and experience in mass spectrometric analysis of these types of structures. These tools may prove to be very useful for prediction of possible *N*-linked structures on mammalian glycopeptides containing the N–X–T/S consensus sequence.

12.7.3 Linkage Analysis Software Tools

Determining the isomeric and anomeric configurations of a particular oligosaccharide sequence, however, is still a difficult task. Each six-carbon saccharide in a chain has potentially five linkage positions. There is an argument that this knowledge is not always required but in some cases (e.g. the cell membrane receptor of the influenza virus), the monosaccharide linkage determines function, so there still remains a need to assign linkage to an oligosaccharide sequence, particularly in glycomics, in which the experiments are designed to look for both sequence and structural differences.

The “best” method for this purpose is NMR spectroscopy [61], as all the oligosaccharide structural parameters can be deduced from a spectrum; however, in practice, there is rarely sufficient material or availability of instrumentation and expert interpretation to make this approach commonplace. The first attempt at the automated interpretation of both *O*- and *N*-linked glycan linkages was commercialized by Oxford GlycoSciences using computer-generated analysis of exoglycosidase arrays and liquid chromatographic separations [62]. This approach was limited by the specificity and availability of the enzyme-arrays and the complexity of the chromatography.

An alternative interpretation of mass fragmentation data is the “bottom-up” approach, in which standard small oligomers with known linkages are MS^n fragmented by ion trap mass spectrometers and their spectra stored in a library (Table 12.5C). These are then used to collate the substructure mass ions that are present in mass spectra of larger oligosaccharides. This approach has been used manually to interpret the MS^n data of released alditols [63] or permethylated standards [64], and in a more automated approach, using the fragmentation patterns of standard, small, permethylated oligomers [65, 66]. The match of the component spectra is reliant upon being able to compile the fragment blocks back into the original structure, but shows promise of being applicable to unknown oligosaccharide structure determination.

The alternative “top-down” approach (Table 12.5D) may provide a *de novo* structural determination process in which the glycan parent mass constrains the possible monosaccharide compositions, and the masses of possible connected branching topologies are generated computationally for matching to the experimental MS^n data. The STAT algorithm [67] supports the fragmentation pathways of underivatized glycans, StrOligo [58] is based on aminated derivatives, whereas OSCAR [68] uses the knowledge of permethylated MS^n glycan fragmentation pathways to reassemble the fragments, from all possible glycosidic and cross-ring cleavages of different branched isomers, into the most probable structure.

The informatic tools which are available to date all have limitations (Table 12.5) as to what they are able to deduce from the MS data. Fragmentation mass spectrometry is clearly necessary if more is required than just monosaccharide composition, but the extent of fragmentation carried out is dependent upon the amount of sample available, the complexity of the structures that it contains and the structural details required to be known. Ion trap mass spectrometry at present is the approach most suited to get the fragmentation required for more detailed information on linkage and anomericity. MALDI ionization is most readily adopted by biologists but requires derivatization and clean-up of the sample with consequent losses. Electrospray ionization offers the availability of a chromatographic interface to clean up and separate isomeric structures and retains the sialylation and sulfation information of the oligosaccharides. These different experimental practices all have their pluses and minuses, but the lack of uniformity and a universally applicable technique together with insufficient validated oligosaccharide standards preclude the easy collection of high-quality mass spectra with which to compare the experimental data. As such, the development of the current informatic approaches has depended on what the scientist has perceived as the need. In general, this has involved developing methods for quickly predicting the structural possibilities, which can then be experimentally verified.

The degree of complexity of the developed bioinformatics increases from databases designed to cope with subsets of structures (e.g. mammalian *N*-links), through those which contain previously characterized structures (e.g. published papers) or synthetic structures, to those attempting to carry out *de novo* analysis by either breaking a hypothetical proposed structure into its component fragments or by using generic fragment building blocks.

To date, all these approaches have yet to become generally applicable as they have not been validated sufficiently on different mass spectrometers or with different sample preparation techniques (Table 12.6). As they are further developed, scoring algorithms which take into account such things as specific structural diagnostic fragments, ion intensities from fragmentation by different ionizations, and cleavage pathways of the various sugar derivatives of diverse oligosaccharide types will be able to cope with the method-specific variables and will allow automatic assignments to be made much more generally and accurately.

Table 12.6 Validation of computational tools for determination of oligosaccharide structure.

Name	Input	Output	Validated on			Automated
			Derivative	MS	Oligosaccharide type	
Cartoonist	MS	Composition Inferred sequence	Permethylated	+MALDI-TOF	<i>N</i> -linked mammalian	Yes
GlycoMod	MS	Composition	User entered	±MALDI/ESI	<i>N</i> - and <i>O</i> -linked and glycopeptides	No
GlycoSidIQ	MS ²	Composition Sequence	Reduced	-ESI ion trap/+MALDI-QTOF, TOF/TOF	<i>N</i> - and <i>O</i> -linked/sialylated, sulfated, neutral standards	Yes
GlycoSearch-MS/ GLYCOSCIENCES.de	MS ²	Sequence	User entered	+MALDI TOF/TOF	<i>N</i> - and <i>O</i> -linked/neutral standards	No
MS ⁿ Spectral Library	MS ² , MS ³	Sequence Linkage	PA	+MALDI-QIT	Neutral <i>N</i> -linked/human	Yes
Catalogue Library	MS ¹⁻⁴	Anomericity Sequence Linkage	Reduced	+MALDI-FT-MS	Neutral <i>O</i> -linked	No
GLYCH	MS ²	Sequence Linkage	Permethylated	+MALDI-TOF/TOF	Saccharide standards	No
MS ⁿ FragLib	MS ¹⁻⁵	Composition Sequence Linkage	Permethylated	+ESI ion trap	Neutral <i>N</i> - and <i>O</i> -linked mammalian	Yes
STAT: Saccharide Topology Analysis Tool	MS ²⁻⁴	Anomericity Sequence	Reducing	+ESI ion trap	<i>N</i> -linked standard, bacterial glycolipids	No
StrOligo	MS ²	Composition Sequence	AB, PMP	+MALDI-TOF/QqTOF	Neutral, monosialyl <i>N</i> -linked and glycopeptides/human protein	Yes
OSCAR: Oligosaccharide Subtree Constraint Algorithm	MS ¹⁻⁵	Sequence Linkage Anomericity Isobaric differentiation	Permethylated	+MALDI-QIT +ESI ion trap	<i>N</i> -linked standards, glycolipids	No

12.8 Conclusion

Mass spectrometric analysis, together with the specific bioinformatics tools being developed, promises to have the same major impact in glycomics as it has had in proteomics. We are in the initial phase of providing the tools needed to facilitate the understanding of a biomolecule which has previously been considered to be both too difficult to analyze while at the same time mistakenly not being considered to be of much importance in biological function. Whether mass spectrometry can give us all the information we need will depend on the informatic tools that are developed to extract the information from the large amount of data which can be generated by these instruments.

We must also keep in mind that proteins represent only one of the biomolecules whose activity is affected by glycosylation. Proteoglycans, glycolipids, peptidoglycans, lipopolysaccharides, and polysaccharides all have important roles in cellular function and all present unique challenges in their particular MS analysis in the future context of glycomics. As the analysis and data interpretation become easier, the importance to biology of these molecules is becoming clearer, with many subtle changes in function being found to be attributable to changes in composition or structure of the glycan moiety.

Abbreviations

CID/CAD	collision induced/associated dissociation
dHex	deoxyhexose
ECD	electron capture dissociation
ESI	electrospray ionization
FAB	fast atom bombardment
FT	Fourier transform
GMF	glycofragment mass fingerprinting
Hex	hexose
HexNAc	<i>N</i> -acetylhexosamine
IRMPD	infrared multiphoton dissociation
LSI	liquid secondary ion
MALDI	matrix-assisted laser desorption/ionization
PSD	post-source decay
TFA	trifluoroacetic acid

Since the preparation of this manuscript a new form of mass spectrometric ionisation, known as Electron Transfer Dissociation(ETD), has emerged that can fragment a peptide without loss of the attached oligosaccharide. This new technique may enable information to be obtained on both the sites of glycosylation, and the heterogeneity of structures at each site, in the one experiment.

References

1. Zaia J: Mass spectrometry of oligosaccharides. *Mass Spectrom Rev* 2004, **23**:161–227.
2. Mechref Y, Novotny M: Structural investigations of glycoconjugates at high sensitivity. *Chem Rev* 2002, **102**:321–369.

3. Stellner K, Saito H, Hakomori S: Determination of aminosugar linkages in glycolipids by methylation. *Arch Biochem Biophys* 1973, **155**:464–472.
4. Peter-Katalinic J: Analysis of glycoconjugates by fast atom bombardment mass spectrometry and related MS techniques. *Mass Spectrom Rev* 1994, **13**:77–98.
5. Domon B, Costello C: A systematic nomenclature of carbohydrate fragmentation in FAB-MS/MS spectra of glycoconjugates. *Glycoconj J* 1988, **5**:397–409.
6. Karlsson N, Karlsson H, Hansson G: Sulphated mucin oligosaccharides from porcine small intestine analyzed by four-sector tandem mass spectrometry. *J Mass Spectrom* 1996, **31**:560–572.
7. Wilson N, Karlsson N, Packer N: In *Separation Methods In Proteomics* (eds Smejkal GB, Lazarev A). New York: Marcel Dekker; Ch. 19, pp. 345–359.
8. Zhang H, Li X, Martin D, Aebersold R: Identification and quantification of *N*-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol* 2003, **21**:660–666.
9. Kaji H, Saito H, Yamauchi Y, Shinkawa T, Taoka M, Hirabayashi J, Kasai K, Takahashi N, Isobe T: Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify *N*-linked glycoproteins. *Nat Biotechnol* 2003, **21**:667–672.
10. Yuan J, Hashii N, Kawasaki N, Itoh S, Kawanishi T, Hayakawa T: Isotope tag method for quantitative analysis of carbohydrates by liquid chromatography–mass spectrometry. *J Chromatogr A* 2005, **1067**:145–152.
11. Xie Y, Liu J, Zhang J, Hedrick J, Lebrilla C: Method for the comparative glycomic analyses of *O*-linked, mucin-type oligosaccharides. *Anal Chem* 2004, **76**:5186–5197.
12. Ninonuevo M, An H, Yin H, Killeen K, Grimm R, Ward R, German B, Lebrilla C: Nano-liquid chromatography–mass spectrometry of oligosaccharides employing graphitized carbon chromatography on microchip with a high-accuracy mass analyzer. *Electrophoresis* 2005, **26**:3641–3649.
13. Karlsson N, Wilson N, Wirth H, Dawes P, Joshi H, Packer N: Negative ion graphitised carbon nano-liquid chromatography/mass spectrometry increases sensitivity for glycoprotein oligosaccharide analysis. *Rapid Commun Mass Spectrom* 2004, **18**:2282–2292.
14. Kawasaki N, Itoh S, Ohta M, Hayakawa T: Microanalysis of *N*-linked oligosaccharides in a glycoprotein by capillary liquid chromatography/mass spectrometry and liquid chromatography/tandem mass spectrometry. *Anal Biochem* 2003, **316**:15–22.
15. Thomsson K, Karlsson H, Hansson G: Sequencing of sulfated oligosaccharides from mucins by liquid chromatography and electrospray ionization tandem mass spectrometry. *Anal Chem* 2000, **72**:4543–4549.
16. Wuhrer M, Koeleman C, Deelder A, Hokke C: Normal-phase nanoscale liquid chromatography–mass spectrometry of underivatized oligosaccharides at low-femtomole sensitivity. *Anal Chem* 2004, **76**:833–838.
17. Packer N, Lawson M, Jardine D, Redmond J: A general approach to desalting oligosaccharides released from glycoproteins. *Glycoconj J* 1998, **15**:737–747.
18. Papac D, Briggs J, Chin E, Jones A: A high-throughput microscale method to release *N*-linked oligosaccharides from glycoproteins for matrix-assisted laser desorption/ionization time-of-flight mass spectrometric analysis. *Glycobiology* 1998, **8**:445–454.
19. Küster B, Wheeler S, Hunter A, Dwek R, Harvey D: Sequencing of *N*-linked oligosaccharides directly from protein gels: in-gel deglycosylation followed by matrix-assisted laser desorption/ionization mass spectrometry and normal-phase high performance liquid chromatography. *Anal Biochem* 1997, **250**:82–101.
20. Carlson D: Oligosaccharides isolated from pig submaxillary mucin. *J Biol Chem* 1966, **241**:2984–2986.
21. Küster B, Hunter A, Wheeler S, Dwek R, Harvey D: Structural determination of *N*-linked carbohydrates by matrix-assisted laser desorption/ionization–mass spectrometry following enzymatic release within sodium dodecyl sulfate–polyacrylamide electrophoresis gels: application

- to species-specific glycosylation of alpha1-acid glycoprotein. *Electrophoresis* 1998, **19**:1950–1959.
22. Wilson N, Schulz B, Karlsson N, Packer N: Sequential analysis of *N*- and *O*-linked glycosylation of 2D-PAGE separated glycoproteins. *Proteome Res* 2002, **1**:521–529.
 23. Kui W, Easton R, Panico M, Sutton-Smith M, Morrison J, Lattanzio F, Morris H, Clark G, Dell A, Patankar M: CA125 characterization of the oligosaccharides associated with the human ovarian tumor marker. *J Biol Chem* 2003, **278**:28619–28634.
 24. Manna P, Joshi L, Reinhold V, Aubert M, Sukanuma N, Pettersson K, Huhtaniemi I: Synthesis, purification and structural and functional characterization of recombinant form of a common genetic variant of human luteinizing hormone. *Hum Mol Genet* 2002, **11**:301–315.
 25. Suzuki S, Honda S: Analysis of carbohydrates by capillary electrochromatography. *Chromatography* 2001, **22**:171–179.
 26. Handa S, Nakamura K: Modification of sialic acid carboxyl group of ganglioside. *J Biochem (Tokyo)* 1984, **95**:1323–1329.
 27. Powell A, Harvey D: Stabilization of sialic acids in *N*-linked oligosaccharides and gangliosides for analysis by positive ion matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* 1996, **10**:1027–1032.
 28. Harvey D: Fragmentation of negative ions from carbohydrates: Part 3. Fragmentation of hybrid and complex *N*-linked glycans. *J Am Soc Mass Spectrom* 2005, **16**:647–659.
 29. Kobata A, Takasaki S: In *Glycobiology* (ed. Fukuda M, Kobata A). Oxford: Oxford University Press; 1993, pp. 165–185.
 30. Packer N, Lawson M, Jardine D, Sanchez J, Gooley A: Analyzing glycoproteins separated by two-dimensional gel electrophoresis. *Electrophoresis* 1998, **19**:981–988.
 31. Charlwood J, Skehel J, Camilleri P: Analysis of *N*-linked oligosaccharides released from glycoproteins separated by two-dimensional gel electrophoresis. *Anal Biochem* 2000, **284**:49–59.
 32. Sagi D, Kienz P, Denecke J, Marquardt T, Peter-Katalinic J: Glycoproteomics of *N*-glycosylation by in-gel deglycosylation and matrix-assisted laser desorption/ionization-time of flight mass spectrometry mapping: application to congenital disorders of glycosylation. *Proteomics* 2005, **5**:2689–2701.
 33. Harvey D, Bateman R, Bordoli R, Tyldesley R: Ionization and fragmentation of complex glycans with a quadrupole time-of-flight mass spectrometer fitted with a matrix-assisted laser desorption/ionization ion source. *Rapid Commun Mass Spectrom* 2000, **14**:2135–2142.
 34. Lewandrowski U, Resemann A, Sickmann A: Laser-induced dissociation/high-energy collision-induced dissociation fragmentation using MALDI-TOF/TOF-MS instrumentation for the analysis of neutral and acidic oligosaccharides. *Anal Chem* 2005, **77**:3274–3283.
 35. Mechref Y, Novotny M, Krishnan C: Structural characterization of oligosaccharides using MALDI-TOF/TOF tandem mass spectrometry. *Anal Chem* 2003, **75**:4895–4903.
 36. Harvey D, Martin R, Jackson K, Sutton C: Fragmentation of *N*-linked glycans with a matrix-assisted laser desorption/ionization ion trap time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom* 2004, **18**:2997–3007.
 37. Zhang J, Schubotho K, Li B, Russell S, Lebrilla C: Infrared multiphoton dissociation of *O*-linked mucin-type oligosaccharides. *Anal Chem* 2005, **77**:208–214.
 38. Yamagaki T, Suzuki H, Tachibana K: In-source and postsource decay in negative-ion matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of neutral oligosaccharides. *Anal Chem* 2005, **77**:1701–1707.
 39. Robbe C, Capon C, Flahaut C, Michalski J: Microscale analysis of mucin-type *O*-glycans by a coordinated fluorophore-assisted carbohydrate electrophoresis and mass spectrometry approach. *Electrophoresis* 2003, **24**:611–621.
 40. Schulz B, Packer N, Karlsson N: Small-scale analysis of *O*-linked oligosaccharides from glycoproteins and mucins separated by gel electrophoresis. *Anal Chem* 2002, **74**:6088–6097.

41. Zamfir A, Bindila L, Lion N, Allen M, Girault H, Peter-Katalinic J: Chip electrospray mass spectrometry for carbohydrate analysis. *Electrophoresis* 2005, **26**:3650–3673.
42. An H, Peavy T, Hedrick JL, Lebrilla C: Determination of *N*-glycosylation sites and site heterogeneity in glycoproteins. *Anal Chem* 2003, **75**:5628–5637.
43. Larsen M, Hojrup P, Roepstorff P: Characterization of gel-separated glycoproteins using two-step proteolytic digestion combined with sequential microcolumns and mass spectrometry. *Mol Cell Proteomics* 2005, **4**:107–119.
44. Tajiri M, Yoshida S, Wada Y: Differential analysis of site-specific glycans on plasma and cellular fibronectins: application of a hydrophilic affinity method for glycopeptide enrichment. *Glycobiology* 2005, **15**:1332–1340.
45. Håkansson K, Emmett M, Marshall A, Davidsson P, Nilsson C: Structural analysis of 2D-gel-separated glycoproteins from human cerebrospinal fluid by tandem high-resolution mass spectrometry. *J Proteome Res* 2003, **2**:581–588.
46. Håkansson K, Cooper H, Hudgins R, Nilsson C: High resolution tandem mass spectrometry for structural biochemistry. *Curr Org Chem* 2003, **7**:1503–1525.
47. Renfrow M, Cooper H, Tomana M, Kulhavy R, Hiki Y, Toma K, Emmett M, Mestecky J, Marshall A, Novak J: Determination of aberrant *O*-glycosylation in the IgA1 hinge region by electron capture dissociation Fourier transform-ion cyclotron resonance mass spectrometry. *J Biol Chem* 2005, **280**:19136–19145.
48. Wada Y, Tajiri M, Yoshida S: Hydrophilic affinity isolation and MALDI multiple-stage tandem mass spectrometry of glycopeptides for glycoproteomics. *Anal Chem* 2004, **76**:6560–6565.
49. Rademaker G, Pegantis S, Blok-Tip L, Langridge JJ, Kleen A, Thomas-Oates JE: Mass spectrometric determination of the sites of *O*-glycan attachment with low picomolar sensitivity. *Anal Biochem* 1998, **257**:149–160.
50. Wada Y, Azadi P, Costello C, Dell A, Dwek R, Geyer H, Geyer R, Kakehi K, Karlsson N, Kato K, *et al.*: Comparison of the methods for profiling glycoprotein glycans: HUPO HGPI (Human Proteome Organization Human Disease Glycomics/Proteome Initiative) multi-institutional study. *Glycobiology* 2007, **17**:411–422.
51. Cooper C, Gasteiger E, Packer N: GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* 2001, **1**:340–349.
52. Cooper C, Joshi H, Harrison M, Wilkins M, Packer N: GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res* 2003, **31**:511–513.
53. Varki A, Marth J: Oligosaccharides in vertebrate development. *Semin Dev Biol* 1995, **6**:127–138.
54. Lohmann K, von der Lieth C: GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res* 2004, **32**:W261–W266.
55. Lütteke T, Bohne-Lang A, Loss A, Götz T, Frank M, von der Lieth C: GLYCOCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* 2006, **16**:71R–81R.
56. Lohmann K, von der Lieth C: GLYCO-FRAGMENT: A web tool to support the interpretation of mass spectra of complex carbohydrates. *Proteomics* 2003, **3**:2028–2035.
57. Clerens S, Van den E, Verhaert P, Geenen L, Arckens L: Sweet Substitute: a software tool for *in silico* fragmentation of peptide-linked *N*-glycans. *Proteomics* 2004, **4**:629–632.
58. Ethier M, Saba J, Spearman M, Krokhin O, Butler M, Ens W, Standing K, Perreault H: Application of the StrOligo algorithm for the automated structure assignment of complex *N*-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003, **17**:2713–2720.
59. Ethier M, Saba J, Ens W, Standing K, Perreault H: Automated structural assignment of derivatized complex *N*-linked oligosaccharides from tandem mass spectra. *Rapid Commun Mass Spectrom* 2002, **16**:1743–1754.
60. Goldberg D, Sutton-Smith M, Paulson J, Dell A: Automatic annotation of matrix-assisted laser desorption/ionization *N*-glycan spectra. *Proteomics* 2005, **5**:865–875.

61. van Halbeek H: ^1H nuclear magnetic resonance spectroscopy of carbohydrate chains of glycoproteins. *Methods Enzymol* 1994, **230**:132–168.
62. Edge C, Rademacher T, Wormald M, Parekh R, Butters T, Wing D, Dwek R: Fast sequencing of oligosaccharides: the reagent-array analysis method. *Proc Natl Acad Sci USA* 1992, **89**:6338–6342.
63. Tseng K, Hedrick J, Lebrilla C: Catalog-library approach for the rapid and sensitive structural elucidation of oligosaccharides. *Anal Chem* 1999, **71**:3747–3754.
64. Tang H, Mechref Y, Novotny M: Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* 2005, **21**:i431–i439.
65. Ashline D, Singh S, Hanneman A, Reinhold V: Congruent strategies for carbohydrate sequencing. 1. Mining structural details by MS(*n*). *Anal Chem* 2005, **77**:6250–6262.
66. Zhang H, Singh S, Reinhold V: Congruent strategies for carbohydrate sequencing. 2. FragLib: an MS(*n*) spectral library. *Anal Chem* 2005, **77**:6263–6270.
67. Gaucher S, Morrow J, Leary J: STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal Chem* 2000, **72**:2331–2336.
68. Lapadula A, Hatcher P, Hanneman A, Ashline D, Zhang H, Reinhold V: Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MS(*n*) data. *Anal Chem* 2005, **77**:6271–6279.
69. Joshi H, Harrison M, Schulz B, Cooper C, Packer N, Karlsson N: Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* 2004, **4**:1650–1664.
70. Kameyama A, Kikuchi N, Nakaya S, Ito H, Sato T, Shikanai T, Takahashi Y, Takahashi K, Narimatsu H: A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. *Anal Chem* 2005, **77**:4719–4725.

Software Tools for Semi-automatic Interpretation of Mass Spectra of Glycans

Kai Maass¹ and Alessio Ceroni²

¹*Institute of Biochemistry, University of Gießen, 35392 Gießen, Germany*

²*Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, UK*

13.1 Introduction

Mass spectrometry is the main analytical technique currently used to address the challenges of glycomics as it offers unrivalled levels of sensitivity and the ability to handle complex mixtures of different glycans. Modern mass spectrometric (MS) techniques are capable of producing mass spectra of fragmented carbohydrates and this information can be exploited to resolve the structure of a glycan molecule. However, interpretation of mass spectra is a major bottleneck in the rapid and reliable analysis of MS data in high-throughput glycomics projects, and robust solutions to this bottleneck are of critical importance. Therefore, it is not surprising that various experimentally oriented groups have been developing software solutions and algorithms to bypass this bottleneck.

The previous chapter describing the current status of tools to interpret glycan MS data demonstrates that automated interpretation is still an evolving field compared with what is available in proteomics. Bioinformatics scientists are attempting to develop a tool that would provide complete structural information (i.e. composition, sequence, branching, linkage, and anomeric state) from glycan mass spectra, without the requirement for additional information derived from specific structural experiments. However, up until now, only a few software tools have been available to support experimentalists during the annotation process. The coverage and capability of these existing tools is somewhat varied, encompassing a number of approaches to the analysis. Most of the existing applications allow only data analysis for specific tasks (such as the study of *N*-glycans from mammalian samples [1]). None of them allows a complete workflow (see Figure 13.1) from the recorded experimental data to a fully assigned spectrum or to glycan structure determination.

The development of the two software tools, *Glyco-Peakfinder* [2] and *GlycoWorkbench* [3], both freely provided by the EUROCarbDB initiative (<http://www.eurocarbdb.org/applications/ms-tools>), aims to close this gap. Both tools can be used independently as they focus on different stages of the interpretation process of MS data. However, they

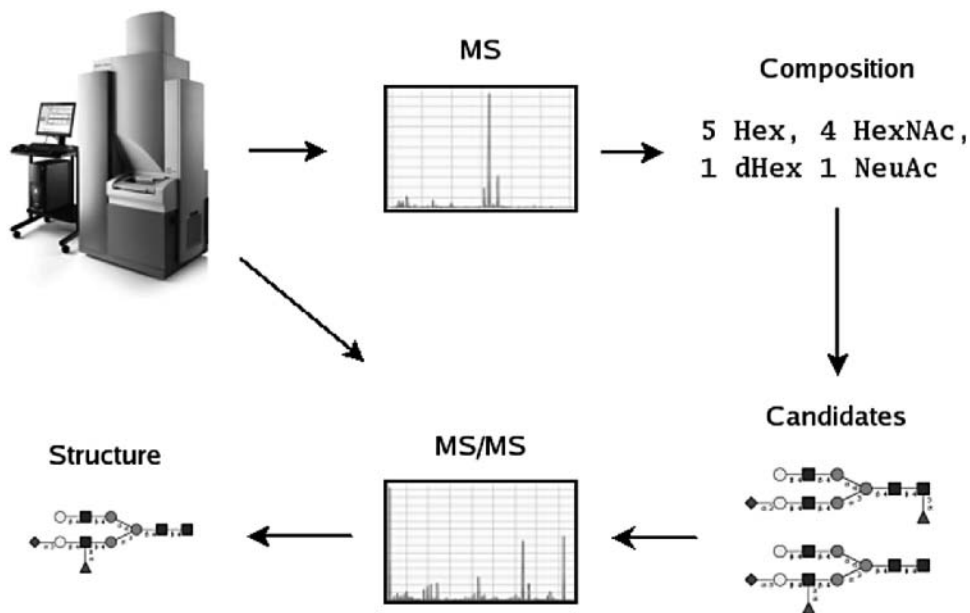


Figure 13.1 Experimental workflow for (semi-)automatic determination of glycan structures from raw data to fully assigned spectrum, via composition analysis (*Glyco-Peakfinder*) and fragment matching (*GlycoWorkbench*).

can interact in a way that allows a complete and smooth workflow from raw data to a completely assigned spectrum. *Glyco-Peakfinder* performs *de novo* composition analysis, allowing the interpretation of MS profiles and also glycan fragment spectra. Its main uses are in the composition analysis of each peak of a mass spectrum of unknown glycans, the identification of glycans in mixtures, and the detection of impurities in samples. Based on the composition suggested by *Glyco-Peakfinder*, a final annotation of each spectrum can be done using the *GlycoWorkbench* software suite, where the theoretical fragmentation patterns of discrete structures can be matched with experimental data.

In the following sections, *Glyco-Peakfinder* and *GlycoWorkbench* are described in detail. The interlocking between the tools, providing a continuous workflow from experimental data to fully assigned spectra, is demonstrated for an example structure taken from a pyridylamino-oligosaccharide Hex₅HexNAc₄dHex-PA fraction obtained from batroxobin, a thrombin-like serine protease from *Bothrops moojeni* venom [4].

13.2 *Glyco-Peakfinder*: Automatic Compositional Analysis of Glycan Mass Spectra

Glyco-Peakfinder is a web application developed for *de novo* composition analysis of glyco-conjugates. It is designed to ease the time-intensive manual annotation of all kinds of mass spectra. Glycan profiles can be analyzed in addition to fragment spectra. *Glyco-Peakfinder* assigns all types of fragmentations including monosaccharide cross-ring cleavages (A-, B-, C-, X-, Y-, Z-fragments, gain and loss of small molecules). The tool provides full user

Table 13.1 Overview of settings for composition analysis in *Glyco-Peakfinder*, Modifications at the reducing end, or of the whole structure, can be considered for composition analysis in *Glyco-Peakfinder*.

Mass options	
Spectrum type	Glycan profile/fragment spectrum
Accuracy of mass	$\Delta m/z \leq 2000$ ppm or $\Delta m/z \leq 5$ u
Mass range	$0 \leq m/z \leq 4000$ (monoisotopic or average)
No. of different residues	31 (3 freely definable)
Charge options	
Charge states	-4 to +4
Charged ions	H ⁺ , Na ⁺ , K ⁺ , Li ⁺ , -H ⁺ , 1 freely definable
Neutral ion exchanges	0 to 3 (Na ⁺ , Li ⁺ , K ⁺ , 1 freely definable)
Fragmentation options	
Number of fragmentations	1 to 4 (combinations of all fragmentation types allowed)
Glycosidic cleavages	B-, C-, Y-, Z-fragmentations
Cross-ring fragmentations	A- and X-fragmentations
Loss or gain of small molecules	E.g. water, carbon dioxide, ketene, methanol
Modification options	
Modifications of the whole structure	Permethylation, peracetylation, perdeuteromethylation, perdeuteroacetylation
Modifications at the reducing end	Reduced end, fluorescent labels (anthranilic acid, 2-aminobenzamide, 2-aminopyridine, etc.), freely definable modification, lipids (selectable from menu), peptides (enter amino acid sequence)

control to handle modified glycans (see Table 13.1), including modifications at the reducing end, or of the whole structure. The option to calculate multiply charged ions increases the range of application to techniques other than matrix-assisted laser desorption/ionization (MALDI) (see Table 13.1). Although the derived information for each entered m/z value is completely independent from the results of neighboring peaks, cross-linking of the results for several peaks provides additional sequence information. To provide access to known carbohydrate structures, a subsequent search for the derived compositions in open-access databases can be performed.

13.2.1 Input Interface

The web interface allows the user full control over all search constraints (see Table 13.1) according to their needs and allows the inclusion of all pre-existing knowledge about the composition of the sample. In this way, the computation of unrealistic compositions can be avoided and the calculation time can be decreased. Consideration of properties of the sample, such as biological background or experimental modifications from the purification and separation process, can be included by constraining the number and kind of monosaccharides to be searched. This decreases computation time considerably.

The first input page, the mass page, allows upload of peak lists from mass spectrometric experiments and specification of the estimated accuracy of the mass signals, and also the choice between monoisotopic and average mass calculation – both representing the experimental technique used. Depending on the experimental conditions, the calculations

in *Glyco-Peakfinder* can be optimized towards a selection of either glycan profile or fragment spectrum data.

Further inputs at the residue page allow the specification of monosaccharides quantities, including their proposed minimum and maximum occurrence in the structure. By default, possibly occurring monosaccharides for calculation of a mammalian *N*-glycan are set. Nevertheless, other residues can be chosen and up to three new monosaccharide types can be entered additionally.

The settings at the ion/charge page refer to the experimental technique and fragmentation style. Choice of the possibly occurring charge states is made with reference to the ionization technique used, while the occurrence of different charged ions (H^+ , Na^+ , K^+ , etc.) refers to the solvent or matrix used for the experiment. Different fragmentation techniques result in varying the settings for the fragment types and the maximum number of cleavages used in computation. Finally, settings on the modification page enable the calculation of either manually modified structures (i.e. permethylation, reduction, pyridylamination, etc.) or of naturally occurring glycoconjugates, such as glycolipids or glycopeptides.

13.2.2 Results of Calculation and Database Search

The algorithm uses the previously set constraints to compute the list of compositions matching the mass to charge values. Results of the calculation are displayed in a tabular form (see Figure 13.2) showing a comparison between the experimentally produced and the calculated masses with their difference (in ppm), and the possible compositions.

Up to this stage, the derived compositions fitting the recorded masses were calculated, without reference to their biological relevance. To allow further investigations, a database search in an open-access database (sourced from GLYCOSCIENCES.de) is provided. Activating the check boxes for specific derived compositions, followed by moving to the structures page (see Figure 13.3), provides structure candidates already found in Nature.

For continuation of a complete workflow resulting in a final assignment of the MS data, the main export function provides the results to *GlycoWorkbench* for further analysis of the plausibility of the structure candidates found.

13.3 *GlycoWorkbench*: Computer-assisted Annotation of Fragment Mass Spectra

The previous section demonstrated how the *Glyco-Peakfinder* tool helps to assign composition(s) for a certain mass value. However, an unequivocal determination of a structure sequence can only be obtained by MS^n fragmentation experiments. Manual annotation of glycan fragment spectra involves a series of tedious and repetitive steps which can be easily automated, resulting in a substantial decrease in the time required for data interpretation. *GlycoWorkbench* is a suite of software tools designed to assist the annotation of glycan fragment spectra. The graphical interface of *GlycoWorkbench* (see Figure 13.4) provides an environment in which structure models can be rapidly assembled, automatically matched with MS^n data, and compared to assess the best candidate. In the following sections, the various features of *GlycoWorkbench* are described. The screenshots provided depict a possible assignment using the candidate structures selected in *GlycoPeakfinder* matching the m/z value 1865.

Glyco-Peakfinder

results
structures
settings

Allowed residues :

Hex	0 - 5
HexNAc	0 - 4
dHex	0 - 1

Precursor Peak :

Mass	Intensity	Composition (check for fragment and structure search)	Charged Ions	Ion type	Mass calculated	Deviation [ppm]
1.865,816	2.490,64	<input checked="" type="checkbox"/> dHex1Hex5HexNAc4-PA	H+		1.865,71454	-54,4

Fragmented Spectra:

Mass	Intensity	Composition (check for fragment and structure search)	Charged Ions	Ion type	Mass calculated	Deviation [ppm]
163,116	823,06					
204,173	11.326,2					
222,178	1.225,42					
300,351	4.986,82	<input type="checkbox"/> HexNAc1-PA	H+	Y	300,15530	-652,0
366,388	2.945,61					
407,471	7.399,35					
446,534	13.289,84	<input type="checkbox"/> dHex1HexNAc1-PA	H+	Y	446,21310	-719,2
503,553	1.955,42	<input type="checkbox"/> HexNAc2-PA	H+	Y	503,23457	-632,8
528,532	723,28					
569,557	493,55					
649,668	1.849,37	<input type="checkbox"/> dHex1HexNAc2-PA	H+	Y	649,29238	-578,5
665,772	615,78	<input type="checkbox"/> Hex1HexNAc2-PA	H+	Y	665,28729	-728,6
811,739	867,33	<input type="checkbox"/> dHex1Hex1HexNAc2-PA	H+	Y	811,34510	-485,5
852,705	542,77	<input type="checkbox"/> dHex1HexNAc3-PA	H+	Y	852,37165	-391,1
1.014,665	705	<input type="checkbox"/> dHex1Hex1HexNAc3-PA	H+	Y	1.014,42437	-237,2
1.298,013	1.174,71	<input type="checkbox"/> dHex1Hex4HexNAc2-PA	H+	Y	1.297,50327	-392,9
1.379,965	601,51	<input type="checkbox"/> dHex1Hex2HexNAc4-PA	H+	Y	1.379,55637	-296,2
1.420,815	562,85					
1.460,000	1.073,85	<input type="checkbox"/> dHex1Hex5HexNAc2-PA	H+	Y	1.459,55599	-304,2
1.703,967	1.409,93	<input type="checkbox"/> dHex1Hex4HexNAc4-PA	H+	Y	1.703,66181	-179,1
1.719,917	835,96	<input type="checkbox"/> Hex5HexNAc4-PA	H+	Y	1.719,65673	-151,4
1.865,816	2.490,64	<input type="checkbox"/> dHex1Hex5HexNAc4-PA	H+	Y	1.865,71454	-54,4

Delete selected annotations
Search composition in DB
GLYCOSCIENCES.de ▾

Change settings
New calculation

EuroCarbDB is a Research Infrastructure Design Study Funded by the 6th Research Framework Program of the European Union
(Contract: RIDS Contract number 011952)

Figure 13.2 *Glyco-Peakfinder* – results window for the composition analysis of a laser-induced dissociation (LID) spectrum of a protonated and pyridylaminated (PA) Hex₅HexNAc₄dHex species (accuracy of mass: ±1000 ppm). The Domon–Costello nomenclature [5] is used to assign peaks.

Glyco-Peakfinder					
introduction		results		structures	
fragments		settings		contact	
Searched composition:	Hex5HexNAc4dHex1-PA			Mass:	1.865,715
				Found structures:	75
Linux ID	2D-Plot of structure				
16667	<pre> a-D-Manp-(1-6)+ a-L-Fucp-(1-6)+ a-D-Manp-(1-6)+ D-GlcNAc a-D-Manp-(1-3)+ b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)+ b-D-GalpNAc-(1-4)-b-D-GlcpNAc-(1-2)-a-D-Manp-(1-3)+ </pre>				
	[Linux File] [IUPAC 2D Graph File]				
16672	<pre> a-L-Fucp-(1-6)+ D-GlcNAc b-D-Galp-(1-4)-b-D-GlcpNAc-(1-2)-a-D-Manp-(1-6)+ b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)+ b-D-Galp-(1-3)-b-D-GlcpNAc-(1-2)-a-D-Manp-(1-3)+ </pre>				
	[Linux File] [IUPAC 2D Graph File]				

Figure 13.3 *Glyco-Peakfinder* – structures window for the database search in GLYCO-SCIENCES.de [6] for the composition Hex₅HexNAc₄dHex. In total 75 structures were found, of which at least two were selected for further investigations in *GlycoWorkbench*.

13.3.1 GlycanBuilder

The main component of *GlycoWorkbench* is *GlycanBuilder*, a flexible visual editor specially designed for a user-friendly input of glycan structures. Glycans often exhibit tree-like non-linear structures, and their constituents exhibit great diversity. Because of the tree-like structure, the input of a glycan sequence is not as straightforward as writing a sequence of characters, as for DNA, RNA, and peptide sequences. The lack of a suitable user-friendly graphical interface to input complex carbohydrate structures in a computer-readable format has long been a severe deficiency in the practical application of glyco-related databases. Additionally, numerous alternative notations are commonly adopted to represent glycan structures graphically. Finally, the more comprehensive formats for digital encoding of glycan structures (such as Glyco-CT, <http://www.eurocarbdb.org/recommendations/encoding> – see Chapter 3) are difficult to produce manually. *GlycanBuilder* addresses all these issues: the user can rapidly specify a glycan structure by simply selecting the points of attachment of the residues, the growing structure is displayed using one of the available symbolic notations, and the output is a computer encoding of the structure in Glyco-CT format. The list of structural constituents comprises a comprehensive collection of saccharides, substituents, saccharide modifications, and reducing-end markers. All the stereochemical information about a saccharide, such as anomeric configuration, configuration of each chiral center, ring size, and linkage position, can be specified. The most commonly used symbolic representations for glycans (Consortium for Functional Glycomics [7] and the Oxford Glycobiology Institute [8]) are available. The display of a glycan is dependent only on its structure and the chosen symbolic notation: the appearance and the spatial placement of the residues are

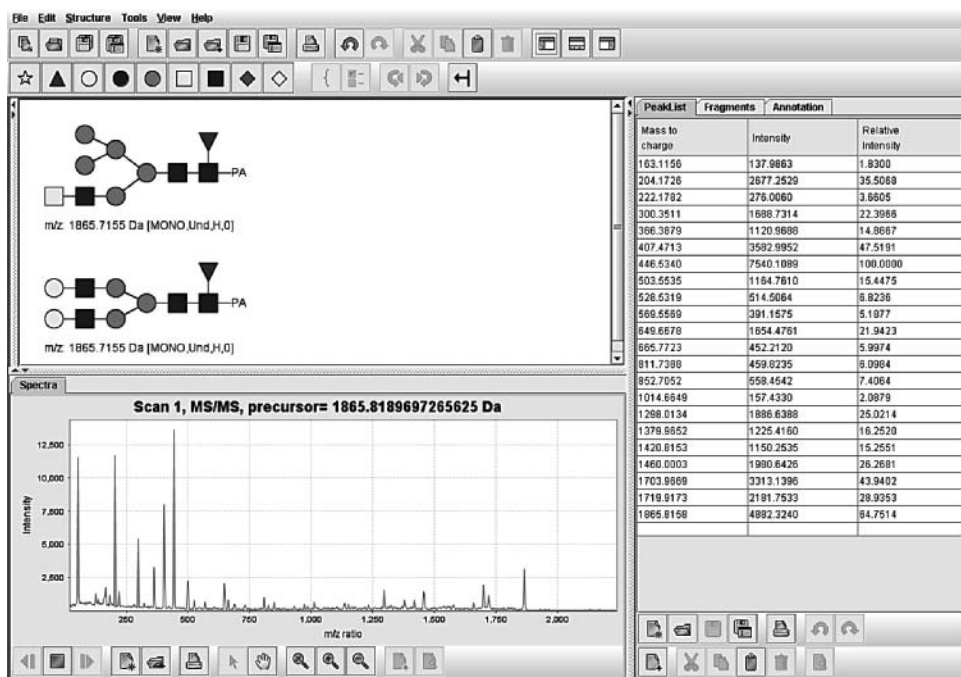


Figure 13.4 *GlycoWorkbench* – main window. *GlycoWorkbench* is an integrated suite of software tools for assisting the annotation of glycan fragment mass spectra. All tools are accessible from a common user interface. Here the *GlycoWorkbench* interface with *GlycanBuilder* (top left), the Spectra view (bottom left) and the PeakList view (right) are shown. The commonly used symbolic representation as recommended by the US Consortium for Functional Glycomics is used to display the oligosaccharide structures. *GlycoWorkbench* can be downloaded from www.eurocarbdb.org/applications/ms-tools/.

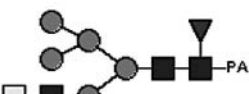
automatically determined according to a set of rules specified by the given notation. The rules are stored in configuration files and new notations can easily be added. The software does not need user interaction to represent a structure, and the symbolic notation can be switched without the need for additional editing. Therefore, the tool can be used both as an editor for drawing structures and as an automated component for generating pictorial representations of computer-encoded glycans.

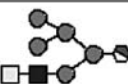

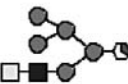
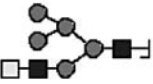
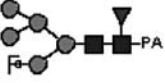
13.3.2 In silico Fragmentation

After the desired input structures have been defined with *GlycanBuilder*, the remaining components of *GlycoWorkbench* can be used to derive their fragments, compute the fragment masses, build a peak list, and annotate it. The computation of fragments and their masses from the intact structure is a central step for the annotation of MS^n spectra. The fragmentation tool creates all topologically possible fragmentations of the precursor molecular ion, applying both multiple glycosidic cleavages and cross-ring fragmentations (see Figure 13.5). The fragments are computed by recursively traversing the tree structure of the glycan and applying all the possible cleavages at each position. Fragmented structures

PeakList Fragments Annotation

List Editor


m/z: 1885.7155 Da [MONO,Und,H,0]

Fragment	Type	m/z	Ions	Neutral Exchanges	Fragment Mass
	$1,4_A$ GlcNAc	1348.4883	H	0	1347.4810
	Y	1379.5571	H	0	1378.5498
	$1,5_A$ GlcNAc	1392.5145	H	0	1391.5073
	B	1420.5095	H	0	1419.5022
	Y	1459.5567	H	0	1458.5494

Navigation icons: back, forward, search, refresh, undo, redo, cut, copy, paste, home.

Figure 13.5 *GlycoWorkbench* – fragmentation window. *GlycoWorkbench* can be used to compute all topologically possible fragmentations of the precursor molecular ion, applying both multiple glycosidic cleavages and cross-ring fragmentations. For a given glycan fragment, the m/z ratio can be calculated both for native and derivatized structures (permethylated/peracetylated) taking into account several types and quantities of ion adducts.

are then subjected to the same process to produce multiple cleavages. For a given glycan fragment, the m/z ratio can be calculated both for native and derivatized structures (permethylated/ peracetylated) taking into account several types and quantities of ion adducts. A visual editor of glycan fragments is also available, where the user can specify in which positions the cleavages are occurring on the displayed structure in order to reproduce an already known fragment molecule.

13.3.3 Annotation of Peaks

The next step in the annotation process is the assignment of possible fragments to each m/z value in a given peak list. In *GlycoWorkbench* a peak list either can be loaded from a tab-separated text file, thus allowing for import from peak-picking software, or it can be created by typing mass and intensity values directly into the application. Alternatively, the raw spectrum can be loaded from several standard XML or application-specific data formats. The mass spectrum is displayed and can be panned or zoomed as in a normal spectrum viewer. The user can select m/z values directly from the spectrum and add them to the peak list. Once the peak list is ready, the fragment m/z values from the *in silico* fragmentation are matched with a given accuracy to each peak in the list. In the positive ion mode, for each fragment all the possible combinations of single or multiple adducts are generated from a list of allowed ions (H^+ , Na^+ , K^+ , Li^+). Alternatively, in the negative ion mode, all possible losses of protons are tested. Neutral exchanges of charges can also be computed and the tool automatically determines the number of available charges for a molecule.

The annotated peak list can be displayed using various panels that show its different aspects. Each panel is based around a spreadsheet-like table view, whose cell values can be sorted by each column, and can be copied into spreadsheet applications. The detailed view (see Figure 13.6) allows the user to assess the list of fragment peak matches for each candidate structure, showing the fragment structure, mass, m/z value, distribution of charges, and annotation accuracy. The proposed annotation can be modified by removing the matches that are not satisfactory given expert knowledge of the fragmentation pathway. The summary view (see Figure 13.7) allows the user to compare the annotations for the different structures back-to-back in the same table. The matching fragments from different structures are shown in adjacent columns, with each row corresponding to a single peak. In this way, signals that clearly distinguish the correct annotation from the other hypothetical models can be identified. The statistics view allows the user to perform a quantitative comparison between the annotations, by showing the number of assigned peaks at different thresholds of relative peak intensity, the root mean square deviation between peak and fragment m/z values and the average intensity of assigned peaks. Finally, a calibration graph shows the annotation accuracies at the various m/z values, allowing the user to check the correct calibration of the mass spectrum. The annotated peak list can be stored in a specific XML format for later consultation.

13.4 Conclusion

Interpretation of mass spectra is the major bottleneck in the rapid and reliable analysis of MS data in high-throughput glycomics projects, and robust solutions to this bottleneck are of critical importance. However, up until now, only a few software tools have been available to support experimentalists during the annotation process. The coverage and capability of these existing tools are somewhat varied, encompassing a number of approaches to the analysis, but allowing data analysis only for specific tasks.

The tools described here, *Glyco-Peakfinder* and *GlycoWorkbench*, focus on different stages of the interpretation of experimental data and can interact in a way which permits a complete and smooth workflow from raw data to fully assigned spectra. Both tools have

Mass to charge	Intensity	Relative Intensity	Fragment	Type	Accuracy	Accuracy PPM	Fragment m/z	Ions	Neutral Exchanges
407.4713	3582.9952	47.5191		B	-0.3052	-749.0085	407.1661	H	0
446.5340	7540.1089	100.0000		Y	-0.3207	-718.2937	446.2133	H	0
503.5535	1164.7610	15.4475		YY	-0.3187	-632.9394	503.2348	H	0
528.5319	514.5064	6.8236			0.0000	0.0000	0.0000	0	0
569.5569	391.1575	5.1877		B	-0.3381	-593.5523	569.2189	H	0
649.6678	1654.4761	21.9423		Y	-0.3751	-577.3511	649.2927	H	0
649.6678	1654.4761	21.9423		BY	-0.4493	-691.5634	649.2185	H	0
665.7723	452.2120	5.9974			0.0000	0.0000	0.0000	0	0

Figure 13.6 *GlycoWorkbench* – annotation detailed view. The Details view enables users to assess for each individual structure the list of fragment peak matches, showing the fragment structure, theoretical mass, m/z value, distribution of charges and annotation accuracy. In this panel, the annotation can also be modified by removing the matches that are not satisfactory given expert knowledge of the fragmentation pathway.

been implemented as part of the EUROCarbDB design study aiming to develop publicly available resources that can be used to further the advancement of glycomics research. The EUROCarbDB will provide resources for assignment, storage, and retrieval of analytical data related to glycan structure determination. *Glyco-Peakfinder* and *GlycoWorkbench* are available through the MS-tools section of applications pages on the EUROCarbDB website. The capability of *GlycoWorkbench* to allow a direct comparison of the scoring of multiple candidate structures retrieved from *Glyco-Peakfinder* can be regarded as a significant step forward in supporting the routine interpretation of mass spectra of glycans.

The integration of tools such as *Glycan-Peakfinder* and *GlycoWorkbench* with a database of glycan structures and analytical data will provide a highly valuable resource for the advancement of the field of glycomics. Bioinformatics analysis of previously acquired annotated MS data will permit the extraction of expert knowledge in a form that could be automatically applied to the interpretation of new data. The tools would then be able to offer more precise answers by using the previous information to filter out results which are

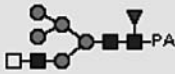
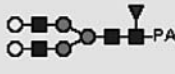
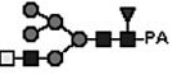
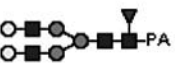

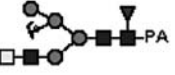
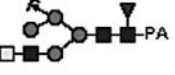
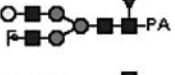
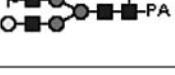





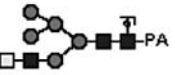

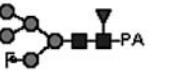
PeakList		Fragments		Annotation	
Stats	Details	Summary	Calibration		
Mass to charge	Intensity	Relative intensity			
446.5340	7540.1089	100.0000			
1865.8158	4882.3240	64.7514			
407.4713	3582.9952	47.5191			
1703.9669	3313.1396	43.9402	 	 	
204.1726	2677.2529	35.5088	  	 	
1719.9173	2181.7533	28.9353			
1460.0003	1980.6426	26.2681			

Figure 13.7 *GlycoWorkbench* – Annotation summary view. The Summary view permits a direct comparison of the peak assignments to fragments for all selected candidate structures back-to-back in the same table. The matching fragments from different structures are shown in adjacent columns, with each row corresponding to a single peak. In this screenshot, the output is ordered based on the relative intensity of the peaks in the fragmentation spectrum. The user can thus identify signals which clearly distinguish the correct annotation from the other hypothetical models.

inconsistent with the acquired knowledge. A comprehensive collection of glycomics data in publicly accessible databases is the only route that would permit a complete automation of MS data interpretation as in the proteomics field.

References

1. Goldberg D, Sutton-Smith M, Paulson J, Dell A: Automatic annotation of matrix-assisted laser desorption/ionization *N*-glycan spectra. *Proteomics* 2005, **5**: 865–875.
2. Maass K, Ranzinger R, Geyer H, von der Lieth C-W, Geyer R: “Glyco-Peakfinder” – *de novo* composition analysis of glycoconjugates. *Proteomics* 2007, **7**: 4435–4444.
3. Ceroni A, Maass K, Geyer H, Geyer R, Dell A, Haslam SM: GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J Proteome Res* 2008, **7**: 1650–1659.
4. Lochnit G, Geyer R: Carbohydrate structure analysis of batroxobin, a thrombin-like serine protease from *Bothrops moojeni* venom. *Eur J Biochem* 1995, **228**: 805–816.
5. Domon B, Costello CE: A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconj J* 1988, **5**: 397–409.
6. Lütteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, von der Lieth C-W: GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* 2006, **16**: 71R–81R.
7. Varki A, Cummings R, Esko J, Freeze H, Hart G, Marth J (eds): *Essentials of Glycobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1999.
8. Royle L, Dwek RA, Rudd PM: Unit 12.6 Determining the structure of oligosaccharides *N*- and *O*-linked to glycoproteins. In *Current Protocols in Protein Science*. Edited by Coligan JE, Dunn BM, Speicher DW, Wingfield PT: John Wiley and Sons; 2006.

Informatics Concepts to Decode Structure-Function Relationships of Glycosaminoglycans

Rahul Raman, S. Raguram and Ram Sasisekharan

Biological Engineering Division, Harvard-MIT Division of Health Sciences and Technology, Center for Biomedical Engineering, Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

14.1 Introduction

Extracellular modulation of phenotype is an emerging paradigm in this current post-genomics age of molecular and cell biology. Glycosaminoglycans (GAGs) are complex polysaccharides found in abundance at the cell–extracellular matrix (ECM) interface of eukaryotic cells. Advances in the technology to analyze GAGs, and in whole-organism genetics have led to a dramatic transformation in understanding the importance of the dynamic nature of the cell–ECM interactions (historically the ECM was considered an inert material that hydrated the cells) in influencing phenotype at cellular and higher order tissue and organ levels. The specific interactions between the chemically heterogeneous GAGs and numerous proteins in the extracellular environment are at the heart of this paradigm. Thus understanding the structure–function relationship of GAGs has gained importance both as a fundamental field and for its potential in identification of novel therapeutic targets. However, progress towards this goal has been limited in the past due to the polydispersity and chemical heterogeneity of GAGs that posed numerous challenges for their isolation and characterization. This chapter discusses the various analytical methods for structurally characterizing GAGs and the need for informatics-based approaches to decode GAG sequence and structure–function relationships.

14.1.1 Glycosaminoglycans

Glycosaminoglycans are complex acidic polysaccharides that are present both on the cell surface and in the extracellular matrix. GAG chains are attached to specific consensus sequences on a core protein. The core protein along with the GAG chains at different attachment sites is collectively known as a proteoglycan. Historically, the ECM was considered inert wherein its sole purpose was to provide a scaffold and hydrate the cells. Thus, GAGs were treated as contaminants during DNA or protein isolation from cells and tissues. The past couple of decades have witnessed a remarkable transformation in the understanding

of the importance of the cell–ECM interactions and how these interactions govern the phenotype at the cell, tissue, and higher levels. Recent advances, specifically in developmental biology and cancer biology, have resulted in a dramatic increase in the number of important roles attributed to GAGs in critical biological processes. These include modulation of developmental processes [1–5], angiogenesis [6–8], axonal growth [9–12], cancer progression [5, 13–19], microbial pathogenesis [20–23], and anticoagulation [24–31].

Based on their location at the cell–ECM interface, GAGs interact with numerous biological agents such as growth factors, cytokines, chemokines, enzymes, morphogens, and microbial pathogens (summarized in [32–35]). An emerging paradigm is that specific GAG–protein interactions modulate the activity of the protein and thus critically impinge on the numerous known biological functions of GAGs [36–38]. GAGs play a critical role in assembling protein–protein complexes such as growth factor–receptor or enzyme–inhibitor aggregates on the cell surface and in the extracellular matrix which are directly involved in initiating cell signaling events or inhibiting biochemical pathways. Thus, the positioning of the protein-binding oligosaccharide motifs along the GAG chain determines if an active signaling complex is assembled at the cell surface or an inactive complex is sequestered in the matrix.

14.1.2 Chemical Structure of GAGs

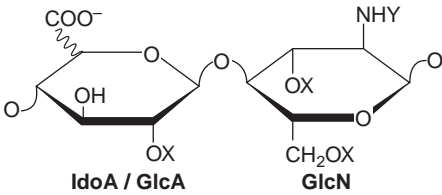
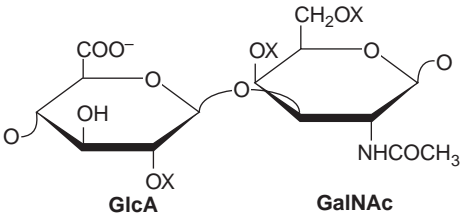
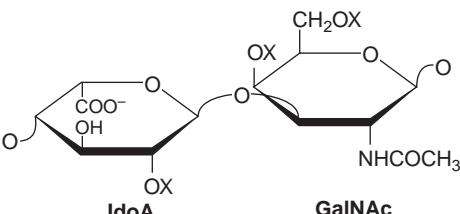
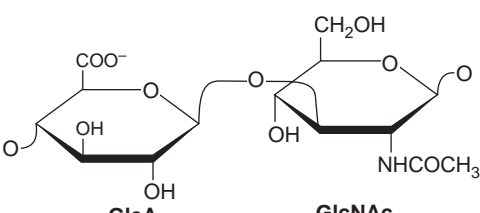
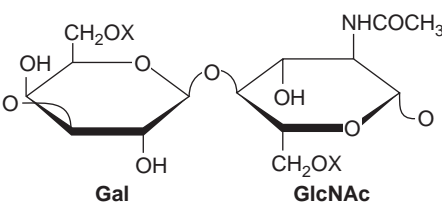
The chemical structure of GAGs is comprised of a disaccharide repeat unit of a hexuronic acid linked to a hexosamine. These disaccharides constitute the basic building blocks of GAGs analogous to amino acids for proteins and nucleotides for DNA. There are five major categories of GAGs classified based on the chemical structure of the disaccharide building block (Table 14.1). The variation in the disaccharide building block primarily arises from sulfation modifications at different positions and the epimeric state of the uronic acid. Among these different categories of GAGs, heparan sulfate GAGs (HSGAGs) have the maximum diversity in their disaccharide building blocks. The primary sequence of GAGs is defined by the backbone chemical structure of the disaccharide repeat units and the sulfation pattern. Another way to look at primary sequence is to consider it as a linear arrangement of the different disaccharide building blocks.

The biosynthesis of GAGs is a complex process that is not template driven, unlike the template-based synthesis of DNA and proteins [39, 40]. The two main events involved in biosynthesis are chain elongation and modification at specific sites along the chain. The chain elongation involves the action of multidomain glycosyltransferases that successively transfer hexosamine and hexuronic acid. The modifications are brought about by several sulfotransferases, where each modification site involves a distinct type of sulfotransferase and its isoforms [41, 42]. It is therefore challenging to understand how the biosynthetic machinery influences the large diversity in the primary sequence of GAGs. As a result of their complex biosynthesis, GAGs derived from even a specific cell type are chemically heterogeneous. The chemical heterogeneity of GAGs has made it challenging to develop analytical techniques for their structural characterization.

14.1.3 Development of Methodologies for Structural Characterization of GAGs

Over the years, there have been significant advances in the development of biochemical tools and sensitive analytical methods to address the challenges in isolation and fine structure

Table 14.1 Disaccharide building blocks of GAGs.

GAG family	Disaccharide building block ^a
Heparin/heparan sulfate	 <p style="text-align: center;">IdoA / GlcA GlcN</p>
Chondroitin sulfate	 <p style="text-align: center;">GlcA GalNAc</p>
Dermatan sulfate	 <p style="text-align: center;">IdoA GalNAc</p>
Hyaluronic acid	 <p style="text-align: center;">GlcA GlcNAc</p>
Keratan sulfate	 <p style="text-align: center;">Gal GlcNAc</p>

^aX = H or SO₃⁻; Y = H or COCH₃, or SO₃⁻. The glycosidic linkages within the building block and between building blocks shown for different GAG families are implied throughout the text.

characterization of GAGs [43–48]. Broadly, the major steps involved in the characterization of GAGs are degradation, separation, and analysis. Degradation is accomplished using chemical and enzymatic methods. The ability to depolymerize or modify a GAG chain in a predictable fashion is an important step in decoding the sequence of a GAG oligosaccharide. Traditionally, separation of GAGs was a challenging process owing to their high sulfation content and presence of structural isomers. These challenges have been circumvented to a great extent by development of sensitive electrophoretic and chromatographic methods [43, 45–48]. Furthermore, there have also been advances in the analysis of small amounts of GAGs using mass spectrometric techniques [49] (MS techniques are discussed in detail in Chapter 12). The development of these tools is beginning to unravel important attributes of GAG oligosaccharide mixtures such as monosaccharide and disaccharide composition, enzymatic degradation maps (analogous to restriction maps for DNA), molecular weight, and chain length distributions. These attributes are critical for understanding structure–function relationships of GAGs in their numerous biological functions.

14.2 Tools for Depolymerization/Modification of GAGs

An important aspect in the characterization and/or sequencing of GAGs is the ability to achieve predictable cleavage (at a specific glycosidic linkage) or modification (addition or removal of a sulfate or acetate group at a specific site) of the oligosaccharide chain. Chemical depolymerization is one of the earliest methods used to characterize GAGs. The best known example is that of the characterization of a pentasaccharide motif in heparin that plays a critical role in its anticoagulant activity [26]. The randomness of cleavage by chemical methods imposes limitations in applying these methods to decode the sequence diversity of GAGs isolated from different cell types. These challenges are addressed by the development of recombinant GAG-degrading enzymes from different bacterial sources. Biochemical characterization of these enzymes permitted their development as tools for sequencing GAGs [36]. Furthermore, these enzymes were also used to investigate directly the role of GAGs in different biological processes such as tumor growth [13, 16, 50], angiogenesis [8], and neural regeneration [9, 10, 51].

14.2.1 Chemical Methods

A summary of different chemical methods used to depolymerize and desulfate GAG chains is given in Table 14.2. Different chemical methods provide distinct depolymerization of the GAG chains. Nitrous acid cleavage can be used to cleave selectively either after GlcNS, or after GlcN (or GalN), under different conditions [52]. Periodate oxidation typically results in splitting of the pyranose ring at the C2–C3 linkage of an unsulfated GlcA or IdoA. The ring-opened residue containing aldehyde groups at C2 and C3 is labile to base treatment, which results in cleavage of the GAG chain. On the other hand, ring splitting can also be achieved for 2-*O*-sulfated IdoA by selective chemical desulfation, which results in the formation of L-galacturonic acid (L-GalA), which subsequently undergoes ring splitting [53]. Chemical β -elimination cleaves at any linkage between hexosamine and uronic acid. The uronic acid is converted into a Δ 4,5-unsaturated uronic acid upon chemical β -elimination and thus it loses its original stereochemical identity.

Table 14.2 Chemical methods for depolymerization/modification of GAGs

Method	Reaction	Product	Comments
Nitrous acid cleavage [52]	Cleaves bond between GlcNS and uronic acid (IdoA/GlcA)	Reducing end GlcNS is modified to 2,5-anhydro-D-mannitol with loss of <i>N</i> -sulfate group	GlcNAc and GalNAc are not reactive. However, GAGs containing the deacetylated free amine form of these sugars are cleaved by nitrous acid at pH \approx 4
Periodate oxidation [53]	Cleaves bond between C2–C3 of unsulfated IdoA and GlcA	Splitting of uronic acid pyranose ring resulting in cleavage of the GAG chain	Controlled oxidation results in split ring uronic acids without cleavage, enhancing the conformational flexibility of the GAG chains
β -Elimination	Esterification of carboxylate group of uronic acid to lower pK_a of C5 proton followed by β -eliminative cleavage of the hexosamine–uronic acid bond	Non-reducing end uronic acid of product has C4–C5 double bond	Stereochemical identity of uronic acid in the parent chain is lost due to formation of Δ 4–5 double bond

Modification of GAGs via chemical desulfation (and resulfation), deacetylation (and reacylation), and enzymatic methods (using biosynthetic enzymes) has been used to investigate their structure–function relationships [54–57]. Model compounds that have a similar chemical structure to GAGs have been generated by chemical and enzymatic modification of bacterial capsular polysaccharides [56, 58]. The chemical backbone of K5 polysaccharide is similar to that of the nascent HSGAG backbone, wherein it is comprised of the \rightarrow 4GlcA β 1 \rightarrow 4GlcNAc α 1 \rightarrow repeat unit. Similarly, the chemical backbone of the K4 polysaccharide is comprised of \rightarrow 4GlcA β 1 \rightarrow 3GalNAc β 1 \rightarrow repeat unit, which is similar to CSGAG backbone.

14.2.2 Enzymes as Tools for Characterization of GAGs

The application of chemical methods to explore the vast sequence diversity of GAGs is limited owing to the randomness in their cleavage and the potential side reactions. To allow a rigorous structural characterization of GAGs, it is important to have a system that can be engineered to depolymerize GAGs at specific linkages. The GAG depolymerizing enzymes produced by bacterial and mammalian cells provided a means to develop such a system. Soil bacteria such as *Pedobacter heparinus* and *Proteus vulgaris* produce heparinases and chondroitinases that depolymerize HS- and CSGAGs, respectively. These enzymes metabolize GAGs in dead animals so that the components may be utilized as a major nutrient source for the bacteria. The bacterial depolymerizing enzymes are lyases that cleave the glycosidic linkage between a hexosamine and a uronic acid through the β -elimination mechanism. This results in the eliminative cleavage of the hexosamine–uronic acid glycosidic bond and results in the Δ 4,5-unsaturated linkage on

the uronic acid. This unsaturated linkage has a characteristic absorbance at 232 nm, which facilitates direct measurement of the enzymatic activity.

The overall sequence of GAG degradation in bacteria involves the depolymerization of GAG chains by the heparinases and chondroitinases down to di- and tetrasaccharide fragments. These small fragments are then acted upon by exolytic enzymes (i.e. enzymes cleaving from the non-reducing terminus) such as Δ UA 2-*O*-sulfatase, Δ 4,5-glycuronidase, *N*-sulfamidase (HSGAG), GlcNAc 3-*O*-sulfatase, GlcNAc 6-*O*-sulfatase, GalNAc 4-*O*-sulfatase, and GalNAc 6-*O*-sulfatase. The activity of these enzymes is summarized in Table 14.3. The development of recombinant GAG-degrading enzymes such as the heparinases [59–63], chondroitinases [64–67], and the exolytic enzymes [68, 69] allows their use as valuable tools for sequencing GAGs [36]. Defining the substrate specificities of these enzymes is valuable for providing experimental constraints that are coupled with different analytical methods to develop strategies for rapid sequencing of GAGs, as outlined in the following sections.

The mammalian GAG-degrading enzymes including heparanases and hyaluronidases are glycosidases, which hydrolyze the glycosidic linkages and retain the epimeric state of the uronic acid in the products formed [8, 70]. The human heparanase I enzyme cleaves at the GlcA–GlcNS,3S,6S linkage within the oligosaccharide motif -GlcNAc,6S-GlcAGlcNS,3S,6S–IdoA2S–GlcNS,6S– (the cleavable linkage is underlined), where the 2-*O*-sulfation on the adjacent IdoA has been implicated to play a key role in the activity of the enzyme [71]. The hyaluronidases have broad substrate specificity (analogous to bacterial chondroitinase ABC) in that they cleave hyaluronic acid and CSGAGs.

The GAG catabolic pathway involves depolymerization of the endocytosed GAG chains by heparanases and hyaluronidases and desulfation of the smaller fragments by lysosomal sulfatases, which are exolytic enzymes that specifically cleave sulfate groups at different positions [72]. Two novel endolytic sulfatases, discovered in different organisms such as quail, mouse, and humans, have been isolated and characterized [73, 74]. These sulfatases cleave the 6-*O*-sulfate group from 6-*O*-sulfated glucosamine within the HSGAGs on the cell surface and the cell–ECM interface [75, 76]. The 6-*O*-sulfate on *N*-sulfated glucosamine (GlcNS,6S) is preferentially cleaved in comparison with that on *N*-acetylated glucosamine (GlcNAc,6S). These endosulfatases represent a novel family of enzymes that dynamically remodel the cell surface GAG chains [13] and hence provide a complementary set of tools for sequencing GAG chains in addition to bacterial and lysosomal enzymes.

14.3 Methods and Techniques for Analysis of GAGs

Most of the methods for the analysis of GAGs can be classified broadly into four categories – chromatographic, electrophoretic, mass spectrometric, and nuclear magnetic resonance (NMR) based methods. An important component in utilizing chromatographic and electrophoretic techniques for the analysis of GAGs is the detection of GAG oligosaccharides. Nitrous acid cleavage and subsequent reduction with tritiated sodium borohydride (NaB^3H_4) introduces a radiolabel on the reducing end ring-contracted 2,5-anhydro-D-mannitol ($[^3\text{H}]\text{aManR}$). The use of bacterial GAG depolymerizing lyases introduces the Δ 4,5-uronic acid at the non-reducing end, which has a characteristic absorbance at 232 nm. Chemical derivatization strategies to introduce fluorescent labels and other mass-based

Table 14.3 Bacterial GAG-degrading enzymes as tools for characterization of GAGs

Enzyme	Species	Substrate specificity ^a	Comments
Heparinase I (hep I)	<i>Pedobacter heparinus</i>	GlcNS,6X-Ido2S	Predominantly exolytic and progressively depolymerizes highly sulfated regions of HSGAGs [59, 62]
Heparinase II (hep II)	<i>Pedobacter heparinus</i>	Any linkage between glucosamine and uronic acid in HSGAGs	Broad HSGAG substrate specificity [60, 63]
Heparinase III (hep III)	<i>Pedobacter heparinus</i>	GlcNY,6X-GlcA/IdoA	Orthogonal substrate specificity to heparinase I that depolymerizes regions of relatively low sulfation in HSGAGs [61, 77]
Chondroitinase AC	<i>Pedobacter heparinus</i>	GalNAc,4X,6X-GlcA2X	Cleaves 4- and 6- <i>O</i> -sulfated GalNAc-containing chondroitin substrates. Inhibited by dermatan sulfate [66, 78]
Chondroitinase AC	<i>Arthrobacter aurescens</i>	GalNAc,4X,6X-GlcA2X	Predominantly exolytic enzyme relative to its analog from <i>P. heparinus</i> with similar substrate specificity [79]
Chondroitinase B	<i>Pedobacter heparinus</i>	GalNAc,4X,6X-IdoA2X	Cleaves IdoA-containing dermatan sulfate as its only substrate [65, 80]
Chondroitinase ABC I	<i>Proteus vulgaris</i>	Any linkage between GalNAc and uronic acid in chondroitin and dermatan sulfate	Broad chondroitin and dermatan sulfate substrate specificity [64, 67, 81]
Chondroitinase ABC II	<i>Proteus vulgaris</i>	Any linkage between GalNAc and uronic acid in chondroitin and dermatan sulfate	Predominantly exolytic enzyme that preferentially cleaves shorter substrates (tetrahexasaccharide) relative to chondroitinase ABC I [67]
Δ4,5-Uronic acid 2- <i>O</i> -sulfatase	<i>Pedobacter heparinus</i>	Removes 2- <i>O</i> -sulfate from non-reducing end ΔUA2S- formed by heparinase cleavage	Order of preference of substrates ΔUA2S-GlcNAc,6S > ΔUA2S-GlcNS,6S > ΔUA2S-GlcNS > ΔUA2S-GlcNAc [69]
Δ4,5-Glycuronidase	<i>Pedobacter heparinus</i>	Cleaves at linkage between unsulfated ΔUA and hexosamine releasing ΔUA	Inhibited by presence of 2- <i>O</i> -sulfate in ΔUA2S [68]
Glucosamine 6-sulfate sulfatase	<i>Pedobacter heparinus</i>	Removes 6- <i>O</i> -sulfate group from GlcNY,6S at non-reducing end	6- <i>O</i> -Sulfated glucosamine is exposed at non-reducing end after cleavage by Δ4,5-glycuronidase
Glucosamine <i>N</i> -sulfamidase	<i>Pedobacter heparinus</i>	Removes <i>N</i> -sulfate group from GlcNS at non-reducing end	Acts after glucosamine 6-sulfate sulfatase since GlcNS,6S inhibits its activity

^aX= sulfated or unsulfated; Y = sulfated or acetylated.

labels have been successfully utilized in chromatographic and electrophoretic, and also in mass spectrometric approaches to analyze GAGs [82].

Chromatographic methods are used in both the isolation and preparation of GAG oligosaccharides for subsequent analysis, and also for the direct analysis, quantification, and sequencing of GAG oligosaccharides [83–86]. The separation of the oligosaccharides is based on size (size-exclusion chromatography) and charge (ion-exchange chromatography). Fine structural characterization of GAGs typically involves depolymerization (and modification) of the GAG oligosaccharide by chemical and enzymatic methods followed by separation and analysis of the resulting fragments using different chromatographic techniques. The characterization of the resulting fragments is performed either by comparing their elution profile with that of internal standards or by monitoring the shifts in the elution profile after treatment with different exolytic enzymes [85, 86]. The sequence of the oligosaccharide is reconstructed based on analysis of the fragments resulting from depolymerization. The use of additional exolytic enzymes to remove sulfate groups and monosaccharides selectively from the non-reducing end provides further constraints to determination of the GAG sequence. Another approach in the chromatographic analysis of GAGs is the coupling of liquid chromatography (LC) with electrospray ionization mass spectrometry (ESI-MS). The LC–MS methodology is used to separate oligosaccharides and simultaneously characterize and quantify them using tandem MS–MS approaches [87–89].

Electrophoretic methods separate GAG oligosaccharides based on the differences in their mobility depending on their mass and charge (governed by their sulfation pattern). Many techniques that utilize polyacrylamide gel electrophoresis (PAGE) have been developed to analyze and characterize GAGs [90–93]. These methods rely on the end labeling of GAGs with a fluorescent tag. Capillary electrophoresis (CE) [43, 45, 46, 48, 94–96] can handle extremely small amounts (femto- to picomolar) of GAG oligosaccharides. The resolving capability of CE is such that structural isomers of short GAG oligosaccharides (di- and tetrasaccharides) can be distinguished. The high resolving power of CE has been exploited in the disaccharide compositional analysis of GAGs. This method involves the treatment of GAGs with heparinases or chondroitinases such that each GAG chain is completely depolymerized into its disaccharide building blocks. The disaccharide building blocks are characterized by comparing their migration times with those of internal standards. CE-derived disaccharide compositions proved to be highly valuable information when used in combination with PEN-MALDI and PEN-NMR sequencing strategies (see below).

Mass spectrometry (MS) is the most often used detection technique in sequencing a single GAG oligosaccharide and also profiling a GAG mixture. A detailed discussion of various MS techniques for analyzing *N*- and *O*-linked glycans and GAGs is covered in Chapter 12. An important MS approach that has been used extensively in GAG analysis is matrix-assisted laser desorption/ionization (MALDI) MS. The MALDI-MS approach [96–101] has been used in conjunction with enzymatic depolymerization of GAGs to detect the fragments formed as non-covalent complexes with the basic peptide (Arg–Gly)₁₅ [96]. The peptide complexes were used to overcome problems caused by the large number of negative charges associated with HSGAG oligosaccharides [96]. The main advantages of the MALDI-MS approach are the high accuracy (<1 Da) of determining oligosaccharide masses and the requirement of low sample amounts (femto- to picomolar). The most recently developed methodology to analyze GAGs is electrospray ionization (ESI) MS [44, 89, 102–109]. The instrumentation for ESI-MS can be configured to control the ionization of the molecular ions and thus prevent the loss of sulfate groups. Furthermore,

the use of ESI-MS with an ion trap allows the fragmentation of specific parent molecular ions by collision-induced dissociation to perform tandem MS characterization of that ion. This methodology is also amenable to on-line coupling with liquid chromatography and capillary electrophoresis. Methodologies that couple reversed-phase ion pairing HPLC with ESI-MS provide better mass accuracy (<0.1 Da) than MALDI-MS and also allow quantification of the oligosaccharides. ESI-MS and tandem MS have also been developed for the rapid, accurate compositional analysis of HSGAG disaccharides [102, 105, 108]. The ESI-MS-based compositional analysis had advantages over CE-based compositional analysis in that it was able to quantify accurately disaccharide building blocks, which are completely unsulfated, and those with completely unsubstituted glucosamine.

NMR spectroscopy has been used in conjunction with other analytical tools to validate the sequence of many GAG oligosaccharides [83, 110–115]. The anomeric proton and carbon atoms of each monosaccharide (with a specific sulfation pattern) have a distinct chemical shift. These characteristic anomeric chemical shifts have been identified for the commonly occurring monosaccharide units in GAGs [116–118]. The anomeric chemical shifts of non-reducing and reducing end monosaccharides have also been used to characterize specific modifications to monosaccharides such as the $\Delta 4,5$ -uronic acid (non-reducing end), ring-contracted aManR (reducing end), and epoxides that result from chemical depolymerization of GAGs [119]. Two-dimensional COSY (COrrrelation SpectroscopY), TOCSY (TOtal Correlation SpectroscopY) and HSQC (Heteronuclear Spin Quantum Coupling) spectra [116, 117] provide an accurate quantification of the different monosaccharides in a GAG oligosaccharide or a GAG mixture. Given that IdoA and GlcA monosaccharides have distinct chemical shifts, NMR spectroscopy provides the best quantification of these monosaccharides. Although the NMR spectra of complex GAG oligosaccharides are sometimes complicated to interpret owing to the overlap of the signals [120], they provide important, detailed structural information. The anomeric chemical shifts of the monosaccharides are also influenced by their neighboring residues. This effect is the most prominent for the anomeric chemical shifts of the hexosamines, which are distinct for hexosamines linked to IdoA, IdoA2S, and GlcA. This distinction permits further quantification of hexosamine–uronic acid linkages in a GAG sample. Sequence determination using NMR often involves additional two-dimensional NOESY (Nuclear Overhauser Effect SpectroscopY) spectra that indicate non-bonded through-space interactions between the protons at the glycosidic linkage. Based on the identity of the monosaccharide, the observed NOEs enable one to walk through the different linkages in the sequence starting from that monosaccharide.

Unfortunately, sequence determination using NMR requires relatively large sample quantities (milligram amounts), which are not easily procured from tissue samples. The sensitivity of this technique is lower than that based on the detection of chromatographic effluents and MS. Specifically, NOESY and ROESY (Rotational nuclear Overhauser Effect SpectroscopY) spectra usually require more sample and high-field instruments for reasonable sensitivity and signal resolution.

Each of the above-mentioned analytical methods used in conjunction with the different biochemical methods for GAG depolymerization provides unique attributes that characterize a GAG sample. For example, disaccharide compositional analysis using CE provides information on the relative molar abundance of different disaccharide building blocks of the form $\Delta\text{UA}2\text{X-GlcNY},3\text{X},6\text{X}$ ($\text{X} = \text{H}$ or SO_3^- , $\text{Y} = \text{H}$, Ac^- , or SO_3^-). MS data provide information on overall chain length and sulfation of a given GAG oligosaccharide. NMR

methods, on the other hand, provide information on monosaccharide composition including IdoA versus GlcA and also the relative abundance of glucosamine–uronic acid linkages of the form –GlcNY,3X,6X–UA2X–. The complementary nature of these different attributes provides orthogonal constraints to decode a GAG sequence rapidly through a combination of different analytical approaches.

14.4 Informatics Approach to Sequencing GAGs

The major information content in a purified homogeneous GAG oligosaccharide is its primary sequence. The diversity of oligosaccharide motifs within a GAG chain is much higher than that of oligonucleotide or oligopeptide motifs of the same chain length due to the substantially larger variations in the disaccharide building blocks of GAGs. For example, there are 48 possible disaccharide building blocks for HSGAGs in comparison with 20 amino acids and four bases for DNA and proteins. In the case of a heterogeneous mixture of GAG oligosaccharides that is typical of biological GAG samples, the major information content is the sequence of each oligosaccharide and its relative abundance in the mixture. The methods and techniques for GAG depolymerization, modification and analysis provide a means to decode the information content in GAGs. However, the application of any single methodology is not sufficient to explore the diversity of GAG oligosaccharide sequences. In many instances a given sequencing approach is not able to determine conclusively the sulfation at a particular site or the epimeric form of the uronic acid. Nevertheless, each methodology can be viewed as a tool that provides unique constraints and/or defines specific attributes such as mass and composition pertaining to a GAG sample. Hence it can be envisioned that a combination of attributes provided by the different tools would enhance the ability to obtain a more comprehensive and systematic definition of GAG information content. Furthermore, incorporation of multiple attributes as constraints facilitates cross-validation of these attributes pertaining to a given GAG sample, and thus provides an unbiased approach to decoding GAG information content. Based on this rationale, a computational framework that permits the combination of constraints to capture and decode the information content in GAGs has been developed [101].

14.4.1 Property-encoded Nomenclature (PEN) for GAGs

The variability in the GAG sequence is defined in terms of the sulfation state and the epimeric form of the uronic acid in the disaccharide building blocks. In the case of HSGAGs and CSGAGs, the sulfation and epimer state is captured as two-state variability. For example, the 2-*O*-position of the uronic acid, and the 3-*O*- and 6-*O*-positions of the GlcN in HSGAGs can be either “sulfated” or “unsulfated” (Figure 14.1a). The *N*- position of GlcN, in the majority of HSGAGs, can be either acetylated or sulfated. The epimeric state of the uronic acid can be either IdoA or GlcA. These two-state variations allow the use of a binary formalism to capture the modifications of a disaccharide building block. By using such a binary formalism, the *property-encoded nomenclature* was developed for HSGAGs (Figure 14.1b) [101]. The four binary digits that capture the sulfation state of the disaccharide building blocks are combined into a single hexadecimal code. The binary digit that captures the epimeric state of the uronic acid is encoded as a “sign” bit so that the + and – of the same hexadecimal code would indicate only a difference in the epimeric state of the uronic acid.

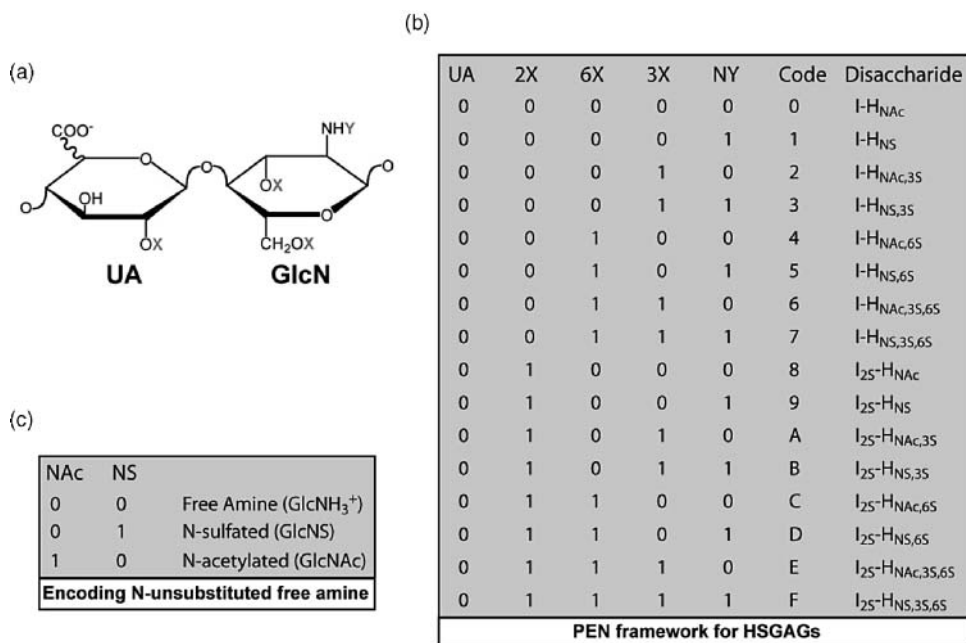


Figure 14.1 PEN framework for HSGAGs. (a) The disaccharide building block of HSGAG (X = site of sulfation, Y = site of acetylation or sulfation). (b) The PEN hexadecimal (base 16) codes for IdoA-containing disaccharides (the short form of the disaccharide structure is used, where I is IdoA and H is GlcN). The UA represents IdoA or GlcA (IdoA = 0, GlcA = 1). The epimeric state of the UA is captured as a sign bit in the final code [IdoA is + (or no sign), GlcA is -] so that + and - of the same hexadecimal code indicate only a difference in the UA. The hexadecimal code is constructed using the binary digits that encode 2X (2-O-position of uronic acid), 6X, 3X, NY (positions on glucosamine). The digit 0 encodes unsulfated (or acetylated for N-position) and 1 encodes sulfation. (c) The scheme used to handle rare modifications such as the unsubstituted N-position.

For example, IdoA–GlcNS,6S is represented using +5 (or just 5), whereas GlcA–GlcNS,6S is represented as -5. The rare occurrence of an unsubstituted N-position in the glucosamine in HSGAGs is further encoded by using an extra bit for the N-position (Figure 14.1c). The addition of the extra bit is easily accommodated due to the numerical nature of the PEN framework. The variability in the disaccharide building blocks of CSGAGs is similar to that of HSGAGs. The sulfation at the 2-O-position of the uronic acid, and 4-O- and 6-O-positions of the GalNAc monosaccharide can be captured using three binary digits. These binary digits can be combined into a single octal code. The epimeric state of the uronic acid can be encoded using a sign bit in a similar way as that of HSGAGs.

The use of the PEN system has many advantages in terms of capturing information content in GAGs. First, the set of bits that encode the sulfation and epimeric states not only define the overall identity of the disaccharide building blocks, but they also capture both the charge distribution and mass of the building blocks. Second, simple binary and mathematical operation on the PEN code allows processing of several binary digits simultaneously. Third, the ability to process several bits simultaneously facilitates incorporation of enzymatic and chemical cleavage of GAGs as constraints since all the cleavage sites can be mapped with a few binary and mathematical operations. Two strategies were developed using the PEN

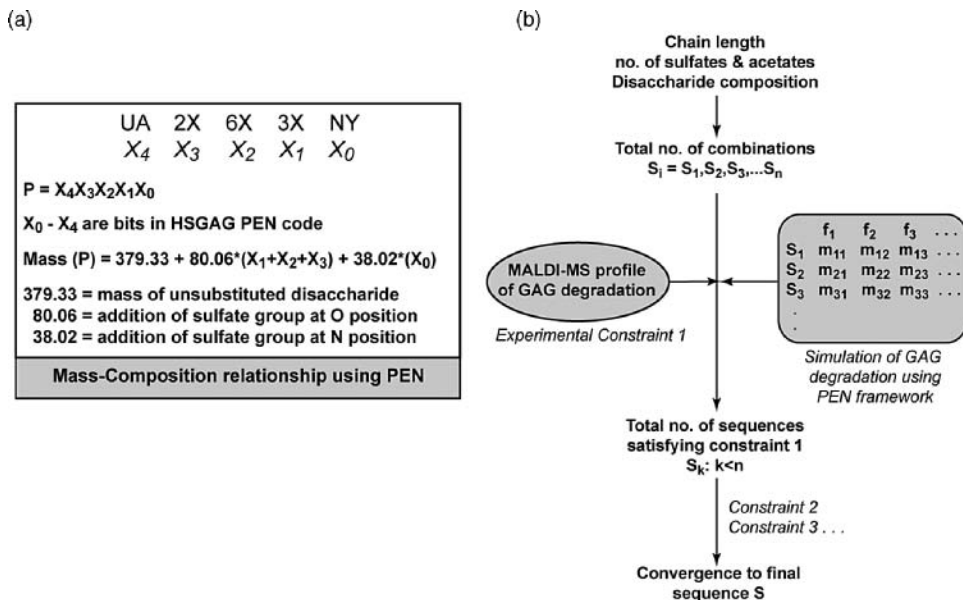


Figure 14.2 PEN-MALDI sequencing strategy. (a) The operations on the PEN code of a disaccharide unit (P shown as a string of binary digits X_0 to X_4) to calculate the mass of a disaccharide (the same logic applies to oligosaccharide) based on its sulfate and acetate pattern. (b) The workflow of the PEN-MALDI sequencing strategy developed based on the mass–composition relationships. The mass of the parent oligosaccharide along with its disaccharide composition obtained from CE is used to create a master list of all the possible sequences. From this master list, sequences are eliminated by iteratively applying the MALDI-MS profile of enzymatic degradation as constraints using the PEN framework until convergence to the final sequence. The rules that govern the specificity and action pattern of GAG degradation process are critical in simulating the theoretical fragments and computing their masses. These theoretical masses are compared with the MALDI-MS profile and those sequences in the master list that do not satisfy the experimental profile are discarded.

framework for a rapid and unbiased sequencing of HSGAGs [101, 120]: PEN-MALDI for MS- and PEN-NMR for NMR-derived sequencing. These strategies are outlined with specific examples in the following.

14.4.2 PEN-MALDI Sequencing Strategy

The mass of a disaccharide repeat unit is directly obtained by performing a single set of mathematical operations on the PEN code (Figure 14.2). The direct mapping of a physical property such as mass to the PEN code enhances searching for GAG sequences based on mass without having to use lookup tables to correlate mass with a specific unit. This relationship was used to construct a theoretical mass–composition map that uniquely assigned the length and overall sulfate/acetate composition of a GAG oligosaccharide (up to a tetradecasaccharide) assuming a mass accuracy of less than 1 Da. The determination of the mass–composition relationship using the PEN framework allowed the development of the PEN-MALDI approach to sequence HSGAG oligosaccharides.

In the PEN-MALDI sequencing methodology for a homogeneous HSGAG oligosaccharide, the sample is subjected to CE-based disaccharide compositional analysis and its mass is obtained by MALDI-MS. Based on the mass the chain length and overall number of sulfates and acetates of the chain are determined. Using this information in conjunction with the disaccharide composition (from CE), a master list of all possible sequences that satisfies these constraints is constructed using the PEN framework. In this manner, no sequences are excluded from the analysis, no matter how unusual a given sequence may be. The masses of oligosaccharide fragments generated from enzymatic and chemical degradation of HSGAGs are applied as additional experimental constraints (using the PEN framework) and sequences that do not satisfy these constraints are eliminated (Figure 14.2). In an iterative manner, moving from experimental constraints to the ever-decreasing master list of possible sequences, rapid convergence to a unique sequence is achieved.

The PEN-MALDI methodology has been successfully applied to sequence many HSGAG oligosaccharides [27, 101]. Importantly, using this methodology, it was shown for the first time that an important antithrombin (AT-III) binding decasaccharide previously derived from heparin [121] contained only a part of the actual pentasaccharide motif [122]. This finding provided valuable insights into the enzymatic generation of low molecular weight heparins with enhanced anticoagulation properties, which were further validated by other studies [28, 122]. The PEN-MALDI sequencing of the AT-III binding decasaccharide (AT-10) derived from heparin is briefly summarized in the following.

The mass of AT-10 determined by MALDI-MS data is 2769 Da, which corresponds to a decasaccharide with 13 sulfate groups and one acetate group. Compositional analysis of AT-10 indicated the presence of three building blocks, corresponding to Δ UA2S–GlcNS,6S (\pm D), Δ UA–GlcNAc,6S (\pm 4) and Δ UA–GlcNS,3S,6S (\pm 7) in the relative ratio 3:1:1 respectively. There were 320 possible combinations of decasaccharide sequences (with 13 sulfates and a single acetate group) with these disaccharide building blocks. Depolymerization of this decasaccharide with hep I (over a short time period) resulted in four distinct mass peaks. Only 52 sequences out of the 320 combinations satisfied hep I depolymerization data. Hep I treatment was also carried out over a long time period to cleave all the hep I cleavable linkages. This resulted in only two mass peaks, corresponding to a trisulfated disaccharide and an octasulfated hexasaccharide, and further reduced the list to 28 sequences. To converge further on the final sequence, a semi-carbazide “mass-tag” ($\Delta = 56.1$ Da) was attached to the reducing end of the decasaccharide and the sample was treated with hep II. This allowed the identification of the saccharide sequence close to and at the reducing end. Subsequent treatment with nitrous acid and the exoenzymes provided further constraints that led the convergence of AT-10 to Δ UA2S–GlcNS,6S–IdoA2S–GlcNS,6S–IdoA2S–GlcNS,6S–IdoA–GlcNAc,6S–GlcA–GlcNS,3S,6S (part of the AT-III binding pentasaccharide motif is underlined).

14.4.3 PEN-NMR Sequencing Strategy

NMR spectroscopy provides both monosaccharide composition and the abundance of the glucosamine–uronic acid (GlcNY,3X,6X–UA2X) linkages. The linkage abundance information from NMR is complementary to the disaccharide composition analysis by CE, which provides the abundance of the UA2X–GlcNY,3X,6X linkages. Thus a combination of these complementary attributes from NMR and CE using the PEN framework permits the systematic deduction of HSGAG oligosaccharide sequences.

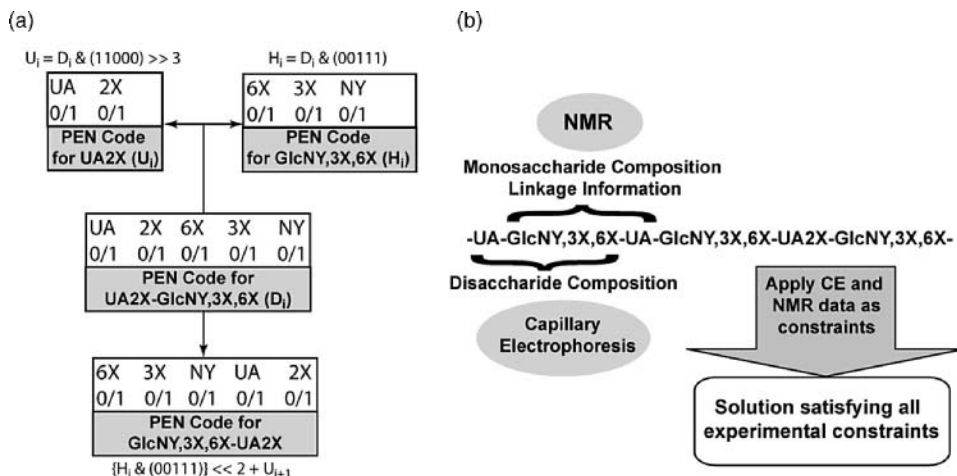


Figure 14.3 PEN-NMR sequencing strategy. (a) The conversion of the PEN code for UA2X–GlcNY,3X,6S disaccharide unit to monosaccharide codes for the uronic acid (U_i) and glucosamine (H_i) and to linkage between GlcNY,3X,6X and the adjacent UA2X (towards its reducing end). These conversions are used to capture the monosaccharide composition and linkage abundance information from NMR spectroscopy. The binary operators “ \gg ”, “ \ll ”, and “ $\&$ ” stand for bitwise shift right, bitwise shift left, bitwise AND, respectively. (b) A schematic of the PEN-NMR sequencing strategy where disaccharide composition (using CE and other methods) along with monosaccharide composition and GlcNY,3X,6X–UA2X linkage abundance are applied as constraints in the PEN framework to converge rapidly on the GAG sequence.

The PEN for HSGAGs encodes a UA2X–GlcNY,3X,6X disaccharide building block. To capture the monosaccharide-level information, this code was converted into two numerical base-4 codes (Figure 14.3). The first base-4 code U_{i4} ($i = i$ th monosaccharide) encodes epimeric state and sulfation pattern of the uronic acid. The second base-4 code H_{i4} encodes the sulfation pattern of the glucosamine. To capture the GlcNY,3X,6X–UA2X linkage abundance provided by NMR, binary operations were used to transpose the uronic acid and glucosamine information in PEN code. The identity and relative abundance of the different monosaccharide units and the GlcNY,3X,6X–UA2X linkage abundance obtained from the different NMR spectroscopic methods were used to generate all possible combinations that satisfy these constraints. Incorporation of the disaccharide composition information (i.e. relative abundance of abundance of UA2X–GlcNY,3X,6X) from CE-based analysis provided additional constraints to converge to the final oligosaccharide sequence. The proof-of-concept of the PEN-NMR strategy was provided by sequencing the AT-10 deca-saccharide using this method [120], which is briefly outlined in the following.

Given the complexity of the AT-10 sequence, there were significant overlaps in the chemical shifts of the protons attached to C6 of the glucosamines in the proton NMR spectra. As a result, it was difficult to distinguish the abundance of the 6-*O*-sulfated and unsulfated glucosamines. However, as described above, disaccharide compositional analysis of AT-10 using CE indicated the presence of three major disaccharide components, Δ UA2S–GlcNS,6S, Δ UA–GlcNAc,6S, and Δ UA–GlcNS,3S,6S, in the ratio 3:1:1. Thus the constraint provided by the CE data facilitated the distinction between the abundances of the 6-*O*-sulfated and unsulfated glucosamine monosaccharides. The relative abundance of the glucosamine

monosaccharides calculated based on NMR data was GlcNS,6S:GlcNS,3S,6S:GlcNAc,6S = 3:1:1 (in agreement with the CE data). In a similar way, the relative abundance of the uronic acid monosaccharides was calculated to be IdoA2S:IdoA:GlcA: Δ UA = 2:1:1:1. The anomeric chemical shifts of the glucosamines were further resolved to provide the linkage abundance GlcNS,6S-IdoA2S:GlcNAc,6S-GlcA:GlcNS,6S-IdoA = 2:1:1. Applying the monosaccharide composition and the linkage abundance information from NMR as constraints using the PEN framework, 12 possible combinations were obtained. Applying the disaccharide compositional analysis as further constraints directly led the convergence to the correct sequence of AT-10. By combining orthogonal linkage abundance information from CE, the PEN-NMR enhances the utility of NMR for GAG analysis in many ways. It assists in the interpretation of complex NMR spectra based on the disaccharide composition information. Furthermore, it minimizes the use of 2D NOESY to determine linkage information.

14.5 Decoding GAG Sequence Diversity – Databases and Bioinformatics Tools

The chemical diversity in the sugar building blocks and the limitations in determination of the complete structures have made it challenging to represent glycan structures systematically in databases [123]. Significant efforts are being made by large-scale international collaborative initiatives such as the Consortium for Functional Glycomics (CFG; www.functionalglycomics.org), EUROCarbDB (www.eurocarbdb.org), Complex Carbohydrate Research Center (CCRC; www.ccrcc.uga.edu), and KEGG Glycan (<http://www.genome.jp/kegg/glycan/>) to address these challenges and develop databases for glycans [124]. In the case of GAGs, online bioinformatics resources (accessible via the Internet similar to those for *N*-linked glycans) are still under development and are not yet publicly available. The following outlines the important aspects of developing databases and bioinformatics tools for GAGs.

14.5.1 Development of a Relational Database to Capture GAG Information Content

In GAGs, the linkages between the uronic acids and hexosamines are fixed and the primary sequence is defined by the sulfation pattern of each monosaccharide and epimeric state of the uronic acid (IdoA versus GlcA). However, it is not possible at present to isolate, purify, and completely determine the sequence of large GAG chains (length >16-mer). GAGs isolated from different cells and tissues are often characterized as mixtures of chains with statistical properties such as composition of individual disaccharide building blocks that make up the chains and relative percentage of specific sulfation patterns that are predominantly represented in the chains. In many cases, structural characterization of purified GAG oligosaccharides results in partially defined sequences wherein there are ambiguities in assigning the epimeric state of specific uronic acids or specific sulfation patterns. A comprehensive approach to capturing the information content in these partial sequences is to construct all the possible combinations of the ambiguities and relate them to the partial sequence. In some cases, the “most likely” sequences can be selected from these combinations based on knowledge of the biological source and the biosynthesis rules. Thus,

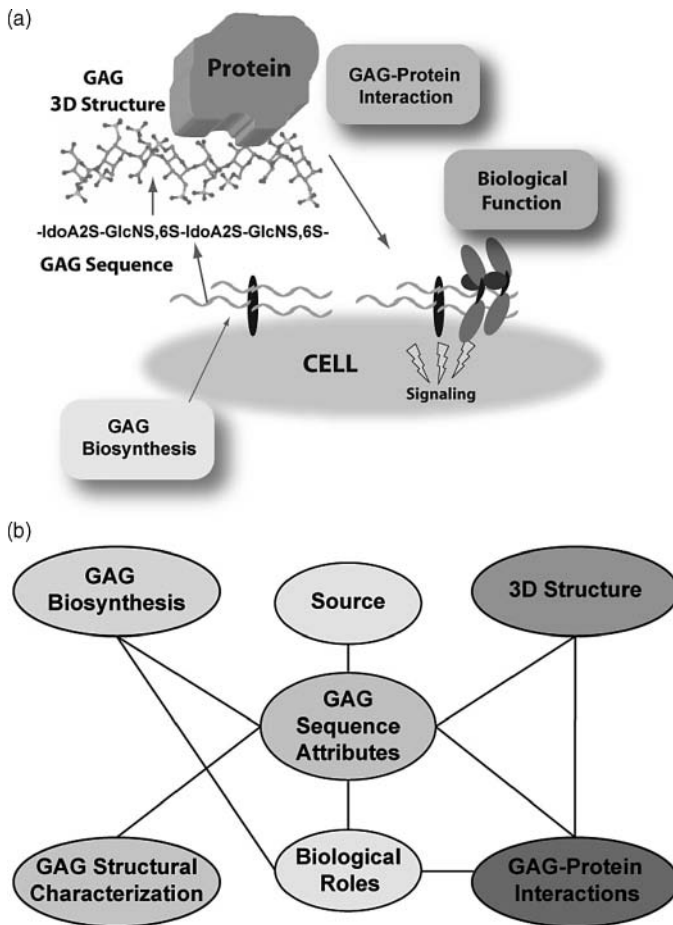


Figure 14.4 Bioinformatics platform for GAGs. (a) A schematic representation of the critical components that govern the structure–function relationships in the numerous biological functions of GAGs. (b) The schema that captures these important components and their interrelationships, and that forms a blueprint for the development of the relational database and bioinformatics tools for GAGs.

an important aspect of GAG database development is the systematic representation of the different sequence attributes ranging from completely determined sequence to statistical distribution of sequence attributes over different chains.

The sequence attributes defining a GAG sample would be the primary object in the GAG database. Based on the overall scheme of the biological roles of GAGs starting from their biosynthesis to GAG–protein interactions modulating a specific function (Figure 14.4a), the important data objects in the GAG database are shown in Figure 14.4b. Examples of information pertaining to these different objects that needs to be captured from an informatics standpoint are shown below. The inherent interrelationships between the different objects (shown using connecting lines in Figure 14.4b) are used to construct the database schema that outlines the data entities and the relationships between the entities. An important

aspect of integrating GAG-related datasets is the development of structured vocabularies and ontologies to capture metadata and interrelationships [125]. For example, it is important to define standard protocols and internal standards, for mass spectrometric, chromatographic, and NMR analysis of GAG samples. Furthermore, it is also important to use standardized vocabularies to define the source of the GAG sample. Implementing these standards would permit comparisons of analytical data sets and GAG sample information across different laboratories involved in GAG research.

14.5.2 Development of Bioinformatics Tools

The development of online tools to query the GAG databases and parse through the important relationships between the different datasets pertaining to GAGs would facilitate understanding GAG structure–function relationships. The PEN informatics framework enhances the development of these tools since mathematical operations on the numerical PEN code allow rapid processing of GAG sequence information. Based on the diversity and interrelationships in the data pertaining to the characterization of GAGs, several bioinformatics tools that permit querying and abstraction of information from the database can potentially be developed. The development of key bioinformatics tools and their utility in the characterization of GAG structure and structure–function relationships are discussed in the following.

14.5.2.1 Mass and Composition Search. The mass of an oligosaccharide or a mass range is passed as a query to the database and it results in a list of sequences satisfying both the mass and mass profiles (containing that particular mass range) of GAG mixtures along with information on their source and other analytical data. The query mass is also directly translated into possible oligosaccharide chain lengths and compositions (number of sulfates and acetates) based on the mass–composition relationships derived using the PEN framework. This information facilitates the decoding of a new sequence that is not present in the database. Mass-based searches provide a valuable resource for analytical methods that utilize different MS and tandem MS approaches to characterize GAGs [109]. Searching by composition involves specifying the monosaccharide or disaccharide composition of a GAG. Using this information to query the database results in both sequences and mixtures of GAGs (where composition information is available) that satisfy the search criteria. This search feature is another valuable resource for rapidly uncovering novel GAG sequences and also new biological sources for existing sequences in the database.

14.5.2.2 Motif Search. A motif as it pertains to GAG sequence is defined in many ways. It could be a part of a GAG oligosaccharide sequence or a specific sulfation pattern over a chain length (irrespective of the backbone chemical structure). Searching the database for motifs provides a handle on GAG oligosaccharides with partial sequence information. In such cases, the query motif would include the unambiguous part of the partial sequence. This query would retrieve complete GAG sequences that contain this motif, which would facilitate the design of additional experiments to assign the sequence conclusively. Another important application of the motif search is to improve the understanding of specificity in GAG–protein interactions. Specificity in GAG–protein interactions is typically defined using an oligosaccharide motif with a specific sulfation pattern. For example, the –GlcNS,6X–IdoA2S–GlcNS,6X– motif binds with high affinity to acidic and basic

fibroblast growth factors [126]. Searching the database for this motif results in sequences that predictably bind to these growth factors with high affinity.

14.6 Significance and Future Directions

Understanding the structure–function relationships of GAGs is a fundamental area of research that contributes to the missing links in the overall paradigm of how genotype governs phenotype. The chemical heterogeneity and polydispersity of GAGs arise from their non-template-driven complex biosynthetic machinery. This has complicated the development of tools for the characterization of GAGs, which has in turn confounded attempts to characterize GAG structure–function relationships. The tools for the characterization of GAGs have come a long way in addressing these challenges, as discussed in this chapter. These tools are paving the way to an understanding of the critical relationships between the sequences of GAGs, or attributes that define a GAG mixture, and their functional endpoints that are achieved via specific GAG–protein interactions. The technological advances in the characterization of GAGs permit an integrated *glycomics* approach to GAG structure–function relationships. In the light of these developments, there are two main areas that lie ahead for advancing the glycomics approach to GAGs, which are summarized in the following.

14.6.1 Decoding Diversity of GAGs

This is the most fundamental aspect, analogous to understanding the genome of entire organisms. There is significant diversity in the GAG sequences among different cell types in a specific tissue within an organism. Hence it is important to define and decode the sequence diversity of GAGs in the context of their biosynthetic pathways. Furthermore, it is important to capture the different attributes such as mono- and disaccharide composition in databases to provide a high-quality repository of information on GAG sequences and their biological sources. This provides a vital link to understanding the structure–function relationships of GAGs in the numerous biological processes.

14.6.2 Specificity of GAG–Protein Interactions

Given that proteins recognize and bind to specific oligosaccharide motifs along a GAG chain, the characterization of GAGs is also important in defining the specificity of GAG–protein interactions. Typically, GAGs are involved in assembling multi-meric protein–protein complexes resulting in protein oligomerization or stabilization of protein–receptor and enzyme–inhibitor complexes. Therefore, in addition to determining the oligosaccharide motifs that are recognized by specific proteins, it is also important to understand the relative positioning of these motifs within a chain that would lead to the assembling of active or inactive protein–protein complexes. From the standpoint of understanding the biological activity of a GAG sequence, it is important to characterize longer GAG chains that contain multiple protein binding oligosaccharide motifs. Recently, microarrays containing hundreds of glycan structures were developed by the CFG and other groups [127]. These glycan microarrays are proving to be valuable resources for investigating the glycan binding specificity of numerous proteins in a high throughput

fashion. Similar microarrays containing well-characterized GAG oligosaccharides would be of tremendous utility in understanding the specificity of GAG–protein interactions.

Abbreviations

The anomeric configuration of the sugars shown against each abbreviation is specific to the GAGs covered in this chapter

ΔUA	uronic acid with Δ4,5-double (unsaturated) bond
2S	2- <i>O</i> -sulfate group on uronic acid
3S	3- <i>O</i> -sulfate group on glucosamine
4S	4- <i>O</i> -sulfate group on <i>N</i> -acetylgalactosamine
6S	6- <i>O</i> -sulfate group on glucosamine or galactosamine
NS	<i>N</i> -sulfate group on <i>N</i> -sulfated glucosamine
aManR	anhydro- <i>D</i> -mannitol
CE	capillary electrophoresis
CSGAG	chondroitin sulfate glycosaminoglycan
ECM	extracellular matrix
ESI-MS	electrospray ionization mass spectrometry
GAG	glycosaminoglycan
GalNAc	<i>N</i> -acetyl-β- <i>D</i> -galactosamine
GlcA	β- <i>D</i> -glucuronic acid
GlcN	α- <i>D</i> -glucosamine
GlcNAc	<i>N</i> -acetyl-α- <i>D</i> -glucosamine (β configuration in hyaluronic acid and keratan sulfate)
hep I	heparinase I
hep II	heparinase II
hep III	heparinase III
HSGAG	heparin/heparan sulfate glycosaminoglycan
IdoA	α- <i>L</i> -iduronic acid
MALDI-MS	matrix-assisted laser desorption/ionization mass spectrometry
PEN	property-encoded nomenclature

References

1. Hacker U, Nybakken K, Perrimon N: Heparan sulphate proteoglycans: the sweet side of development. *Nat Rev Mol Cell Biol* 2005, **6**:530–541.
2. Hwang HY, Olson SK, Esko JD, Horvitz HR: *Caenorhabditis elegans* early embryogenesis and vulval morphogenesis require chondroitin biosynthesis. *Nature* 2003, **423**:439–443.
3. Lin X: Functions of heparan sulfate proteoglycans in cell signaling during development. *Development* 2004, **131**:6009–6021.
4. Mizuguchi S, Uyama T, Kitagawa H, Nomura KH, Dejima K, Gengyo-Ando K, Mitani S, Sugahara K, Nomura K: Chondroitin proteoglycans are involved in cell division of *Caenorhabditis elegans*. *Nature* 2003, **423**:443–448.
5. Sugahara K, Mikami T, Uyama T, Mizuguchi S, Nomura K, Kitagawa H: Recent advances in the structural biology of chondroitin sulfate and dermatan sulfate. *Curr Opin Struct Biol* 2003, **13**:612–620.

6. Casu B, Guerrini M, Guglieri S, Naggi A, Perez M, Torri G, Cassinelli G, Ribatti D, Carminati P, Giannini G, Penco S, Pisano C, Belleri M, Rusnati M, Presta M: Undersulfated and glycol-split heparins endowed with antiangiogenic activity. *J Med Chem* 2004, **47**:838–848.
7. Iozzo RV: Basement membrane proteoglycans: from cellar to ceiling. *Nat Rev Mol Cell Biol* 2005, **6**:646–656.
8. Vlodavsky I, Goldshmidt O, Zcharia E, Atzmon R, Rangini-Guatta Z, Elkin M, Peretz T, Friedmann Y: Mammalian heparanase: involvement in cancer metastasis angiogenesis and normal development. *Semin Cancer Biol* 2002, **12**:121–129.
9. Bradbury EJ, Moon LD, Popat RJ, King VR, Bennett GS, Patel PN, Fawcett JW, McMahon SB: Chondroitinase ABC promotes functional recovery after spinal cord injury. *Nature* 2002, **416**:636–640.
10. Chau CH, Shum DK, Li H, Pei J, Lui YY, Wirthlin L, Chan YS, Xu XM: Chondroitinase ABC enhances axonal regrowth through Schwann cell-seeded guidance channels after spinal cord injury. *FASEB J* 2004, **18**:194–196.
11. Holt CE, Dickson BJ: Sugar codes for axons? *Neuron* 2005, **46**:169–172.
12. Inatani M, Irie F, Plump AS, Tessier-Lavigne M, Yamaguchi Y: Mammalian brain morphogenesis and midline axon guidance require heparan sulfate. *Science* 2003, **302**:1044–1046.
13. Dai Y, Yang Y, MacLeod V, Yue X, Rapraeger AC, Shriver Z, Venkataraman G, Sasisekharan R, Sanderson RD: HSulf-1 and HSulf-2 are potent inhibitors of myeloma tumor growth *in vivo*. *J Biol Chem* 2005, **280**:40066–40073.
14. Denholm EM, Lin YQ, Silver PJ: Anti-tumor activities of chondroitinase AC and chondroitinase B: inhibition of angiogenesis proliferation and invasion. *Eur J Pharmacol* 2001, **416**:213–221.
15. Mousa SA: Emerging links between thrombosis inflammation and cancer: role of heparin. *Acta Chir Belg* 2005, **105**:237–248.
16. Sanderson RD, Yang Y, Kelly T, MacLeod V, Dai Y, Theus A: Enzymatic remodeling of heparan sulfate proteoglycans within the tumor microenvironment: growth regulation and the prospect of new cancer therapies. *J Cell Biochem* 2005, **96**:897–905.
17. Sasisekharan R, Shriver Z, Venkataraman G, Narayanasami U: Roles of heparan-sulphate glycosaminoglycans in cancer. *Nat Rev Cancer* 2002, **2**:521–528.
18. Trowbridge JM, Gallo RL: Dermatan sulfate: new functions from an old glycosaminoglycan. *Glycobiology* 2002, **12**:117R–125R.
19. Turnbull J, Powell A, Guimond S: Heparan sulfate: decoding a dynamic multifunctional cell regulator. 2001, *Trends Cell Biol* **11**, 75–82.
20. Fry EE, Lea SM, Jackson T, Newman JW, Ellard FM, Blakemore WE, Abu-Ghazaleh R, Samuel A, King AM, Stuart DI: The structure and function of a foot-and-mouth disease virus–oligosaccharide receptor complex. *EMBO J* 1999, **18**:543–554.
21. Ganesh VK, Smith SA, Kotwal GJ, Murthy KH: Structure of vaccinia complement protein in complex with heparin and potential implications for complement regulation. *Proc Natl Acad Sci USA* 2004, **101**:8924–8929.
22. Mardberg K, Trybala E, Tufaro F, Bergstrom T: Herpes simplex virus type 1 glycoprotein C is necessary for efficient infection of chondroitin sulfate-expressing gro2C cells. *J Gen Virol* 2002, **83**:291–300.
23. Shukla D, Liu J, Blaiklock P, Shworak NW, Bai X, Esko JD, Cohen GH, Eisenberg RJ, Rosenberg RD, Spear PG: A novel role for 3-*O*-sulfated heparan sulfate in herpes simplex virus 1 entry. *Cell* 1999, **99**:13–22.
24. Casu B, Guerrini M, Torri G: Structural and conformational aspects of the anticoagulant and anti-thrombotic activity of heparin and dermatan sulfate. *Curr Pharm Des* 2004, **10**:939–949.
25. Fareed J, Hoppensteadt DA, Bick RL: An update on heparins at the beginning of the new millennium. *Semin Thromb Hemost* 2000, **26**:5–21.
26. Petitou M, Casu B, Lindahl U: 1976–1983, a critical period in the history of heparin: the discovery of the antithrombin binding site. *Biochimie* 2003, **85**:83–89.

27. Shriver Z, Raman R, Venkataraman G, Drummond K, Turnbull J, Toida T, Linhardt R, Biemann K, Sasisekharan R: Sequencing of 3-*O* sulfate containing heparin decasaccharides with a partial antithrombin III binding site. *Proc Natl Acad Sci USA* 2000, **97**:10359–10364.
28. Sundaram M, Qi Y, Shriver Z, Liu D, Zhao G, Venkataraman G, Langer R, Sasisekharan R: Rational design of low-molecular weight heparins with improved *in vivo* activity. *Proc Natl Acad Sci USA* 2003, **100**:651–656.
29. Johnson DJ, Li W, Adams TE, Huntington JA: Antithrombin-S195A factor Xa-heparin structure reveals the allosteric mechanism of antithrombin activation. *EMBO J* 2006, **25**:2029–2037.
30. Carter WJ, Cama E, Huntington JA: Crystal structure of thrombin bound to heparin. *J Biol Chem* 2005, **280**:2745–2749.
31. Li W, Johnson DJ, Esmon CT, Huntington JA: Structure of the antithrombin-thrombin-heparin ternary complex reveals the antithrombotic mechanism of heparin. *Nat Struct Mol Biol* 2004, **11**:857–862.
32. Capila I, Linhardt RJ: Heparin–protein interactions. *Angew Chem Int Ed* 2002, **41**:391–412.
33. Powell AK, Yates EA, Fernig DG, Turnbull JE: Interactions of heparin/heparan sulfate with proteins: appraisal of structural factors and experimental approaches. *Glycobiology* 2004, **14**:17R–30R.
34. Raman R, Sasisekharan V, Sasisekharan R: Structural insights into biological roles of protein–glycosaminoglycan interactions. *Chem Biol* 2005, **12**:267–277.
35. Tumova S, Woods A, Couchman JR: Heparan sulfate proteoglycans on the cell surface: versatile coordinators of cellular functions. *Int J Biochem Cell Biol* 2000, **32**:269–288.
36. Sasisekharan R, Raman R, Prabhakar V: Glycomics approach to structure–function relationships of glycosaminoglycans. *Annu Rev Biomed Eng* 2006, **8**:181–231.
37. Esko JD, Selleck SB: Order out of chaos: assembly of ligand binding sites in heparan sulfate. *Annu Rev Biochem* 2002, **71**:435–471.
38. Sasisekharan R, Venkataraman G: Heparin and heparan sulfate: biosynthesis structure and function. *Curr Opin Chem Biol* 2000, **4**:626–631.
39. Silbert JE, Sugumaran G: Biosynthesis of chondroitin/dermatan sulfate. *IUBMB Life* 2002, **54**:177–86.
40. Sugahara K, Kitagawa H: Heparin and heparan sulfate biosynthesis. *IUBMB Life* 2002, **54**:163–175.
41. Habuchi H, Habuchi O, Kimata K: Sulfation pattern in glycosaminoglycan: does it have a code? *Glycoconj J* 2004, **21**:47–52.
42. Kusche-Gullberg M, Kjellen L: Sulfotransferases in glycosaminoglycan biosynthesis. *Curr Opin Struct Biol* 2003, **13**:605–611.
43. Linhardt RJ, Toida T: Characterization of glycosaminoglycans by capillary electrophoresis. *Methods Mol Biol* 2003, **213**:131–144.
44. Mahoney DJ, Aplin RT, Calabro A, Hascall VC, Day AJ: Novel methods for the preparation and characterization of hyaluronan oligosaccharides of defined length. *Glycobiology* 2001, **11**:1025–1033.
45. Volpi N, Maccari F: Electrophoretic approaches to the analysis of complex polysaccharides. *J Chromatogr B Anal Technol Biomed Life Sci.* 2006, **834**:1–13.
46. Vynios DH, Karamanos NK, Tsiganos CP: Advances in analysis of glycosaminoglycans: its application for the assessment of physiological and pathological states of connective tissues. *J Chromatogr B Anal Technol Biomed Life Sci* 2002, **781**:21–38.
47. Whitelock JM, Iozzo RV: Isolation and purification of proteoglycans. *Methods Cell Biol* 2002, **69**:53–67.
48. Zamfir A, Peter-Katalinic J: Capillary electrophoresis-mass spectrometry for glycoscreening in biomedical research. *Electrophoresis* 2004, **25**:1949–1963.
49. Zaia J: Mass spectrometry of oligosaccharides. *Mass Spectrom Rev* 2004, **23**:161–227.

50. Liu D, Shriver Z, Venkataraman G, El Shabrawi Y, Sasisekharan R: Tumor cell surface heparan sulfate as cryptic promoters or inhibitors of tumor growth and metastasis. *Proc Natl Acad Sci USA* 2002, **99**:568–573.
51. Moon LD, Asher RA, Rhodes KE, Fawcett JW: Regeneration of CNS axons back to their target following treatment of adult rat brain with chondroitinase ABC. *Nat Neurosci* 2001, **4**:465–466.
52. Conrad HE: Nitrous acid degradation of glycosaminoglycans. *Curr Protoc Mol Biol* 2001, Chapter 17, Unit17 22A.
53. Naggi A, Casu B, Perez M, Torri G, Cassinelli G, Penco S, Pisano C, Giannini G, Ishai-Michaeli R, Vlodavsky I: Modulation of the heparanase-inhibiting activity of heparin through selective desulfation graded *N*-acetylation and glycol splitting. *J Biol Chem* 2005, **280**:12103–12113.
54. Chen J, Jones CL, Liu J: Using an enzymatic combinatorial approach to identify anticoagulant heparan sulfate structures. *Chem Biol* 2007, **14**:986–993.
55. Karst NA, Linhardt RJ: Recent chemical and enzymatic approaches to the synthesis of glycosaminoglycan oligosaccharides. *Curr Med Chem* 2003, **10**:1993–2031.
56. Lindahl U, Li JP, Kusche-Gullberg M, Salmivirta M, Alaranta S, Veromaa T, Emeis J, Roberts I, Taylor C, Oreste P, Zoppetti G, Naggi A, Torri G, Casu B: Generation of “neoheparin” from *E. coli* K5 capsular polysaccharide. *J Med Chem* 2005, **48**:349–352.
57. Naggi A, De Cristofano B, Bisio A, Torri G, Casu B: Generation of anti-factor Xa active 3-*O*-sulfated glucosamine-rich sequences by controlled desulfation of oversulfated heparins. *Carbohydr Res* 2001, **336**:283–290.
58. Volpi N: Milligram-scale preparation and purification of oligosaccharides of defined length possessing the structure of chondroitin from defructosylated capsular polysaccharide K4. *Glycobiology* 2003, **13**:635–640.
59. Ernst S, Rhomberg AJ, Biemann K, Sasisekharan R: Direct evidence for a predominantly exolytic processive mechanism for depolymerization of heparin-like glycosaminoglycans by heparinase I. *Proc Natl Acad Sci USA* 1998, **95**:4182–4187.
60. Rhomberg AJ, Shriver Z, Biemann K, Sasisekharan R: Mass spectrometric evidence for the enzymatic mechanism of the depolymerization of heparin-like glycosaminoglycans by heparinase II. *Proc Natl Acad Sci USA* 1998, **95**:12232–12237.
61. Pojasek K, Shriver Z, Hu Y, Sasisekharan R: Histidine 295 and histidine 510 are crucial for the enzymatic degradation of heparan sulfate by heparinase III. *Biochemistry* 2000, **39**:4012–4019.
62. Godavarti R, Sasisekharan R: Heparinase I from *Flavobacterium heparinum*. Role of positive charge in enzymatic activity. *J Biol Chem* 1998, **273**:248–255.
63. Shaya D, Tocilj A, Li Y, Myette J, Venkataraman G, Sasisekharan R, Cygler M: Crystal structure of heparinase II from *Pedobacter heparinus* and its complex with a disaccharide product. *J Biol Chem*. 2006, **281**:15525–15535.
64. Prabhakar V, Raman R, Capila I, Bosques CJ, Pojasek K, Sasisekharan R: Biochemical characterization of the chondroitinase ABC I active site. *Biochem J* 2005, **390**:395–405.
65. Michel G, Pojasek K, Li Y, Sulea T, Linhardt RJ, Raman R, Prabhakar V, Sasisekharan R, Cygler M: The structure of chondroitin B lyase complexed with glycosaminoglycan oligosaccharides unravels a calcium-dependent catalytic machinery. *J Biol Chem* 2004, **279**:32882–32896.
66. Huang W, Boju L, Tkalec L, Su H, Yang HO, Gunay NS, Linhardt RJ, Kim YS, Matte A, Cygler M: Active site of chondroitin AC lyase revealed by the structure of enzyme–oligosaccharide complexes and mutagenesis. *Biochemistry* 2001, **40**:2359–2372.
67. Hamai A, Hashimoto N, Mochizuki H, Kato F, Makiguchi Y, Horie K, Suzuki S: Two distinct chondroitin sulfate ABC lyases. An endoeliminase yielding tetrasaccharides and an exoeliminase preferentially acting on oligosaccharides. *J Biol Chem* 1997, **272**:9123–9130.
68. Myette JR, Shriver Z, Kiziltepe T, McLean MW, Venkataraman G, Sasisekharan R: Molecular cloning of the heparin/heparan sulfate delta 4,5 unsaturated glycuronidase from *Flavobacterium heparinum* its recombinant expression in *Escherichia coli* and biochemical determination of its unique substrate specificity. *Biochemistry* 2002, **41**:7424–7434.

69. Raman R, Myette JR, Shriver Z, Pojasek K, Venkataraman G, Sasisekharan R: The heparin/heparan sulfate 2-*O*-sulfatase from *Flavobacterium heparinum*. A structural and biochemical study of the enzyme active site and saccharide substrate specificity. *J Biol Chem* 2003, **278**:12167–12174.
70. Stern R: Hyaluronan catabolism: a new metabolic pathway. *Eur J Cell Biol* 2004, **83**:317–325.
71. Pikas DS, Li JP, Vlodavsky I, Lindahl U: Substrate specificity of heparanases from human hepatoma and platelets. *J Biol Chem* 1998, **273**:18770–18777.
72. Ernst S, Langer R, Cooney CL, Sasisekharan R: Enzymatic degradation of glycosaminoglycans. *Crit Rev Biochem Mol Biol* 1995, **30**:387–444.
73. Dhoot GK, Gustafsson MK, Ai X, Sun W, Standiford DM, Emerson CP Jr: Regulation of Wnt signaling and embryo patterning by an extracellular sulfatase. *Science* 2001, **293**:1663–1666.
74. Morimoto-Tomita M, Uchimura K, Werb Z, Hemmerich S, Rosen SD: Cloning and characterization of two extracellular heparin-degrading endosulfatases in mice and humans. *J Biol Chem* 2002, **277**:49175–49185.
75. Ai X, Do AT, Kusche-Gullberg M, Lindahl U, Lu K, Emerson CP Jr: Substrate specificity and domain functions of extracellular heparan sulfate 6-*O* endosulfatases QSulf1 and QSulf2. *J Biol Chem*. 2006, **281**:4969–4976.
76. Saad OM, Ebel H, Uchimura K, Rosen SD, Bertozzi CR, Leary JA: Compositional profiling of heparin/heparan sulfate using mass spectrometry: assay for specificity of a novel extracellular human endosulfatase. *Glycobiology* 2005, **15**:818–826.
77. Godavarti R, Davis M, Venkataraman G, Cooney C, Langer R, Sasisekharan R: Heparinase III from *Flavobacterium heparinum*: cloning and recombinant expression in *Escherichia coli*. *Biochem Biophys Res Commun* 1996, **225**:751–758.
78. Pojasek K, Shriver Z, Kiley P, Venkataraman G, Sasisekharan R: Recombinant expression purification and kinetic characterization of chondroitinase AC and chondroitinase B from *Flavobacterium heparinum*. *Biochem Biophys Res Commun* 2001, **286**:343–351.
79. Lunin VV, Li Y, Linhardt RJ, Miyazono H, Kyogashima M, Kaneko T, Bell AW, Cygler M: High-resolution crystal structure of *Arthrobacter aureescens* chondroitin AC lyase: an enzyme–substrate complex defines the catalytic mechanism. *J Mol Biol* 2004, **337**:367–386.
80. Pojasek K, Raman R, Kiley P, Venkataraman G, Sasisekharan R: Biochemical characterization of the chondroitinase B active site. *J Biol Chem* 2002, **277**:31179–31186.
81. Huang W, Lunin VV, Li Y, Suzuki S, Sugiura N, Miyazono H, Cygler M: Crystal structure of *Proteus vulgaris* chondroitin sulfate ABC lyase I at 1.9 Å resolution. *J Mol Biol* 2003, **328**:623–634.
82. Lamari FN, Kuhn R, Karamanos NK: Derivatization of carbohydrates for chromatographic electrophoretic and mass spectrometric structure analysis. *J Chromatogr B Anal Technol Biomed Life Sci* 2003, **793**:15–36.
83. Yamada S, Sakamoto K, Tsuda H, Yoshida K, Sugiura M, Sugahara K: Structural studies of octasaccharides derived from the low-sulfated repeating disaccharide region and octasaccharide serines derived from the protein linkage region of porcine intestinal heparin. *Biochemistry* 1999, **38**:838–847.
84. West LA, Roughley P, Nelson FR, Plaas AH: Sulphation heterogeneity in the trisaccharide (GalNAcSbeta1, 4GlcAbeta1, 3GalNAcS) isolated from the non-reducing terminal of human aggrecan chondroitin sulphate. *Biochem J* 1999, **342** (Pt 1): 223–229.
85. Vives RR, Pye DA, Salmivirta M, Hopwood JJ, Lindahl U, Gallagher JT: Sequence analysis of heparan sulphate and heparin oligosaccharides. *Biochem J* 1999, **339**:767–773.
86. Merry CL, Lyon M, Deakin JA, Hopwood JJ, Gallagher JT: Highly sensitive sequencing of the sulfated domains of heparan sulfate. *J Biol Chem* 1999, **274**:18455–18462.
87. Karlsson NG, Schulz BL, Packer NH, Whitelock JM: Use of graphitised carbon negative ion LC–MS to analyse enzymatically digested glycosaminoglycans. *J Chromatogr B Anal Technol Biomed Life Sci* 2005, **824**:139–147.

88. Thanawiroon C, Rice KG, Toida T, Linhardt RJ: Liquid chromatography/mass spectrometry sequencing approach for highly sulfated heparin-derived oligosaccharides. *J Biol Chem* 2004, **279**:2608–2615.
89. Kuberan B, Lech M, Zhang L, Wu ZL, Beeler DL, Rosenberg RD: Analysis of heparan sulfate oligosaccharides with ion pair-reverse phase capillary high performance liquid chromatography–microelectrospray ionization time-of-flight mass spectrometry. *J Am Chem Soc* 2002, **124**:8707–8718.
90. Karousou EG, Militopoulou M, Porta G, De Luca G, Hascall VC, Passi A: Polyacrylamide gel electrophoresis of fluorophore-labeled hyaluronan and chondroitin sulfate disaccharides: application to the analysis in cells and tissues. *Electrophoresis* 2004, **25**:2919–2925.
91. Plaas AH, West L, Midura RJ, Hascall VC: Disaccharide composition of hyaluronan and chondroitin/dermatan sulfate. Analysis with fluorophore-assisted carbohydrate electrophoresis. *Methods Mol Biol* 2001, **171**:117–128.
92. Calabro A, Midura R, Wang A, West L, Plaas A, Hascall VC: Fluorophore-assisted carbohydrate electrophoresis (FACE) of glycosaminoglycans. *Osteoarthritis Cartilage* 2001, **9** Suppl A: S16–S22.
93. Turnbull JE, Hopwood JJ, Gallagher JT: A strategy for rapid sequencing of heparan sulfate and heparin saccharides. *Proc Natl Acad Sci USA* 1999, **96**:2698–2703.
94. Ruiz-Calero V, Puignou L, Galceran MT: Determination of glycosaminoglycan monosaccharides by capillary electrophoresis using laser-induced fluorescence detection. *J Chromatogr B Anal Technol Biomed Life Sci* 2003, **791**:193–202.
95. Mao W, Thanawiroon C, Linhardt RJ: Capillary electrophoresis for the analysis of glycosaminoglycans and glycosaminoglycan-derived oligosaccharides. *Biomed Chromatogr* 2002, **16**:77–94.
96. Rhomberg AJ, Ernst S, Sasisekharan R, Biemann K: Mass spectrometric and capillary electrophoretic investigation of the enzymatic degradation of heparin-like glycosaminoglycans. *Proc Natl Acad Sci USA* 1998, **95**:4176–4181.
97. Juhasz P, Biemann K: Utility of non-covalent complexes in the matrix-assisted laser desorption/ionization mass spectrometry of heparin-derived oligosaccharides. *Carbohydr Res* 1995, **270**:131–147.
98. Mechref Y, Novotny MV: Matrix-assisted laser desorption/ionization mass spectrometry of acidic glycoconjugates facilitated by the use of spermine as a co-matrix. *J Am Soc Mass Spectrom* 1998, **9**:1293–1302.
99. Schiller J, Arnhold J, Benard S, Reichl S, Arnold K: Cartilage degradation by hyaluronate lyase and chondroitin ABC lyase: a MALDI-TOF mass spectrometric study. *Carbohydr Res* 1999, **318**:116–122.
100. Sturiale L, Naggi A, Torri G: MALDI mass spectrometry as a tool for characterizing glycosaminoglycan oligosaccharides and their interaction with proteins. *Semin Thromb Hemost* 2001, **27**:465–472.
101. Venkataraman G, Shriver Z, Raman R, Sasisekharan R: Sequencing complex polysaccharides. *Science* 1999, **286**:537–542.
102. Behr JR, Matsumoto Y, White FM, Sasisekharan R: Quantification of isomers from a mixture of twelve heparin and heparan sulfate disaccharides using tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2005, **19**:2553–2562.
103. Naggar EF, Costello CE, Zaia J: Competing fragmentation processes in tandem mass spectra of heparin-like glycosaminoglycans. *J Am Soc Mass Spectrom* 2004, **15**:1534–1544.
104. Pope RM, Raska CS, Thorp SC, Liu J: Analysis of heparan sulfate oligosaccharides by nano-electrospray ionization mass spectrometry. *Glycobiology* 2001, **11**:505–513.
105. Saad OM, Leary JA: Compositional analysis and quantification of heparin and heparan sulfate by electrospray ionization ion trap mass spectrometry. *Anal Chem* 2003, **75**:2985–2995.

106. Saad OM, Leary JA: Heparin sequencing using enzymatic digestion and ESI-MSⁿ with HOST: a heparin/HS oligosaccharide sequencing tool. *Anal Chem* 2005, **77**:5902–5911.
107. Yang HO, Gunay NS, Toida T, Kuberan B, Yu G, Kim YS, Linhardt RJ: Preparation and structural determination of dermatan sulfate-derived oligosaccharides. *Glycobiology* 2000, **10**:1033–1039.
108. Zaia J, Costello CE: Compositional analysis of glycosaminoglycans by electrospray mass spectrometry. *Anal Chem* 2001, **73**:233–239.
109. Zaia J, Li XQ, Chan SY, Costello CE: Tandem mass spectrometric strategies for determination of sulfation positions and uronic acid epimerization in chondroitin sulfate oligosaccharides. *J Am Soc Mass Spectrom* 2003, **14**:1270–1281.
110. Kinoshita A, Yamada S, Haslam SM, Morris HR, Dell A, Sugahara K: Isolation and structural determination of novel sulfated hexasaccharides from squid cartilage chondroitin sulfate that exhibits neuroregulatory activities. *Biochemistry* 2001, **40**:12654–12665.
111. Ueoka C, Nadanaka S, Seno N, Khoo KH, Sugahara K: Structural determination of novel tetra- and hexasaccharide sequences isolated from chondroitin sulfate H (oversulfated dermatan sulfate) of hagfish notochord. *Glycoconj J* 1999, **16**:291–305.
112. Kinoshita A, Yamada S, Haslam SM, Morris HR, Dell A, Sugahara K: Novel tetrasaccharides isolated from squid cartilage chondroitin sulfate E contain unusual sulfated disaccharide units GlcA(3-*O*-sulfate)beta1–3GalNAc(6-*O*-sulfate) or GlcA(3-*O*-sulfate)beta1–3GalNAc. *J Biol Chem* 1997, **272**:19656–19665.
113. Sugahara K, Tanaka Y, Yamada S, Seno N, Kitagawa H, Haslam SM, Morris HR, Dell A: Novel sulfated oligosaccharides containing 3-*O*-sulfated glucuronic acid from king crab cartilage chondroitin sulfate K. Unexpected degradation by chondroitinase ABC. *J Biol Chem* 1996, **271**:26745–26754.
114. Chai W, Kogelberg H, Lawson AM: Generation and structural characterization of a range of unmodified chondroitin sulfate oligosaccharide fragments. *Anal Biochem* 1996, **237**:88–102.
115. Chai W, Hounsell EF, Bauer CJ, Lawson AM: Characterization by LSI-MS and ¹H NMR spectroscopy of tetra-, hexa-, and octa-saccharides of porcine intestinal heparin. *Carbohydr Res* 1995, **269**:139–156.
116. Guerrini M, Naggi A, Guglieri S, Santarsiero R, Torri G: Complex glycosaminoglycans: profiling substitution patterns by two-dimensional nuclear magnetic resonance spectroscopy. *Anal Biochem* 2005, **337**:35–47.
117. Guerrini M, Bisio A, Torri G: Combined quantitative ¹H and ¹³C NMR spectroscopy for characterization of heparin preparations. *Semin Thromb Hemost* 2001, **274**:100–123.
118. Yates EA, Santini F, Guerrini M, Naggi A, Torri G, Casu B: ¹H and ¹³C NMR spectral assignments of the major sequences of twelve systematically modified heparin derivatives. *Carbohydr Res* 1996, **294**:15–27.
119. Casu B, Torri G: Structural characterization of low molecular weight heparins. *Semin Thromb Hemost* 1999, **25**:17–25.
120. Guerrini M, Raman R, Venkataraman G, Torri G, Sasisekharan R, Casu B: A novel computational approach to integrate NMR spectroscopy and capillary electrophoresis for structure assignment of heparin and heparan sulfate oligosaccharides. *Glycobiology* 2002, **12**:713–719.
121. Toida T, Hileman RE, Smith AE, Vlahova PI, Linhardt RJ: Enzymatic preparation of heparin oligosaccharides containing antithrombin III binding sites. *J Biol Chem* 1996, **271**:32040–32047.
122. Shriver Z, Sundaram M, Venkataraman G, Fareed J, Linhardt R, Biemann K, Sasisekharan R: Cleavage of the antithrombin III binding site in heparin by heparinases and its implication in the generation of low molecular weight heparin. *Proc Natl Acad Sci USA* 2000, **97**:10365–10370.
123. von der Lieth CW, Bohne-Lang A, Lohmann KK, Frank M: Bioinformatics for glycomics: status methods requirements and perspectives. *Brief Bioinform* 2004, **5**:164–178.

124. Raman R, Raguram S, Venkataraman G, Paulson JC, Sasisekharan R: Glycomics: an integrated systems approach to structure–function relationships of glycans. *Nat Methods* 2005, **2**:817–824.
125. Packer NH, von der Lieth CW, Aoki-Kinoshita KF, Lebrilla CB, Paulson JC, Raman R, Rudd P, Sasisekharan R, Taniguchi N, York WS: Frontiers in glycomics: Bioinformatics and biomarkers in disease An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda, MD (September 11–13, 2006). *Proteomics* 2007, **8**:8–20.
126. Raman R, Venkataraman G, Ernst S, Sasisekharan V, Sasisekharan R: Structural specificity of heparin binding in the fibroblast growth factor family of proteins. *Proc Natl Acad Sci USA* 2003, **100**:2357–2362.
127. Blixt O, Head S, Mondala T, Scanlan C, Huflejt ME, Alvarez R, Bryan MC, Fazio F, Calarese D, Stevens J, Razi N, Stevens DJ, Skehel JJ, van Die I, Burton DR, Wilson IA, Cummings R, Bovin N, Wong CH, Paulson JC: Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc Natl Acad Sci USA* 2004, **101**:17033–17038.

NMR Databases and Tools for Automatic Interpretation of Spectra of Carbohydrates

Claus-Wilhelm von der Lieth

*Formerly at Deutsches Krebsforschungszentrum (German Cancer Research Centre),
Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany*

15.1 Introduction

This section summarizes the concepts of how NMR data are stored, and assigned to atoms in glycan structures, in databases. It describes several approaches to how the stored data can be used for the automatic assignment of NMR spectra, and the estimation of chemical shifts for given structures, in addition to automatic identification of glycan structures using library searches and artificial neural networks. One intriguing property of NMR resonances is that each single chemical shift can be assigned unambiguously to exactly one atom in a given structure. Additionally, the exact value of the chemical shift depends on the atom's chemical environment and is essentially influenced by the type of bonds formed with the directly adjacent atoms. Several computational approaches that make use of these fundamental properties of NMR will be discussed.

15.2 Advantages and Disadvantages of NMR Approaches

In comparison with other analytical methods used for the structure determination of complex carbohydrates, NMR measurements have two major advantages:

- They allow a complete and unambiguous assignment of all structural features of glycans – stereochemistry of monosaccharide units, the type of linkage between connected units and even their conformational preferences – using the same experimental setup. Only the absolute configuration cannot be determined.
- NMR measurements are non-destructive, which has the advantage that the same sample can be used for a variety of NMR experiments and, provided that the carbohydrate does not degrade, produce the same spectrum even after a long period of cold storage.

Since all structural features can be directly detected, glycan structures which have been determined by NMR measurements are generally regarded as highly reliable, although wrong assignments and misinterpretation of NMR spectra obviously occur and are also found in published data. Nevertheless, structural assignments based on NMR measurements can be regarded as quality criteria for the administration and maintenance of glyco-related databases.

The immense wealth of information provided by NMR methods, however, has to be weighed against the intrinsic insensitivity of NMR compared with, for example, mass spectrometry and HPLC approaches. Considerably larger amounts of pure material are required, normally in the micromolar range, which means that glycans at their physiological concentration cannot be determined. Since the biosynthesis of glycans is a non-template-driven process, no biological amplification procedure such as the polymerase chain reaction (PCR) – a molecular biological technique creating multiple copies of DNA without using a living organism – is applicable to glycans. This has the consequence that time-consuming expression systems have to be established and additional enrichment and purification steps are required to obtain sufficient amounts of pure samples that are suitable to produce good NMR spectra. Therefore, the determination of glycan structures using NMR methods is currently only routinely used in areas where it is relatively easy to obtain sufficient amounts of pure material from natural sources and where a complete assignment of all structural features is not possible using other analytical methods. This is often the case for polysaccharides isolated from bacteria and plants, which exhibit a large diversity in their monosaccharide building blocks as well as unusual linkages. Even today, new residues are frequently detected and a detailed structural analysis is required, which can best be done using NMR techniques.

15.3 NMR is Often Used to Determine Bacterial Polysaccharides

Bacteria are often cultured in fermenters containing several tens of liters of medium, where the cells are allowed to grow for days. Depending on the type of polysaccharide, it may be isolated either from the growth medium by precipitation or extracted from the bacterial cell. Often hundreds of milligrams of polysaccharide can be isolated. Since bacteria produce a large variety of glycoconjugates and polysaccharides as part of their cell walls, and since they contain many monosaccharides not found in vertebrates, NMR is the analytical method of choice since it is able to cope with this enormous structural diversity and complexity. In contrast to the identification of *N*-glycans in mammals, for example, where the knowledge of the biosynthetic pathways considerably restricts the potential number of residues occurring and also the number of possible linkages, a complete assignment of all structural features is required in the case of bacterial polysaccharides.

NMR techniques are also intensively applied for the structural determination of plant oligo- and polysaccharides, which also often exhibit a large variety of complex structural features such as microheterogeneity.

15.4 NMR and Informatics Approaches

In the early days of computational chemistry during the 1970s, it was recognized that NMR chemical shifts are well suited for the automated identification of chemical compounds using

library searches, and also for estimation of spectra based on the structure of a compound [1, 2]. The reason for this suitability of NMR data for computational approaches is, on the one hand, that each NMR resonance – the so-called chemical shift (given in parts per million relative to an internal standard) – can often be assigned to exactly one atom in a given structure. However, magnetically equivalent nuclei will give a single resonance but doubling in intensity. Additionally, a single “atom” may give more than one resonance if there is “slow” conformational (or chemical) averaging.

On the other hand, the exact value of the chemical shift depends on its chemical surroundings and is mainly determined by the type of bonds formed with the directly adjacent atoms. The influence of remote atoms decreases with their distance. This knowledge led to the concept of the so-called HOSE (Hierarchical Organization of Spherical Environments) like codes [1], which are chiefly used to predict the NMR spectra of organic molecules. A HOSE code is a canonical string describing the spherical environment of each atom in a given structure. Normally, HOSE codes are automatically generated from connection tables of molecules, which encode the topology of atoms. Typically, HOSE codes with a depth of between four and six spheres are stored in relational databases together with their assigned chemical shift values.

The effort to include all measured and assigned NMR spectra of organic molecules into databases started at the end of the 1970s and continues today. The larger, commercially available collections currently contain several hundred thousand of mainly ^{13}C , and to a lesser extent also ^1H , ^{15}N , ^{17}O , ^{19}F and ^{31}P , NMR spectra, mainly of synthetic organic molecules but also of some natural products. NMRShiftDB [3, 4], an open-content database for chemical structures and assigned NMR shifts, is now freely available on the Internet (www.nmrshiftdb.org). It currently contains more than 20 000 ^{13}C and ^1H NMR spectra. Although the basic philosophy of NMRShiftDB is based on open-access, open-content and open-source principles implemented using modern computational concepts, scientifically its main applications follow the lines worked out by the various predecessors in this field: it offers a spectrum similarity search where a list of experimental peaks is compared with all spectra – or a user-definable subset – of the database and as a result a list of the spectra with best matches is displayed. Alternatively, chemical shifts are predicted for a given structure on the basis of the assigned HOSE code.

15.4.1 *Encoding of Stereochemistry is Required for Carbohydrates*

Nearly all published approaches to predicting NMR shifts encode only the constitution of a molecule and neglect its stereochemistry. The use of a stereo HOSE code has only been described for the CSEARCH approach [5]. The authors described a remarkable improvement between measured and predicted shifts for stereocenters. Unfortunately, the stereo-enhanced version of CSEARCH is not publicly available.

It is well known that stereochemistry has a strong influence on chemical shifts (see Table 15.1). Therefore, as two or more different configurations will have the same HOSE code, the NMR shift estimation of natural products, which normally exhibit several stereocenters, show broad error bars. For carbohydrates, a spherical code which does not take into account stereochemistry is useless. It cannot distinguish between α and β linkages, and corresponding atoms in hexoses (galactose, glucose, mannose, etc.) would always receive the same HOSE code despite their different stereochemical environments. However, as

Table 15.1 ^1H and ^{13}C NMR chemical shifts (in ppm) for atoms in non-reducing end monosaccharides of selected disaccharides. Shift values predicted with CASPER (<http://www.casper.org.au/casper/>) [6, 7] are displayed.

	H1	H2	H3	C1	C2	C3
<i>Anomeric difference</i>						
a-D-Manp-(1-2)-a-D-Manp	5.05	4.07	3.85	102.99	70.87	71.33
b-D-Manp-(1-2)-a-D-Manp	4.75	4.05	3.66	100.37	70.58	73.71
<i>Linkage difference</i>						
a-D-Manp-(1-3)-a-D-Manp	5.11	4.06	3.87	102.97	70.97	71.37
a-D-Manp-(1-4)-a-D-Manp	5.30	3.99	3.84	102.60	71.94	71.36
a-L-Fucp-(1-2)-b-D-Galp	5.20	3.81	3.88	100.74	69.46	70.63
a-L-Fucp-(1-3)-b-D-Galp	5.16	3.80	3.94	101.52	69.40	70.51
a-L-Fucp-(1-6)-b-D-Galp	4.94	3.79	3.84	99.82	69.02	70.62
<i>Configurational difference (at C-2 or C-4)</i>						
a-D-Glcp-(1-3)-b-D-Glcp	5.32	3.58	3.76	99.94	72.60	73.91
a-D-Manp-(1-3)-b-D-Glcp	5.24	4.05	3.86	101.62	71.22	71.42
a-D-Galp-(1-3)-b-D-Glcp	5.31	3.82	3.85	100.13	69.48	70.26
<i>Monosaccharide residues</i>						
a-D-Glcp	5.23	3.54	3.72	92.99	72.47	73.78
a-D-Manp	5.18	3.94	3.86	94.94	71.69	71.25
a-D-Galp	5.22	3.78	3.81	93.18	69.35	70.13
a-L-Fucp	5.20	3.77	3.86	93.12	69.09	70.30

demonstrated in Table 15.1, the chemical shifts vary considerably for the various stereoisomers, with the anomeric configuration, and with linkage type.

A comparison [8] of nine different ^{13}C NMR shift estimation tools revealed, for the anti-tumor agent taxol, a complex natural compound having 13 stereocenters, that the best result is obtained with the traditional spherical encoding approach based on HOSE codes and a large database. For the 47 carbon atoms of taxol, an overall standard deviation of 1.5 ppm between the estimated data and the experimental data was observed. The maximum deviation was 4.8 ppm. This high divergence between estimated and measured chemical shifts can be largely attributed to the complete neglect of any encoding of the stereochemistry, which is especially important for a molecule where more than one-quarter of all carbon atoms represent stereocenters. Other factors may be the neglect of solvents, pH values, and temperature under which the spectra were recorded.

15.4.2 Special Requirements for NMR Databases of Carbohydrates

It is obvious that the main reason why carbohydrates – although being organic molecules – were not included in the small-molecule NMR databases was because the basic design of the underlying data model did not foresee any encoding of stereochemistry. Another obstacle was that the NMR databases offer no suitable way to handle the concept of residues and sequences, which is the normal way to describe biological macromolecules such as DNA, RNA, proteins, and oligosaccharides. A complete topological description for all atoms of a molecule was required as input. This has the consequence that for each individual stereocenter the orientations of the connected atoms would have to be assigned, which is laborious and error-prone when done manually. For each hexose, five stereocenters have to

be input. Molecule builders, which use monosaccharide building blocks for the construction and assignment of carbohydrates (such as GlycoWorkbench; Chapter 13), were not available for a long time. It is therefore not surprising that mainly carbohydrates attached to organic molecules were included in the small-molecule NMR databases. Glycoscientists therefore worked to set up their own specialized database, which fits the nomenclature of glycans more closely.

15.5 SugaBase

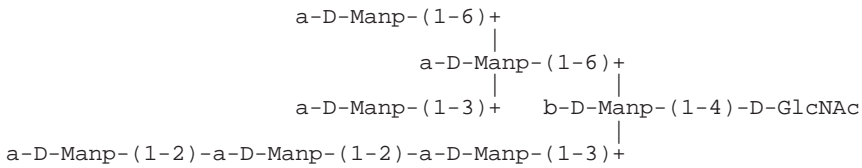
SugaBase was the first attempt to establish an NMR database of carbohydrate structures. It was organized by the Bijvoet Center of the University of Utrecht, The Netherlands, during the early 1990s [9, 10]. SugaBase was thought of as an extension of the Complex Carbohydrate Structure Database (CCSD; often called CarbBank [11, 12]) with NMR data. The link between the two data collections was accomplished using the CCSD ID number. The description of the carbohydrate residue topology was taken directly from CarbBank. SugaBase provides ASCII files where all residues given in the CarbBank representation are subsequently listed as a table. To perform an assignment of the measured resonances to the corresponding atoms in each monosaccharide residue, the commonly used nomenclature to name atoms (e.g. H1 is the H atom covalently connected to C1) is listed together with its measured chemical shift (in ppm) and – if available – to which atom it couples and the coupling constant.

To make the assignment of atoms canonical within a carbohydrate sequence, SugaBase makes two additional assumptions. First, the sequence according to which the monosaccharide residues are listed starts from the reducing end of the carbohydrate chain; this means from the right side of the sugar notation since IUPAC rules clearly state that carbohydrates should be presented from left to right going from the non-reducing to the reducing end. Second, SugaBase introduced a so-called linkage descriptor to permit an unambiguous assignment of each residue. The linkage descriptor lists the linkage types – that means the number of the connected atoms – following the path from the residue under examination towards the reducing end (see Figure 15.1). In such a way, residues in different branches and also at various positions within the chain can be easily and unambiguously identified.

SugaBase was maintained during the first half of the 1990s. It contains about 1300 ^1H and ^{13}C NMR spectra, mainly of *N*-glycan structures found in mammals. NMR peak lists and their assignments to atoms in a carbohydrate structure were extracted from the literature and input manually into the database. SugaBase was first implemented on a local PC, and later made available through a web interface, which remains active today (www.boc.chem.uu.nl/sugabase/sugabase.html). It can be used either to perform simple searches for carbohydrate structures or to retrieve the best-matching spectra when given a list of chemical shifts. However, SugaBase met the same fate as CarbBank: when public funding stopped, the development and maintenance of SugaBase were also reduced to the lowest possible level, which essentially meant keeping the web interface active. Since that time, no attempts – except Bacterial Carbohydrate Structure Database (www.glyco.ac.ru/bcsdb) – have been made to enter systematically the NMR spectra of carbohydrates in databases. Unfortunately, the publishers of scientific journals do not require glycoscientists to deposit their NMR data, prior to publication, in databases where they would be accessible by everyone without any barrier. In effect, this means that most of the primary NMR data are

CC: CCSD:2819
 AU: Spellman MW; Basa LJ; Leonard CK; Chakel JA; O'Connor JV; Wilson S;
 van Halbeek H
 TI: Carbohydrate structures of human tissue plasminogen activator
 expressed in chinese hamster ovary cells
 CT: J Biol Chem (1989) 264: 14100-14111
 FC: 279b5ad9
 SC: B3
 AM: 1H-NMR
 BS: (CN) Chinese hamster, (OT) CHO cells
 SB: van Halbeek H
 DA: 12-02-1990
 MT: N-linked glycoprotein; recombinant glycoprotein
 DB: PIR1:UKHUT; RN:74399-85-2
 PM: tPA, tissue plasminogen activator, human, recombinant
 SI: CBank:17333

 structure:



H#: N-0803-002819
 CC: CCSD:2819
 MHz 500
 Temp 300
 Solv D2O

Residue	Linkage	Proton PPM	
D-GlcNAc		H-1a	5.245
		H-1b	4.72
		NAc	2.043
b-D-Manp	4	H-1	4.77
		H-2(a)	4.244
		H-2(b)	4.232
a-D-Manp	6,4	H-1	4.874
		H-2	4.144
a-D-Manp	6,6,4	H-1	4.911
		H-2	3.98
a-D-Manp	3,6,4	H-1(a)	5.083
		H-1(b)	5.108
		H-2	4.069
a-D-Manp	3,4	H-1	5.345
		H-2	4.118
a-D-Manp	2,3,4	H-1	5.302
		H-2	4.110
a-D-Manp	2,2,3,4	H-1	5.054
		H-2	4.069

Figure 15.1 Display of a SugaBase entry.

lost to the scientific community. This a major difference to the genomics and proteomics area, where scientists are required to deposit their sequences as part of the publication process.

15.5.1 Integration of NMR data of SugaBase into the GLYCOSCIENCES.de Portal

At the end of the 1990s, the NMR data provided by SugaBase were integrated into the GLYCOSCIENCES.de portal (www.glycosciences.de) – the former SWEET-DB [6, 13, 14]. The basic idea of this portal is to make available all experimental data for a given carbohydrate (sub)structure using the same query language and a consistent way of representing the results. About 1000 new NMR spectra extracted from the literature were entered manually following the SugaBase philosophy and concepts. Currently, more than 27 000 ^1H NMR shifts and 14 000 ^{13}C NMR shifts can be recalled [6]. In addition to spectral and structure searches, which are implemented in a similar way to those in SugaBase, the option to estimate ^1H and ^{13}C NMR shifts was made available as a new feature through the GLYCOSCIENCES.de portal [14].

15.6 Spectral Search

To search for spectra, a list of NMR shifts is needed as input. A straightforward comparison of the input shift list with all shifts of all spectra contained in the database is performed. All chemical shifts inside a specified tolerance (normally 0.005 for ^1H and 0.01 for ^{13}C shifts are used) are regarded as hits. A score is calculated, which takes into account the number of matched peaks per spectrum and also the difference between the experimental and library peaks. The search results in a list of hits ordered by decreasing scores and the corresponding spectra and structures are displayed. The matched peaks of a spectrum and the number of matched peaks per residue are indicated by color (Figure 15.2).

Spectral searches work fairly efficiently provided that the spectra of the molecules to be found – or at least a similar structure – are contained in the database. However, if this is not the case, spectral searches will produce less meaningful results. The performance of the matching algorithm depends on the number of peaks of the entered spectrum and will increase linearly with the number of library spectra. With current computer power, more than 1000 library spectra can be compared in less than 1 s of computing time.

15.7 Chemical Shift Estimation

The chemical shifts of glycosyl residues in disaccharides differ from those in the free monosaccharides in a predictable manner. These differences depend mainly on the configuration of the atoms near the glycosidic linkage. These findings have led to the implementation of rule-based approaches to chemical shift estimation. CASPER [7, 15–17] is the most developed program for automatic spectrum interpretation using rule-based approaches (see the following chapter). It is now available through a web interface and is directly integrated into the GLYCOSCIENCES.de portal [6].

Following the general philosophy in glycobiology to describe carbohydrate structures through the topology of their monosaccharide building blocks rather than through an explicit encoding of the topology of all atoms, a residue-based spherical code was developed and implemented in GLYCOSCIENCES.de which is used for NMR shift estimation. Since each monosaccharide describes implicitly the stereochemistry of all stereocenters, the resulting

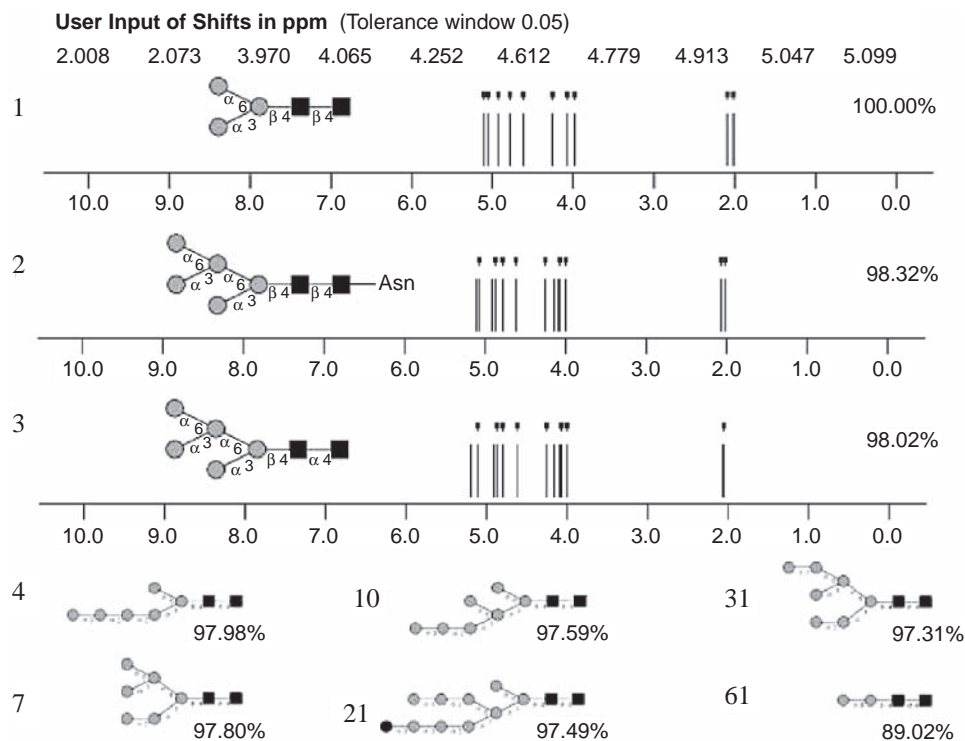


Figure 15.2 Example of a ^1H NMR spectral search as implemented in GLYCOSCIENCES.de [14]. The input shifts are given at the top. They are compared with all shifts in all spectra of the database. A peak hit is assigned whenever the input shift matches a chemical shift of a library spectrum within the user-definable tolerance window. A similarity score for each spectrum is calculated based on the number of matched shifts and their difference in ppm between input and library shift. The library spectra exhibiting a high score are displayed together with the associated glycan structures and the details of the comparison in descending order of the similarity score. Here, the ^1H NMR shifts for the *N*-glycan core as deposited in the database are taken for search. The best-matching spectrum received a similarity of 100%. A schematic presentation of the spectrum is displayed and matched shifts are indicated by small dots above the line indicating a peak. The number within the hit list is given for each retrieved entry.

code reflects well the structural specialities of complex carbohydrates as required for NMR shift estimation.

Tables 15.2 and 15.3 list the environment code used to encode each atom of the *D*-Manp residues of the *N*-glycan given in Figure 15.1. The code's basic principle is discussed for the atoms of β -*D*-Manp (linkage 4) at a branching point of the *N*-glycan. To reflect the fact that closely connected atoms have a larger impact on the chemical shift of the atom looked at, two rules were applied to order the list of attached residues:

1. The connected monosaccharides are ordered according to their increasing distance (in terms of number of bonds) from the atom to be encoded.
2. If two distances are equal, the one connecting to the smaller ring-atom number receives the higher priority (see code for atom H2: the *D*-Glc_pNac at position 1 receives a higher priority than the α -*D*-Manp at position 3).

Table 15.2 Estimation of ^1H NMR chemical shifts for the D-Manp residues in the *N*-glycan given in Figure 15.1. The estimation is based on the database of shifts assigned to a specific structural environment as implemented in the GLYCOSCIENCES.de database [6, 14]. A complete list of estimated shifts is given only for the first two D-Manp residues (linkage 4, and 3, 4). For the other residues, the estimated shifts are given only for those atoms where experimental data are available in GLYCOSCIENCES.de.

Residue	Linkage	Atom	Exp. (ppm)	Est (ppm)	Estimated chemical shift (ppm)			No. of retrieved environment codes	Environment code
					Min.	Max.	SD		
B-D-Manp	4	H1	4.77	4.77	4.760	4.780	0.07	20	H1:(1-4)D-GLCNAC:(3+1)A-D-MANP:(6+1)A-D-MANP
B-D-Manp	4	H2	4.24	4.77	4.135	4.236	0.034	14	H2:(1-4)D-GLCNAC:(3+1)A-D-MANP:(6+1)A-D-MANP
B-D-Manp	4	H3		3.78	3.775	3.748	0.005	2	H3:(3+1)A-D-MANP:(1-4)D-GLCNAC:(6+1)A-D-MANP
B-D-Manp	4	H4		3.79				1	H4:(3+1)A-D-MANP:(6+1)A-D-MANP:(1-4)D-GLCNAC
B-D-Manp	4	H5		3.60				1	H5:(6+1)A-D-MANP:(1-4)D-GLCNAC:(3+1)A-D-MANP
B-D-Manp	4	H6		3.79				1	H6:(6+1)A-D-MANP:(1-4)D-GLCNAC:(3+1)A-D-MANP
A-D-Manp	3,4	H1	5.34	5.34	5.323	5.390	0.009	54	H1:(1-3)B-D-MANP:(2+1)A-D-MANP
A-D-Manp	3,4	H2	4.12	4.10	4.078	4.130	0.014	43	H2:(2+1)A-D-MANP:(1-3)B-D-MANP

(Continued)

Table 15.2 (Continued)

Residue	Linkage	Atom	Exp. (ppm)	Est (ppm)	Estimated chemical shift (ppm)		No. of retrieved environment codes	Environment code
					Min.	Max.		
A-D-Manp	3,4	H3		4.00			1	H3:(2+1)A-D-MANP:(1-3)B-D-MANP
A-D-Manp	3,4	H4		3.69	3.690	3.690	2	H4:(2+1)A-D-MANP:(1-3)B-D-MANP
A-D-Manp	3,4	H5		3.75	3.748	3.760	7	H5:(1-3)B-D-MANP:(2+1)A-D-MANP
A-D-Manp	3,4	H6		3.62	3.625	3.623	5	H6:(1-3)B-D-MANP:(2+1)A-D-MANP
A-D-Manp	2,3,4	H1	5.30	5.30	5.290	5.341	47	H1:(1-2)A-D-MANP:(2+1)A-D-MANP
A-D-Manp	2,3,4	H2	4.11	4.07	4.072	4.115	36	H2:(2+1)A-D-MANP:(1-2)A-D-MANP
A-D-Manp	2,2,3,4	H1	5.05	5.05	5.002	5.067	94	H1:(1-2)A-D-MANP
A-D-Manp	2,2,3,4	H2	4.07	4.05	2.061	4.083	83	H2:(1-2)A-D-MANP
A-D-Manp	6,4	H1	4.87	4.88	4.863	4.918	57	H1:(1-6)B-D-MANP:(3+1)A-D-MANP:(6+1)A-D-MANP
A-D-Manp	6,4	H2	4.14	4.15	4.130	4.180	46	H2:(1-6)B-D-MANP:(3+1)A-D-MANP:(6+1)A-D-MANP
A-D-Manp	3,6,4	H1	5.08	5.11	5.046	5.164	46	H1:(1-3)A-D-MANP
A-D-Manp	3,6,4	H2	4.07	4.07	4.049	4.224	52	H2:(1-3)A-D-MANP
A-D-Manp	6,6,4	H1	4.91	4.91	4.886	4.931	38	H1:(1-6)A-D-MANP
A-D-Manp	6,6,4	H2	3.98	3.97	3.974	4.006	31	H2:(1-6)A-D-MANP

Table 15.3 Estimation of the ^{13}C NMR chemical shifts for the D-Manp residues in the *N*-glycan given in Figure 15.1. Note that the length of some of the environment codes may differ from those given in Table 15.2 because appropriate ^{13}C data for full-length codes were missing from the database and therefore the code was reduced until a hit was found.

Residue	Linkage	Atom	Est. (ppm)	Estimated chemical shift (ppm)			No. of retrieved environment codes	Environment code
				Min	Max	SD		
A-D-Manp	3,4	C1	101.5				1	C1:(1-3)B-D-MANP
A-D-Manp	3,4	C2	79.1	78.80	79.30	0.15	13	C2:(2+1)A-D-MANP
A-D-Manp	3,4	C3	75.6	70.20	71.21	0.38	13	C3:(2+1)A-D-MANP
A-D-Manp	3,4	C4	67.6	67.30	68.01	0.32	13	C4:(2+1)A-D-MANP
A-D-Manp	3,4	C5	77.7				1	C5:(1-3)B-D-MANP
A-D-Manp	3,4	C6	62.4				1	C5:(1-3)B-D-MANP
A-D-Manp	2,3,4	C1	101.1	100.70	102.10	0.49	12	C1:(1-2)A-D- MANP:(2+1)A-D-MANP
A-D-Manp	2,3,4	C2	79.0	78.80	79.2	0.16	10	C2:(2+1)A-D- MANP:(1-2)A-D-MANP
A-D-Manp	2,2,3,4	C1	102.6	102.10	103.00	0.29	7	C1:(1-2)A-D-MANP
A-D-Manp	2,2,3,4	C2	70.8	70.60	71.70	0.37	7	C2:(1-2)A-D-MANP
A-D-Manp	6,4	C1	102.1	101.80	102.2	0.16	6	C1:(3+1)A-D- MANP:(6+1)A-D-MANP
A-D-Manp	6,4	C2	71.0	70.80	71.30	0.21	6	C1:(3+1)A-D- MANP:(6+1)A-D-MANP
A-D-Manp	3,6,4	C1	102.6				1	C1:(1-3)A-D-MANP
A-D-Manp	3,6,4	C2	70.3				1	C2:(1-3)A-D-MANP
A-D-Manp	6,6,4	C1	100.3				1	C1:(1-6)A-D-MANP
A-D-Manp	6,6,4	C2	70.8				1	C2:(1-6)A-D-MANP

For each atom in all structures in the database, the corresponding codes are generated and stored together with the assigned chemical shifts in a new shift-environment table. To perform shift estimation, the corresponding codes for each atom of the input molecule are generated and looked up in the shift-environment table. Depending on the completeness of stored codes for each atom of the input structure, a number of hits with differing shift values can be retrieved from the shift-environment table. The result is reported as the mean value of the stored database entries, including some statistics such as the standard deviation and the highest and lowest shift value for the given code. If no match is found for a complete query environment code, residues are subsequently removed from the end of the environment code until hits are found. The quality of the estimation depends on the completeness of the code for which data have been found, on the number of hits found for each environment code, and of course on the structural diversity and comprehensiveness of the database. Clearly, a growth in the number of assigned spectra in the database and a corresponding increase in the shift-environment table will improve the quality of the shift estimation. Table 15.2 displays the results of the estimation of the ^1H shifts for the D-Manp residues of the N-glycan given in Figure 15.1. Table 15.3 shows the respective ^{13}C shifts.

The NMR shift estimation tool implemented in GLYCOSCIENCES.de allows all data used for the prediction to be viewed. For each environment code, all assigned data can be recalled: a histogram (see Figure 15.3a) displays the shift distribution, the source structure from which each individual shift in the histogram originates is given, and the reference from which the structure and assignment are extracted can be recalled.

An analysis [6] revealed that chemical shift estimation can calculate accurately ^1H and ^{13}C NMR shifts of glycans. In many cases the discrepancy between calculated and experimental chemical shifts is as low as 0.05 ppm/resonance for ^1H and 0.4 ppm/resonance for ^{13}C , which is comparable to the differences between measurements from different laboratories resulting from slightly dissimilar experimental conditions. Such a predictive ability may be sufficient to establish the structure of many oligo- and polysaccharides and is in many cases sufficiently accurate to be used for an automatic assignment of NMR spectra.

15.7.1 Possible Use of NMR Shift Estimation

One obvious use of NMR shift estimation tools will be to monitor the data integrity and quality of data that will enter an NMR database. Since the estimation of single shifts is very fast, all assigned peaks of newly added spectra can be estimated, outliers can be easily detected and checked, and reports can be prepared for the database curator. Since the number of peaks for a given structure is clearly defined, it is straightforward to check also the completeness of a newly entered spectrum. Based on both criteria – does the assigned shift belong to well-populated shift environment-shift distribution, and are shifts for all atoms reported? – rational arguments are available which allow a quality score to be defined for each spectrum.

NMR shift estimation may also be a useful tool for the identification of unknown carbohydrates when the spectra matching approach fails because of the lack of reference spectra in the database. Assuming that the monosaccharide composition of the carbohydrate to be analyzed is known, the combinatorial space of all possible isomers should be reduced considerably, so that within a few seconds of computational time the peaks of all atoms of

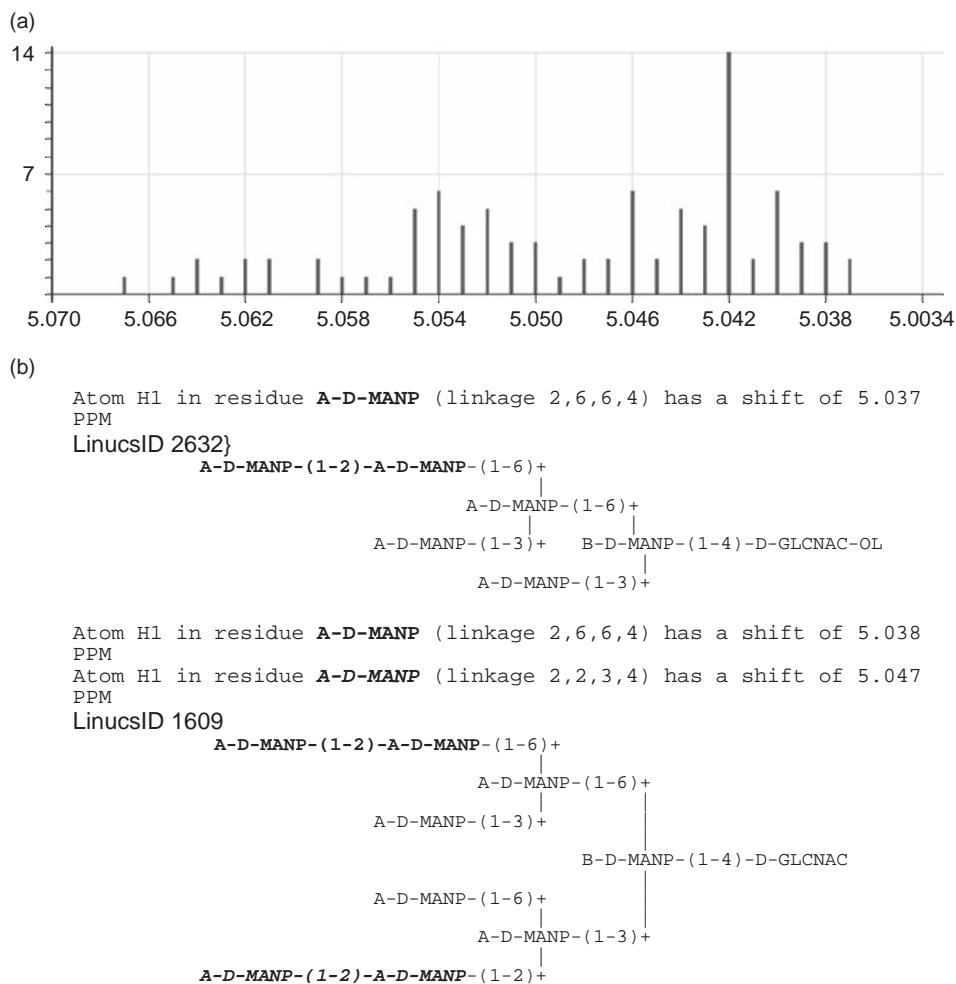


Figure 15.3 (a) Histogram of source values for the estimation of A-D-MANP [H1:(1-2)A-D-MANP] (total number of shifts reported for this code: 94). (b) Two examples of shifts deposited for the structural code A-D-MANP [H1:(1-2)A-D-MANP].

all possible isomers should be estimated and compared with the spectrum of the unknown compound. A list of carbohydrate structures whose estimated spectra best match the experimental spectrum will be the result. In the case of a high congruence of experimental and estimated shifts, an automatic assignment may be possible. A score describing the degree of reliance for each peak may be introduced for this purpose.

It is obvious that such a combinatorial approach will work best for smaller carbohydrates, where the number of possible isomers does not become too large. However, if the unknown carbohydrate can be assigned to one of the well-known classes of compounds, such as *N*- and *O*-glycans or glycolipids, knowledge of biological pathways can also be taken into account to reduce the number of possible isomers.

15.8 Use of Spectral Matching and Shift Estimation in Automatic Procedures

Spectral matching and shift estimation have in common that they perform well; thousands of library spectra can be compared and thousands of shifts can be estimated within a few seconds. Both algorithms are simple and robust and need no, or only moderate, human interference when applied to a broad variety of carbohydrate structures. Both algorithms can be used as part of automatic procedures. A major limitation for the spectral matching approach is that reliable results can only be expected when spectra of similar compounds are contained in the database. Consequently, one has to ascertain that the expected structures are indeed present before applying automatic library search procedures. Also, the range of scores, which are indicative for a certain class of compounds, should be known.

The shift estimation also depends to some (although to a lesser) extent on the structural comprehensiveness and diversity of the available environment codes and assigned shifts. If no accurate structural code is available, less specific ones will be used and the predictions will become less reliable.

References

1. Bremser W: HOSE – a novel substructure Code. *Anal Chim Acta* 1978, **103**:355–365.
2. Bremser W: Expectation ranges of ^{13}C NMR chemical shifts. *Magn Reson Chem* 1985, **23**:271–275.
3. Steinbeck C, Krause S, Kuhn S: NMRShiftDB – constructing a free chemical information system with open-source components. *J Chem Inf Comput Sci* 2003, **43**:1733–1739.
4. Steinbeck C, Kuhn S: NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* 2004, **65**:2711–2717.
5. Schütz V, Purtuc V, Felsing S, Robien W: CSEARCH-STEREO: a new generation of NMR database systems allowing three-dimensional spectrum prediction. *Fresenius' J Anal Chem* 1997, **359**:33–41.
6. Loss A, Stenutz R, Schwarzer E, von der Lieth C: GlyNest and CASPER: two independent approaches to estimate ^1H and ^{13}C NMR shifts of glycans available through a common web-interface. *Nucleic Acids Res* 2006, **34** (Web Server issue): W733–W737.
7. Jansson P, Stenutz R, Widmalm G: Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel web-based version of the computer program CASPER. *Carbohydr Res* 2006, **341**:1003–1010.
8. Meiler J, Maier W, Will M, Meusinger R: Using neural networks for ^{13}C NMR chemical shift prediction – comparison with traditional methods. *J Magn Reson* 2002, **157**:242–252.
9. van Kuik J, Vliegthart J: Databases of complex carbohydrates. *Trends Biotechnol* 1992, **10**:182–185.
10. van Kuik J, Hard K, Vliegthart J: A ^1H NMR database computer program for the analysis of the primary structure of complex carbohydrates. *Carbohydr Res* 1992, **235**:53–68.
11. Doubet S, Bock K, Smith D, Darvill A, Albersheim P: The Complex Carbohydrate Structure Database. *Trends Biochem Sci* 1989, **14**:475–477.
12. Doubet S, Albersheim P: CarbBank. *Glycobiology* 1992, **2**:505.
13. Loss A, Bunsmann P, Bohne A, Loss A, Schwarzer E, Lang E, von der Lieth C: SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res* 2002, **30**:405–408.

14. Lütteke T, Bohne-Lang A, Loss A, Götz T, Frank M, von der Lieth C: GLYCOCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* 2006, **16**:71R–81R.
15. Jansson P, Kenne L, Widmalm G: CASPER – a computerised approach to structure determination of polysaccharides using information from n.m.r. spectroscopy and simple chemical analyses. *Carbohydr Res* 1987, **168**:67–77.
16. Jansson P, Kenne L, Widmalm G: Computer-assisted structural analysis of polysaccharides with an extended version of CASPER using proton and carbon-13 NMR data. *Carbohydr Res* 1989, **188**:169–191.
17. Jansson P, Kenne L, Widmalm G: CASPER: a computer program used for structural analysis of carbohydrates. *J Chem Inf Comput Sci* 1991, **31**:508–516.

16 Automatic Spectrum Interpretation Based on Increment Rules: CASPER

Roland Stenutz

Department of Organic Chemistry, Stockholm University, 106 91 Stockholm, Sweden

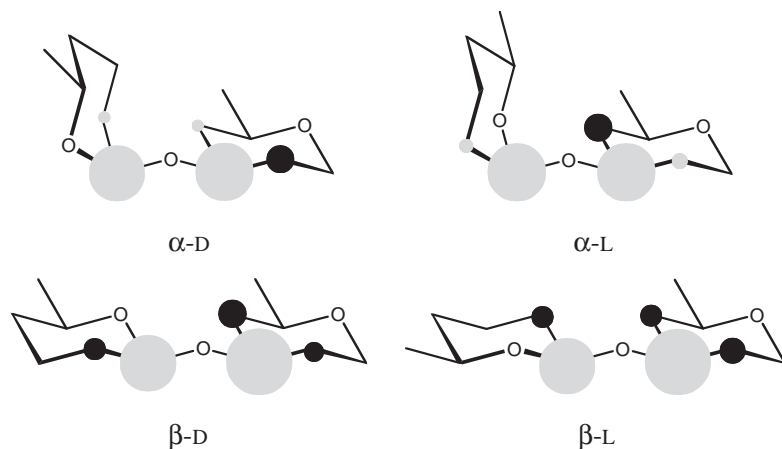
16.1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is indispensable for the structure determination of natural products. A complete structure elucidation will normally include several different 1D and 2D NMR experiments and a lengthy interpretation. Different computerized methods have been developed to speed up the process by extracting structural information from 1D spectra. One of these approaches is the prediction of NMR chemical shifts from a structure of the molecule under investigation. Methods based on topological (two-dimensional [1]) descriptors are frequently used to predict the NMR spectra for “small” organic compounds [2–4]. Often stereochemistry, which is one of the most important aspects of carbohydrates, is neglected. Recent attempts to include information about geometric isomers have improved the performance for unsaturated and cyclic compounds [5, 6]. Other methodologies base their calculations on the three-dimensional structure of a molecule [7–9] and, thus, take into account stereochemistry, but this requires an all-atom model and is not practical for polymers and other large structures. It has also been noted that a single 3D structure in many cases is insufficient to allow accurate predictions of chemical shifts and that conformational averaging has to be taken into account [10]. Knowledge about the conformation of the molecule under investigation is often unavailable for flexible molecules such as glycans. Finally, solvation, which is neglected in these approaches, can be of importance for the conformation of glycans and hence also affects their NMR properties. There are some reports of methods based on 3D models, quantum mechanical or empirical, applied to mono- [11], di- [12], and trisaccharides [13], but they have not found wider use.

The most striking difference between biopolymers, such as glycans, and other natural products is that they consist of monomers that can be linked only in a limited number of ways. It is therefore convenient to describe them as a sequence of monomers and to calculate the NMR spectra of the polymers using displacements from the chemical shifts of the corresponding monomers. In the case of proteins, the changes in chemical shifts are sensitive to the secondary structure [14] and, using sequence data obtained by other methods, may be used to determine secondary structure. In contrast, the chemical shift changes in

Table 16.1 Comparison between ^{13}C glycosylation shifts in some α -D-, β -D-, α -L- and β -L-linked disaccharides.

	C1'	C2'	C2	C3	C4
α -D-Glcp-(1 \rightarrow 3)- α -D-GlcpOMe	6.96	0.19	-1.40	7.41	0.15
α -L-Fucp-(1 \rightarrow 3)- α -D-GlcpOMe	7.30	0.24	0.35	7.55	-1.48
β -D-Glcp-(1 \rightarrow 3)- α -D-GlcpOMe	6.83	-0.78	-0.62	9.64	-1.55
β -L-Fucp-(1 \rightarrow 3)- α -D-GlcpOMe	6.91	-0.82	-1.31	9.55	-0.80



The area of the circles is proportional to the glycosylation shift. Gray circles indicate positive and black circles negative values

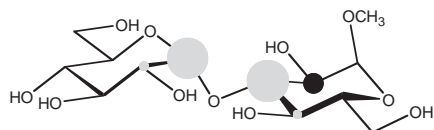
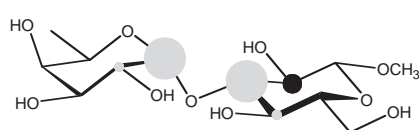
carbohydrate residues are mainly determined by the stereochemistry at the glycosidic linkage and thus the sequence of the residues (Table 16.1). The trends in substituent-induced chemical shifts have been studied in glycosides[15] and in disaccharides [16]. Together with algorithms for structure generation and comparison with experimental data, this provides a means to identify unknown compounds [17–19].

16.2 Chemical Shift Calculation

The chemical shifts of glycosyl residues in an oligo- or polysaccharide differ from those of the free monosaccharides in a predictable manner. These differences, referred to as *glycosylation shifts*, depend mainly on the configuration of the atoms near the glycosidic linkage. Differences in configuration at positions remote to the glycosidic linkage have only a small influence on the ^{13}C glycosylation shifts and hence similar glycosylation shifts are found, for example, for α -D-Glcp-(1 \rightarrow 3)- α -D-GlcpOMe and α -D-Fucp-(1 \rightarrow 3)- β -D-GlcpOMe (Table 16.2). Thus, the glycosylation shifts are transferable between similar linkages. They are also additive provided that there are no steric interactions between residues further removed in sequence. However, such interactions can occur in 2-linked aldoses and vicinally, that is, 2,3- or 3,4- disubstituted residues [18]. Since these interactions can cause significant (up to 6 ppm in ^{13}C NMR), mainly upfield, deviations from additivity, corrections should be made if possible. Proton chemical shifts are more sensitive to changes in the chemical environment than ^{13}C chemical shifts and sometimes different chemical

Table 16.2 ^{13}C glycosylation shifts for $\alpha\text{-D-Glcp-(1}\rightarrow\text{3)-}\alpha\text{-D-GlcpOMe}$ and $\alpha\text{-D-Fucp-(1}\rightarrow\text{3)-}\beta\text{-D-GlcpOMe}$.

	C1'	C2'	C2	C3	C4
$\alpha\text{-D-Glcp-(1}\rightarrow\text{3)-}\alpha\text{-D-GlcpOMe}$	6.96	0.19	-1.40	7.41	0.15
$\alpha\text{-D-Fucp-(1}\rightarrow\text{3)-}\beta\text{-D-GlcpOMe}$	7.24	0.20	-1.20	7.14	0.24

 $\alpha\text{-D-Glcp-(1}\rightarrow\text{3)-}\alpha\text{-D-GlcpOMe}$  $\alpha\text{-D-Fucp-(1}\rightarrow\text{3)-}\beta\text{-D-GlcpOMe}$

The area of the circles is proportional to the glycosylation shift. Gray circles indicate positive and black circles negative values

shifts can be observed for the α - and β -anomers of reducing oligosaccharides even at sites remote in sequence to the anomeric center.

To calculate the chemical shifts of a glycan, one starts with the chemical shifts of the individual monosaccharides (Table 16.3). Then the glycosylation shifts (substituent-induced chemical shifts) for all linkages are added. If there are linkages at neighboring positions then a correction, if available, is added.

Although approaches based on glycosylation shifts may appear unsophisticated compared with atom-based methods, they do provide a mechanism for the implicit inclusion of effects on the chemical shifts caused by conformational factors and differences in solvation. In favorable cases they are as accurate as databases and give errors no larger than those commonly caused by differences in experimental conditions, about 0.2 ppm per ^{13}C chemical shift – this is about 5–10 times better than predictions made by topology-based approaches [5]. The performance for ^1H NMR chemical shifts is less impressive (~ 0.06 ppm/chemical shift), perhaps because of the greater sensitivity to changes in remote parts of the glycan and also to differences in experimental conditions. This may also be due to the smaller amount of work that has been done on improving the performance of the ^1H NMR prediction and the smaller dispersion of the chemical shifts. From experience, an average error of about

Table 16.3 Calculation of the ^{13}C chemical shifts of $\beta\text{-D-Glcp}$ in $\beta\text{-3-O-fucosyllactose}$, $\alpha\text{-L-Fucp-(1}\rightarrow\text{3)-}\beta\text{-D-Galp-(1}\rightarrow\text{4)-}\beta\text{-D-Glcp}$.

1. Start with the chemical shifts of the corresponding monosaccharide:

	C1	C2	C3	C4	C5	C6
$\beta\text{-D-Glcp}$	96.84	75.20	76.76	70.71	76.76	61.84

2. Add glycosylation shifts for each linkage:

$\alpha\text{-L-Fucp-(1}\rightarrow\text{3)-}\beta\text{-D-Glcp}$	-0.21	0.31	7.19	-1.38	-0.02	0.01
$\beta\text{-D-Galp-(1}\rightarrow\text{4)-}\beta\text{-D-Glcp}$	0.17	-0.22	-1.46	9.19	-1.14	-0.56
Uncorrected result	96.46	75.29	82.49	78.52	75.60	61.29

3. Add corrections for vicinal substitution (if available):

	0.27	1.21	-4.20	-4.60	0.73	-0.37
Final chemical shifts	96.73	76.50	78.29	73.92	76.33	60.92

0.4 ppm in the prediction of ^{13}C chemical shifts is still acceptable whereas the ^1H data have to be accurate to better than 0.05 ppm to be useful.

16.3 Structure Generation

Manual structure elucidation using spectroscopic methods normally starts with the interpretation of experimental data to give successively larger fragments. These are then finally joined to form a structure proposal. Computational approaches, on the other hand, often proceed in the opposite direction starting with a large set of possible structures and reduce this set until only a few structures remain. Since the number of glycan structures is unlimited, some *a priori* constraints, such as the number and nature of the glycosyl residues, have to be imposed. Depending on the type of glycan, these constraints may vary.

CASPER [17, 20–23] was originally conceived as an aid for the structure elucidation of bacterial polysaccharides composed of repeating units. Both capsular polysaccharides and the *O*-antigens of lipopolysaccharides belong to this category. They have a backbone consisting of about 4–5 residues and branches consisting of a single residue are common. However, the scope of the program has been extended several times and there are no longer any restrictions on the number of branches. Even the highly branched *N*-linked oligosaccharides of glycoproteins can be handled.

The algorithm used for structure generation in CASPER is very flexible and allows any chemically reasonable, that is, fully connected, structure to be generated. In order to limit the number of structures produced, and hence the time required, a number of restrictions can be imposed (see below). Trial structures are generated by selecting sets of mutually exclusive residues, often the α - and β -anomers of a residue, together with information about the positions available for bonding. For each combination of residues, all different linkage variations are made. This is achieved by creating linkages until all anomeric positions are used or until there are no more positions that can be linked (Figure 16.1). Normally, information from component and linkage analyses is used to limit the number of possible glycosyl residues and linkage positions. These are the two factors that have the greatest impact on the number of generated structures since the number of possible residues exceeds 100 and each has 3–4 potential linkage positions.

The anomeric coupling constants, both $^1J_{\text{C,H}}$ values and $^3J_{\text{H,H}}$ values, can also be used as restraints. The $^1J_{\text{C,H}}$ values can be defined as being small (<169 Hz) or large (>169 Hz) corresponding to axial or equatorial H1 in pyranoses. The $^3J_{\text{H,H}}$ values are divided into small (<3 Hz), medium (~4 Hz) and large (>7 Hz), corresponding to the couplings observed in α/β -manno-, α -gluco- and β -gluco-pyranoses, respectively. The linewidth in the NMR spectra of polysaccharides often makes it difficult to distinguish between α - and β -manno-pyranoses (1.6 versus 0.9 Hz) and therefore their $^3J_{\text{H,H}}$ values have been grouped together. Since there may be problems in obtaining coupling constants that can be unambiguously classified and since some compounds such as furanoses and ketoses do not fit into these groups, this information is used to set a lower limit on the number of residues with certain coupling constants. It should be noted that the difference between the chemical shifts in α - and β -linked residues alone is sufficient to determine the anomeric configuration. Hence the main advantage of using coupling constants is to reduce the number of structures generated and cut the simulation time. A typical bacterial polysaccharide (five residues) with known components (but without known anomeric configurations) and linkage positions

1) Sets of mutually exclusive residues (often α - and β -forms of the same glycosyl residue) are created.



2) One residue is selected from each set (16 possible combinations)



3) The residues are connected in all possible ways (12 different sequences)



4) Repeat 3 and 4 until all possibilities have been exhausted (total of $16 \times 12 = 192$ structures)

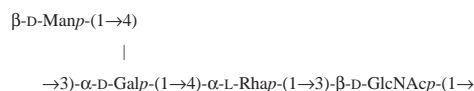
Figure 16.1 Generation of structures from components (e.g. *E. coli* O75 O-antigen).

will generate about 1000 structures, but the number of possibilities increases rapidly with the number of residues (Figure 16.2).

A slightly different approach to the generation of trial structures is taken by the BIOPSEL program [24]. There, all the allowed topologies are hardwired and restricted to polysaccharides. A maximum of six residues is allowed and the correct anomeric configuration of each residue must also be supplied, whereas the linkage position of the residues may be omitted.

16.4 Comparing Computed and Experimental Data

With methods for the generation of glycan structures and for the calculation of NMR spectra at hand, only a method for ranking is needed to give a working system for automated structure elucidation.



Constraints	Generated structures
All residues, linkage positions & $^1J_{\text{C1,H1}}$ values	12
All residues & linkage positions	192
All residues; linkage positions of GlcNAc undefined	3360
All residues; linkage positions of GlcNAc and Rha undefined	15360

Figure 16.2 Influence of constraints on the number of structures generated for the repeating unit of the *E. coli* O75 O-antigen.

Simulated and experimental 1D NMR data can be compared in different ways. The method originally used by CASPER was a simple one-to-one comparison between sorted chemical shifts. The rank was then based on the absolute value of the difference between the calculated and measured chemical shifts. This approach is very rapid but only works if the simulated and experimental spectra contain the same number of resonances. Often it is difficult to obtain experimental data of sufficient quality to identify all resonances clearly. In such cases, it is possible to let the program find the ‘best’ assignment of the chemical shifts by omitting one or more of the computed chemical shifts. This procedure takes a significantly longer time since many different assignments are tested. In either case, the error calculated is the smallest sum of the absolute differences between experimental and calculated chemical shifts. The actual error may be larger since the assignments are always made to reduce the total error and are not necessarily correct. The resulting order of the structures is similar to that obtained when root mean square deviations are used (as in BIOPSEL [24]). Spectral databases often use a different method for matching where several experimental values may be matched by the same database value and no penalty is added for unmatched database peaks. It was found that this approach made less distinction between structures.

16.5 Databases

The information that is required by CASPER for generating structures and calculating NMR chemical shifts is stored in external files that are loaded at run time. For every type of residue there is a description of the spin system and allowed linkage positions. The connectivity of the carbon and hydrogen atoms, which is required for the simulation of 2D spectra, is also stored in the same data structure. Separate tables contain one-to-one correspondences of atoms in different residue types that, for example, allow chemical shift differences from a 6-deoxyhexose to be used for a 2-acetamido-2-deoxyhexose. The chemical shifts, coupling constants related to the anomeric protons and stereochemistry are stored individually for each glycosyl residue/monosaccharide. The residue data also contain links to relevant glycosylation shifts and to the description of the connectivity. Links to other residues with similar stereochemistry at the linkage position are also stored. For example, the values for α -D-Glcp may be used for C1 of α -D-Galp (C4 epimer) or for C6 of α -D-Manp (C2 epimer). This information is needed to find approximate values for missing glycosylation shifts. Because of the large number of possible steric corrections, used to compensate for steric interactions between adjacent residues, they are stored based on the relation between the absolute configuration of the three residues, the axial/equatorial disposition of the glycosidic linkages, and the linkage positions. If more than one correction set is found, then the one having the most similar glycosidic linkage is used.

Since it is not practical to obtain experimental values for all glycosylation shifts and steric corrections, an exact match might not be found and approximations are often used. When an exact match for a fragment or its enantiomer is not found, then the database is searched for linkages that have one or both residues replaced by a default residue. If more than one default value is found then the set of glycosylation shifts is used that is judged from structural similarity and data quality to have the highest accuracy. The data for glycosylation shifts and steric corrections can be stored as such, or as chemical shifts for di- and trisaccharides, and

Table 16.4 Glycosyl residues supported by the CASPER web interface. All residues are in the pyranose form.

Hexoses	D-Galp, D-Glcp, D-Manp
6-Deoxyhexoses	D-Fucp, L-Fucp, D-Rhap, L-Rhap, D-Quip
3,6-Dideoxyhexoses	Abep, Colp, Parp, Tyyp, Asc p
2-Acetamido-2-deoxyhexoses	L-FucpNAc, D-GalpNAc, D-GlcpNAc, D-ManpNAc, MurpNAc, L-RhapNAc, D-QuipNAc
Uronates	D-GalpA, D-GalpANA, D-GlcpA, D-ManpA, D-ManpANA

extracted at run time. This facilitates the maintenance of the databases since the additional step of manually extracting glycosylation shifts and steric corrections is avoided. Changes in the monosaccharide data are also automatically passed on to glycosylation shifts and changes in disaccharide data are passed on to the steric corrections, which simplify database updates. The current database contains approximately 75 glycosyl residues belonging to 30 different types, 150 glycosylation shifts and 70 steric corrections. Twenty-five residues are supported by the web interface (www.casper.org.au.se/casper) [22] (Table 16.4). The content of the databases has either been recorded at our laboratory under standardized conditions (D₂O solvent, alkaline pD, 70 °C, dioxane $\delta_C = 67.40$ and TSP $\delta_H = 0.00$) or has been taken from the literature and adjusted to the same scale to ensure consistency.

16.6 The CASPER Program

The program is written in the C programming language and is approximately 14 000 lines long. Currently the database consists of 70 files. A command line interface allows the user to access all the functions of the program including the databases. This interface is rather complex, even though the loading of the database files and other repetitive tasks is automated through the use of scripts, and therefore a web interface was designed [22] to handle simpler tasks (<http://www.casper.org.au.se/casper>). Two separate forms are available for the simulation of the NMR data of a known structure, and for sequence analysis using information from component and linkage analysis, and NMR data. User input is checked by client-side procedures for consistency before it is submitted to the CASPER server. To prevent misuse, the data are checked again on the server, converted into a script for CASPER and then executed. The results are then converted into HTML and returned. Since a rather simple interface is presented to the user and the design itself prevents some common errors, productivity is greatly increased. Using the web interface also allows users at other locations to access CASPER without the need to install the program.

The CASPER program can be used in a variety of ways. The simplest application is to use it as a collection of NMR data for mono- and disaccharides. Although this does not make use of any of the program's computational capabilities, the usefulness of such a database should not be underestimated. One need only consider the wide use of chemical shift collections such as that of Bock and Pedersen [25], which has been cited in more than 1000 publications.

In cases where the structure, or a likely structure, of a glycan is known, the program can be used to calculate the NMR spectrum. The glycan structure can be entered either residue by residue or by using a line notation similar to the IUPAC-IUBMB extended

form. If an experimental spectrum is supplied, this can be assigned by CASPER. In this way, a hypothetical structure or fragment can be tested rapidly and a starting point for the manual interpretation of a spectrum is obtained. Finally, all of the program features can be used and the sequence of glycosyl residues in an oligo- or polysaccharide determined from information about the components and linkages.

16.7 Performance

The performance of the web version of CASPER has been extensively tested [22]. The published ^{13}C NMR data for 200 different oligo- and polysaccharides could be simulated with an average error of 0.3 ppm/resonance. A more detailed analysis of 44 LPS spectra from *Escherichia coli* showed that the correct structure was ranked first or second in 50% of the cases. With a set of oligosaccharides of mammalian origin, the performance was even better and the correct structure was always among the two structures with highest rank. Presumably the glycosyl residues in these oligosaccharides are better represented in the database, resulting in higher accuracy in the simulation.

The results using ^1H NMR spectra are less encouraging, in part because of the greater sensitivity of protons to experimental conditions and steric interactions with residues remote in the sequence, but also because of the smaller dispersion of the chemical shifts. The use of $^3J_{\text{H,H}}$ values is currently restricted to the anomeric positions but might be useful as an additional restraint in the assignment of ^1H resonances.

It is not possible to give any general rule as to when the best CASPER structure is correct since this depends on how similar the spectra are which different structures give rise to. No proposed structure should be excluded unless there is a significant difference in the calculated fit to the next ranked structure.

Although there is room for increasing the accuracy of the spectral simulations and extending the set of residues that can be handled, the most urgent area for improvement is the generation of trial structures and interfacing with other glycomics tools. Our knowledge of the biosynthetic pathways, in particular of *N*-glycans, has reached a point where it should be possible to generate only structures that are compatible with known pathways [26, 27]. It should also be possible to make an initial selection of structures either from a database, or from structures generated by other automated methods of analysis, such as mass spectrometry. Such developments require a standardized description of carbohydrate structure and also of the experimental results to allow easy transfer of data. This may increase the accuracy in the structure predictions and allow structures that are difficult to determine by any one method to be elucidated. An advantage of more accurate structure prediction is that chemical analyses might be avoided altogether, which would greatly reduce the time and amount of sample required.

16.8 Conclusion

The ^{13}C NMR spectra of carbohydrates can be calculated accurately from glycosylation shifts and corrections for steric effects. In many cases, the discrepancy between calculated and experimental chemical shifts is as low as 0.2 ppm/resonance, which is comparable to the difference between measurements from different laboratories resulting from slightly dissimilar experimental conditions. This is sufficient to establish the structures of many

oligo- and polysaccharides. Programs that combine chemical shift calculations with algorithms for the generation of trial structures provide a powerful tool for structure elucidation. However, it should be emphasized that any previously unknown structures should be verified by additional spectroscopic and, amount of sample permitting, chemical analyses. A significant number of structures in the literature have been subject to revision and great care has to be taken before using published structures as, for example, a starting point for synthesis.

The addition of a web interface and also the ongoing work to connect CASPER to other glycomics databases and tools on the Internet are likely to make the program a valuable tool.

References

1. Bremser W: Expectation ranges of ^{13}C NMR chemical shifts. *Magn Resonan Chem* 1985, **23**: 271–275.
2. Bürgin Schaller R, Munk ME, Pretsch E: Spectra estimation for computer-aided structure determination. *J Chem Inf Comput Sci* 1996, **36**: 239–243.
3. Pretsch E, Fürst A, Badertscher M, Bürgin R: C13Shift: a computer program for the prediction of ^{13}C NMR spectra based on an open set of additivity rules. *J Chem Inf Comput Sci* 1992, **32**: 291–295.
4. Steinbeck C, Kuhn S: NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* 2004, **65**:2711–2717.
5. Satoh H, Koshino H, Uzawa J, Nakata T: CAST/CNMR: highly accurate ^{13}C NMR chemical shift prediction system considering stereochemistry. *Tetrahedron* 2003, **59**:4539–4547.
6. Schütz V, Purtuc V, Felsinger S, Robien W: CSEARCH-STEREO: a new generation of NMR database systems allowing three-dimensional spectrum prediction. *Fresenius' J Anal Chem* 1997, **359**:33–41.
7. Abraham R: A model for the calculation of proton chemical shifts in non-conjugated organic compounds. *Prog Nucl Magn Reson Spectrosc* 1999, **35**:85–152.
8. de Dios A: *Ab initio* calculations of the NMR chemical shift. *Prog Nucl Magn Reson Spectrosc* 1996, **29**:229–278.
9. Helgaker T, Jaszunski M, Ruud K: *Ab initio* methods for the calculation of NMR shielding and indirect spin–spin coupling constants. *Chem Rev* 1999, **99**:293–352.
10. Mazeau K, Taravel F, Tvaroska I: Angular dependence of the C-6 chemical shift and the conformation of the hydroxymethyl group in carbohydrates. *Chem Pap Chem Zvesti* 1996, **50**:77–83.
11. McIntyre M, Small G: Carbon-13 nuclear magnetic resonance spectrum simulation methodology for the structure elucidation of carbohydrates. *Anal Chem* 1987, **59**:1805–1811.
12. Small G, McIntyre M: Structure elucidation methodology for disaccharides based on carbon-13 nuclear magnetic resonance spectrum simulation. *Anal Chem* 1989, **61**:666–674.
13. Clouser D, Jurs P: Simulation of the ^{13}C nuclear magnetic resonance spectra of trisaccharides using multiple linear regression analysis and neural networks. *Carbohydr Res* 1995, **271**: 65–77.
14. Szilágyi L: Chemical shifts in proteins come of age. *Prog Nucl Magn Reson Spectrosc* 1995, **27**:325–443.
15. Lemieux R, Koto S: The conformational properties of glycosidic linkages. *Tetrahedron* 1974, **30**:1933–1944.
16. Kochetkov N, Chizhov O, Shashkov A: Dependence of ^{13}C chemical shifts on the spatial interaction of protons, and its application in structural and conformational studies of oligo- and polysaccharide. *Carbohydr Res* 1984, **133**:173–185.

17. Jansson P, Kenne L, Widmalm G: Casper – a computerised approach to structure determination of polysaccharides using information from n.m.r. spectroscopy and simple chemical analyses. *Carbohydr Res* 1987, **168**:67–77.
18. Lipkind G, Shashkov A, Knirel Y, Vinogradov E, Kochetkov N: A computer-assisted structural analysis of regular polysaccharides on the basis of carbon-13 NMR data. *Carbohydr Res* 1988, **175**:59–75.
19. Toukach F, Knirel Y: New database of bacterial carbohydrate structures. In *XVIII International Symposium on Glycoconjugates, Florence, Italy; 2005*, pp. 216–217.
20. Jansson P, Kenne L, Widmalm G: Computer-assisted structural analysis of polysaccharides with an extended version of casper using proton and carbon-13 NMR data. *Carbohydr Res* 1989: 169–191.
21. Jansson P, Kenne L, Widmalm G: CASPER: a computer program used for structural analysis of carbohydrates. *J Chem Inf Comput Sci* 1991, **31**:508–516.
22. Jansson P, Stenutz R, Widmalm G: Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel web-based version of the computer program CASPER. *Carbohydr Res* 2006, **341**:1003–1010.
23. Stenutz R, Jansson P, Widmalm G: Computer-assisted structural analysis of oligo- and polysaccharides: an extension of CASPER to multibranched structures. *Carbohydr Res* 1998, **306**:11–17.
24. Toukach F, Shashkov A: Computer-assisted structural analysis of regular glycopolymers on the basis of ¹³C NMR data. *Carbohydr Res* 2001, **335**:101–114.
25. Bock K, Pedersen C: Carbon-13 nuclear magnetic resonance spectroscopy of monosaccharides. *Adv Carbohydr Chem Biochem* 1983, **41**:27–66.
26. Kornfeld R, Kornfeld S: Assembly of asparagine-linked oligosaccharides. *Annu Rev Biochem* 1985, **54**:631–664.
27. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006, **34**: D354–D357.

Interpretation of ^{13}C NMR Spectra by Artificial Neural Network Techniques (NeuroCarb)

Andreas Stoeckli, Matthias Studer, Brian Cutting and Beat Ernst

Institute of Molecular Pharmacy, Pharmacenter of the University of Basel, 4056 Basel, Switzerland

17.1 Introduction

An increasing number of biopharmaceuticals (e.g. recombinant proteins [1–4] and monoclonal antibodies [5–7]) are glycoproteins, for which the correct glycosylation pattern is of critical importance for both their structure and function. Therefore, a major challenge in the biotechnological manufacturing of recombinant therapeutic glycoproteins is to achieve and maintain the correct glycosylation pattern in production. In this regard, the commonly used expression systems, such as mammalian cell lines, bacteria, yeast, or insect cells, exhibit an inherent disadvantage, because they do not reproduce the oligosaccharide substitution pattern reliably under the conditions of large-scale production [8]. As a result, the therapeutic glycoproteins are not obtained in the required quality [9]. In order to achieve higher production yields of correctly glycosylated proteins and to fulfill the quality standards required by health authorities, fast, accurate, and preferably inexpensive in-process controls of the glycosylation pattern are necessary.

Glycosylation is an important post-translational modification of proteins. The analysis of the oligosaccharides includes the identification of the individual monosaccharides, their substitution pattern, their anomeric configurations, and their position in a linear or branched oligomer. The most widely used methods for the structural characterization of the oligosaccharide components of glycoproteins are mass spectrometry [10] and fluorescent tagging of monosaccharides released enzymatically by exoglycosidase digestion. For the latter method, digested components are identified by either high-performance liquid chromatography or capillary electrophoresis [11, 12]. This application of exoglycosidase digestion is an extremely valuable tool, particularly when arrays of glycosidases are used [13].

An alternative approach for oligosaccharide analysis is offered by ^{13}C NMR spectroscopy. In general, the chemical shifts of two adjacent monosaccharide moieties are not, or only moderately, influenced by one another. Therefore, an incremental approach is feasible. This makes ^{13}C NMR spectroscopy an excellent tool for the analysis of oligosaccharides. However, severe drawbacks of the ^{13}C NMR approach are the long data acquisition times and the relatively large amounts of analyte required.

For the interpretation of ^1H and ^{13}C NMR spectra of carbohydrates, various computer-assisted approaches have already been reported [14–19]. Our approach to simplify and accelerate the assignment of ^{13}C NMR spectra is based on the application of artificial neural networks (ANNs).

17.2 Neural Networks

The original work on ANNs was started over 60 years ago by McCulloch and Pitts [20] and Hebb [21]. Since the original ANNs had severe limitations in solving simple problems such as learning the logical XOR (exclusive-or) operations, progress in this research was limited until 1982, when Hopfield [22] introduced new perspectives by adding a non-linear character to each single neuron. It is noteworthy that in the 1970s and early 1980s, preceding the work of Hopfield, several important contributions were achieved by Kohonen [23] and Grossberg [24].

ANNs have been designed to imitate the processing of information in the human brain. It is important to recognize that in the phrase *neural network* the emphasis is on the word *network* rather than *neural*. This distinction highlights the fact that the way in which the artificial neurons are connected or networked is much more important than the way in which each neuron performs the single operation for which it is designed.

The function of an artificial neuron is to mimic the action of a biological neuron (Figure 17.1a). It must accept many different input signals, \mathbf{X} , from many neighboring neurons and process them in a predefined way (Figure 17.1b). Depending on the processing, the neuron decides either to fire an output signal \mathbf{Y} or not, and how intense the output signal should be. ANNs can be composed of different numbers of neurons, which can be placed in one or multiple layers (Figure 17.1c and d).

Analogous to the wide range of tasks performed by the human brain, ANNs have been developed for a broad range of applications in chemistry and drug design [26–35]. However, in the field of automated carbohydrate structure elucidation assisted by neural networks, only one approach has been published so far. Meyer and co-workers [36–38] developed ANNs for the interpretation of oligosaccharides derived from alditols.

The following characteristics qualify neural networks for the analysis of spectroscopic data [39]. ANNs have the ability to build models of data by capturing the most important features. In this process, input and output data are used to assign the weights of the connections within the network. Hence, no explicit mathematical knowledge of relationships that are implicitly present in the network is required. The use of ANNs consists of two separate steps: training and prediction. Whereas the former is usually fairly slow, especially with large data sets, the latter is close to instantaneous. Finally, the quality of a trained neural network can be further refined by the addition of new experimental data without the necessity to repeat the entire training process.

The first consideration when employing an ANN is the nature of the problem. Does the problem require a supervised or an unsupervised approach [26] A *supervised* problem means that a set of experiments with known outcomes for each specific input is at hand, whereas an *unsupervised* problem means that one deals with a set of experimental data with no specific answers attached to them. Typically, a supervised problem is the generation of a mathematical model for prediction of properties, namely to find a correlation between the inputs and outputs. Unsupervised learning is distinguished from supervised learning by the fact that there is no *a priori* output. A data set of input objects is typically treated

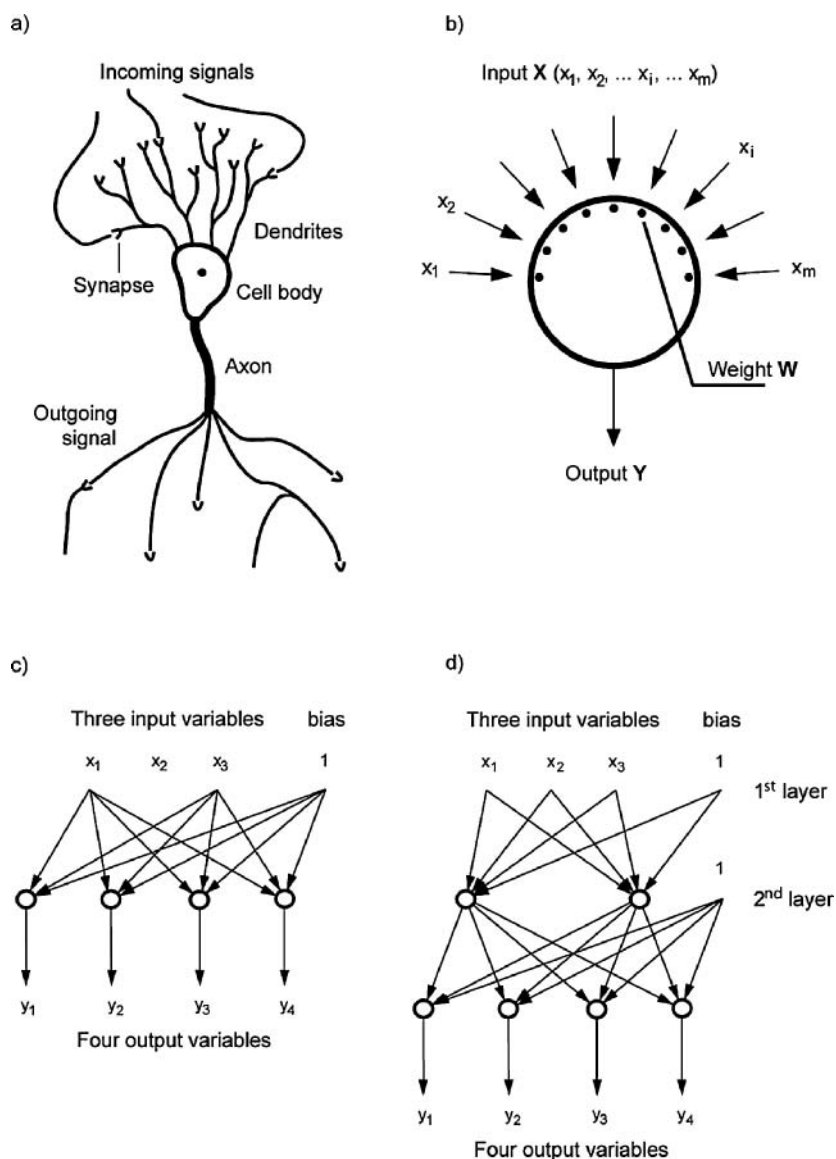
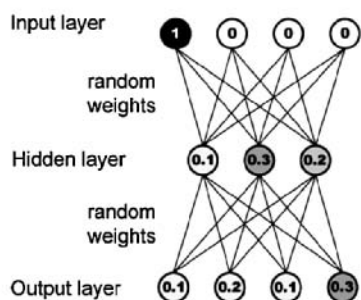


Figure 17.1 Biological (a) and artificial (b) neurons are highly analogous. Incoming signals or inputs are received and interpreted to determine what outgoing signal or output should be produced. With artificial neural networks, inputs interpreted through various weights assigned in single (c) or multiple (d) hidden layers. Adapted from [25].

as a set of random variables and a joint density model thereof is built. Unsupervised methods are used for projection or display of the data in the most informative way. Most unsupervised ANN applications fall within the domain of estimation problems such as statistical modeling, compression, filtering, blind source separation, and clustering. In general, problems associated with data handling at an early stage require unsupervised methods [40]. In the following, the two learning methods are briefly presented.

a) Initial state of network



b) Trained state of network

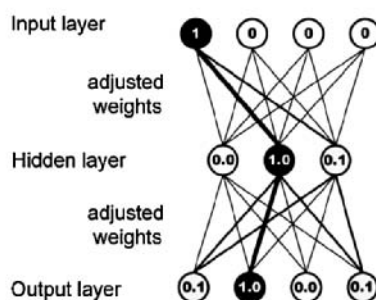


Figure 17.2 Prior to training, the neural network is unable to map the input data properly into the desired output (a). The training procedure allows adjustments of the weights connecting the input, hidden, and output layers, to map known input data accurately into a known output (b).

17.2.1 Supervised Learning (*Back-propagation Algorithm*)

In supervised learning, one set of observations, called inputs, is assumed to be the cause of another set of observations, called outputs. For the training of a neural network applied for supervised learning, data pairs consisting of input objects (e.g. chemical shifts) and the corresponding target values (e.g. anomeric configuration or type of monosaccharide) are therefore required. In the training phase, a neural network, consisting of two or more layers of neurons, which are connected by randomly assigned weights, processes the inputs. The weights of the connections among the neurons are adjusted, until the output error drops below a predefined threshold. With a neural network trained in such a way, the attribution of input objects, for example the chemical shifts of an unknown oligosaccharide, to its chemical structure, becomes possible (Figure 17.2).

A typical example of supervised learning is the back-propagation algorithm, first proposed by Paul Werbos in 1974 [41]. It is usually applied in multilayer feed-forward networks [42]. As the algorithm's name implies, the errors (and therefore the learning) propagate backwards from the output nodes to the inner nodes. Hence, technically speaking, back-propagation is used to calculate the gradient of the error of the network with respect to the network's modifiable weights.

17.2.2 Unsupervised Learning

From the theoretical point of view, supervised and unsupervised learning differ only in the causal structure. In unsupervised learning, it is assumed that all the observations are caused by latent variables, that is, the observations are assumed to be at the end of a causal chain. With unsupervised learning, it is possible to learn larger and more complex models than with supervised learning. This is because in supervised learning one is trying to find the connections between two sets of observations, namely the input and the output observations. Consequently, the difficulty of the learning task increases exponentially with the number of steps between the two sets. In unsupervised learning, however, the learning can proceed hierarchically from the observation into ever more abstract levels of representation. Each additional hierarchy needs to learn only one step and therefore the learning time increases

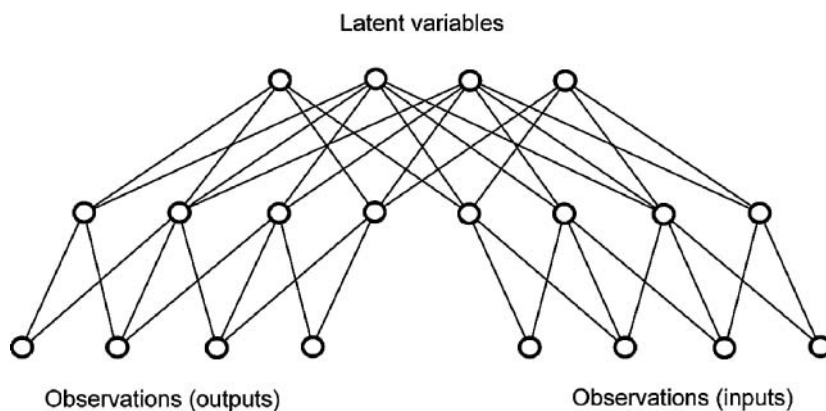


Figure 17.3 Unsupervised learning can be used for bridging the causal gap between input and output observations. The latent variables in the higher levels of abstraction are the causes for both sets of observations and mediate the dependence between inputs and outputs.

linearly with the number of levels in the model hierarchy. If the causal relation between the input and output observations is complex – in a sense that there is a large causal gap – unsupervised learning is preferentially applied. Instead of finding the functions which describe the pathway from inputs to outputs, unsupervised learning tries building a model upwards from both sets of observations in the hope that at higher levels of abstraction, the gap is easier to bridge (see Figure 17.3).

The self-organizing map (SOM) – often referred as a Kohonen map [40] – uses unsupervised learning to produce a low-dimensional representation of the training samples while preserving the topological properties of the input space (see Figure 17.4). This makes SOMs reasonable for visualizing low-dimensional views of high-dimensional data. SOMs are mainly used for dimensionality reduction. In other words, the objective is to map input patterns of arbitrary dimension N on to a discrete map with one or two dimensions. Patterns close to one another in the input space should be close to one another in the SOM.

In our project, supervised and unsupervised learning was combined. The classification of the monosaccharide moieties established by the Kohonen feature maps (unsupervised learning) was used for the training of back-propagation networks (supervised learning).

17.3 NeuroCarb

For the realization of NeuroCarb, Kohonen feature maps and multiple back-propagation neural networks [42, 43] were applied. NeuroCarb consists of an ensemble of ANNs trained to identify disaccharides consisting of glucose, galactose, and mannose, their substitution patterns and their anomeric configurations based on ^{13}C NMR peak lists.

For a proof of concept, we initially used unsupervised learning (Kohonen networks) to demonstrate that neural networks are able to classify monosaccharide moieties according to the following features: (i) their anomeric configuration (α/β), (ii) the linkage position of appended sugar moieties (linkage position), and (iii) the type of monosaccharide (group affiliation). The classification performance is presented in Table 17.1.

The workflow of the training, validation, and test phase and of the work phase of NeuroCarb are presented in Figures 17.5 and 17.6, respectively.

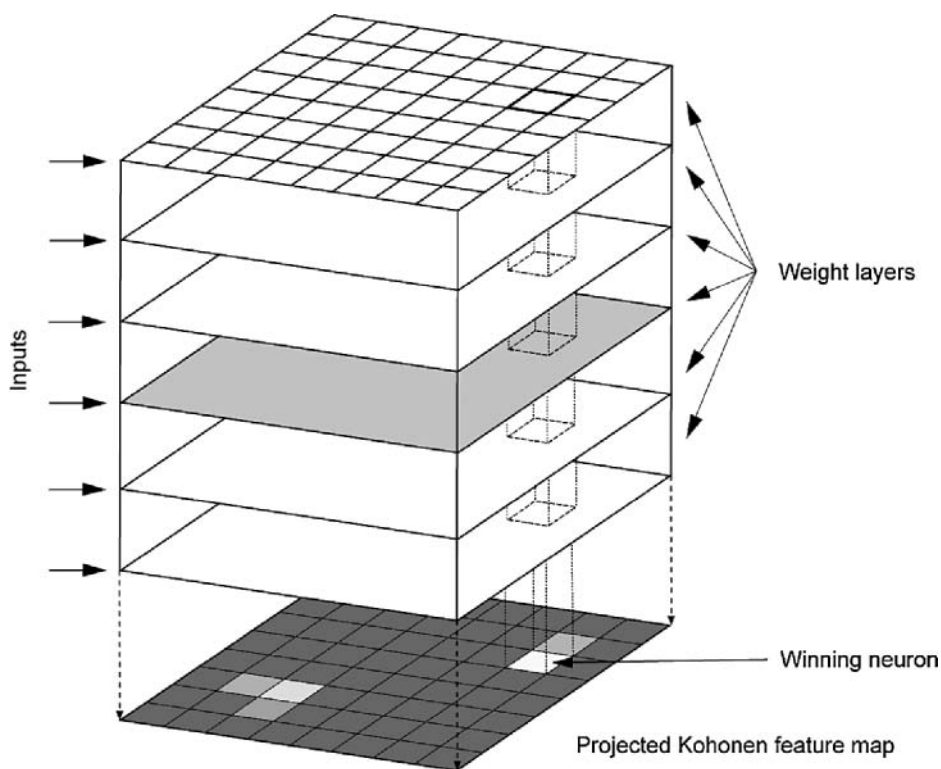


Figure 17.4 Kohonen networks are an example of unsupervised learning. Incoming data are compressed by mapping the presented input into an output space of significantly smaller number of dimensions; thus, in a multidimensional input objects are projected on to a two-dimensional map. On this feature map, similar input objects are grouped together.

17.3.1 Training, Validation, and Test Phase

17.3.1.1 Data Preprocessing. The ^{13}C NMR peak lists of over 2600 monosaccharide moieties deduced from reported mono- and oligosaccharides were stored in an in-house database. The groups of specific monosaccharides were divided into subgroups according to their substitution pattern (Table 17.2). As an example, the subgroup $\alpha\text{-D-Glcp-1R}$ contains all D-glucose derivatives with an α -anomeric configuration, substitution at the C1 position

Table 17.1 Average classification performance of the three Kohonen networks trained to categorize the test compounds according to their anomeric configuration (α/β), the position of the type of monosaccharide (group affiliation), and the appended group (linkage position).

	α/β	Group affiliation	Linkage position
True (%)	100	99.7	94.4
Unknown (%)	0.0	0.0	1.9
False (%)	0.0	0.3	3.7

Training, Validation and Test Phase

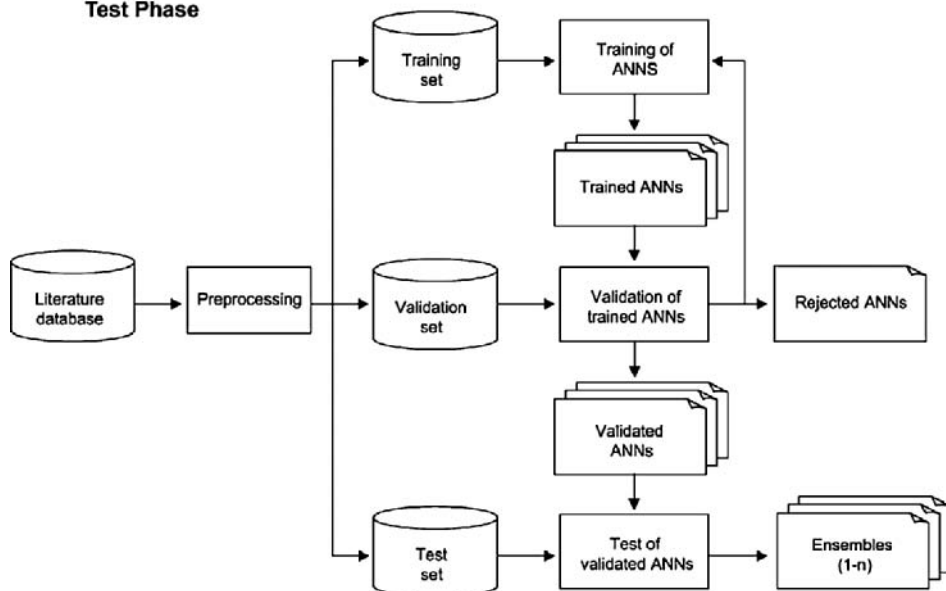


Figure 17.5 Workflow of the training, validation, and test phase of NeuroCarb.

Work Phase

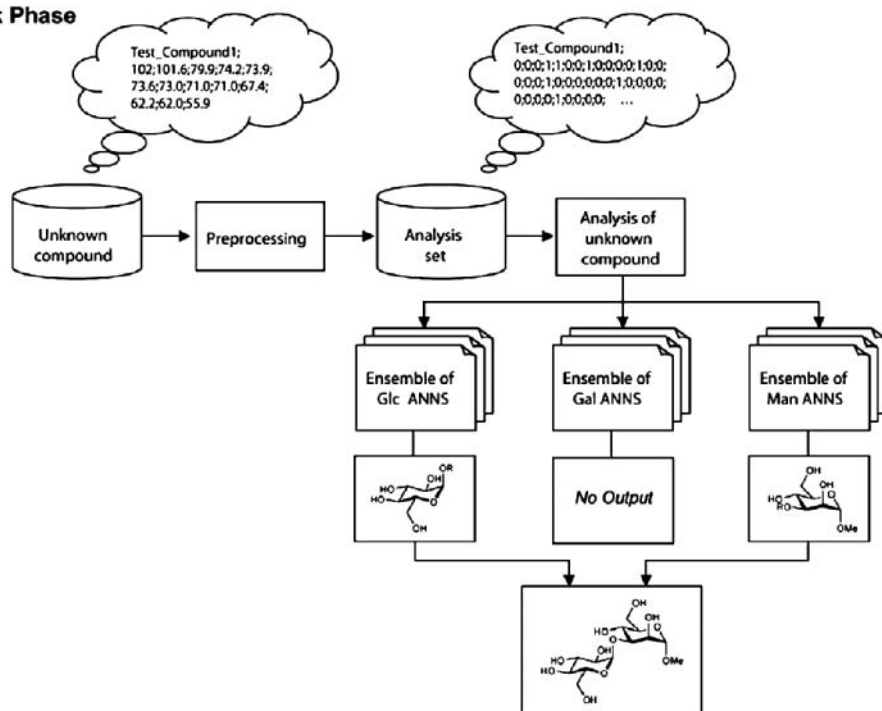


Figure 17.6 Workflow of the work phase of NeuroCarb: structure elucidation of $\alpha\text{-D-Glc}-(1\rightarrow3)\text{-}\alpha\text{-D-Manp-OMe}$.

Table 17.2 Averaged literature ^{13}C NMR chemical shift data ($\bar{\delta}_{1-n}$ in ppm) of substituted monosaccharide moieties used for preprocessing and training.

	$\bar{\delta}_1$	$\bar{\delta}_2$	$\bar{\delta}_3$	$\bar{\delta}_4$	$\bar{\delta}_5$	$\bar{\delta}_6$	$\bar{\delta}_7$	$\bar{\delta}_8$
α -D-Glcp-1R	99.5	73.8	72.9	72.3	70.3	61.3		
α -D-Glcp-OH	92.8	73.5	72.2	72.2	70.5	61.6		
α -D-Glcp-OH-2R	92.4	81.4	72.4	71.7	70.3	61.5		
α -D-Glcp-OH-3R	92.1	83.1	71.5	71.1	68.6	61.5		
α -D-Glcp-OH-4R	92.6	79.3	72.4	72.0	70.9	61.0		
α -D-Glcp-OH-6R	92.8	73.8	72.3	70.7	70.3	67.6		
α -D-Glcp-OMe	100.0	73.9	72.3	72.0	70.4	61.4	55.7	
α -D-Glcp-OMe-2R	99.2	80.6	72.8	72.1	70.5	61.5	55.6	
α -D-Glcp-OMe-3R	100.0	81.8	72.2	71.4	69.7	61.4	56.3	
α -D-Glcp-OMe-4R	99.6	78.3	73.1	71.7	70.8	60.9	55.8	
α -D-Glcp-OMe-6R	100.0	74.0	71.9	70.7	70.2	66.3	55.8	
β -D-Glcp-1R	103.2	76.6	76.3	74.0	70.3	61.5		
β -D-Glcp-OH	96.7	76.6	76.6	75.0	70.5	61.7		
β -D-Glcp-OH-2R	97.0	79.4	76.6	75.3	70.6	61.5		
β -D-Glcp-OH-3R	96.4	85.8	76.4	73.8	68.8	61.6		
β -D-Glcp-OH-4R	96.5	79.4	75.6	75.1	74.7	60.9		
β -D-Glcp-OH-6R	96.8	76.5	75.2	74.9	70.3	67.2		
β -D-Glcp-OMe	104.0	76.7	76.7	73.9	70.6	61.7	57.9	
β -D-Glcp-OMe-2R	104.0	79.3	76.7	75.7	70.5	61.8	58.0	
β -D-Glcp-OMe-3R	103.8	85.5	76.3	73.1	69.0	61.5	57.8	
β -D-Glcp-OMe-4R	103.5	78.5	75.4	74.8	73.2	60.7	57.7	
β -D-Glcp-OMe-6R	104.2	76.6	75.7	73.8	70.2	69.3	58.0	
α -D-Galp-1R	99.9	71.8	70.0	69.8	69.1	61.7		
α -D-Galp-OH	93.3	71.4	70.4	70.2	69.5	62.2		
α -D-Galp-OH-3R	93.1	78.1	71.3	69.5	68.6	61.9		
α -D-Galp-OH-4R	93.7	78.3	72.7	70.4	69.8	62.2		
α -D-Galp-OH-6R	92.9	69.9	69.7	69.6	69.2	68.6		
α -D-Galp-OMe	100.2	71.4	70.3	70.0	69.0	61.9	55.8	
α -D-Galp-OMe-2R	98.4	76.0	71.8	70.9	69.5	61.8	55.4	
α -D-Galp-OMe-3R	99.9	80.1	70.9	69.4	67.9	61.7	56.3	
α -D-Galp-OMe-4R	100.2	79.4	71.4	70.2	69.3	61.3	55.9	
α -D-Galp-OMe-6R	100.2	70.1	70.0	69.7	69.2	68.2	56.0	
β -D-Galp-1R	104.1	75.9	73.3	71.7	69.3	61.6		
β -D-Galp-OH	97.4	76.0	73.9	73.0	69.8	62.0		
β -D-Galp-OH-3R	97.2	81.6	76.0	71.9	68.5	61.8		
β -D-Galp-OH-4R	97.7	80.0	76.1	72.2	71.0	61.6		
β -D-Galp-OH-6R	97.1	74.1	73.3	72.4	69.3	68.0		
β -D-Galp-OMe	104.6	75.9	73.7	71.6	69.5	61.8	57.9	
β -D-Galp-OMe-2R	103.1	78.9	75.7	73.3	69.3	61.4	57.6	
β -D-Galp-OMe-3R	104.5	80.4	75.3	71.5	69.6	61.3	57.4	
β -D-Galp-OMe-4R	104.7	75.7	73.5	71.4	69.2	61.4	57.8	
β -D-Galp-OMe-6R	104.5	74.3	73.4	71.6	69.6	68.3	57.8	
α -D-Manp-1R	101.7	73.9	71.6	71.0	67.6	61.8		
α -D-Manp-OH	94.9	73.3	71.6	71.2	67.9	62.0		
α -D-Manp-OH-2R	92.4	78.9	73.3	70.2	67.4	61.5		
α -D-Manp-OH-4R	94.8	77.9	71.9	71.3	70.1	61.7		
α -D-Manp-OH-6R	95.2	73.4	71.8	71.5	70.3	68.0		
α -D-Manp-OMe-2R	99.8	79.2	73.5	70.9	67.9	61.9	55.7	
α -D-Manp-OMe-3R	101.4	79.1	73.6	70.3	66.9	61.6	55.5	
α -D-Manp-OMe-4R	101.8	74.8	71.6	71.6	70.9	61.5	55.3	

Table 17.2 (Continued)

	$\bar{\delta}_1$	$\bar{\delta}_2$	$\bar{\delta}_3$	$\bar{\delta}_4$	$\bar{\delta}_5$	$\bar{\delta}_6$	$\bar{\delta}_7$	$\bar{\delta}_8$
α -D-Manp-OMe-6R	101.4	71.6	71.4	70.7	67.8	66.6	55.6	
β -D-Manp-1R	101.3	77.2	73.9	71.4	67.7	61.9		
β -D-ManpNAc-1R	174.5	99.4	75.3	73.4	72.9	60.7	53.6	22.2
β -D-Manp-OH	94.5	77.0	74.0	72.1	67.6	61.9		
β -D-Manp-OH-2R	94.1	81.4	77.0	73.4	67.7	61.6		
β -D-Manp-OH-4R	94.8	77.8	75.8	72.8	71.5	61.7		
β -D-Manp-OH-6R	94.9	76.2	74.2	72.3	70.3	67.8		
β -D-Manp-OMe-2R	101.5	78.6	77.0	72.7	68.0	61.5	57.4	
β -D-Manp-OMe-4R	101.7	77.5	75.8	72.5	70.6	61.2	57.6	

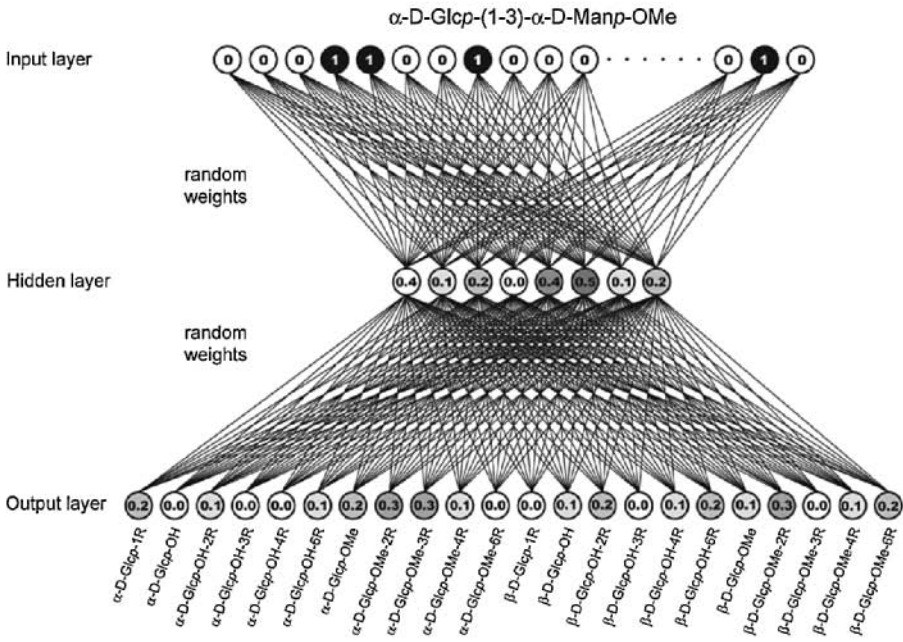
and no further substitution on the remaining positions. The chemical shifts of all members of a subgroup were adjusted according to Gottlieb *et al.* [44], averaged, and the corresponding standard deviations calculated.

Next, each average value was randomly shifted 240 times within the range of its standard deviation to account for the chemical shift deviation. The whole data set was then converted into a binary input pattern file by the ANN PFG [45]. Neural networks can suffer from either underfitting – not sufficiently complex to detect the pattern correctly in a noisy data set – or overfitting – too complex so it reacts on noise in the data. The best way to avoid overfitting is to use several sets for training the network. Therefore, the entire data set was divided equally into three data sets (*training set*, *validation set* and *test set*; see Figure 17.5). These data sets were used at different stages of the subsequent process.

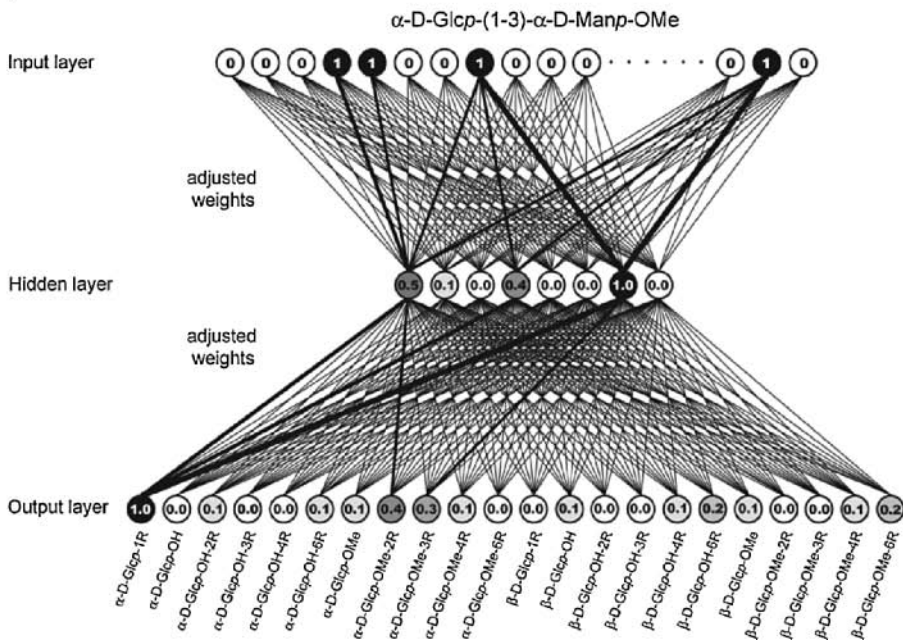
17.3.1.2 Training and Validation Phase. After the preprocessing of the literature data and classification into three groups according to the type of monosaccharide, individual back-propagation networks were trained using the software package Statistica [46]. For each individual problem, the optimal network architecture and the ideal learning parameters had to be determined by an iterative process. During the training process, all patterns of the *training set* were presented to each individual network. In order to check for over- or underfitting, the appropriateness of a network was estimated by means of the *validation set* after each training cycle. In the final step, the deviation between the network output and the experimental data was expressed as an error value. In order to minimize this error, the weights of the junctions between the individual neurons were adjusted. The training cycles were repeated until the error dropped below a predefined threshold. A network was regarded as validated and ready to be used in the test phase when the predefined abort criterion was reached. Otherwise, the network was rejected.

17.3.1.3 Test Phase. During the test phase, all validated networks were further evaluated with the *test set*, which has never been presented during the training process and is therefore unknown to the networks. The “winners”, defined as the 15–20 best performing networks, were then combined in an ensemble file. This process was accomplished independently for the three groups of saccharides, namely glucose, galactose, and mannose.

a) Initial state of Glc-network



b) Trained state of Glc-network



17.3.2 Work Phase

At the current stage of the project, the peak list of the unknown disaccharide is manually transformed and presented to each network of the three ensembles. Each ensemble computes a single independent output. The consensus output of all three ensembles represents the result of the structure elucidation.

The work phase is illustrated in Figure 17.6, where the structure of an unknown disaccharide [α -D-Glcp-(1-3)- α -D-Manp-OMe] is elucidated (Figure 17.7). After entering its chemical shifts (peak list) manually into NeuroCarb, they are transformed into a pattern file by the ANN PFG and presented to each network of the three ensembles. Each ensemble generates its output as described above. In the final step, these outputs are evaluated and combined to a structure suggestion. Since the subgroups representing the reducing and non-reducing monosaccharides have been identified, there is only one possible alignment for the disaccharide.

17.3.3 Performance of NeuroCarb

The trained ensembles of networks were tested using a set of 65 unknown disaccharides, each containing at least one glucose, galactose, or mannose moiety. The ensembles recognized correctly 83% (Gal, 82.8%; Glc, 75.5%; Man, 91.2%) of the individual monosaccharide moieties.

The current version of NeuroCarb is limited to the analysis of disaccharides composed of the mammalian saccharide moieties glucose, galactose, and mannose. However, it is under constant development and is currently being expanded to recognize Fuc, NeuNAc, GlcNAc, GalNAc, GlcA, and Xyl.

In comparison with classical reference table procedures, methods based on neural networks such as NeuroCarb have the advantage of being more stable towards incomplete and poor quality data, as could be shown with incomplete data sets.

Due to the limited number of NMR data from our own laboratory, the *training set* and *test set* were predominantly composed from literature data. Unfortunately, the quality of these data was found to be highly variable. This might be one reason for the 17% failure.

Figure 17.7 Glc network. (a) Feed-forward network for glucose in its initial state. A feed-forward network consists of neurons, which are organized in layers. Each neuron of one layer is connected to every neuron of the following layer. Each connection between two neurons initially has a randomly assigned weight. Neurons of the first layer obtain a preprocessed input, whereas neurons of the following layers add up all weighted inputs from all neurons of the former layer. This information is processed within the neuron and sent to the next layer. This is repeated through all layers of the network. With randomized weights, the output differs logically from the target output. During the training process, all patterns are repeatedly presented to the network and the weights are iteratively adjusted in order to minimize the error of prediction. Training stops when the error of prediction drops below a predefined threshold. (b) Feed-forward network in its trained state for glucose. Network weights are no longer adjusted. All descriptors of a new input object are processed through all layers with fixed weights. The calculated network output is evaluated and the result is presented to the user. In our example, given a specific input [α -D-Glcp-(1-3)- α -D-Manp-OMe] a trained Glc-ANN recognizes the subgroup α -D-Glcp-1R.

17.4 Conclusion and Outlook

In the near future, the major focus of NeuroCarb will be on the automation of the entire analysis procedure and its integration into a web interface. The next version of NeuroCarb will be trained to recognize all mammalian saccharides in linear oligosaccharides (e.g. trisaccharides or tetrasaccharides). However, a major challenge will be the structure elucidation of branched oligosaccharides, since the chemical shifts of vicinally disubstituted saccharides are heavily influenced by steric factors [47] and can only be predicted with an appropriate training set.

In summary, it has been shown that the neural network approach applied to NMR 1D shift data for disaccharides is in principle able to identify the anomeric configurations (α/β), the substitution pattern, and the monosaccharide composition of disaccharides. The extension to complex oligosaccharides, such as branched oligosaccharides consisting of all mammalian monosaccharides, is currently under investigation.

References

1. Schmidt FR: Recombinant expression systems in the pharmaceutical industry. *Appl Microbiol Biot* 2004, **65**:363–372.
2. Novotny MV, Mechref Y: New hyphenated methodologies in high-sensitivity glycoprotein analysis. *J Sep Sci* 2005, **28**:1956–1968.
3. Brooks SA: Appropriate glycosylation of recombinant proteins for human use: Implications of choice of expression system. *Mol Biotechnol* 2004, **28**:241–256.
4. Prete M, Perosa, F, Favoino E, Dammacco F: Biological therapy with monoclonal antibodies: a novel treatment approach to autoimmune disease. *Clin Exp Med* 2005, **5**:141–160.
5. Jine Y, Zonghui Y: Production and application of anti-drug monoclonal antibodies. *Jiangxi Nongye Daxue Xuebao* 2005, **27**:305–308.
6. Adams GP, Weiner LM: Monoclonal antibody therapy of cancer. *Nat Biotechnol* 2005, **23**:1147–1157.
7. Kannagi R, Hakomori S: A guide to monoclonal antibodies directed to glycotopes. *Adv Exp Med Biol* 2001, **491**:587–630.
8. Warner TG: Enhancing therapeutic glycoprotein production in Chinese hamster ovary cells by metabolic engineering endogenous gene control with antisense DNA and gene targeting. *Glycobiology* 1999, **9**:841–850.
9. Gomord V, Sourrouille C, Fitchette A-C, Bardor M, Pagny S, Lerouge P, Faye L: Production and glycosylation of plant-made pharmaceuticals: the antibodies as a challenge. *Plant Biotechnol J* 2004, **2**:83–100.
10. Dell A, Morris HR: Glycoprotein structure determination by mass spectrometry. *Science* 2001, **291**:2351–2356.
11. Mechref Y, Novotny MV: Structural investigations of glycoconjugates at high sensitivity. *Chem Rev* 2002, **102**:321–369.
12. Shilova NV, Bovin NV: Fluorescent labels for the analysis of mono- and oligosaccharides. *Russ J Bioorg Chem* 2003, **29**:309–324.
13. Edge CJ, Rademacher TW, Wormald MR, Parekh RB, Butters TD, Wing DR, Dwek RA: Fast sequencing of oligosaccharides: the reagent-array analysis method. *Proc Natl Acad Sci USA* 1992, **89**:6338–6342.
14. Egli H, Smith DH, Djerassi C: Application of artificial intelligence for chemical inference. Part XLI. Computer-assisted structural interpretation of proton NMR spectral data. *Helv Chim Acta* 1982, **65**:1898–1920.

15. Hounsell EF, Wright DJ: Computer-assisted interpretation of proton NMR spectra in the analysis of the structure of oligosaccharides. *Carbohydr Res* 1990, **205**:19–29.
16. Bot DSM, Cleij P, Van't Klooster HA, Van Halbeek H, Veldink GA, Vliegthart JFG: Identification and substructure analysis of oligosaccharide chains derived from glycoproteins by computer retrieval of high-resolution proton-NMR spectra. *J Chemom* 1988, **2**:11–27.
17. Anderson DR, Grimes WJ: Application of microcomputers to the interpretation of high-resolution nuclear magnetic resonance spectra of asparagine-linked oligosaccharides: evaluation of high-mannose structures. *Anal Biochem* 1985, **146**:13–22.
18. Lipkind GM, Shashkov AS, Knirel YA, Vinogradov EV, Kochetkov NK: A computer-assisted structural analysis of regular polysaccharides on the basis of carbon-13 NMR data. *Carbohydr Res* 1988, **175**:59–75.
19. Cumming DA, Helleqvist C, Touster O: On the utility of carbon-13 NMR spectroscopy in the identification of the primary structures of manno-oligosaccharides and glycopeptides. *Carbohydr Res* 1988, **179**:369–380.
20. McCulloch WS, Pitts W: A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 1943, **5**:115–133; Pitts W, McCulloch WS, How we know universals. The perception of auditory and visual forms. *Bull Math Biophys* 1947, **9**:127–147.
21. Hebb DO: *The Organization of Behaviour*. New York: Wiley; 1949, pp. 60–78.
22. Hopfield JJ: Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 1982, **79**:2554–2558.
23. Kohonen T: Correlation matrix memories. *IEEE Trans Comput* 1972, **C-21**:353–359.
24. Grossberg S: Adaptive pattern classification and universal recording: 1. Parallel development and coding of neural feature detectors. *Biol Cybernet* 1976, **23**:121–134.
25. Zupan J: In *Handbook of Chemoinformatics*, 1st edn (ed. Gasteiger J). Weinheim: Wiley-VCH Verlag GmbH; 2003, pp. 1167–1215.
26. Gasteiger J, Zupan J: Neural networks in chemistry. *Angew Chem Int Ed Engl* 1993, **32**:503–527.
27. Rufino AR, Brant AJC, Santos JBO, Ferreira MJP, Emerenciano VP: Simple method for identification of skeletons of aporphine alkaloids from ^{13}C NMR data using artificial neural networks. *J Chem Inf Model* 2005, **45**:645–651.
28. Meiler J, Köck M: Novel methods of automated structure elucidation based on ^{13}C NMR spectroscopy. *Magn Reson Chem* 2004, **42**:1042–1045.
29. Munk ME, Madison MS, Robb EW: The application of neural networks to the integrated interpretation multispectral data. In *Book of Abstracts, 216th ACS National Meeting*, Boston, MA, 23–27 August 1998; PHYS-397.
30. Munk ME, Madison MS, Robb EW: The neural network as a tool for multispectral interpretation. *J Chem Inf Comput Sci* 1996, **36**:231–238.
31. Stenutz R, Jansson P-E, Widmalm G: Computer-assisted structural analysis of oligo- and polysaccharides: an extension of CASPER to multibranched structures. *Carbohydr Res* 1998, **306**:11–17.
32. Stenutz R, Erbing B, Widmalm G, Jansson, P-E, Nimmich W: The structure of the capsular polysaccharide from *Klebsiella* type 52, using the computerized approach CASPER and NMR spectroscopy. *Carbohydr Res* 1997, **302**:79–84.
33. Jansson P-E, Kenne L, Widmalm G: Computer-assisted structural analysis of oligosaccharides using CASPER. *Anal Biochem* 1991, **199**:11–17.
34. Jansson P-E, Kenne L, Widmalm G: Computer-assisted structural analysis of polysaccharides with an extended version of Casper using proton and carbon-13 NMR data. *Carbohydr Res* 1989, **188**:169–191.
35. Jansson P-E, Kenne L, Widmalm G: Casper – a computerised approach to structure determination of polysaccharides using information from N.M.R. spectroscopy and simple chemical analyses. *Carbohydr Res* 1987, **168**:67–77.

36. Meyer B, Hansen T, Nute D, Albersheim P, Darvill A, York W, Sellers J: Identification of the ^1H -NMR spectra of complex oligosaccharides with artificial neural networks. *Science* 1991, **251**:542–544.
37. Radomski JP, van Halbeek H, Meyer B: Neural network-based recognition of oligosaccharide ^1H -NMR spectra [letter]. *Nat Struct Biol* 1994, **1**:217–218.
38. Thomsen JU, Meyer B: Pattern recognition of the ^1H NMR spectra of sugar alditols using a neural network. *J Magn Reson* 1989, **84**:212–217.
39. Gasteiger J, Teckentrup A, Terfloth L, Spycher S: Neural networks as data mining tools in drug design. *J Phys Org Chem* 2003, **16**:232–245.
40. Kohonen T: Self-organized formation of topologically correct feature maps. *Biol Cybernet* 1982, **43**:59–69.
41. Werbos P: *Beyond Regression – New Tools for Prediction and Analysis in the Behavioral Sciences*. Cambridge, MA: Harvard University; 1974.
42. Rumelhart DE, Hinton GE, Williams RJ: Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1 (ed. Rumelhart DE). Cambridge, MA: MIT Press; 1986, pp. 318–362.
43. Zupan J, Gasteiger J: In *Kohonen Network in Neural Networks in Chemistry and Drug Design*, 2nd edn (eds Zupan J, Gasteiger J). Weinheim: Wiley-VCH Verlag GmbH; 1999, pp. 81–100.
44. Gottlieb HE, Kotlyar V, Nudelman A: NMR chemical shifts of common laboratory solvents as trace impurities. *J Org Chem* 1997, **62**:7512–7515.
45. Studer MD: NeuroCarb; artificial neural networks for NMR structure elucidation of oligosaccharides. PhD Thesis, Faculty of Science, University of Basel; 2006.
46. StatSoft: *STATISTICA, Version 6*. Bedford: StatSoft; 2004.
47. Hermansson K, Jansson P-E, Kenne L, Widmalm G, Lindh F: A ^1H and ^{13}C NMR study of oligosaccharides from human milk. Application of the computer program CASPER. *Carbohydr Res* 1992, **235**:69–81.

Section 5:
3D Structures of Complex
Carbohydrates

Conformational Analysis of Carbohydrates – A Historical Overview

Martin Frank

Deutsches Krebsforschungszentrum (German Cancer Research Center), Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany

18.1 Introduction

Complex carbohydrates represent a particularly challenging class of molecules in terms of describing their three-dimensional (3D) structure. Due to their inherent flexibility, these molecules very often exist in solution as an ensemble of conformations rather than as a single, well-defined structure. Recent advances in experimental and theoretical methods have begun to yield further insight into their conformational behavior; however, general rules governing their conformational preferences are still difficult to derive. In fibers, polysaccharides are frequently found to form (double) helical structures that are stabilized by hydrogen bonds or ion bridges. However, in solution, the formation of stable secondary or tertiary structures – as is frequently the case for proteins (polypeptides) – seems to be very rare for polysaccharides. In solution, polysaccharides generally exist in disordered or random coil conformations [1].

X-Ray and neutron diffraction are very powerful methods for determining the atomic positions of carbohydrates in a crystal [2, 3]. The 3D structures of a variety of carbohydrates (most of them smaller than tetrasaccharides) have been determined using these methods and the atomic coordinates are available from the Cambridge Structural Database (CSDB) [4]. For a variety of reasons, only a limited number of crystallographic structures have been reported in which a complex oligosaccharide is well resolved. Unfortunately, larger carbohydrates fairly often persist as syrups or gels and refuse to crystallize, or they are too flexible to yield sufficient electron density and their 3D structure consequently cannot be derived using X-ray crystallography. How difficult it is to obtain a crystal suitable for X-ray diffraction studies, even for a trisaccharide that is generally considered as “rigid”, can be seen in the case of the histo-blood group antigen Lewis^X. It took almost 2 years to obtain a crystal of sufficient quality to solve the structure [5]. With the exception of cyclic carbohydrates such as cyclodextrins and cycloamyloses, almost no crystal structures are available of pure oligosaccharides consisting of more than four monosaccharide units. However, some polysaccharides can form fibers and therefore X-ray fiber diffraction studies [6–8] are playing an important role in the determination of the 3D structures of such

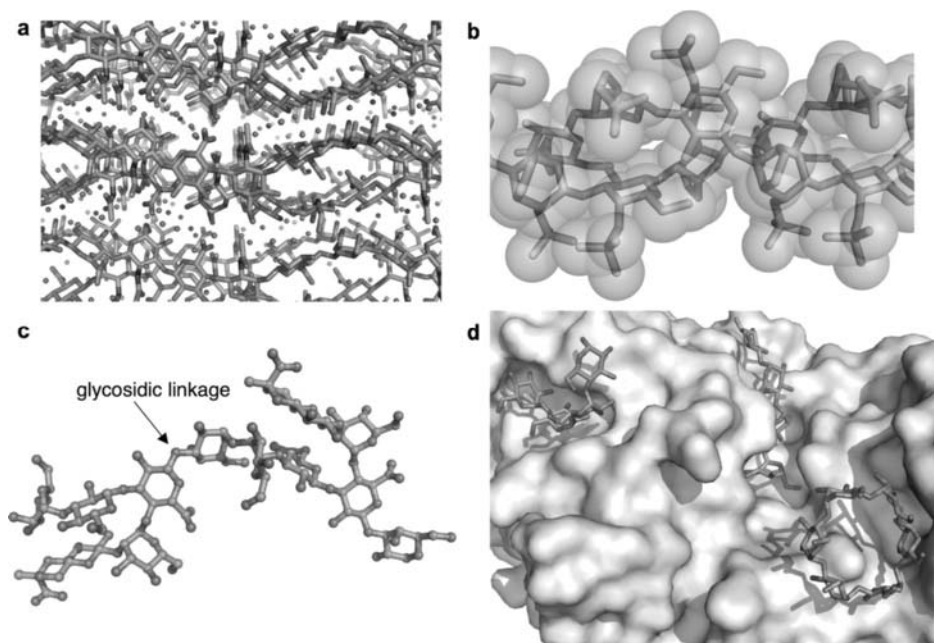


Figure 18.1 Some polysaccharide structures available from the PDB: (a) hyaluronic acid (PDB code 1HYA) [205]; (b) iota-carrageenan (PDB code 1CAR) [206]; (c) capsular polysaccharide from *E. coli* (PDB code 1CAP) [207]; (d) α -D-Glcp polymers bound to α -amylase (PDB code 3BCD) [208].

polysaccharides. Since the beginning of the 1990s, more and more crystal structures have been reported where carbohydrates are covalently attached to a (glyco)protein or constitute the ligand in a protein–carbohydrate complex [9, 10]. These experimentally determined 3D structures are freely accessible from the Protein Data Bank (PDB) [11] and some examples are shown in Figure 18.1 (see also Chapter 20).

Molecules in a crystal or fiber are subject to packing constraints and to the optimization of intramolecular and intermolecular interactions. Therefore, it should be kept in mind that the conformation adopted in the solid state might be different from conformations preferred in solution. Nevertheless, crystallographic studies on small carbohydrates have been very useful for providing information on bond lengths, bond angles, and torsion angles in the crystal lattice. These data have been used as initial parameters for the development of force fields to calculate 3D models of low-energy structures of carbohydrates using molecular mechanics [12–14] (see Chapter 19).

In recent years, NMR methods, especially nuclear Overhauser effect (NOE) measurements, have been widely used to study oligosaccharide conformation in solution [15]. Unfortunately, many oligosaccharide NOEs cannot be resolved or are difficult to assign. Additionally there are often too few inter-residue NOEs to make an unambiguous 3D structure determination possible. In general, the interpretation of structural experimental data frequently needs to be supported by molecular modeling methods. One of the main aims of computer modeling of carbohydrates is to generate reasonable 3D models (Figure 18.2) that can be used to rationalize experimentally derived observations. Conformational analysis by computational methods consequently plays a key role in the determination of 3D structures of complex carbohydrates.

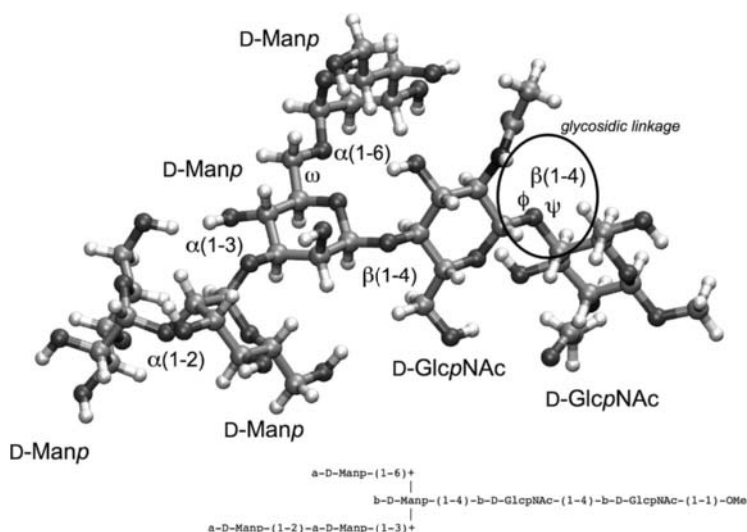


Figure 18.2 A 3D representation of a low-energy conformation of an (extended) *N*-glycan core ($\text{Man}_4\text{GlcNAc}_2$) highlighting the various linkage types, $\beta(1-4)$, $\alpha(1-2)$, $\alpha(1-3)$, and $\alpha(1-6)$, that are occurring in this structure. The glycosidic torsions are named ϕ ($\text{O}_5\text{-C}_1\text{-O}_x\text{-C}_x$), ψ ($\text{C}_1\text{-O}_x\text{-C}_x\text{-C}_{x+1}$) [and ω ($\text{O}_5\text{-C}_5\text{-C}_6\text{-O}_6$) in 1–6 linkages] (x = attachment position on the adjacent monosaccharide). The central mannose residue constitutes a branch point. This structural motif is frequently found in glycoproteins. For a more detailed description of features of carbohydrate structures, see Chapter 19.

Complex carbohydrates are built up from monosaccharides (usually cyclic structures) that are connected by a variety of glycosidic linkage types (Figure 18.2). In addition, carbohydrate chains can contain branches, in contrast to linear protein and DNA chains. The influence of branching on 3D structure and flexibility is of particular interest because all factors that modulate the shape of the carbohydrate could have an impact on its biological function. Therefore, key questions in conformational analysis of carbohydrates are: what are the preferred conformations of the monosaccharide rings; which are the most likely orientations of the glycosidic linkages; and how flexible are these linkages? A large variety of computational methods have been applied to study the conformations and flexibility of carbohydrates; however, it is beyond the scope of this chapter to describe all of them in detail. This chapter presents a review of the history of (computational) conformational analysis of carbohydrates, which is intended not only to highlight how the whole field has evolved over the last 60 years but also to give the interested reader appropriate literature references to the methods used. In addition, a variety of reviews and book chapters on conformational analysis of carbohydrates have been published and are recommended for further reading [1, 16–29].

18.2 Conformational Analysis of (Poly)saccharides – the Pioneering Years

In the 1940s, one of the first studies of pyranose ring conformation of aldoses (see Chapter 2) was undertaken by Hassel and Ottar [30], who suggested that the pyranose ring can exist in two possible chair conformations, 1C_4 and 4C_1 (Figure 18.3). In considering the relative

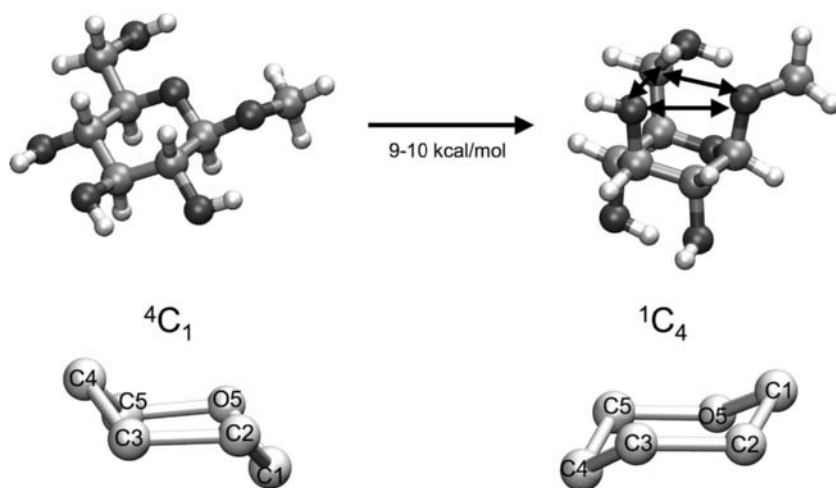


Figure 18.3 Chair conformations of the pyranose ring of β -D-GlcpOMe. The 1C_4 conformer has about 9–10 kcal mol $^{-1}$ higher energy than the 4C_1 conformer (calculated using the TINKER/MM3 force field [209]). This is mainly due to unfavorable interactions of the bulky axial substituents in the 1C_4 conformer (Hassel–Ottar effect [30]), some of which are indicated by arrows. “ 4C_1 ” means that the C4 atom is above and the C1 atom is below the ring plane formed by the remaining ring atoms C2, C3, C5, and O5.

stability of these conformations, it was assumed that the conformation placing both the CH₂OH group and OH group(s) in axial orientations on the same side of the ring is energetically unfavorable (“Hassel–Ottar effect”) [31–33].

In the 1960s, the availability of 3D structures derived from high-quality crystal structures of glucose and the disaccharide cellobiose [β -D-Glcp-(1–4)- β -D-Glcp] [34] made it possible to study the conformations of carbohydrates systematically using theoretical methods. Based on the pioneering work of Ramachandran *et al.* [35], systematic variations of the glycosidic linkage torsions (ϕ and ψ) (Figure 18.1) were performed to study the conformational preferences of disaccharides [36, 37]. The rings were treated as rigid (*rigid residue approximation*) and the distances of atom pairs were calculated for each orientation of the two monomeric units. Initially, interatomic distance calculations were done “by hand”. Relatively simple models (hard-sphere contact criteria) were used to judge whether the calculated interatomic distances represented a conformation that would be classified as “allowed” or “disallowed.” The geometries derived from crystal structures (Figure 18.4) were used to describe the rings, so the quality of the atomic coordinates had a major influence on the calculation results. To refine the model further, the (non-bonded) conformational energy of all “allowed” conformations was calculated using van der Waals or Buckingham potentials and plotted as a function of the glycosidic torsions ϕ and ψ in a Ramachandran-type plot (*conformational map*) (Figure 18.5). This approach made it possible to determine the theoretically most favorable orientation(s) of the glycosidic linkage.

In the late 1960s, the availability of computers for the systematic exploration of theoretical polysaccharide conformations made it possible to increase the number of atoms that could be included in the calculations, and to cover the conformational map more completely [37–39]. Using a computer, Sundararajan and Rao studied systematically the

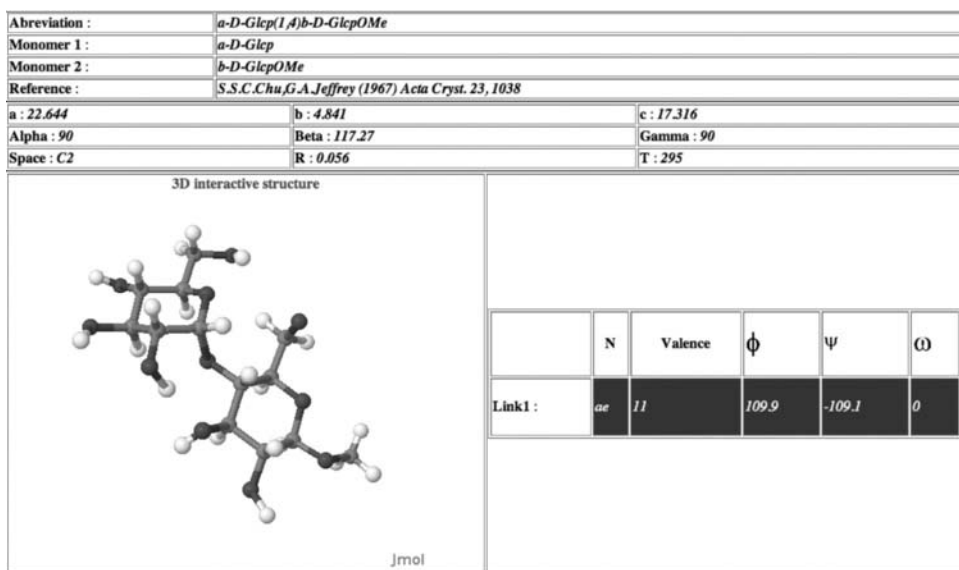


Figure 18.4 Experimentally determined X-ray 3D structure of maltose. Screenshot of the entry in the database of X-ray structures of oligosaccharides (<http://www.cermav.cnrs.fr/cgi-bin/oligos/oligos.cgi>).

influence of the orientation of substituents (axial versus equatorial, Hassel–Ottar effect), on the pyranose ring conformation [33]. Initially they used only a Kitaigorodsky type of function to estimate the non-bonded interaction energy of atoms. Later they also included coulomb interactions in their calculation of the conformational energy [40]. It was found

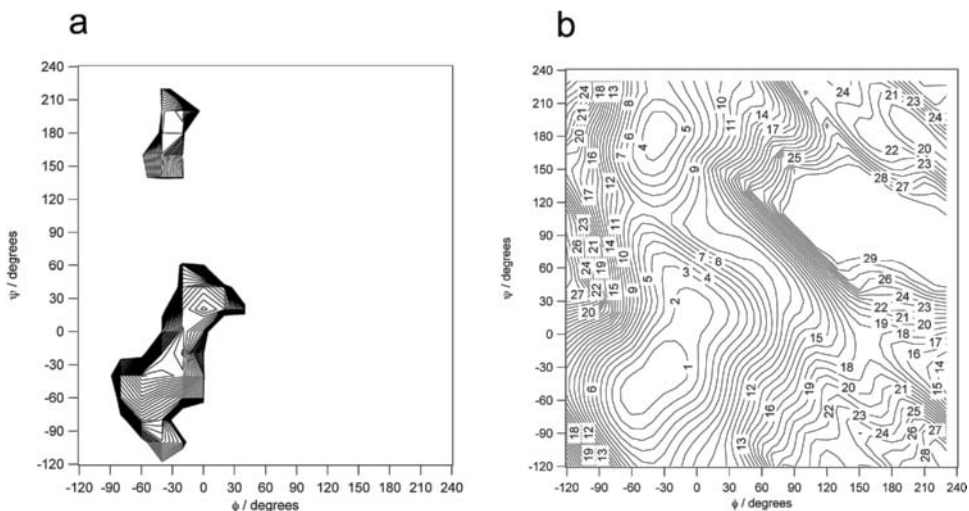


Figure 18.5 Conformational energy maps of maltose [α -D-Glcp-(1–4)- α -D-GlcpOMe] calculated with TINKER/MM3 [210] and CAT [211]: (a) rigid-residue map; (b) adiabatic map. Contours are drawn between 0 and 30 kcal mol⁻¹ in steps of 1 kcal mol⁻¹.

that, as suggested by earlier studies, the 4C_1 chair conformation is the energetically most favorable one for most pyranose rings (Figure 18.3).

During the 1970s the amount of computational work on carbohydrate structures slowly increased. Brant and Goebel published fundamental work on configuration statistics of polysaccharides [41]. Tvaroska *et al.* [42] improved energy functions to calculate *rigid energy maps* (Figure 18.5a) by including van der Waals, torsional, and hydrogen bonding terms. Systematic investigation of the properties of the glycosidic linkage by comparing results from hard-sphere calculations with experimental findings from NMR experiments and optical rotation measurements were performed by Lemieux and Koto [43]. They concluded that the so called *exo-anomeric effect* [44] has a significant influence on the conformational preference of the glycosidic torsion ϕ (defined by the atoms O5–C1–O_x–C_x, see Section 19.2). Due to the *exo-anomeric effect*, a *gauche* orientation ($\phi = \pm 60^\circ$) is more stable than an *anti* orientation ($\phi = 180^\circ$). Quantum-mechanical (*ab initio*) calculations to study the anomeric and *exo-anomeric* effects of simple carbohydrate models were already being performed in the early 1970s [45]. In 1978, Perez and Marchessault published a survey on the ϕ/ψ torsions found in 40 crystal structures of glycopyranosides [46]: it was found that the torsion ϕ varies only slightly and shows exclusively values in agreement with the *exo-anomeric effect*. They suggested that an extra potential function addressing this effect should be developed and included in the energy functions (force field) used to calculate the conformational energy of polysaccharides.

Important milestones in the 1970s with respect to molecular modeling in general were the publication of the MM2 force field by Allinger [47], and the first molecular dynamics (MD) simulations (see Section 19.3.4) of proteins on a picosecond timescale by Karplus and McCammon [48]. Gelin and Karplus also improved the concept of calculating conformational maps to study the conformational energy as a function of two dihedral angles: by relaxing all bonds, angles, and torsion angles (except the two dihedral angles of interest) at each grid point, using energy minimization methods, they calculated the first flexible-geometry maps (*relaxed maps*) [49]. For their study they used a general empirical force field based on the work of Lifson and Warshel [50]. It took another 4 years before energy minimization was introduced into carbohydrate modeling by Melberg and Rasmussen in the late 1970s [51]. The prerequisite for this was the development of a complete *empirical force field for carbohydrates (FF300)*. Using this force field, Melberg and Rasmussen performed energy minimizations of β -maltose [α -D-Glcp-(1–4)- β -D-Glcp] using a steepest descent method and a modified Newton algorithm, and finally they used the minimized structure as input geometry for quantum mechanics calculations [52].

18.3 The *exo-Anomeric Effect* – To Be or Not To Be . . .

In the 1980s, the theoretical models to describe conformations of carbohydrates were further refined by including solvent effects in semiempirical calculations [53], and by developing functions to describe the contributions of the *exo-anomeric effect* [54]. The HSEA (Hard-Sphere-Exo-Anomeric) program [55, 56] was frequently used to calculate 3D models of carbohydrates, in particular to support structure determination by NMR. The agreement between HSEA models and ${}^1\text{H}$ NMR results was remarkably good for the oligosaccharides studied by Lemieux and coworkers [57]. On the other hand, Lipkind *et al.* [58] found that calculation results based on the HSEA method did not give good

agreement with experimental results for cellobiose [β -D-Glcp-(1–4)- β -D-GlcpOMe], and for maltose [α -D-Glcp-(1–4)- β -D-GlcpOMe] the inclusion of the *exo*-anomeric effect was no improvement. Therefore, they concluded that it is unnecessary to take the *exo*-anomeric effect into account, and that the conformational equilibrium of disaccharides could be described sufficiently by force field energy functions that take into account non-bonded interactions and torsional contributions only [59]. The discussion about the importance of the *exo*-anomeric effect on carbohydrate structure was a topic in many publications during the following years. Independently, it became obvious that a weakness of the “rigid-residue” approach to calculating conformational maps was that the results depend heavily on the starting molecular geometry used, and that the relative energies of side minima generally were found to be too high. Tvaroska and Perez [60] suggested therefore to generally relax the geometries of the conformations investigated by energy minimization before evaluating and comparing their energies. Finally, in the late 1980s, the first complete relaxed conformational maps of carbohydrates were calculated by Homans *et al.* [61], using semiempirical methods, and by French [62], who used the MM2 force field for the energy minimization of each conformation on the ϕ/ψ grid. Initially, the calculation of such maps was laborious and, on the available computers, very time consuming. The extra effort required to calculate relaxed maps was justified by the hope of escaping the limitations of the rigid-residue methods.

18.4 Carbohydrates on the Move . . .

During the 1980s, the application of high-field NMR to determine the 3D structure of oligosaccharides in solution [63–66] and to study carbohydrate–protein complexes [67, 68] greatly increased. The analysis of 3D structures of carbohydrates by NMR proved to be a difficult task; therefore, several computational tools and methods had to be developed to assist in the interpretation of the experimental results. It became increasingly clear that oligosaccharides are more flexible than was originally thought. It was found that some experimental results could only be explained by more than one conformation interconverting rapidly on the NMR time-scale and the measured NMR spectrum may therefore represent an average of all the conformations in the ensemble. The concept of “virtual conformations” was introduced to the conformational analysis of oligosaccharides [64]. To judge whether NMR results could be explained by only one low-energy conformation or were a result of a mixture of several low-energy conformations, it was necessary to perform conformational energy calculations that allowed all possible conformational energy minima to be determined accurately. Therefore, there was an urgent need for improved conformational maps that took into account side-chain flexibility (*adiabatic maps*) (Figure 18.5b).

At the end of the 1980s, the progress in improving the quality of the conformational maps was fairly rapid: Tran *et al.* [69] calculated an (almost) adiabatic conformational map of maltose by combining six relaxed maps calculated with different orientations of the primary exocyclic groups using the MM2CARB method (MM2CARB is the MM2 force field modified with the acetal-segment parameters of Jeffrey and Taylor [70] to take into account the *exo*-anomeric effect). In a parallel study, Ha *et al.* [71] calculated an improved, partly adiabatic, conformational map of maltose using a revised CHARMM-type force field. In both studies, it was found that the use of flexible residues had no effect on the general location of the low-energy regions of the maps, but that the accessible area increased and

energy barriers separating the local minima were significantly lowered, making it more likely that conformational transitions occur. Further support for using a flexible-residue approach came from French [72], who compared rigid and relaxed conformational maps of cellobiose and maltose (calculated using MM2) with ϕ/ψ values derived from crystal structures. He found that allowing the monosaccharide units to relax resulted in much better agreement with experimental data and concluded that flexible residues should be used instead of rigid residues when calculating conformational maps. The quality of the conformational maps was further improved by the RAMM (RANdom Molecular Mechanics) method [73], where the pendant groups of the carbohydrate rings were varied using a random-walk technique in order to find the conformer having the lowest energy for each ϕ/ψ grid point. The drawback remained that, even for a disaccharide, the calculation of a complete adiabatic conformational map was not feasible with the computer power available in the late 1980s due to the huge number of conformations that have to be minimized and evaluated (multi-minimum problem) (see Section 19.3.1).

Mapping of calculated NOE intensities [74] or H–H distance constraints [75] as a function of the glycosidic torsions ϕ and ψ , and using the energy landscape of the pseudo-adiabatic maps as additional (background) information in the plots, helped greatly to determine the linkage conformation from 2D NMR experiments. In order to determine the glycosidic linkage conformation by measuring vicinal proton–carbon coupling constants, new Karplus-type equations [76] were developed (see Chapter 20). CASPER [77], a program for computer-assisted structural sequence analysis of polysaccharides using chemical shift data, was released and helped to speed up this time-consuming task.

One of the major observations made during the 1980s was that theoretical results were in better agreement with experimental results when not just one particular rigid conformation of a carbohydrate but the whole conformational distribution was taken into account [58]. Consequently, a new chapter in the conformational analysis of carbohydrates was opened in 1986 when Brady published the first molecular dynamics (MD) simulation of α -D-glucose [78] and Post *et al.* simulated the dynamics of (GlcNAc)₆ bound to lysozyme [79]. Both MD simulations were performed using the general molecular mechanics program CHARMM [80] and the PEF422 parameters developed by Rasmussen [81]. Shortly afterwards, MD simulations of crystalline α -cyclodextrin hexahydrate [82] and oligomannose-type disaccharides [83] using GROMOS [84] were published. For the first time, not only did the flexibility of carbohydrates become “visible”, but also two force fields (and programs), originally developed for the simulation of proteins, were extended with parameters specific for carbohydrates. A revised CHARMM-type force field for carbohydrates was developed by Ha *et al.* [71] and was used to calculate an improved, partly adiabatic, conformational map of maltose and to perform MD simulations in order to understand and describe more precisely the flexibility of the glycosidic linkages of disaccharides [85].

Two more highlights in carbohydrate modeling appeared at the end of the 1980s, namely the release of the MM3 force field by Allinger *et al.* [86] and the first MD simulation of α -D-glucose in explicit water conducted by Brady [87].

18.5 Carbohydrate Meets Protein

In the 1990s, the development of carbohydrate-specific parameters for use with force fields that were originally developed for simulations of proteins (GROMOS [88, 89], OPLS [90]), nucleic acids (AMBER [91–96]), or small molecules (TRIPOS [97]) continued.

General force fields such as CVFF [98] were successfully used in the conformational analysis of oligosaccharides in explicit solvent [99]. MM3 was used to calculate adiabatic conformational maps for a variety of disaccharides [100]. Relaxed quantum mechanical conformational maps [101] and maps calculated in explicit solvent [102] were published. New approaches appeared that addressed the multi-minimum problem in conformational analysis of carbohydrates either by treating OH groups as united atoms (CHEAT [103]), or by improving the conformational search strategy (CICADA [104]). Despite their limitations, HSEA calculations were continuously applied successfully in the context of NMR studies. The HSEA force field for carbohydrates was combined with the ECEPP/2 (empirical conformation energy program for peptides) force field in the GEGOP [105] program for the calculation of glycopeptides. In addition, during this period, the first databases were developed to compile systematically carbohydrate-related information [106–109].

During the 1990s, a variety of methods were applied to the conformational analysis of carbohydrates: Monte Carlo (MC) simulations [110, 111], restrained MD simulations [112], simulated annealing [113], stochastic dynamics (SD) simulations [114], mixed-mode MC/SD simulations [115], free energy calculations [116], adaptive umbrella sampling [117], *ab initio* MD simulations [118], and forced MD simulations (for comparison with atomic force microscopy experiments) [119]. Calculations were performed in the gas phase [120, 121], in explicit solvent [122–124], and by applying a continuum dielectric water solvent model (GB/SA) [125]. The simulation times increased from a few picoseconds to nanoseconds even when using explicit solvent [126]. Initial studies of the interaction of carbohydrates with explicit solvent molecules were also carried out [122, 127–130] and modeling strategies were applied to glycolipids [131].

Also in the 1990s, a number of developments led to a dramatic increase in the understanding of factors important for carbohydrate–protein binding. These included the increasing availability of high-resolution crystal structures of enzymes [132], and lectins containing even large oligosaccharides as ligands [133–137], the application of transfer NOE experiments to study carbohydrate–protein complexes in detail in solution [138], NMR titration studies [139], and titration microcalorimetry studies [140]. MD simulations of carbohydrate–protein complexes [141] and glycoproteins [142] were performed on the picosecond time-scale. Improved strategies for docking of carbohydrates into the binding sites of proteins were developed [97, 143–150], and homology modeling was applied to model the three-dimensional structure of lectins [145]. Interactions between carbohydrates and proteins were studied in detail using computational methods [151]. Using rational, computer-assisted drug design methods and the X-ray crystal structure of influenza virus neuraminidase [152], potent carbohydrate-based inhibitors of the enzyme were designed [153], one of which (zanamivir) is now marketed as the anti-influenza drug Relenza. Finally, SWEET, a web-based molecular builder for carbohydrates [154], appeared at the end of the 1990s, heralding one of the environments for future developments – the Internet.

18.6 Modern Times

With the increasing power and data storage capacity of modern computers, the new millennium brought applications of free energy perturbation (FEP) simulations to protein–carbohydrate complexes [155], and extension of MD simulations in explicit solvent to time-scales above 10 ns [156]. Even *ab initio* molecular dynamics simulation [157]

became feasible. Studies of the energetics of carbohydrate–protein complexes using MD-based approaches such as MM-GB/SA [158] or linear interaction energy (LIE) [159] were performed. The catalytic mechanism of glycosyl transferases was studied using high-level *ab initio* methods [160]. MD simulations of glycolipids embedded into a membrane [161] and of glycolipids forming micelles [162, 163] were also performed. Detailed MD studies of the specific interaction of carbohydrates with water, in the free and bound state, has become one of the hot topics in recent years [164–168]. In order to describe more accurately non-bonded interactions between atoms, efforts to develop polarizable force fields for solvents, proteins, and carbohydrates increased [169–171]. The recent publication of the MM4 force field for carbohydrates [172] and coarse grain force fields for polysaccharide simulations in water [173, 174] shows that also in the field of molecular mechanics there is still ongoing development.

Empirical free energy models for automated docking of carbohydrates to proteins [175, 176] have been developed to improve the results from virtual screening of carbohydrate-based compound libraries. Recent attempts to predict carbohydrate binding sites on proteins [177–180] are encouraging. A number of new experimental techniques, in addition to traditional NMR methods such as transferred nuclear Overhauser enhancement (TR-NOE) [181] or chemically induced dynamic nuclear polarization (CIDNP) [182], brought a deeper insight into the carbohydrate–protein binding event at an atomic level. These include group epitope mapping (GEM) by saturation transfer difference (STD) NMR [183], surface plasmon resonance (SPR), and atomic force microscopy (AFM) experiments [8, 184]. Methods which use the OH protons of carbohydrates as additional conformational sensors in NMR experiments have been used to determine the 3D structures of carbohydrates more precisely [185, 186]. In a recent investigation, the three-dimensional structure of a very large oligosaccharide (containing 30 monosaccharide residues) was partly determined by a combined modeling and NMR approach [187].

In the past 5 years, easy-to-use and freely available web-based tools [188] to perform MD simulations of carbohydrates over the Internet [189], *in silico* glycosylation of proteins [190], and NMR shift prediction [191–193] have all been established. A library-based automated calculation of many thousands of conformational maps [194] is a first example of high-throughput conformational analysis of carbohydrates. The challenges that have been faced just recently are the accurate modeling of the structure and dynamics of glycoproteins under physiological conditions [195] and the development of efficient and accurate methods to estimate the binding affinity of protein–carbohydrate complexes [176, 196–198].

18.7 Conclusion and Outlook

Over the last 60 years, our understanding of carbohydrate structure has greatly increased. However, some fundamental experimental problems still exist: first, it is often difficult to obtain oligosaccharides in large enough quantities to allow 3D structure determination by NMR or X-ray crystallography. In addition, due to their inherent flexibility, it is very difficult to obtain crystals or well-defined electron density that allow unambiguous structure determination. The lack of sufficient inter-residue NOE data can also result in some ambiguity in the 3D structures determined by NMR. In general, NMR studies of carbohydrates also suffer from difficulties in shift assignment or coupling constant determination due to overlapping peaks. Despite these difficulties, the 3D structures of

many poly- and oligosaccharides have been solved by NMR methods. The availability of an automatic carbohydrate synthesizer [199] will further help to drive the field forward.

Due to the tremendous increase in computer power over the years, it has now become feasible to simulate the dynamics of large oligosaccharides, glycoproteins, and protein-carbohydrate complexes in explicit solvent on the 100 ns time-scale. However, some conformational transitions need much longer time-scales in order to be sampled adequately (see Section 19.3). The increase in the simulation time to the microsecond time-scale is difficult since the required CPU time would be very high at the moment (many months) and the requirement on hard disk space for storing the simulation trajectory would also be enormous (many tens of gigabytes). Consequently, the problem of adequate sampling still exists for MD simulations of carbohydrates in solution, and the simulation of the conformational equilibrium is still difficult to achieve in routine simulations.

Although significant limitations still exist, the use of MD simulations has turned out to be an excellent methodology to study the conformational properties of carbohydrates and other biomolecules [200]. In order to make the outcome of an MD simulation more reliable, the theoretical results should always be compared with experimental results if possible. It has to be kept in mind that disagreement between computational and experimental results does not necessarily mean that the force field used is inappropriate for the simulation of carbohydrate structure. It can also mean that other simulation parameters used are not appropriate (e.g. simulation time, solvent model). Obviously, experimental results are very important for validating the quality of theoretical calculations (see also Chapter 20). However, the quality of the experimental data also needs to be taken into account. For example, the 3D structures of carbohydrates that are contained in the PDB are not always reliable. It has been recognized that in order to improve the quality of the 3D structures contained in the PDB, theoretical validation procedures for carbohydrates have to be established [194, 201, 202].

Taking into account the limited number of scientists working in the field of conformational analysis of carbohydrates, the progress that has been made over the years in this difficult field is remarkable. While it has been known for many years that probably the majority of proteins are glycosylated [203], and it has been shown that carbohydrates play an important role in health and disease, surprisingly the worldwide efforts to decipher carbohydrate structure and function are still very limited. Consequently, the field of structural glycobiology clearly lags behind the developments in the protein sciences field. In recent years, this unfortunate situation has been realized by the scientific community and an NIH White Paper was published that emphasizes the need for glycomics approaches in biomarker discovery and drug development [204]. A recent MIT technology review nominated glycomics as one of the 10 emerging technologies that will change the world. Conformational analysis of carbohydrates clearly will play a key role in revealing the function of these fascinating molecules on a molecular level.

References

1. Rao VSR, Qasba PK, Balaji PV, Chandrasekaran R: *Conformation of Carbohydrates*. Amsterdam: Harwood Academic Publishers; 1998.
2. Jeffrey GA: Crystallographic studies of carbohydrates. *Acta Crystallogr Sect B* 1990, **46** (Pt 2): 89–103.

3. Perez S: Oligosaccharide and polysaccharide conformations by diffraction methods. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Oxford: Elsevier; 2007, pp. 193–219.
4. Allen FH, Taylor R: Research applications of the Cambridge Structural Database (CSD). *Chem Soc Rev* 2004, **33**:463–475.
5. Perez S, MouhousRiou N, Nifantev NE, *et al.*: Crystal and molecular structure of a histoblood group antigen involved in cell adhesion: the Lewis x trisaccharide. *Glycobiology* 1996, **6**:537–542.
6. Millane RP: Polysaccharide structures: X-ray fiber diffraction studies. In *Computer Modeling of Carbohydrate Molecules* (eds French AD, Brady JW), ACS Symposium Series, Vol. 430. Washington, DC: American Chemical Society; 1990, pp. 315–331.
7. Nishiyama Y, Sugiyama J, Chanzy H, Langan P: Crystal structure and hydrogen bonding system in cellulose I(alpha) from synchrotron X-ray and neutron fiber diffraction. *J Am Chem Soc* 2003, **125**:14300–14306.
8. Sletmoen M, Maurstad G, Sikorski P, *et al.*: Characterization of bacterial polysaccharides: steps towards single-molecular studies. *Carbohydr Res* 2003, **338**:2459–2475.
9. Qasba PK, Ramakrishnan B: X-ray crystal structures of glycosyltransferases. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Oxford: Elsevier; 2007, pp. 251–281.
10. Buts L, Loris R, Wyns L: X-ray crystallography of lectins. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Amsterdam: Elsevier; 2007, pp. 221–249.
11. Berman H, Henrick K, Nakamura H, Markley JL: The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 2007, **35**:D301–D303.
12. Engler EM, Andose JD, Schleyer R: Critical evaluation of molecular mechanics. *J Am Chem Soc* 1973, **95**:8005–8025.
13. Weiner SJ, Kollman PA, Case DA, *et al.*: A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* 1984, **106**:765–776.
14. Allinger NL, Rahman M, Lii JH: A molecular mechanics force-field (MM3) for alcohols and ethers. *J Am Chem Soc* 1990, **112**:8293–8307.
15. Widmalm G: General NMR spectroscopy of carbohydrates and conformational analysis in solution. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Oxford: Elsevier; 2007, pp. 101–132.
16. Carver JP: Experimental structure determination of oligosaccharides. *Curr Opin Struct Biol* 1991, **1**:716–720.
17. Rice KG, Wu PG, Brand L, Lee YC: Experimental determination of oligosaccharide 3-dimensional structure. *Curr Opin Struct Biol* 1993, **3**:669–674.
18. Vanhalbeek H: NMR developments in structural studies of carbohydrates and their complexes. *Curr Opin Struct Biol* 1994, **4**:697–709.
19. Brady JW: Theoretical studies of oligosaccharide structure and conformational dynamics. *Curr Opin Struct Biol* 1991, **1**:711–715.
20. Tvaroska I: Theoretical aspects of structure and conformation of oligosaccharides. *Curr Opin Struct Biol* 1992, **2**:661–665.
21. Perez S: Theoretical aspects of oligosaccharide conformation. *Curr Opin Struct Biol* 1993, **3**:675–680.
22. Woods RJ: Computational carbohydrate chemistry: what theoretical methods can tell us. *Glycoconj J* 1998, **15**:209–216.
23. Woods RJ: Three-dimensional structures of oligosaccharides. *Curr Opin Struct Biol* 1995, **5**:591–598.
24. Imberty A, Perez S: Structure, conformation, and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations. *Chem Rev* 2000, **100**:4567–4588.

25. Wormald MR, Petrescu AJ, Pao YL, *et al.*: Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modelling. *Chem Rev* 2002, **102**:371–386.
26. French AD, Brady JW (eds): *Computer Modeling of Carbohydrate Molecules*. Washington, DC: American Chemical Society; 1990.
27. Jimenez-Barbero J, Peters T (eds): *NMR Spectroscopy of Glycoconjugates*. Weinheim: Wiley-VCH Verlag GmbH; 2003.
28. Vliegthart JFG, Woods RJ (eds): *NMR Spectroscopy and Computer Modeling of Carbohydrates*. Washington, DC: American Chemical Society; 2006.
29. Kamerling JP (ed.): *Comprehensive Glycoscience – from Chemistry to Systems Biology*. Oxford: Elsevier; 2007.
30. Hassel O, Ottar B: The structure of molecules containing cyclohexane or pyranose rings. *Acta Chem Scand* 1947, **1**:929–942.
31. Reeves RE: The shape of pyranoside rings. *J Am Chem Soc* 1950, **72**:1499–1506.
32. Rao VSR, Foster JF: On the conformation of the D-glucopyranose ring in maltose and in higher polymers of D-glucose. *J Phys Chem* 1963, **67**:951–952.
33. Sundararajan PR, Rao VSR: Theoretical studies on the conformation of aldopyranoses. *Tetrahedron* 1968, **24**:289–295.
34. Chu SSC, Jeffrey GA: The refinement of the crystal structures of β -D-glucose and cellobiose. *Acta Crystallogr Sect B* 1968, **24**:830–838.
35. Ramachandran GN, Ramakrishnan C, Sasisekharan V: *Aspects of Protein Structure*. New York: Academic Press; 1963, p. 121.
36. Rao VS, Sundararajan PR, Ramakrishnan C, Ramachandran GN: *Conformation of Biopolymers*, Vol. 2. New York: Academic Press; 1967, pp. 721–737.
37. Rees DA, Skerrett RJ: Conformational analysis of cellobiose, cellulose, and xylan. *Carbohydr Res* 1968, **7**:334–348.
38. Sathyanarayana BK, Rao VS: Conformational studies on β -D-(1–3)-linked xylan. *Carbohydr Res* 1970, **15**:137–145.
39. Sathyanarayana BK, Rao VS: Conformational studies on β -glucans. *Biopolymers* 1971, **10**:1605–1615.
40. Rao VSR, Vijayalakshmi KS, Sundararajan PR: Theoretical studies on the conformation of aldohexopyranoses. *Carbohydr Res* 1971, **17**:341–352.
41. Brant DA, Goebel KD: A general treatment of the configuration statistics of polysaccharides. *Macromolecules* 1975, **8**:522–530.
42. Tvaroska I, Perez S, Marchessault RH: Conformational analysis of (1–6)- α -D-glucan. *Carbohydr Res* 1978, **61**:97–106.
43. Lemieux RU, Koto S: The conformational properties of glycosidic linkages. *Tetrahedron* 1974, **30**:1933–1944.
44. Lemieux RU, Pavia AA, Martin JC, Watanabe KA: Solvation effects on conformational equilibria. Studies related to the conformational properties of 2-methoxytetrahydropyran and related methyl glycopyranosides. *Can J Chem* 1969, **47**:4427.
45. Jeffrey GA, Pople JA, Radom L: The application of *ab initio* molecular orbital theory to the anomeric effect. A comparison of theoretical predictions and experimental data on conformations and bond lengths in some pyranoses and methyl pyranosides. *Carbohydr Res* 1972, **25**:117–131.
46. Perez S, Marchessault RH: The *exo*-anomeric effect: experimental evidence from crystal structures. *Carbohydr Res* 1978, **65**:114–120.
47. Allinger NL: Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *J Am Chem Soc* 1977, **99**:8127–8134.
48. Karplus M, McCammon JA: Protein structural fluctuations during a period of 100 ps. *Nature* 1979, **277**:578.

49. Gelin BR, Karplus M: Role of structural flexibility in conformational calculations. Application to acetylcholine and beta-methylacetylcholine. *J Am Chem Soc* 1975, **97**:6996–7006.
50. Lifson S, Warshel A: Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. *J Chem Phys* 1968, **49**:5116–5129.
51. Melberg S, Rasmussen K: Conformations of disaccharides by empirical force-field calculations: Part I, β -maltose. *Carbohydr Res* 1979, **69**:27–38.
52. Melberg S, Rasmussen K: The non-bonded interactions in D-glucose and β -maltose: an *ab initio* study of conformations produced by empirical force-field calculations. *Carbohydr Res* 1979, **76**:23–37.
53. Tvaroska I, Kozar T: Theoretical studies on the conformation of saccharides. 3. Conformational properties of the glycosidic linkage in solution and their relation to the anomeric and exoanomeric effects. *J Am Chem Soc* 1980, **102**:6929–6936.
54. Tvaroska I: An attempt to derive the potential function for evaluation of the energy associated with the *exo*-anomeric effect. *Carbohydr Res* 1984, **125**:155–160.
55. Lemieux RU, Bock K, Delbaere LTJ, *et al.*: The conformations of oligosaccharides related to the ABH and Lewis human blood group determinants. *Can J Chem* 1980, **58**:631–653.
56. Thogersen H, Lemieux RU, Bock K, Meyer B: Further justification for the *exo*-anomeric effect. Conformational analysis based on nuclear magnetic resonance spectroscopy of oligosaccharides. *Can J Chem* 1982, **60**:44–57.
57. Lemieux RU, Bock K: The conformational analysis of oligosaccharides by $^1\text{H-NMR}$ and HSEA calculation. *Arch Biochem Biophys* 1983, **221**:125–134.
58. Lipkind GM, Verovsky VE, Kochetkov NK: Conformational states of cellobiose and maltose in solution: a comparison of calculated and experimental data. *Carbohydr Res* 1984, **133**:1–13.
59. Scott RA, Scheraga HA: Conformational analysis of macromolecules. II. The rotational isomeric states of the normal hydrocarbons. *J Chem Phys* 1966, **44**:3054–3069.
60. Tvaroska I, Perez S: Conformational-energy calculations for oligosaccharides: a comparison of methods and a strategy of calculation. *Carbohydr Res* 1986, **149**:389–410.
61. Homans SW, Dwek RA, Rademacher TW: Tertiary structure in N-linked oligosaccharides. *Biochemistry* 1987, **26**:6553–6560.
62. French AD: Rigid- and relaxed-residue conformational analysis of cellobiose using the computer program MM2. *Biopolymers* 1988, **27**:1519–1525.
63. Homans SW, Dwek RA, Rademacher TW: Solution conformations of N-linked oligosaccharides. *Biochemistry* 1987, **26**:6571–6578.
64. Cumming DA, Carver JP: Virtual and solution conformations of oligosaccharides. *Biochemistry* 1987, **26**:6664–6676.
65. Poppe L, Dabrowski J, von der Lieth C-W, *et al.*: Solution conformation of sialosylcerebroside (Gm4) and its NeuAc(α 2–3)Gal- β sugar component. *Eur J Biochem* 1989, **180**:337–342.
66. Dewaard P, Leeftang BR, Vliegthart JFG, *et al.*: Application of 2D and 3D NMR experiments to the conformational study of a diantennary oligosaccharide. *J Biomol NMR* 1992, **2**:211–226.
67. Kronis KA, Carver JP: Specificity of isolectins of wheat germ agglutinin for sialyloligosaccharides: a 360-MHz proton nuclear magnetic resonance binding study. *Biochemistry* 1982, **21**:3050–3057.
68. Kronis KA, Carver JP: Thermodynamics of wheat germ agglutinin–sialyloligosaccharide interactions by proton nuclear magnetic resonance. *Biochemistry* 1985, **24**:834–840.
69. Tran V, Buleon A, Imberty A, Perez S: Relaxed potential-energy surfaces of maltose. *Biopolymers* 1989, **28**:679–690.
70. Jeffrey GA, Taylor R: MM1Carb. *J Comput Chem* 1980, **1**:99–109.
71. Ha SN, Giammona A, Field M, Brady JW: A revised potential-energy surface for molecular mechanics studies of carbohydrates. *Carbohydr Res* 1988, **180**:207–221.
72. French AD: Comparisons of rigid and relaxed conformational maps for cellobiose and maltose. *Carbohydr Res* 1989, **188**:206–211.

73. Kozar T, Petrak F, Galova Z, Tvaroska I: RAMM – a new procedure for theoretical conformational-analysis of carbohydrates. *Carbohydr Res* 1990, **204**:27–36.
74. Brisson JR, Carver JP: Solution conformation of α -D-(1–3)- and α -D-(1–6)-linked oligomannosides using proton nuclear magnetic resonance. *Biochemistry* 1983, **22**:1362–1368.
75. Poppe L, von der Lieth C-W, Dabrowski J: Conformation of the glycolipid globoside head group in various solvents and in the micelle-bound state. *J Am Chem Soc* 1990, **112**:7762–7771.
76. Tvaroska I, Hricovini M, Petrakova E: An attempt to derive a new Karplus-type equation of vicinal proton carbon coupling-constants for C–O–C–H segments of bonded atoms. *Carbohydr Res* 1989, **189**:359–362.
77. Jansson PE, Kenne L, Widmalm G: Computer-assisted structural-analysis of polysaccharides with an extended version of Casper using H-1-NMR and C-13-NMR data. *Carbohydr Res* 1989, **188**:169–191.
78. Brady JW: Molecular dynamics simulations of α -D-glucose. *J Am Chem Soc* 1986, **108**:8153–8160.
79. Post CB, Brooks BR, Karplus M, *et al.*: Molecular dynamics simulations of native and substrate-bound lysozyme. A study of the average structures and atomic fluctuations. *J Mol Biol* 1986, **190**:455–479.
80. Brooks BR, Bruccoleri RE, Olafson BD, *et al.*: CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 1983, **4**:187–217.
81. Rasmussen K: PEF422. *Acta Chem Scand* 1982, **36**:323.
82. Koehler JEH, Saenger W, van Gunsteren WF: A molecular dynamics simulation of crystalline alpha-cyclodextrin hexahydrate. *Eur Biophys J* 1987, **15**:197–210.
83. Homans SW, Pastore A, Dwek RA, Rademacher TW: Structure and dynamics in oligomannose-type oligosaccharides. *Biochemistry* 1987, **26**:6649–6655.
84. van Gunsteren WF, Berendsen HJC: GROMOS. *Mol Phys* 1977, **34**:1311.
85. Ha SN, Madsen LJ, Brady JW: Conformational analysis and molecular dynamics simulations of maltose. *Biopolymers* 1988, **27**:1927–1952.
86. Allinger NL, Yuh YH, Lii JH: Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J Am Chem Soc* 1989, **111**:8551–8566.
87. Brady JW: Molecular-dynamics simulations of alpha-D-glucose in aqueous solution. *J Am Chem Soc* 1989, **111**:5155–5165.
88. Kouwijzer MLCE, Vaneijck BP, Kroon J: An extension of the Gromos force-field for carbohydrates, resulting in improvement of the crystal-structure determination of alpha-D-galactose. *Acta Crystallogr Sect B* 1995, **51**:209–220.
89. Spieser SAH, van Kuik JA, Kroon-Batenburg LMJ, Kroon J: Improved carbohydrate force field for GROMOS: ring and hydroxymethyl group conformations and *exo*-anomeric effect. *Carbohydr Res* 1999, **322**:264–273.
90. Damm W, Frontera A, TiradoRives J, Jorgensen WL: OPLS all-atom force field for carbohydrates. *J Comput Chem* 1997, **18**:1955–1970.
91. Weiner PK, Kollman PA: AMBER: assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J Comput Chem* 1981, **2**:287–303.
92. Homans SW: A molecular mechanical force field for the conformational analysis of oligosaccharides: comparison of theoretical and crystal structures of Man α (1–3)Man β (1–4)GlcNAc. *Biochemistry* 1990, **29**:9110–9118.
93. Edge CJ, Singh UC, Bazzo R, *et al.*: 500-picosecond molecular-dynamics in water of the Man-alpha-1–2-Man-alpha glycosidic linkage present in Asn-linked oligomannose-type structures on glycoproteins. *Biochemistry* 1990, **29**:1971–1974.
94. Woods RJ, Dwek RA, Edge CJ, Fraserreid B: Molecular mechanical and molecular dynamical simulations of glycoproteins and oligosaccharides. 1. Glycam-93 parameter development. *J Phys Chem* 1995, **99**:3832–3846.

95. Senderowitz H, Parish C, Still WC: Carbohydrates: united atom AMBER* parameterization of pyranoses and simulations yielding anomeric free energies. *J Am Chem Soc* 1996, **118**:2078–2086.
96. Senderowitz H, Still WC: A quantum mechanically derived all-atom force field for pyranose oligosaccharides. AMBER parameters and free energy simulations. *J Org Chem* 1997, **62**:1427–1438.
97. Imberty A, Hardman KD, Carver JP, Perez S: Molecular modelling of protein–carbohydrate interactions. Docking of monosaccharides in the binding site of concanavalin A. *Glycobiology* 1991, **1**:631–642.
98. Hagler AT, Lifson S, Dauber P: Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 2. A benchmark for the objective comparison of alternative force fields. *J Am Chem Soc* 1979, **101**:5122–5130.
99. Siebert HC, Reuter G, Schauer R, *et al.*: Solution conformations of GM3 gangliosides containing different sialic acid residues as revealed by NOE-based distance mapping, molecular mechanics, and molecular dynamics calculations. *Biochemistry* 1992, **31**:6962–6971.
100. Dowd MK, Zeng J, French AD, Reilly PJ: Conformational analysis of the anomeric forms of kojibiose, nigerose, and maltose using Mm3. *Carbohydr Res* 1992, **230**:223–244.
101. Mazeau K, Tvaroska I: Pcilo quantum-mechanical relaxed conformational energy map of methyl 4-thio- α -maltoside in solution. *Carbohydr Res* 1992, **225**:27–41.
102. Naidoo KJ, Brady JW: Calculation of the Ramachandran potential of mean force for a disaccharide in aqueous solution. *J Am Chem Soc* 1999, **121**:2244–2252.
103. Grootenhuis PDJ, Haasnoot CAG: A Charmm based force-field for carbohydrates using the Cheat approach – carbohydrate hydroxyl-groups represented by extended atoms. *Mol Simul* 1993, **10**:75–95.
104. Koca J, Perez S, Imberty A: Conformational analysis and flexibility of carbohydrates using the Cicada approach with MM3. *J Comput Chem* 1995, **16**:296–310.
105. Stuikeprill R, Meyer B: A new force-field program for the calculation of glycopeptides and its application to a heptacosapeptide-decasaccharide of immunoglobulin-G1 – importance of 1–6-glycosidic linkages in carbohydrate–peptide interactions. *Eur J Biochem* 1990, **194**:903–919.
106. Doubet S, Bock K, Smith D, *et al.*: The Complex Carbohydrate Structure Database. *Trends Biochem Sci* 1989, **14**:475–477.
107. Imberty A, Gerber S, Tran V, Perez S: Data-bank of 3-dimensional structures of disaccharides, a tool to build 3-D structures of oligosaccharides. 1. Oligo-mannose type *N*-glycans. *Glycoconj J* 1990, **7**:27–54.
108. Perez S, Delage MM: A Database of 3-dimensional structures of monosaccharides from molecular-mechanics calculations. *Carbohydr Res* 1991, **212**:253–259.
109. van Kuik JA, Vliegthart JF: Databases of complex carbohydrates. *Trends Biotechnol* 1992, **10**:182–185.
110. Peters T, Meyer B, Stuikeprill R, *et al.*: A Monte-Carlo method for conformational-analysis of saccharides. *Carbohydr Res* 1993, **238**:49–73.
111. Poppe L, Stuikeprill R, Meyer B, Vanhalbeek H: The solution conformation of sialyl- α (2–6)-lactose studied by modern Nmr techniques and Monte-Carlo simulations. *J Biomol NMR* 1992, **2**:109–136.
112. Rutherford TJ, Homans SW: Restrained vs free dynamics simulations of oligosaccharides: application to solution dynamics of biantennary and bisected biantennary *N*-linked glycans. *Biochemistry* 1994, **33**:9606–9614.
113. Homans SW, Forster M: Application of restrained minimization, simulated annealing and molecular-dynamics simulations for the conformational-analysis of oligosaccharides. *Glycobiology* 1992, **2**:143–151.
114. Hardy BJ, Gutierrez A, Lesiak K, *et al.*: Structural analysis of the solution conformation of methyl 4-*O*- β -D-glucopyranosyl- α -D-glucopyranoside by molecular mechanics and *ab*

- initio* calculation, stochastic dynamics simulation, and NMR spectroscopy. *J Phys Chem* 1996, **100**:9187–9192.
115. Bernardi A, Raimondi L, Zanferrari D: Conformational analysis of saccharides with Monte Carlo stochastic dynamics simulations. *THEOCHEM* 1997, **395**:361–373.
 116. Ha SH, Gao JL, Tidor B, *et al.*: Solvent effect on the anomeric equilibrium in D-glucose – a free-energy simulation analysis. *J Am Chem Soc* 1991, **113**:1553–1557.
 117. Vanejck BP, Hooft RWW, Kroon J: Molecular-dynamics study of conformational and anomeric equilibria in aqueous D-glucose. *J Phys Chem* 1993, **97**:12093–12099.
 118. Molteni C, Parrinello M: Glucose in aqueous solution by first principles molecular dynamics. *J Am Chem Soc* 1998, **120**:2168–2171.
 119. Rief M, Oesterhelt F, Heymann B, Gaub HE: Single molecule force spectroscopy on polysaccharides by atomic force microscopy. *Science* 1997, **275**:1295–1297.
 120. Qasba PK, Balaji PV, Rao VS: Molecular dynamics simulations of oligosaccharides and their conformation in the crystal structure of lectin–carbohydrate complex: importance of the torsion angle psi for the orientation of alpha 1,6-arm. *Glycobiology* 1994, **4**:805–815.
 121. Widmalm G, Venable RM: Molecular-dynamics simulation and NMR-study of a blood group-H trisaccharide. *Biopolymers* 1994, **34**:1079–1088.
 122. Brady JW, Schmidt RK: The role of hydrogen bonding in carbohydrates: molecular dynamics simulations of maltose in aqueous solution. *J Phys Chem* 1993, **97**:958–966.
 123. Woods RJ, Fraserreid B, Dwek RA, Edge CJ: Role of nonbonded interactions in determining solution conformations of oligosaccharides. In *Modeling the Hydrogen Bond* (ed. Smith DA), ACS Symposium Series, Vol. 569. Washington, DC: American Chemical Society; 1994, pp. 252–268.
 124. Bagley S, Odelius M, Laaksonen A, Widmalm G: Molecular-dynamics simulation of sucrose in aqueous and dimethyl-sulfoxide solution. *Acta Chem Scand* 1994, **48**:792–799.
 125. Bernardi A, Raimondi L: Conformational-analysis of GM1 oligosaccharide in water solution with a new set of parameters for the Neu5Ac moiety. *J Org Chem* 1995, **60**:3370–3377.
 126. Engelsen SB, Dupenhoat CH, Perez S: Molecular relaxation of sucrose in aqueous-solution – how a nanosecond molecular-dynamics simulation helps to reconcile NMR data. *J Phys Chem* 1995, **99**:13334–13351.
 127. Leeftang BR, Vliegthart JFG, Kroonbatenburg LMJ, *et al.*: A H-1-NMR and MD study of intramolecular hydrogen-bonds in methyl beta-cellobioside. *Carbohydr Res* 1992, **230**:41–61.
 128. Liu Q, Brady JW: Anisotropic solvent structuring in aqueous sugar solutions. *J Am Chem Soc* 1996, **118**:12276–12286.
 129. Wang CX, Chen WZ, Tran V, Douillard R: Analysis of interfacial water structure and dynamics in alpha-maltose solution by molecular dynamics simulation. *Chem Phys Lett* 1996, **251**:268–274.
 130. Vishnyakov A, Widmalm G, Kowalewski J, Laaksonen A: Molecular dynamics simulation of the alpha-D-Manp-(1–3)-beta-D-Glcp-OMe disaccharide in water and water DMSO solution. *J Am Chem Soc* 1999, **121**:5403–5412.
 131. Nyholm PG, Pascher I: Orientation of the saccharide chains of glycolipids at the membrane-surface – conformational-analysis of the glucose ceramide and the glucose glyceride linkages using molecular mechanics (MM3). *Biochemistry* 1993, **32**:1225–1234.
 132. Norris GE, Stillman TJ, Anderson BF, Baker EN: The three-dimensional structure of PNGase F, a glycosylasparaginase from *Flavobacterium meningosepticum*. *Structure* 1994, **2**:1049–1059.
 133. Weis WI, Brown JH, Cusack S, *et al.*: Structure of the influenza virus hemagglutinin complexed with its receptor, sialic acid. *Nature* 1988, **333**:426–431.
 134. Bourne Y, Rouge P, Cambillau C: X-ray structure of a biantennary octasaccharide-lectin complex refined at 2.3 Å resolution. *J Biol Chem* 1992, **267**:197–203.
 135. Weis WI, Drickamer K, Hendrickson WA: Structure of a C-type mannose-binding protein complexed with an oligosaccharide. *Nature* 1992, **360**:127–134.

136. Wright CS: Crystal structure of a wheat germ agglutinin/glycophorin–sialoglycopeptide receptor complex. Structural basis for cooperative lectin–cell binding. *J Biol Chem* 1992, **267**:14345–14352.
137. Bourne Y, Bolgiano B, Liao DI, *et al.*: Crosslinking of mammalian lectin (galectin-1) by complex biantennary saccharides. *Nat Struct Biol* 1994, **1**:863–870.
138. Bevilacqua VL, Kim YM, Prestegard JH: Conformation of β -methylmelibiose bound to the ricin B-chain as determined from transferred nuclear Overhauser effects. *Biochemistry* 1992, **31**:9339–9349.
139. Asensio JL, Canada FJ, Bruix M, *et al.*: The interaction of hevein with *N*-acetylglucosamine-containing oligosaccharides – solution structure of hevein complexed to chitobiose. *Eur J Biochem* 1995, **230**:621–633.
140. Williams BA, Chervenak MC, Toone EJ: Energetics of lectin–carbohydrate binding – a microcalorimetric investigation of concanavalin α -oligomannoside complexation. *J Biol Chem* 1992, **267**:22907–22911.
141. Asensio JL, Canada FJ, Jimenezbarbero J: Studies of the bound conformations of methyl alpha-lactoside and methyl beta-allolactoside to ricin-B chain using transferred NOE experiments in the laboratory and rotating frames, assisted by molecular mechanics and dynamics calculations. *Eur J Biochem* 1995, **233**:618–630.
142. Naidoo KJ, Brady JW: Molecular dynamics simulations of a glycoprotein: the lectin from *Erythrina corallodendron*. *THEOCHEM* 1997, **395**:469–475.
143. Carver JP, MacKenzie AE, Hardman KD: Molecular model for the complex between concanavalin A and a biantennary-complex class glycopeptide. *Biopolymers* 1985, **24**:49–63.
144. Biswas M, Sekharudu YC, Rao VS: The conformation of glycans of the oligo-D-mannosidic type, and their interaction with concanavalin A: a computer-modelling study. *Carbohydr Res* 1987, **160**:151–170.
145. Gohier A, Espinosa JF, JimenezBarbero J, *et al.*: Knowledge based modeling of a legume lectin and docking of the carbohydrate ligand: the *Ulex europaeus* lectin I and its interaction with fucose. *J Mol Graphics Modell* 1996, **14**:322–327.
146. Smith JA, GomezPaloma L, Case DA, Chazin WJ: Molecular dynamics docking driven by NMR-derived restraints to determine the structure of the calicheamicin gamma(I)(1) oligosaccharide domain complexed to duplex DNA. *Magn Reson Chem* 1996, **34**:S147–S155.
147. Coutinho PM, Dowd MK, Reilly PJ: Automated docking of isomaltose analogues in the glucoamylase active site. *Carbohydr Res* 1997, **297**:309–324.
148. Rao VS, Lam K, Qasba PK: Architecture of the sugar binding sites in carbohydrate binding proteins – a computer modeling study. *Int J Biol Macromol* 1998, **23**:295–307.
149. Minke WE, Diller DJ, Hol WGJ, Verlinde CLMJ: The role of waters in docking strategies with incremental flexibility for carbohydrate derivatives: heat-labile enterotoxin, a multivalent test case. *J Med Chem* 1999, **42**:1778–1788.
150. Visegrady B, Than NG, Kilar F, *et al.*: Homology modelling and molecular dynamics studies of human placental tissue protein 13 (galectin-13). *Protein Eng* 2001, **14**:875–880.
151. von der Lieth CW, Kozar T: Towards a better semiquantitative estimation of binding constants: molecular dynamics exploration of the conformational behavior of isolated sialyllactose and sialyllactose complexed with influenza A hemagglutinin. *THEOCHEM* 1996, **368**:213–222.
152. Varghese JN, McKimm-Breschkin JL, Caldwell JB, *et al.*: The structure of the complex between influenza-virus neuraminidase and sialic-acid, the viral receptor. *Proteins: Struct Funct Genet* 1992, **14**:327–332.
153. von Itzstein M, Wu WY, Kok GB, *et al.*: Rational design of potent sialidase-based inhibitors of influenza-virus replication. *Nature* 1993, **363**:418–423.
154. Bohne A, Lang E, von der Lieth C-W: W3-Sweet: carbohydrate modeling by Internet. *J Mol Model* 1998, **4**:33–43.

155. Pathiaseril A, Woods RJ: Relative energies of binding for antibody–carbohydrate–antigen complexes computed from free-energy simulations. *J Am Chem Soc* 2000, **122**:331–338.
156. Almond A, Bunkenborg J, Franch T, *et al.*: Comparison of aqueous molecular dynamics with NMR relaxation and residual dipolar couplings favors internal motion in a mannose oligosaccharide. *J Am Chem Soc* 2001, **123**:4792–4802.
157. Qian X, Nimlos MR, Davis M, *et al.*: *Ab initio* molecular dynamics simulations of beta-D-glucose and beta-D-xylose degradation mechanisms in acidic aqueous solution. *Carbohydr Res* 2005, **340**:2319–2327.
158. Bryce RA, Hillier IH, Naismith JH: Carbohydrate-protein recognition: molecular dynamics simulations and free energy analysis of oligosaccharide binding to concanavalin A. *Biophys J* 2001, **81**:1373–1388.
159. Wall ID, Leach AR, Salt DW, *et al.*: Binding constants of neuraminidase inhibitors: an investigation of the linear interaction energy method. *J Med Chem* 1999, **42**:5142–5152.
160. Tvaroska I, Andre I, Carver JP: *Ab initio* molecular orbital study of the catalytic mechanism of glycosyltransferases: description of reaction pathways and determination of transition-state structures for inverting *N*-acetylglucosaminyltransferases. *J Am Chem Soc* 2000, **122**:8762–8776.
161. Zuegg J, Gready JE: Molecular dynamics simulation of human prion protein including both N-linked oligosaccharides and the GPI anchor. *Glycobiology* 2000, **10**:959–974.
162. Bogusz S, Venable RM, Pastor RW: Molecular dynamics simulations of octyl glucoside micelles: structural properties. *J Phys Chem B* 2000, **104**:5462–5470.
163. Vasudevan SV, Balaji PV: Dynamics of ganglioside headgroup in lipid environment: molecular dynamics simulations of GM1 embedded in dodecylphosphocholine micelle. *J Phys Chem B* 2001, **105**:7033–7041.
164. Pratap JV, Bradbrook GM, Reddy GB, *et al.*: The combination of molecular dynamics with crystallography for elucidating protein–ligand interactions: a case study involving peanut lectin complexes with T-antigen and lactose. *Acta Crystallogr, Sect D: Biol Crystallogr* 2001, **57**:1584–1594.
165. Lee SL, DeBenedetti PG, Errington JR: A computational study of hydration, solution structure, and dynamics in dilute carbohydrate solutions. *J Chem Phys* 2005, **122**:204511;1–10.
166. Kuttel MM, Naidoo KJ: Free energy surfaces for the alpha(1→4)-glycosidic linkage: implications for polysaccharide solution structure and dynamics. *J Phys Chem B* 2005, **109**:7468–7474.
167. Lerbret A, Bordat P, Affouard F, *et al.*: Influence of homologous disaccharides on the hydrogen-bond network of water: complementary Raman scattering experiments and molecular dynamics simulations. *Carbohydr Res* 2005, **340**:881–887.
168. Momany FA, Appell M, Strati G, Willett JL: B3LYP/6–311++G** study of monohydrates of alpha- and beta-D-glucopyranose: hydrogen bonding, stress energies, and effect of hydration on internal coordinates. *Carbohydr Res* 2004, **339**:553–567.
169. Halgren TA, Damm W: Polarizable force fields. *Curr Opin Struct Biol* 2001, **11**:236–242.
170. Ren P, Ponder JW: Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J Comput Chem* 2002, **23**:1497–1506.
171. Muslim AM, Bryce RA: Carbohydrate conformation in aqueous solution: calculation of a QM/MM potential of mean force. *Chem Phys Lett* 2004, **388**:473–478.
172. Allinger NL, Chen KH, Lii JH, Durkin KA: Alcohols, ethers, carbohydrates, and related compounds. I. The MM4 force field for simple compounds. *J Comput Chem* 2003, **24**:1447–1472.
173. Molinero V, Goddard WA: M3B: a coarse grain force field for molecular simulations of malto-oligosaccharides and their water mixtures. *J Phys Chem B* 2004, **108**:1414–1427.
174. Bathe M, Rutledge GC, Grodzinsky AJ, Tidor B: A coarse-grained molecular model for glycosaminoglycans: application to chondroitin, chondroitin sulfate, and hyaluronic acid. *Biophys J* 2005, **88**:3870–3887.

175. Laederach A, Reilly PJ: Specific empirical free energy function for automated docking of carbohydrates to proteins. *J Comput Chem* 2003, **24**:1748–1757.
176. Hill AD, Reilly PJ: A Gibbs free energy correlation for automated docking of carbohydrates. *J Comput Chem* 2008, **29**:1131–1141.
177. Taroni C, Jones S, Thornton JM: Analysis and prediction of carbohydrate binding sites. *Protein Eng* 2000, **13**:89–98.
178. Neumann D, Kohlbacher O, Lenhof HP, Lehr CM: Lectin–sugar interaction – calculated versus experimental binding energies. *Eur J Biochem* 2002, **269**:1518–1524.
179. Jambon M, Imberty A, Deleage G, Geourjon C: A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins: Struct Funct Genet* 2003, **52**:137–145.
180. Voss C, Eyol E, Frank M, *et al.*: Identification and characterization of riproximin, a new type II ribosome-inactivating protein with antineoplastic activity from *Ximenia americana*. *FASEB J* 2006, **20**:1194–1196.
181. Otting G: Experimental NMR techniques for studies of protein–ligand interactions. *Curr Opin Struct Biol* 1993, **3**:760–768.
182. Siebert HC, Kaptein R, Beintema JJ, *et al.*: Carbohydrate–protein interaction studies by laser photo CIDNP NMR methods. *Glycoconj J* 1997, **14**:531–534.
183. Mayer M, Meyer B: Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. *J Am Chem Soc* 2001, **123**:6108–6117.
184. Dettmann W, Grandbois M, Andre S, *et al.*: Differences in zero-force and force-driven kinetics of ligand dissociation from beta-galactoside-specific proteins (plant and animal lectins, immunoglobulin G) monitored by plasmon resonance and dynamic single molecule force microscopy. *Arch Biochem Biophys* 2000, **383**:157–170.
185. Siebert HC, Andre S, Asensio JL, *et al.*: A new combined computational and NMR-spectroscopical strategy for the identification of additional conformational constraints of the bound ligand in an aprotic solvent. *ChemBioChem* 2000, **1**:181–195.
186. Sandstrom C, Baumann H, Kenne L: The use of chemical shifts of hydroxy protons of oligosaccharides as conformational probes for NMR studies in aqueous solution. Evidence for persistent hydrogen bond interaction in branched trisaccharides. *J Chem Soc, Perkin Trans 2* 1998: 2385–2393.
187. Rodriguez-Carvajal MA, Herve du Penhoat C, Mazeau K, *et al.*: The three-dimensional structure of the mega-oligosaccharide rhamnogalacturonan II monomer: a combined molecular modeling and NMR investigation. *Carbohydr Res* 2003, **338**:651–671.
188. Berteau O, Stenutz R: Web resources for the carbohydrate chemist. *Carbohydr Res* 2004, **339**:929–936.
189. Frank M, Gutbrod P, Hassayoun C, von der Lieth C-W: Dynamic molecules: molecular dynamics for everyone. An Internet-based access to molecular dynamic simulations: basic concepts. *J Mol Model* 2003, **9**:308–315.
190. Bohne-Lang A, von der Lieth C-W: GlyProt: *in silico* glycosylation of proteins. *Nucleic Acids Res* 2005, **33**:W214–W219.
191. Lutteke T, Bohne-Lang A, Loss A, *et al.*: GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* 2006, **16**:71R–81R.
192. Loss A, Stenutz R, Schwarzer E, von der Lieth C-W: GlyNest and CASPER: two independent approaches to estimate ¹H and ¹³C NMR shifts of glycans available through a common web-interface. *Nucleic Acids Res* 2006, **34**:W733–W737.
193. Jansson PE, Stenutz R, Widmalm G: Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel web-based version of the computer program CASPER. *Carbohydr Res* 2006, **341**:1003–1010.
194. Frank M, Lutteke T, von der Lieth C-W: GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res* 2007, **35**:287–290.

195. Blanchard V, Frank M, Leeﬂang BR, *et al.*: The structural basis of the difference in sensitivity for PNGase F in the de-*N*-glycosylation of the native bovine Pancreatic ribonucleases B and BS. *Biochemistry* 2008, **47**:3435–3446.
196. Neumann D, Lehr CM, Lenhof HP, Kohlbacher O: Computational modeling of the sugar–lectin interaction. *Adv Drug Deliv Rev* 2004, **56**:437–457.
197. Kerzmann A, Neumann D, Kohlbacher O: SLICK – scoring and energy functions for protein–carbohydrate interactions. *J Chem Inf Model* 2006, **46**:1635–1642.
198. DeMarco ML, Woods RJ: Structural glycobiology: a game of snakes and ladders. *Glycobiology* 2008, **18**:426–440.
199. Seeberger PH: Automated oligosaccharide synthesis. *Chem Soc Rev* 2008, **37**:19–28.
200. Karplus M, McCammon JA: Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 2002, **9**:646–652.
201. Lutteke T, Frank M, von der Lieth C-W: Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr Res* 2004, **339**:1015–1020.
202. Crispin M, Stuart DI, Jones EY: Building meaningful models of glycoproteins. *Nat Struct Mol Biol* 2007, **14**:354; discussion 354–355.
203. Apweiler R, Hermjakob H, Sharon N: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999, **1473**:4–8.
204. Packer NH, von der Lieth C-W, Aoki-Kinoshita KF, *et al.*: Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda, MD (September 11–13, 2006). *Proteomics* 2008, **8**:8–20.
205. Winter WT, Smith PJ, Arnott S: Hyaluronic acid: structure of a fully extended 3-fold helical sodium salt and comparison with the less extended 4-fold helical forms. *J Mol Biol* 1975, **99**:219–235.
206. Arnott S, Scott WE, Rees DA, McNab CG: Iota-carrageenan: molecular structure and packing of polysaccharide double helices in oriented fibres of divalent cation salts. *J Mol Biol* 1974, **90**:253–267.
207. Moorhouse R, Winter WT, Arnott S, Bayer ME: Conformation and molecular organization in fibers of the capsular polysaccharide from *Escherichia coli* M41 mutant. *J Mol Biol* 1977, **109**:373–391.
208. Tan TC, Mijts BN, Swaminathan K, *et al.*: Crystal structure of the polyextremophilic alpha-amylase AmyB from *Halothermothrix orenii*: details of a productive enzyme-substrate complex and an N domain with a role in binding raw starch. *J Mol Biol* 2008, **378**:850–868.
209. Stortz CA: Comparative performance of MM3(92) and two TINKER MM3 versions for the modeling of carbohydrates. *J Comput Chem* 2005, **26**:471–483.
210. Ponder J: Tinker 4; <http://dasher.wustl.edu/tinker/>.
211. Frank M: Conformational Analysis Tools (CAT); <http://www.md-simulations.de/CAT/>.

19

Predicting Carbohydrate 3D Structures Using Theoretical Methods

Martin Frank

Deutsches Krebsforschungszentrum (German Cancer Research Centre), Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany

19.1 Introduction

Because of the scarcity of experimental data on carbohydrate 3D structure, computational methods have often played an important role in the analysis of carbohydrate conformation. In the “early days” of computer modeling of carbohydrates, starting in the 1960s, calculations and method developments were mainly focused on supporting polysaccharide structure determination and the development of carbohydrate force fields that could reproduce the conformations found in crystal structures of smaller carbohydrates. During the 1980s, glycoproteins and protein–carbohydrate complexes attracted greater attention and carbohydrates were more and more studied in the context of being ligands for proteins, or covalently attached to a protein as *N*- or *O*-glycans. This required improved methods, especially force fields, to be developed that were able to describe accurately the structure and dynamics of proteins and carbohydrates under “physiological conditions.”

Over the years, a great variety of computational methods to study carbohydrates have been developed (see Chapter 18) and are now applied routinely to support the interpretation of experimental results (see Chapter 20). User-friendly, web-based programs are now available to build up reasonable 3D models of carbohydrates [1]. Recently, easy-to-use open-access web interfaces have also been developed to perform molecular dynamics simulations of carbohydrates [2]. The availability of such tools generates the impression that carbohydrate modeling has become an easy task and that one can readily perform a conformational analysis within a short time just by making a few mouse clicks. In practice, this can be true for the modeling of small carbohydrates *in vacuo*, but a detailed simulation of the conformational properties of complex biomolecules such as *N*-glycans, especially when attached to proteins, under “physiological conditions”, often requires the use of sophisticated “academic” software tools [3–9]. These modeling software packages, unfortunately, can be far from “easy-to-use”, especially when building blocks or parameters are required that are not predefined in the template library. Normally no graphical user interface (GUI) is available; therefore performing molecular simulations of carbohydrates frequently requires low-level setup in a UNIX terminal. In addition, the available tools for

data analysis with respect to carbohydrate-specific properties are currently far from being as well developed as they are for the study of proteins. The decision as to which modeling strategy would be the most appropriate to be applied to the study of a specific scientific question is not always straightforward and therefore often requires “expert knowledge” in order to avoid pitfalls. As a result, only a few research groups worldwide are continuously publishing scientific papers dealing with carbohydrate modeling.

A variety of modeling methods have been applied to the conformational analysis of carbohydrates (see Chapter 18). Of these, the calculation of adiabatic maps for disaccharides using systematic search methods, and molecular dynamics simulations of oligosaccharides in explicit solvent, are by far the most popular methods in modeling of carbohydrate 3D structures at the moment. Although quantum mechanical (*ab initio*) methods are frequently used in carbohydrate modeling, these methods are still computationally too demanding to be used routinely to study complex carbohydrates. Quantum mechanics methods [10] are mainly used to study enzymatic reactions [11] and to calculate force constants and atom charges to be used as force field parameters [12, 13]. The development of carbohydrate force fields in itself is a challenging task and is still in progress [8, 9, 13–17]. Since carbohydrates are polar molecules, the proper treatment of atom charges is likely to be of significant importance, particularly for modeling of intermolecular interactions. The discussion about including extra terms for (*exo*) anomeric effects into force fields has a long tradition in carbohydrate modeling (see Chapter 18). Another, not yet completely solved, problem in carbohydrate modeling is the difficulty of general force fields to reproduce quantitatively the experimentally derived rotamer distribution of the exocyclic primary alcohol (hydroxymethyl) groups of carbohydrates [15, 18, 19].

Computer modeling of carbohydrates is frequently driven by the question: “What is the 3D structure of carbohydrate X?” A general modeling strategy to answer this question would be to find the conformation with minimum potential energy (the “global minimum”). However, there are two major drawbacks in following this approach. First, alternative conformations (local minima) may exist that have only slightly higher potential energy than the global minimum and that are therefore also populated at room temperature. Consequently, these conformations will contribute to the properties of the carbohydrate and should therefore not be neglected. Second, since the search methods used to find the global minimum are normally applied in vacuum or implicit solvent, the energy ranking of the different local minima might change when taking into account interactions with explicit solvent molecules. Fairly often the global minimum found in vacuum is stabilized by one or more intramolecular hydrogen bonds. When explicit solvent molecules are present, these intramolecular hydrogen bonds may not occur very frequently because solvent molecules may serve as alternative partners for hydrogen bonding [20, 21]. As a result, a different local minimum might be preferred in solution, because the interaction with the solvent might be more favorable, or the solvent itself can adopt a more favorable state and the system as a whole represents a state of lower free energy [22, 23]. In fact, complex carbohydrates are in general flexible, and can easily adopt a conformation that represents a conformational state of higher potential energy than the global minimum of the carbohydrate itself, if this results in a lower free energy of the complete biological system. This occurs frequently when a carbohydrate binds to a protein surface. However, a carbohydrate obviously cannot adopt a conformation where overlapping of atoms occurs. As a conclusion, it is not necessarily of the utmost importance to determine precisely the conformation of the global minimum in a general conformational analysis of complex carbohydrates since, as noted above, the exact conformation of the global minimum depends on the environment. It is much more

important to determine the accessible conformational space so that one can obtain a clear picture of what the possible shapes are that a carbohydrate can adopt. This can have more significance in a biological context as it can indicate the conformations that are accessible for protein–carbohydrate binding, which may trigger a physiological process or an immune response. As a consequence, three-dimensional structures of complex carbohydrates should be discussed as *conformational ensembles* instead of individual conformations.

This chapter gives an overview of the factors that determine the 3D structure of a carbohydrate. Computational methods that are frequently used to explore the conformational space of carbohydrates are briefly discussed and finally a convenient way to generate 3D structures of glycoproteins is described.

19.2 Factors That Have an Influence on the 3D Structure of Carbohydrates

A typical complex carbohydrate structure is built up from monosaccharides that are connected through glycosidic linkages (Figure 19.1). In contrast to amino acids and nucleic acids, the monosaccharides can be linked at multiple positions, and the chains can therefore contain branches. The overall 3D structure of a carbohydrate depends mainly on the type of glycosidic linkage(s) present and on intrinsic properties of the monosaccharide such as stereochemistry and substitution pattern. Branching can also have a significant influence on the flexibility and shape of a complex carbohydrate.

19.2.1 Type and Conformation of the Monosaccharide

The building blocks of carbohydrates – the monosaccharides – occur mostly as five-membered furanose (*f*) or six-membered pyranose (*p*) rings, and can carry bulky (e.g. acetamido) or charged (e.g. sulfate, amino) substituents (Figure 19.2). The six-membered pyranose rings normally exist in a chair conformation, whereas five-membered furanose rings prefer envelope or twist forms. Ring conformation is indicated using the numbers of the ring atoms: 4C_1 denotes a chair conformation in which the C4 atom is above, and the C1 atom is below, the plane formed by the remaining ring atoms C2, C3, C5, and O5 in a clockwise orientation (see Figure 18.3). For pyranose rings, the energy barrier between the two possible chair forms (1C_4 and 4C_1) is relatively high. 4C_1 is energetically much more stable than the 1C_4 form for most common D-monosaccharides (e.g. $\Delta E = 8 \text{ kcal mol}^{-1}$ for β -D-Glcp [24]).¹ As a consequence, most modeling studies do not consider ring puckering if the carbohydrate consists only of rings in the pyranose form. The situation is different if the carbohydrate contains rings in the furanose form. These rings are more flexible and the various ring conformations can easily interconvert due to significantly lower energy barriers [25, 26]. Ring puckering greatly increases the number of possible conformations and this renders performing a conformational analysis of complex carbohydrates that contain monosaccharides in furanose ring form very complicated. Cremer–Pople [27] or Altona–Sundaralingam [28] parameters are generally used to describe ring puckering. Due to the Hassel–Ottar effect [29], the preferred ring conformations are those that put the most exocyclic groups (hydroxymethyl and hydroxyl groups or bulky substituents) in

¹ The situation is reversed for L-monosaccharides (1C_4 is more stable) because they are mirror images of the D-monosaccharides.

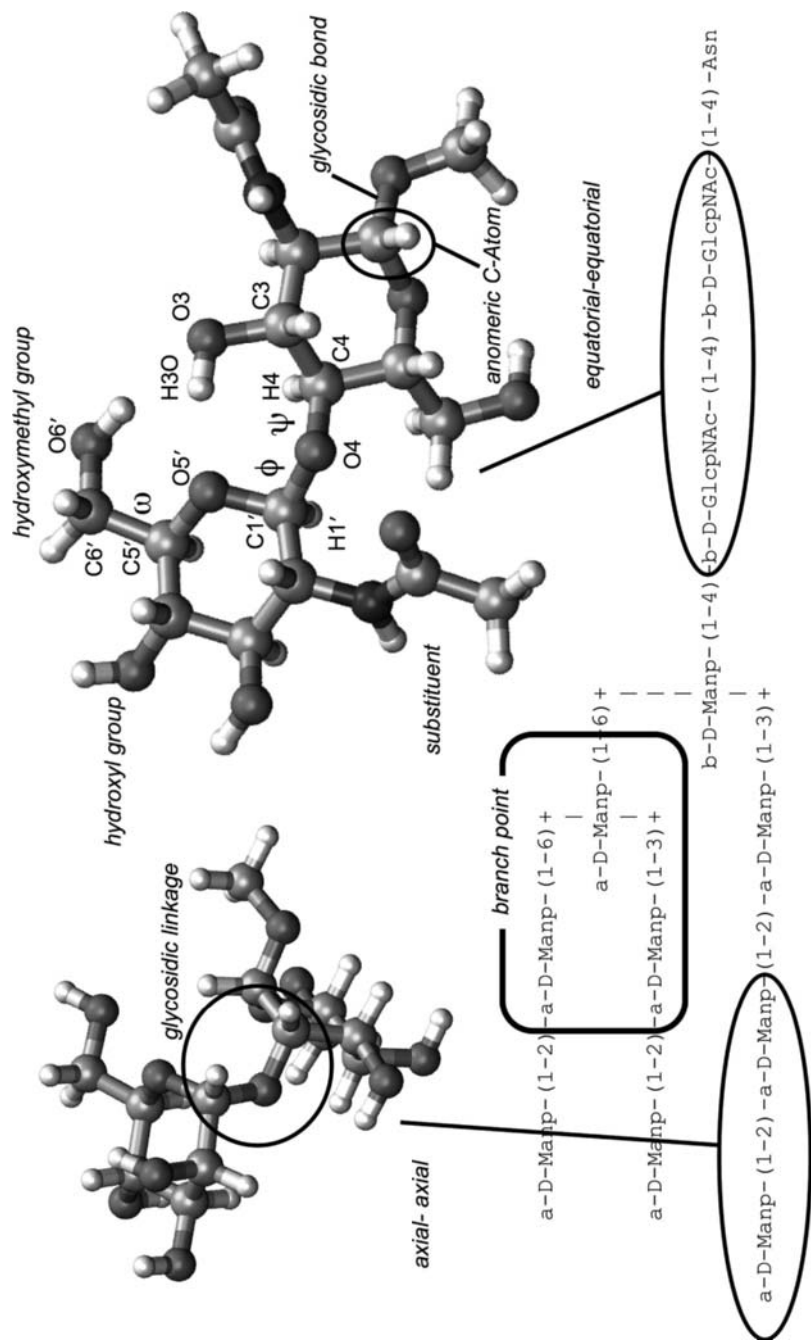


Figure 19.1 A complete high-mannose *N*-glycan structure (GlcNAc₂Man₉) shown as a 2D IUPAC graph representation (textual representation). Structural features and naming definitions that are important for the conformational analysis of carbohydrates are highlighted. To discriminate between the atoms of the different monosaccharides, primes can be appended to the atom names of the second, third, etc., residue (O3', O3'', etc). Residue counting is started from the monosaccharide at the “reducing end” (right).

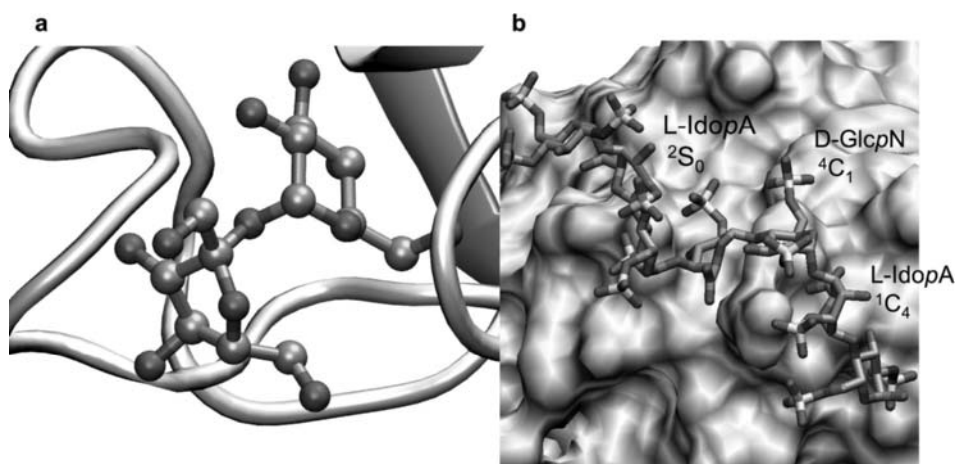


Figure 19.2 X-ray structures highlighting conformations of furanoses and highly substituted pyranoses: (a) sucrose [α -D-Fruf-(2→1)- α -D-Glcp] bound to *Pterocarpus angolensis* lectin (PDB Code 1N3P) [136]; (b) heparin bound to fibroblast growth factor complex FGF1–FGFR2 (PDB Code 1E00) [137].

an equatorial orientation. Depending on the stereochemistry of the monosaccharide base type this can result in a distortion of the ring conformation and in such a case various ring conformations have to be taken into account even in conformational analysis of rings in pyranose form. A prominent example is L-IdopA, a constituent of heparin [30], which can occur in both chair (1C_4 and 4C_1) conformations and in a skew boat (2S_0) conformation [31, 32] (Figure 19.2b).

19.2.2 Orientation of the Glycosidic Linkage

Monosaccharides can be connected by a variety of linkage types (see Figure 19.1). The term “glycosidic bond” is used for the bond between the anomeric C-atom and the glycosidic O- (or N-, S-) atom. The term “glycosidic linkage” is used to define a set of all bonds that connect two monosaccharide rings. The conformational properties of a glycosidic linkage are significantly influenced by the stereochemistry at the anomeric C-atom (α/β). In the 4C_1 chair conformation, α -D-hexoses have the glycosidic bond in an axial orientation with respect to the average ring plane, and for β -D-hexoses the glycosidic bond is equatorial. Depending on the connected monosaccharide and the position through which it is linked, the glycosidic linkages can be axial–axial, axial–equatorial, equatorial–axial or equatorial–equatorial with respect to the ring planes (Figure 19.1). The geometry and torsion angles of the glycosidic linkage are the most important geometric parameters in defining the flexibility and three-dimensional structure of oligosaccharides. Since these properties affect the possible interactions of carbohydrates and therefore their biological function, the stereochemistry of the glycosidic linkages is of crucial importance.

The dihedral angle describing the conformational preferences and rotational flexibility of the glycosidic bond is named ϕ (phi). Two definitions for ϕ are generally used, namely $\phi = \text{O}_{\text{ring}}-\text{C}_{\text{anomeric}}-\text{O}_x-\text{C}_x$ and $\phi_{\text{H}} = \text{H}_{\text{anomeric}}-\text{C}_{\text{anomeric}}-\text{O}_x-\text{C}_x$ (x = attachment position on the adjacent monosaccharide). ϕ_{H} is the preferred definition in the context of NMR

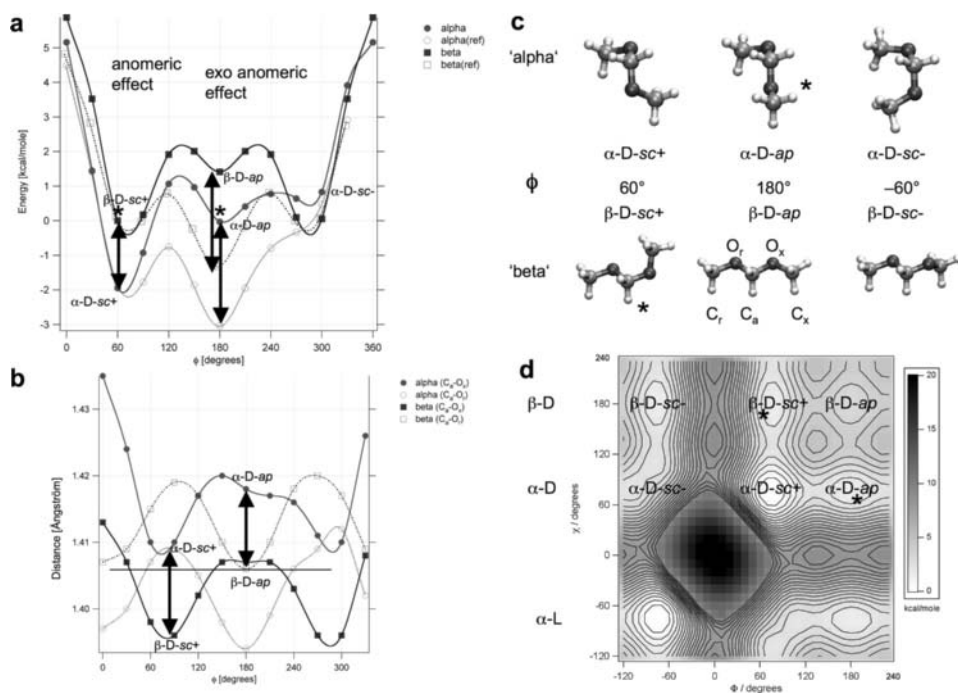


Figure 19.3 Quantum mechanics [DFT (B3LYP)/6–31G*(solvent)] calculations using dimethoxymethane as a model compound to study the (*exo*-) anomeric effect. In the compound atom O_r is analogous to the ring oxygen in a monosaccharide and C_a to the anomeric C atom (see Figure 19.3c for atom labels). (a) Rotation energy profile of ϕ (O_r–C_a–O_x–C_x). In the reference compound 1-methoxypropane O_x is replaced by a CH₂ group. The stabilization of the synclinal (*gauche*) orientation due to the (*exo*-) anomeric effect can be estimated to be about 2 kcal mol⁻¹ (indicated by the arrows). The conformations marked with * are identical due to symmetry. That means for the model compound the transition α -D-sc+ \rightarrow β -D-sc+ is in principle the same as the transition α -D-sc+ \rightarrow α -D-ap. This highlights the close relation of anomeric and *exo*-anomeric effect. The α -D-sc- conformer is higher in energy than the α -D-sc+ conformer because of steric repulsion of the methyl groups. (b) Influence of the (*exo*-) anomeric effect on the length of the C–O bonds. For the β -D-ap conformer the C_a–O_x and C_a–O_r bond length are about equal whereas in the β -D-sc+ conformer the C_a–O_x bond length has decreased and the C_a–O_r bond length has increased due to the *exo*-anomeric effect. The arrows indicate the change in the length of the glycosidic bond due to the anomeric effect. (c) Selected staggered conformations of dimethoxymethane in an arrangement representing an “alpha” or “beta” glycosidic bond in a D-monosaccharide. (d) Conformational map of dimethoxymethane calculated with TINKER/MM3. The locations of the minima displayed in (c) are highlighted. The symmetry relationship between β -D-sc+ and α -D-ap can be seen.

studies, whereas ϕ is mainly used in X-ray crystallography. The symbol ϕ is frequently in use for both definitions. In order to avoid confusion, it is recommended that a torsion angle is always defined by listing (at least once) the four atom names.

The anomeric C-atom is part of an acetal function (–O–C–O–) and the local geometry is greatly influenced by electronic effects from the ring oxygen and the anomeric substituent, most of which are summarized by the term *anomeric effect* (Figure 19.3). In its original definition, the anomeric effect refers to the tendency for an electronegative substituent at the anomeric center of a pyranose ring to adopt, upon ring closure, the axial rather than the equatorial orientation, in contrast to predictions based solely on steric

interactions. In a given glycosidic linkage, ring opening is no longer occurring and the axial versus equatorial orientation of the glycosidic oxygen can only change through ring puckering, that would normally result in a large energy penalty. Consequently, such an inverted ring conformation would not be highly populated even if it were stabilized by a few kcal mol⁻¹ due to a favorable anomeric effect. Therefore, the anomeric effect can be neglected in a routine conformational analysis. However the *exo-anomeric effect*, which has the same stereoelectronic origin as the anomeric effect, has a remarkable influence on the orientational preferences of the torsion ϕ . The *exo-anomeric effect* is the stabilization of the synclinal (sc) or *gauche* conformations ($\phi = \pm 60^\circ$) over the antiperiplanar (*anti*, ap) conformation ($\phi = 180^\circ$) by about 2–3 kcal mol⁻¹. This electronic effect can readily be quantified by analyzing the quantum mechanical torsion energy profiles of dimethoxymethane (CH₃–O–CH₂–O–CH₃) in comparison with those of 1-methoxypropane (CH₃–CH₂–CH₂–O–CH₃) (Figure 19.3), where no “anomeric” C atom is present. There are many publications discussing the origin and size of the (*exo*) anomeric effects [33–37] and the interested reader may use these references as a starting point to delve deeper into this topic. The bottom line with respect to carbohydrate 3D structure is that the (*exo*) anomeric effect stabilizes a \pm synclinal, or *gauche*, orientation of an O–C–O–C torsion angle over an *anti* orientation. Torsions with this atom sequence appear twice in a typical monosaccharide unit: the anomeric torsion C_{ring}–O_{ring}–C_{anomeric}–O_x (anomeric effect) and the glycosidic torsion O_{ring}–C_{anomeric}–O_x–C_x (*exo-anomeric effect*). In terms of the *exo-anomeric effect*, an explanation for this finding could be that in a *gauche* orientation the orbitals representing the electron lone pairs of the oxygen O_x would be aligned with σ^* orbitals of the O_{ring}–C_{anomeric} bond and electron density could be shifted from O_x towards the anomeric C-atom (this effect is sometimes called *negative hyperconjugation*). The bond order of the C_{anomeric}–O_x bond would increase and would result in a shorter C_{anomeric}–O_x bond, a longer O_{ring}–C_{anomeric} bond and the angle O_{ring}–C_{anomeric}–O_x would widen. This behavior can indeed be predicted by quantum mechanical (*ab initio*) calculations (Figure 19.3). Other effects that might have an influence could be the unfavorable alignment of C–O bond dipoles in the *anti* orientation or repulsive interaction of the occupied oxygen lone pairs.

In addition to the glycosidic torsion ϕ , at least one more rotatable bond exists in a glycosidic linkage. This torsion is named ψ (psi) and again two common definitions exist: $\psi = \text{C}_{\text{anomer}}\text{--O}_x\text{--C}_x\text{--C}_{x+1}$ and $\psi_{\text{H}} = \text{C}_{\text{anomer}}\text{--O}_x\text{--C}_x\text{--H}_x$ (x = attachment position on the adjacent monosaccharide). Due to the steric constraints that the attached rings impose on the accessible conformational space of a glycosidic linkage, the linkage torsions do not show a simple threefold rotational energy profile like the hydroxymethyl or OH groups, therefore conformational (ϕ/ψ) maps are used to describe the accessible conformational space of glycosidic linkages (Figure 19.4). It is evident from the maps that there is more than one energy minimum: Typically there is a large low-energy area where ϕ_{H} and ψ_{H} are close to 0° (*syn conformer*) and separated minima with higher energy where either ϕ_{H} or ψ_{H} is 180° (*anti conformer*). That glycosidic linkages can adopt *anti* conformations in solution has been shown experimentally by NMR spectroscopy [38–40]. Due to the *exo-anomeric effect* $\phi_{\text{H}} \approx +40^\circ$ is preferred for β -linkages (Figure 19.4a) and $\phi_{\text{H}} \approx -40^\circ$ for α -linkages (Figure 19.4b, c). A systematic (but not exhaustive) overview of how the accessible conformational space of two linked glucose residues is influenced by linkage type and exocyclic groups is shown in Figure 19.5 (more details about conformational maps can be found in the Section 19.4.1).

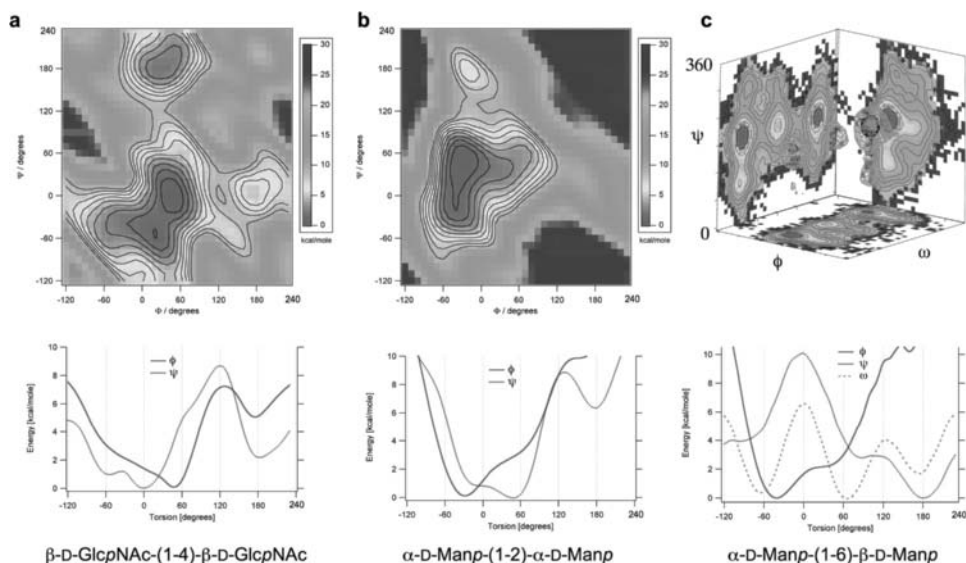


Figure 19.4 Conformational energy maps (top) and rotation profiles (bottom) of the glycosidic torsions of disaccharide fragments that are typically found in *N*-glycans. The adiabatic energy maps were calculated using CAT [71] interfaced to TINKER/MM3 ($\epsilon = 4$). Contours are drawn in steps of 1 kcal mol^{-1} until 10 kcal mol^{-1} above the global minimum. The ϕ_H and ψ_H rotation profiles were derived from the corresponding adiabatic map by applying the reverse Boltzmann equation using a temperature of 300 K, which results in a 2D histogram or weighting matrix, from which 1D histograms for ϕ_H and ψ_H can be calculated simply by adding up the columns and rows, respectively. The histograms can then be back-transferred to an energy profile by applying the Boltzmann equation. (a) Adiabatic map of $\beta\text{-D-GlcNAc-(1-4)-}\beta\text{-D-GlcNAc}$ (*N,N'*-diacetylchitobiose) with a global minimum at $\phi_H/\psi_H \approx 45^\circ/0^\circ$ and secondary minima at $\phi_H/\psi_H \approx 20^\circ/-50^\circ$; $30^\circ/185^\circ$; $180^\circ/5^\circ$. From the ψ_H profile it can be concluded that the transition to the *anti* state ($\psi_H \approx 180^\circ$) will probably occur across the transition state at $\psi_H \approx -120^\circ$ (5 kcal mol^{-1}) because the energy barrier is lower than at $\psi_H \approx +120^\circ$ (9 kcal mol^{-1}). (b) Adiabatic map of $\alpha\text{-D-Manp-(1-2)-}\alpha\text{-D-Manp}$ with a global minimum at $\phi_H/\psi_H \approx -30^\circ/40^\circ$ and secondary minima at $\phi_H/\psi_H \approx -40^\circ/0^\circ$; $-20^\circ/170^\circ$. From the ϕ_H and ψ_H profiles it is evident that the energy barriers to the *anti* states are higher in α -(1-2) linkages than β -(1-4) linkages. The preference for $\phi_H \approx -60^\circ$ over $\phi_H \approx +60^\circ$ in α -linkages (the reverse is found for β -linkages) as expected from the *exo*-anomeric effect is nicely reproduced by the MM3 implementation in TINKER. (c) Free energy map of $\alpha\text{-D-Manp-(1-6)-}\beta\text{-D-Manp}$ derived from an MD simulation with a global minimum at $\phi_H/\psi/\omega \approx -40^\circ/180^\circ/60^\circ$ ($\psi = \text{C1-O6-C6-C5}$). Three secondary minima are present, as can be seen from the iso-surface plot (top; wire frame representation). The projections of the 3D map into 2D iso-contour maps (ϕ_H/ψ , ψ/ω , ϕ_H/ω) are also shown. The ϕ_H profile shows a similar shape to that in $\alpha\text{-D-Manp-(1-2)-}\alpha\text{-D-Manp}$. However, the ψ profile has an overall lower energy. The ω profile shows two low-energy states (*gg* and *gt*) and a third minimum (*tg*) at about $1.7 \text{ kcal mol}^{-1}$ higher energy. In summary, the accessible conformational space for a (1-6)-linkage is significantly greater than that for a (1-2)- or (1-4)-linkage, mainly because of the flexibility of the ω torsion. A full-color version of this figure is included in the Plate section of this book.

In (1-6) linkages, the glycosidic linkage consists of three rotatable bonds, namely ϕ , ψ , and ω (normally defined as O5-C5-C6-O6). This means that the conformational map is four-dimensional (three torsions and the conformational energy). These conformational maps can be displayed as iso-surface plots (Figure 19.4c, top). In (2-8) and (2-9) linked

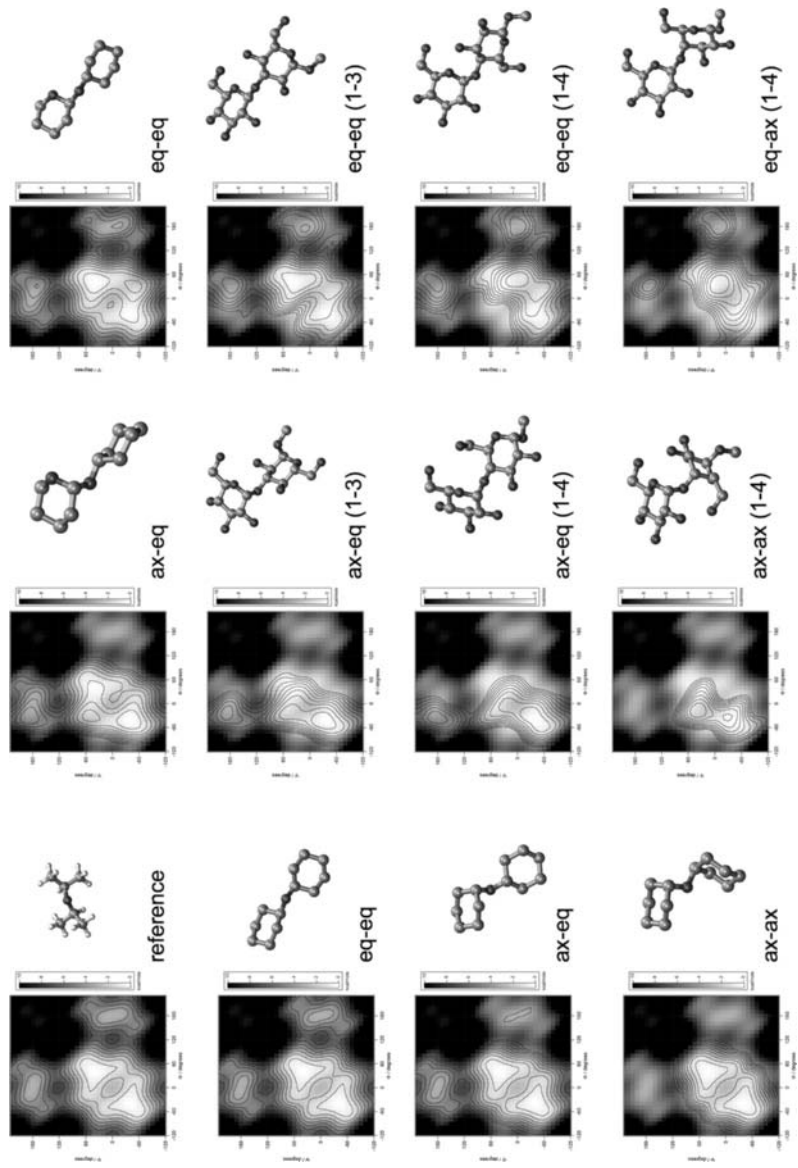


Figure 19.5 Energy maps showing the influence of the linkage type on the accessible conformational space of two glucose residues that are connected through a glycosidic linkage. Maps for a number of model compounds that show the influence of the rings and the *exo-anomeric* effect are also shown. The map of the reference compound (disopropyl ether) representing the maximum accessible space of a linkage is always displayed in the background of the maps as gray shading. Contours are drawn in steps of 1 kcal mol^{-1} until 10 kcal mol^{-1} above the global minimum. Hydrogen atoms are not displayed (except for the reference molecule) for clarity.

sialic acids, the linkages are built by four or five torsions, respectively. This renders the analysis of such glycosidic linkages a very complex task [41, 42] and the graphical display of conformational maps of the complete linkage impossible. However, a convenient method to reduce the dimensionality of conformational maps is the projection of the map into 2D (ϕ/ψ , ω/ψ , or ϕ/ω) (Figure 19.4c, top) or 1D (individual torsions ϕ , ψ , or ω) (Figure 19.4a–c, bottom) using population statistics and the Boltzmann law.

One of the major assumptions in conformational analysis of carbohydrates is that the energy of a conformation adopted by the whole molecule can be described as a sum of linkage energies derived from conformational analysis of each individual glycosidic linkage (*independent linkage approximation*). A detailed discussion about the physical foundations and the limitations of this approach can be found in reference [43]. Following this approach, the oligo- or polysaccharide is “decomposed” into its disaccharide fragments. For each disaccharide fragment, the accessible conformational space is explored by varying the glycosidic linkage torsions (and possibly also the exocyclic groups) in order to find the conformation(s) with lowest energy. The torsion values found for the local energy minima of the disaccharide fragments are then used to build the structure of the complete carbohydrate. This approach is particularly useful for polysaccharides [44], since otherwise chains containing hundreds of residues can be modeled only with great difficulty. The drawback is that all interactions between not directly connected monosaccharide residues are neglected. A method that works fairly satisfactorily for linear polymer chains can therefore be problematic when going to highly branched *N*-glycans.

19.2.3 Exocyclic Groups and Hydrogen Bonding

Hydroxymethyl ($-\text{CH}_2\text{OH}$) and hydroxyl ($-\text{OH}$) groups (Figure 19.1) also have a significant effect on the conformational properties of a carbohydrate, as can be seen from the conformational maps in Figure 19.5. In principle, hydroxyl groups can participate in the formation of interresidue (intramolecular) hydrogen bonds and may therefore have an influence on the conformational equilibrium by stabilizing one local ϕ/ψ minimum over the other. However, intramolecular hydrogen bonds in general are often significantly weakened in explicit water [20] because intermolecular hydrogen bonding to solvent molecules is likely to occur instead. Therefore, the influence of direct intramolecular hydrogen bonds on carbohydrate 3D structure in solution can be questioned. Nevertheless, it has been found that extensive coordination of water molecules in the first solvation shell of carbohydrates exists and water bridges across the glycosidic linkage can occur [22, 45–48] (Figure 19.6). However, strong intramolecular hydrogen bonds may occur even in explicit solvent when the interacting atoms are not accessible for solvent molecules.

Since calculations incorporating explicit solvent molecules are demanding on computer time, alternative approaches such as “gas-phase” calculations using a higher dielectric constant [49] or the use of implicit or continuum solvent models can be employed [50]. It is tempting to set the dielectric constant (ϵ) to 80 (approximately the dielectric constant of water) in a “gas-phase” calculation to simulate the dielectric properties of water. However, this approach may be problematic because this would scale down dramatically all polar (Coulomb) interactions even for short distances, and would therefore practically “switch off” polar interactions such as hydrogen bonding and may result in a different global minimum [51, 52]. Consequently, dielectric constants of about 3–7 (approximately the dielectric constant in a condensed phase) are used frequently to calculate conformational maps using

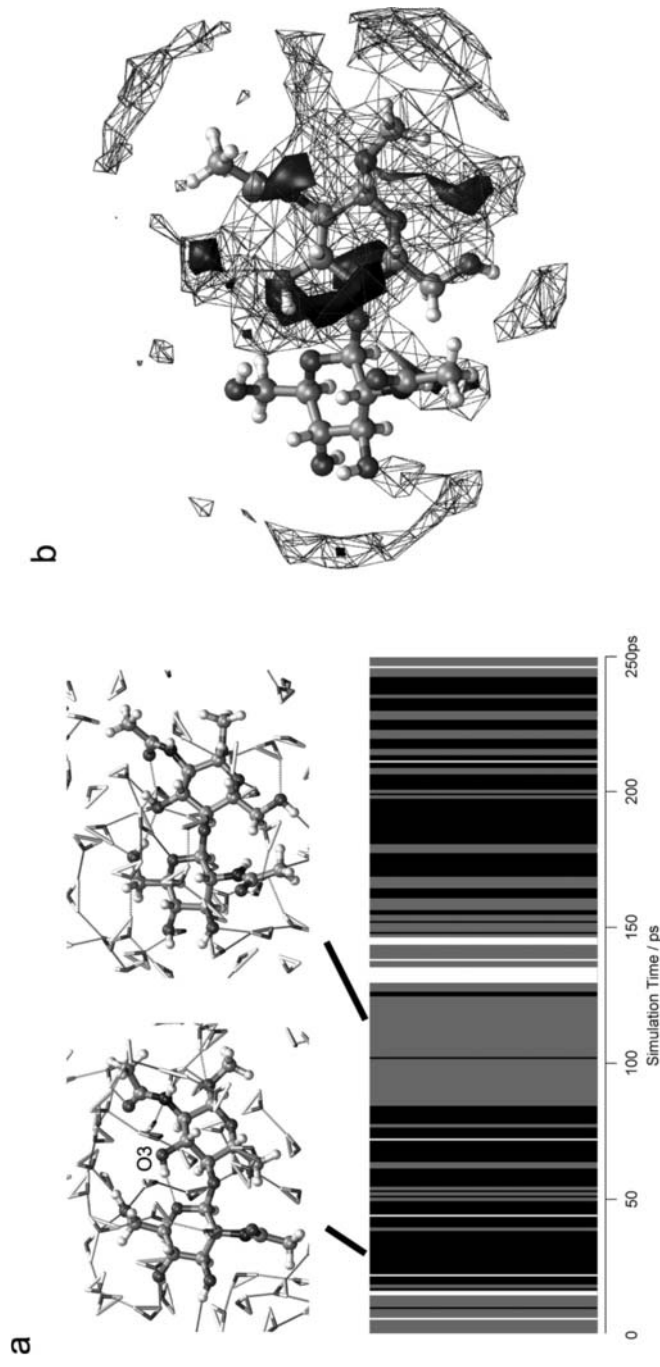


Figure 19.6 Stability of intramolecular hydrogen bonds and ordering of water molecules in the first solvation shell of *N,N'*-diacetylchitobiose as found in an MD simulation in water (AMBER/Glycam [3], 300 K, 5 ns, TIP3 water model) using periodic boundary conditions. (a) Analysis of the participation of the reducing residue O3 atom as a donor in hydrogen bonding. Selected snapshots showing examples of the hydrogen bonding network when the intramolecular hydrogen bond O3–H3O...O5' is present (left) and when the H3O atom is involved in a hydrogen bond to a water (W) molecule (right). The hydrogen bond trajectory (bottom) shows the existence of the hydrogen bond O3–H3O...O5' (black) and O3–H3O...W (gray) as a function of time. It can be seen that the lifetime of a hydrogen bond is in the picosecond time-scale. (b) Probability plot showing the anisotropic solvent structuring [138] in the vicinity of a GlcNAc residue in an *N,N'*-diacetylchitobiose molecule. Volumes with a high probability of finding a water molecule (solid dark surface areas) are caused by interaction with the carbohydrate through either one strong hydrogen bond (NH–W), water bridges (OH–W–HO), or the ordering of water molecules by hydrophobic surfaces (CH₃ groups, hydrophobic sides of the rings). The analysis was performed using the CAT program [71]. VMD [139] was used for molecular graphics.

systematic search methods [53] (Figure 19.4). However, if possible, it is recommended to include explicit solvent molecules especially for the simulation of the conformational properties of larger oligosaccharides since otherwise they tend to “collapse”. In the “gas phase”, the attractive forces (van der Waals and Coulomb) are not counterbalanced by repulsive forces of the solvent molecules, which results in the effect that residues that are normally far apart in space drift towards each other (if the linkages between them are flexible enough) during a minimization or MD simulation, resulting in an unrealistic, very compact structure.

The hydroxymethyl groups are relatively bulky and they can therefore have an influence on the shape of the conformational map, especially in 1–4 linkages (Figure 19.5). In 1–6 linkages, the hydroxymethyl group is part of the glycosidic linkage (ω torsion) and is mainly responsible for the increased flexibility of 1–6 linkages. The torsion ω shows a threefold rotation profile (Figure 19.4c, bottom) with minima at about $-60^\circ = gg$ (*gauche–gauche*), $60^\circ = gt$ (*gauche–trans*), and $180^\circ = tg$ (*trans–gauche*). The term “*trans–gauche*” means that the C6–O6 bond is orientated *trans* to the C5–O5 bond and *gauche* to the C5–C4 bond (see also Figure 20.4a in Chapter 20). The relative energies of the local minima depend on different factors: A ‘*gauche* effect’ favors the *gg* and *gt* orientations. The *gt* orientation is favored over *gg* because of van der Waals repulsion (O6 is close to O5 and C4 in the *gg* orientation). The *tg* orientation is unfavorable in monosaccharides that have an OH group in position 4 in an equatorial orientation (as in D-Glcp) due to a 1–5 repulsion of the oxygen atoms similar to the Hassel–Ottar effect (1–3 diaxial OH groups). If the OH4 group is in an axial position (as in D-Galp), the *gg* orientation is destabilized for the same reason. A hydrogen bond between O6 and O4 can have a stabilizing effect, but in solvent this hydrogen bond is significantly less probable; therefore, a *tg* orientation is not likely to occur for D-Glcp residues [23]. The same is true for *gg* in D-Galp. The experimentally determined population distributions (*gg:gt:tg*) are β -D-Glcp 31:61:8 and β -D-Galp 3:72:25 [54], which correspond to relative energies (in kcal mol⁻¹) of about 0.4:0.0:1.2 and 1.9:0.0:0.6, respectively.

19.3 Exploring the Accessible Conformational Space of Carbohydrates

19.3.1 Conformational Energy Maps

Conformational energy maps of glycosidic linkages display the potential energy surface as a function of the glycosidic torsion angles ϕ and ψ (and ω) (see Figure 19.4 top and Figure 19.5). The energy is normally determined using force field methods, but quantum mechanical (QM) methods have also been used to calculate the energies at each ϕ/ψ grid point [55]. Conformational maps are mainly used to determine preferred ϕ/ψ orientations (location of energy minima) or the flexibility of a glycosidic linkage (size of low-energy ϕ/ψ area). Based on the *independent linkage approximation* [43], the information derived from disaccharide conformational maps can be used to build 3D structures of polysaccharides. A broad range of computational protocols, gradually including an increasing amount of conformational flexibility and reflecting various degrees of accuracy of the computational methods used, have been applied to explore the conformational space accessible to carbohydrates [49, 56–60]. In the classical approach, developed by Ramachandran and co-workers (see Chapter 18), the linkage conformation is studied by systematically varying

the glycosidic torsions ϕ/ψ and calculating the potential energy at each grid point. Until the late 1980s, the available computers were only powerful enough to generate conformational maps routinely following the *rigid-residue approach*: rigid monomeric groups and variable glycosidic linkages. The three-dimensional coordinates for the neutral sugar residues were obtained from appropriate X-ray or neutron diffraction data. Consequently, the quality of the coordinates had a great influence on the results of the calculation. Two programs, namely HSEA (hard-sphere *exo*-anomeric effect) [61] and PFOS (potential function for oligosaccharides) [62] have been widely used to calculate rigid-residue energy maps of carbohydrates. The advantage of the rigid-residue approach is that it is extremely fast. One of the most popular tools that is based on the rigid-residue approach is the web-based carbohydrate builder SWEET-II [63] (www.glycosciences.de).

To give experimental support to the applicability and predictive power of the force fields used, the values of ϕ and ψ derived from crystal structures were plotted with corresponding conformational rigid-residue maps [64, 65]. In spite of the simple model used, the results were encouraging; many of the experimental observations could be explained. However, not allowing the individual residues to optimize (“relax”) at each increment of ϕ and ψ gave unreasonably high energies for some experimentally observed structures. The energies of side minima and energy barriers were also often found to be too high. Finally, in the late 1980s, computers and software were available to perform an energy minimization of the generated structure at each ϕ/ψ grid point [56]. Since only the ϕ/ψ torsions were fixed during the minimization, the rest of the atoms (ring, exocyclic groups) could adopt a more favorable relative orientation, and this resolved many of the problems of the rigid residue maps – the energy surface was found to be much flatter and most of the conformations observed crystallographically were located in areas of low energy. Indeed, it was found that the position of the global energy minimum in a *relaxed map* [66] was similar to the minimum in the rigid-residue map. Consequently, van der Waals interactions seem to have a dominant influence on the conformations of many oligosaccharides. It is relatively easy to rank conformations by increasing or decreasing van der Waals repulsion in order to locate ϕ/ψ combinations of no or minimal overlapping atoms, which explains much of the success of the rigid-residue approach. However, the new relaxed-residue conformational maps were in better agreement with experimental results. Since the early 1990s, mainly the force field (and program) MM2 (later MM3) [67] has been used to calculate relaxed conformational energy maps of disaccharides.

Obviously, the final energy value at each ϕ/ψ combination on the grid of a conformational energy map depends also on the orientation of the exocyclic (hydroxymethyl and hydroxyl) groups. Therefore, it would be desirable also to vary their orientation systematically in a conformational search, in order to find the conformation with the lowest energy at each grid point. Following this approach and plotting the lowest energy value found at each grid point results in an *adiabatic map*. Unfortunately, varying systematically all the exocyclic groups would lead to a combinatorial explosion of the number of conformations that have to be generated. The following calculation will illustrate the problem: a disaccharide such as β -D-GlcpNAc-(1-4)- β -D-GlcpNAc (next to ϕ/ψ) 10 rotatable bonds. Just taking into account the three staggered orientations for each rotatable bond would result in 59 049 (3^{10}) combinations. These all need to be minimized at each ϕ/ψ grid point. About 10^7 energy evaluations (assuming 200 minimization steps for each combination) would be required for each grid point and about 10^{10} for an adiabatic map of resolution 10° . This can be achieved using current computers, but clearly not routinely. One alternative approach to including

the flexibility of the exocyclic torsions is RAMM (RANdom Molecular Mechanics) [68]. Here the pendant groups of the carbohydrate ring are varied using a random-walk technique in order to find the conformer having the lowest energy for each ϕ/ψ grid point.

For most of the “adiabatic” maps that are currently published, only two of the three different staggered orientations of the hydroxymethyl torsion ω (O5–C5–C6–O6) [-60° (*gg*), $+60^\circ$ (*gt*), 180° (*tg*)] were taken into account, and the OH torsions within a ring were all set oriented in a clockwise (*cw*) and subsequently an anticlockwise (*ac*) orientation. Following this approach, 16 relaxed ϕ/ψ maps have to be calculated that are combined to give an “adiabatic” map [51]. A clockwise or anticlockwise orientation of the OH groups is not always as straightforward as it is for D-Glcp, hence the described approaches frequently include manual adjustment of torsions and are therefore very laborious and not very suitable for automation. A modified approach that allows the calculation of “adiabatic” maps fully automatically (including calculation setup and analysis) is as follows. First, all three staggered orientations of the hydroxymethyl groups are included in the search. Since intramolecular hydrogen bonding in a solvent, hydrogen bonds are neglected to a large extent by fixing all OH groups in an *anti* orientation (H–C–O–H = 180°). Following this strategy, fully automatic pseudo-“adiabatic” maps can be generated using the MM3 force field as implemented in the TINKER software [69, 70] with only coordinates as input. Because the TINKER software does not provide a systematic search functionality the Conformational Analysis Tools (CAT) software [71] can be used for all the data handling and analysis. For a typical map with a resolution of 10° there are 11 664 ($36 \times 36 \times 3 \times 3$) structures to be minimized. On a modern computer this takes about 2–3 h, which means that the overall throughput per CPU (central processing unit) is about 10 adiabatic conformational maps per day in batch mode. Since the TINKER software is only used as an energy minimizer in the workflow, automatic calculation of *ab initio* maps would also be possible, in principle, by interfacing the CAT software to an *ab initio* program for the energy minimization, but the calculation time would be several orders of magnitude higher.

An alternative approach to calculating conformational maps is based on population analysis of conformational ensembles derived from molecular dynamics (MD), stochastic dynamics (SD), or Metropolis Monte Carlo (MMC) simulations. The ϕ and ψ values of the linkages found in the conformational ensemble are binned (grouped) into a 2D grid and the population of each bin (grid point) is derived by counting (2D histogram). The bin with the highest population is used as reference population (P_{ref}) and the relative free energy of each bin is derived using the equation $\Delta G = -RT \ln(P_{\text{x}}/P_{\text{ref}})$. This method results in free energy conformational maps; however, it only gives correct values if the conformational equilibrium has been reached and $P_{\text{x}}/P_{\text{ref}}$ has converged to a constant value. In order to achieve conformational equilibrium faster, the conformational space of branched oligosaccharides can be explored using high-temperature (HT) MD simulations (performed at 700–1000 K) [60]. To avoid ring inversion, which occurs frequently at high temperatures, the intra-ring torsions can be constrained to their initial values. Calculating conformational maps using high-temperature molecular dynamics (HTMD) is very efficient since maps of all linkages of the oligosaccharide can be derived simultaneously. Entropic effects and the influence of branching are included in the conformational maps derived from MD simulations.

Databases containing conformational maps are available at www.cermav.cnrs.fr/glyco3d (Glyco3D, relaxed or adiabatic maps) [72] and www.glycosciences.de/glycomapsdb (GlycoMapsDB, free energy maps) [73].

19.3.2 Searching for Local Energy Minima in Conformational Hyperspace

Deriving ϕ/ψ values of local energy minima from conformational maps is one possibility to gain information about possibly stable conformations of carbohydrates. However, CCA (Conformational Clustering Analysis) [74], MCMM (Monte Carlo Multi Minimum) [75, 76], CICADA (Channels In Conformational space Analyzed by Driver Approach) [77, 78], and other approaches [79, 80] can also be applied to search for local minima in conformational space.

In the CCA method [74], the conformational space of the (small) oligosaccharide is explored systematically using a grid search, RAMM, or using MD simulations. When the generated structures are minimized, they assemble into the next closest local minimum. Clustering in torsional space yields a set of conformations that represent directly local energy minima. In the MCMM approach [75, 76], conformational space is explored by a Monte Carlo random walk method. Conducting a random conformational search is fairly simple. A starting structure is chosen, random variations to selected coordinates are applied, the structure is minimized, and the result is compared with minima found during previous conformational search steps. After the resulting structure has been stored as a new, unique conformer, or has been rejected as a duplicate, the cycle is repeated. The CICADA algorithm [77, 78] drives each selected torsion angle in both possible directions with full geometry optimization at each point except for the driven bond. After energy minimization, the new conformers are saved. They are used as starting points for further exploration of the potential energy surface (PES). The calculation stops when no new conformers are found within a given energy window. Traveling over the PES yields paths that end in energy minima. Maxima (“transition states”) can also be identified [81].

19.3.3 Metropolis Monte Carlo (MMC) Simulations

MMC simulations have been applied to access the conformational space of oligosaccharides [57, 82–85]. Starting from any given initial geometry of the molecule, a new conformation is generated by a random displacement along the inner coordinates of the system. A force field is used to calculate the energy content for each new conformation. It can be shown that when repeated statistically often enough, an ensemble of structures is generated where each conformation will occur according to its real population at a given temperature. In order to achieve correct statistical weighting, the Metropolis test is applied [86]. The MMC algorithm, as implemented in the program GEGOP (GEometry of GlycOProteins) [87], comprises several steps that are summarized as follows [57]. (1) A low-energy conformer is chosen as the initial conformation and its energy is calculated. (2) The generation of new conformational states is achieved by adding random step lengths to all variable dihedral angles spanning the conformational space (MMC steps). A maximum step length must be assigned to each direction in conformational space. An MMC step is defined as the move of a single parameter, and a complete sweep through all variable angles is called a macro step or macro move. (3) The energy difference $\Delta E = E_{\text{new}} - E_{\text{old}}$ between the newly generated conformation state and the previous state is usually calculated after each macro step. If ΔE is negative or zero, the move is always accepted, otherwise the Metropolis test is applied. In this test, a random number ξ is generated in the interval $[0, 1]$ and compared with the Boltzmann factor $\exp(-\Delta E/RT)$, where R is the gas constant and T the absolute temperature. If ξ is smaller than the Boltzmann factor, the move is accepted and the altered conformation is counted as the new state. If ξ is greater than the Boltzmann factor, the old conformation

is kept and the old state is counted as a new state, thus ensuring that on average each conformational state is visited with a frequency according to its statistical weight at a given temperature. (4) Steps 2 and 3 are repeated until convergence of the running averages of a property of interest (e.g. NOEs, distances, potential energy) is achieved or the maximum number of steps has been reached. The ensemble of conformational states of the equilibrium phase should approach, in the limit, thermodynamic equilibrium.

The monosaccharide rings are normally kept rigid in MMC simulations of carbohydrates, which makes these calculations very fast. Nevertheless, the rigid-residue MMC approach faces similar problems to those found for the rigid-residue conformational maps: the minima are surrounded by high energy barriers and transitions are not likely to occur. To make transitions more likely during an MMC simulation, the run is performed at high temperatures (1000–2000 K or higher). The linkage torsions and the exocyclic groups are allowed to change their torsion value randomly in an MMC step. A maximum allowed change (“step length”) of 20° per MMC step normally gives a good acceptance ratio of about 30–50% from the Metropolis test. Increasing the step length is likely to result in a significantly lowered acceptance ratio [88]. Large moves have a high probability of being rejected due to large energy differences. The number of MMC steps required to reach convergence depends on various parameters, but about 10^6 steps is generally a good value to start with. It is most likely that only a local conformational equilibrium for ϕ/ψ can be reached in a rigid MMC run. However, MMC is a good technique to sample the conformational space around an energy minimum. This is particularly useful for generating randomized polysaccharide chains [44]. To overcome the problems caused by the rigid residues, a restrained energy minimization of the system can be performed after each macro step. This normally gives much better exploration of conformational space, but it also causes a dramatic increase in the computer time required for the calculation.

19.3.4 *Molecular Dynamics (MD) Simulations*

MD simulations [89–93] directly model the motions of molecules on the atomic scale by numerically solving Newton’s equation of motion for all atoms in a molecular system. In this method, the various atomic forces experienced by these atoms are calculated from the gradient of the conformational energy function (force field). MD simulations provide a complete description of the evolution of a system over time, from which physically observable properties can be computed. Among the advantages of this technique are that it allows efficient exploration of the conformational space around local low-energy structures, and also the study of conformational fluctuations, transitions, and rates. On the other hand, it is very difficult to explore larger regions of the conformational space if the local minima are separated by high energy barriers ($>4\text{--}5\text{ kcal mol}^{-1}$). The system can easily become trapped in a local minimum and computer times of many months may be required until a transition out of this local minimum occurs. Despite these limitations, MD simulations have been used to explore the conformational space of complex oligosaccharides, to help resolve questions in the interpretation of nuclear Overhauser effect (NOE) data, to examine structural flexibility, and to study solvation effects which cannot be probed easily by experiment. MD simulation is currently the most often used theoretical method to explore the conformational space of complex systems on an atomic level. Recently, (short) MD simulations of a complete virus [94] and the ribosome [95] have been performed. The use of MD simulations in the carbohydrate field started in 1986 when Brady performed the

first MD simulation of α -D-glucose in gas phase [96]. Since then, MD simulations have been routinely applied to study the conformational properties of more and more complex carbohydrates [46, 48, 60, 97–114].

In MD simulations (in contrast to rigid-residue MMC simulations), a more realistic approximation exists of the actual energy surface that the carbohydrate experiences during the course of a conformational fluctuation, as the internal coordinates relax dynamically in response to the changes in, for example, the glycosidic torsions. Therefore, MD simulations have clear advantages over MMC simulations. Although effects that are based on population statistics can in principle also be studied by MMC, kinetic properties (correlation times, lifetimes, etc.) can only be derived using MD. In practice, the exploration of the accessible conformational space of a carbohydrate in explicit solvent is more easily performed using MD than MMC, although MMC has been widely used to study liquids [115, 116].

Carbohydrate force-field parameters are available for most of the commonly used MD programs [8, 9, 117]. AMBER [3] can be used conveniently for MD simulations of “standard” carbohydrates in water, simply because a variety of carbohydrate building blocks with high-quality charges are available (www.glycam.com) [12, 13, 118, 119]. CHARMM and GROMOS have also been widely used for MD simulations of carbohydrates, but the number of publicly available building blocks is currently very limited. Generating new building blocks is time consuming and sometimes not very easy, and would require at least *ab initio* calculations to be performed in order to derive the atom charges.

In the early 1990s, MD simulations *in vacuo* (“gas phase”) covered a time-scale of only a few nanoseconds [120, 121]. Simulations in solvent were usually not longer than 1 ns [100, 122–126]. In the intervening period, it has been realized that that these time-scales are probably too short to allow even equilibration of the system [127]. Sampling the conformational space of disaccharide linkages on the microsecond time-scale was challenging, but achievable by using many months of CPU time [128]. Nowadays, microsecond simulations of carbohydrates in the “gas phase” can be performed routinely, using TINKER/MM3, within a few days. A problem that is associated with such long simulations is not the CPU time required, but rather the hard disk space required to store the huge number of snapshots that are sampled during the simulation. In the case of the microsecond simulation of the disaccharide *N,N'*-diacetylchitobiose in the gas phase (Figure 19.7), 10^7 snapshots were recorded and the trajectory in TINKER format occupies about 40 GByte of hard disk space. The size of the data precludes graphical analysis, since this amount of data cannot be loaded into computer memory. Nevertheless, sequential processing of the data is possible. For data analysis, one can use programs that are distributed with the simulation software. However, frequently it is necessary for scientists to write their own analysis routines or scripts. An analysis package that is tailored to analyse large MD trajectories efficiently is Conformational Analysis Tools (CAT) [71]. This software is also used as an analysis engine for the Dynamic Molecules web service (www.md-simulations.de/manager/) [2], where MD simulations of carbohydrates can be performed very conveniently in the “gas phase”.

The typical result of an MD simulation can be a trajectory plot of the change of an atom–atom distance or a torsion angle over time. Figure 19.7a shows the trajectory of the glycosidic linkage torsion ψ of *N,N'*-diacetylchitobiose [β -D-GlcpNAc-(1–4)- β -D-GlcpNAc]. It can immediately be seen that some transitions into the *anti* conformation ($\psi \approx 180^\circ$) are occurring, but they are rare. On average, the lifetime of the *syn* conformer in this simulation is about 80 ns; however, there is a huge variation in the lifetimes seen, so it is not easy to predict when a transition will occur. That such

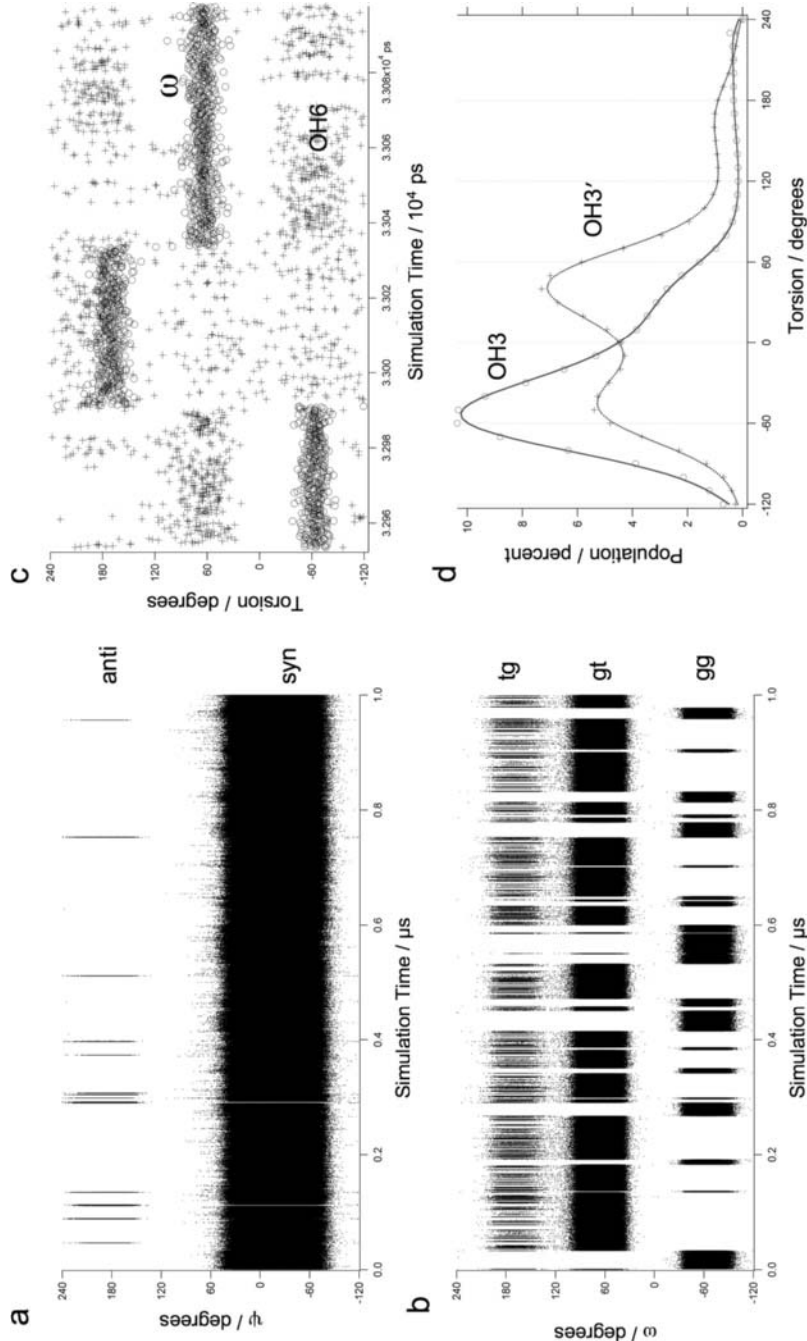


Figure 19.7 Microsecond MD simulation of N,N' -diacetylchitobiose in the gas phase ($\epsilon = 4$) at 300 K using TINKER/MM3. (a) Trajectory of ψ_H showing some transitions between the *syn* and *anti* states. (b) Trajectory of ω showing transitions between all three states (*gg*, *gt*, and *tg*). (c) Trajectory of ω and OH6 showing correlation between the two torsions (due to hydrogen bonding). (d) Population profile (histogram) of the OH3 and OH3' torsion (for details, see text). The analysis of the 40 GByte trajectory (10^7 frames) was performed using CAT [71]. Igor (www.wavemetrics.com) was used for the scientific plots.

transitions into *anti* conformations do occur in reality has been shown by NMR spectroscopy [38–40].

The trajectory plot of the torsion ω (hydroxymethyl group rotation, Figure 19.7b) reveals that there are many conformational transitions during 1 μ s. The *gt* state of this torsion is obviously the most populated and there are more transitions to the *tg* state than to the *gg* state. In spite of this, the lifetime of the *tg* state is very short, whereas the *gg* state can exist for many tens of nanoseconds. This means that the *gg* state is separated from the other two states by high(er) energy barriers. Figure 19.7c displays a section of the ω trajectory together with the trajectory of the corresponding OH6 torsion (C5–C6–O6–H6O). The plot reveals that the two torsions are correlated. The OH6 torsion prefers a different orientation depending on the orientation of ω . This correlation between the two torsions has its origin in the different geometrical requirements to form a hydrogen bond to the ring O (*gg* and *gt*) or the OH4 group (*tg*). The effect of hydrogen bonding can also be explored by calculating population distributions of torsions. Figure 19.7d displays the histogram of the torsions OH3 (H3–C3–O3–H3O) and OH3' (H3'–C3'–O3'–H3O'). The histogram of torsion OH3 shows one large maximum at -60° . This is the orientation that is required to form a hydrogen bond to the ring oxygen (O5') of the next residue (see Figure 19.1). It is obvious that this hydrogen bond dominates very much the conformational preferences of the OH3 group. This hydrogen bond is also present in crystal structures of *N,N'*-diacetylchitobiose. The OH3' group cannot form a hydrogen bond to the adjacent residue; however, intraresidue hydrogen bonds that can be formed with the OH4 or OH2 group favor the two *gauche* states over the *anti* state.

As noted previously, the stability of intramolecular hydrogen bonds depends very much on the dielectric medium (solvent) around the carbohydrate. The conformational behavior of the OH3 group of the β -D-GlcpNAc-(1–4)- β -D-GlcpNAc disaccharide unit in simulations performed in water is slightly different from that seen in the “gas phase” ($\epsilon = 4$). For example, in a 100 ns MD simulation of the heptasaccharide GlcNAc₂Man₅ in explicit solvent at 300 K (Figure 19.8), the OH3 torsion energy profile (OH3 of the first β -D-GlcpNAc residue) shows three minima (Figure 19.8e), indicating that the intramolecular hydrogen bond that exists at -60° is weakened and alternative orientations of the OH3 group are also (meta)stable (Figures 19.6 and 19.8e). That the hydrogen bond O3–H3O...O5' still exists (about 60–70%) even in water can be seen from the hydrogen bond analysis (Figure 19.8f; codes 1_4YB2_O3_20/1_4YB3_O5_35 and 1_4YB3_O3_47/1_VMB4_O5_62).

In general, torsion energy profiles (Figure 19.8e) can be derived from histograms (Figure 19.7d) by applying the Boltzmann law. Similarly, ϕ/ψ 2D histograms can be converted to free energy conformational maps [60] (Figure 19.8b). Since this can be performed for all glycosidic linkages of a complex oligosaccharide simultaneously, it is a very efficient method to calculate all conformational maps of a given carbohydrate (see Section 19.4.1). However, these profiles and maps are only (theoretically) correct if the population statistics have been derived from an ensemble that represents a conformational equilibrium. This is not so easy to achieve; especially the equilibration of the torsions ϕ , ψ , and ω can still be a challenge, particularly when the simulations are performed in explicit solvent at room at 300 K (Figure 19.8a).

In summary, it is recommended to perform MD simulations of complex carbohydrates in explicit solvent and the simulation time should be at least 50 ns at 300 K. Although one has to take into account that such simulations are much more computer time demanding than “gas-phase” simulations and more than 1000 CPU hours are normally required to

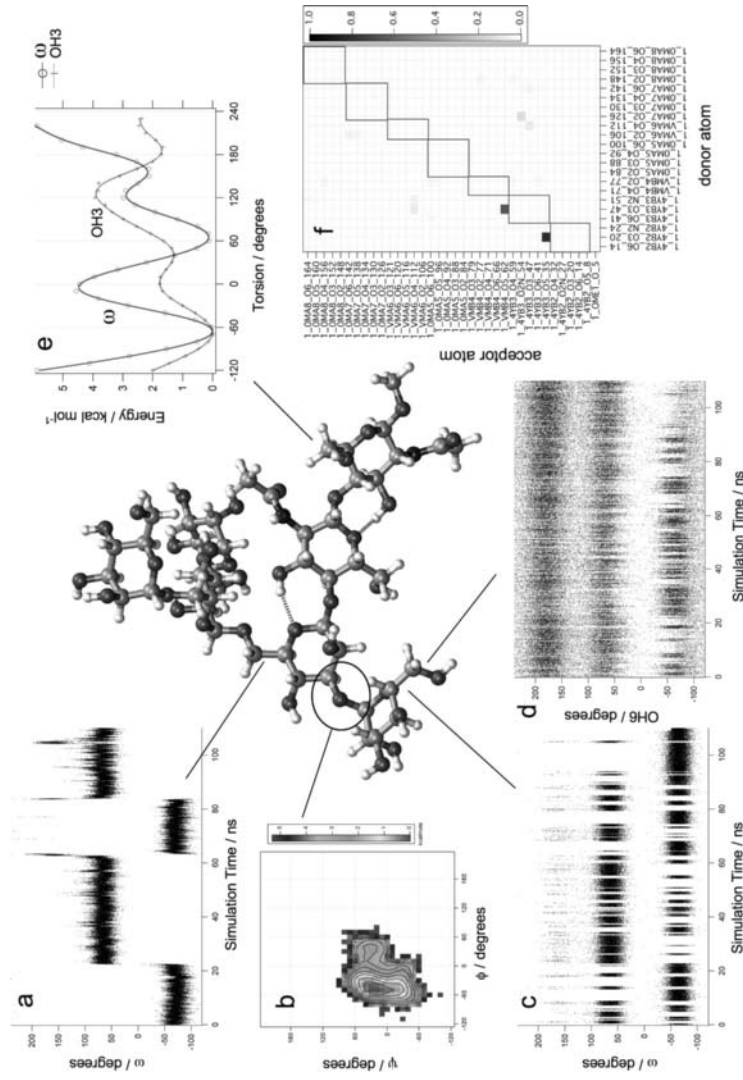


Figure 8

Figure 19.8 100 ns MD simulation of a high-mannose type *N*-glycan (GlcNAc₂Man₅) in explicit solvent at 300 K using AMBER/Glycam [3]. A snapshot is displayed in the center and a variety of analysis results that can be derived are shown. (a) Trajectory of the ω torsion of the 1–6 linkage from the core mannose. (b) Free energy map for the α -D-Mannp-(1–3)- β -D-Mannp linkage. (c) Trajectory of the ω torsion of the 344-terminal mannose (“344” indicates the linkage path towards the reducing end). (d) Trajectory of the OH6 torsion of the 344-terminal mannose. (e) Torsion energy profile of the ω and OH3 torsions of the first GlcNAc residue. (f) Hydrogen bond matrix displaying the probability for a hydrogen bond between a donor atom and an acceptor atom for the entire heptasaccharide. The labels are autogenerated by CAT based on molecule number, residue name + number, atom name, atom index.

perform a 50–100 ns MD simulation (Figure 19.8) in a solvent box using periodic boundary conditions. The huge investment of computer time is justified because the solvent simulations allow the direct exploration of the structure and dynamics of carbohydrates under conditions (almost) identical with those present during an NMR experiment. Therefore, MD simulations are frequently used to rationalize results derived by NMR, that are difficult to interpret otherwise.

19.4 Generation of 3D Structures of Glycoproteins

It is estimated that about 50% of all proteins are glycosylated [129]. However, only about 5% of the X-ray structures deposited in the Protein Data Bank (PDB) [130] have carbohydrate chains attached, and these chains seldom contain more than two or three monosaccharide residues. It is therefore of interest to have computational tools which permit *in silico* glycosylation at specific locations of a given 3D protein structure, where it is known experimentally that a specific type of glycosylation occurs. Two such (web-based) tools are available that facilitate the *in silico* glycosylation of proteins: GlycamWeb (<http://www.glycam.com>) and GlyProt (<http://www.glycosciences.de/glyprot/>) (Figure 19.9) [131].

The prerequisite for both GlycamWeb and GlyProt is that a 3D structure of the protein is available. The coordinates can be uploaded from a local computer in PDB format or can be directly retrieved from the PDB by entering a PDB code into the web interface. Potential *N*-glycosylation sites are automatically detected by searching the amino acid sequence for the sequon Asn–X–Ser/Thr, where X is not proline. In the next step, the surface accessibility of the sequon needs to be checked in order to judge whether a glycan can be attached to the Asn side-chain. One possibility is to calculate the solvent-accessible surface area of the Asn, and if the area is above a certain limit then it is assumed that glycosylation is possible (GlycamWeb). An alternative strategy is to attach an *N*-glycan core to the Asn side-chain and check all reasonable orientations of the Asn–GlcNAc linkage torsions ϕ , ψ , χ_1 , and χ_2 in order to find a structure with no or only minor atom overlapping between the glycan and the protein (GlyProt). The set of torsion values that are checked during the search were derived from a statistical analysis of glycoproteins contained in the PDB using the GlyTorsion tool [132]. When all the spatially accessible *N*-glycosylation sites have been determined, the user can select pre-built glycans from a library that are then attached to the glycosylation sites. Some physicochemical properties such as mass, radius of gyration, and surface area of the glycoprotein are calculated (GlyProt).

Tools such as GlyProt and GlycamWeb are very useful in order to gain an impression of the spatial arrangement of the *N*-glycans and their size in relation to the protein. However, they do not take into account flexibility of the glycan and the glycosylation site. The tools construct one reasonable conformation of the glycoprotein out of a large number of possible structures. In order to obtain a more realistic picture of the accessible conformational space of the *N*-glycans attached and their interactions with the protein surface, the initial models can be output in a format that is suitable to serve as input for a molecular dynamics simulation in explicit solvent (GlycamWeb). Such glycoprotein systems usually contain many tens of thousands of atoms – most of which are solvent (Figure 19.10a) – and MD simulations on the nanosecond time-scale require many weeks of computer time and are therefore computationally very demanding [133, 134]. From the simulation data, one can obtain an estimate of the flexibility of the glycan when attached to a protein (Figure 19.10b).

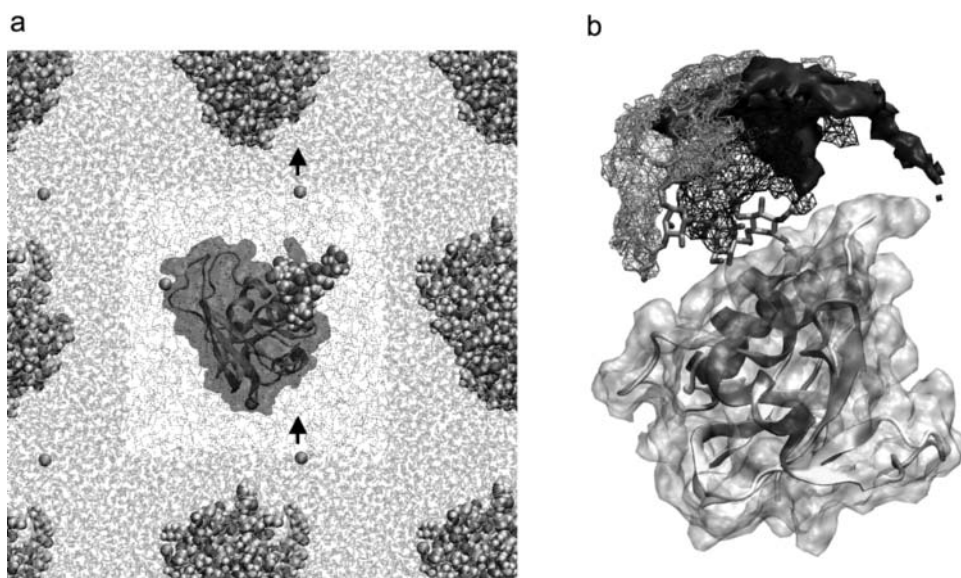


Figure 19.10 5 ns MD simulation of ribonuclease B [134], with the glycan $\text{GlcNAc}_2\text{Man}_5$ at Asn34, in explicit solvent at 300 K using AMBER/Glycam [3]. (a) Simulations in explicit solvent are generally performed using periodic boundary conditions (PBC) and particle-mesh Ewald (PME) for proper treatment of the long-range electrostatics. PBC means that the water box and the solute are copied in all three dimensions so that an infinite periodic system is generated. If a particle leaves the central box on one side, the copy of this particle enters the box from the other side (indicated by the arrow). In the central box, the protein is represented by its (transparent) solvent-accessible surface and the protein backbone is displayed in a cartoon representation. The glycan atoms and the chloride ions are displayed as van der Waals spheres (CPK). The dotted lines represent the hydrogen bond network of the water. The copies of the central box show the water molecules as lines and the glycoprotein in CPK. (b) Iso-surface plots of the 3D population analysis showing the space occupation (flexibility) of the three terminal Man residues of the *N*-glycan (3-branch, solid; 63- and 66-branch, wire frame).

Possibly even more important, an analysis of the interaction of carbohydrate residues with the protein surface can be performed. This gives information on which parts of the protein are protected by the carbohydrate, for example against proteolytic degradation, and which parts of the carbohydrate are exposed to the solvent and can therefore be recognized by a lectin or an enzyme.

19.5 Conclusion and Outlook

During the last decade, it has become obvious that approaches based on molecular dynamics simulations have become the method of choice to study the conformational properties of carbohydrates under physiological conditions. Although a 50 ns simulation of large molecules in explicit solvent may still require many weeks of calculation time, even when using modern parallelized computers, the several gigabytes of data calculated contain an enormous amount of information on the dynamics and intermolecular interaction properties of the system studied which can be extracted and compared with experimental data. The use of force fields that take into account the polarizability of molecules and that include

proper parameterization for carbohydrates will further improve the quality of the results. Coarse-grain force fields [135] will also open new routes in the simulation of carbohydrate properties. The simulation of glycoproteins under “physiological conditions” on the nanosecond time-scale has become feasible recently [134]. This may open up new possibilities to rationalize the connection between the 3D structure and biological function of glycoproteins.

Abbreviations

CAT	Conformational Analysis Tools
CCA	Conformational Clustering Analysis
CICADA	Channels In Conformational space Analyzed by Driver Approach
CPU	Central Processing Unit
DFT	Density Functional Theory
GEGOP	GEometry of GlycOProteins
HSEA	Hard-Sphere <i>Exo</i> -Anomeric
HTMD	High-Temperature Molecular Dynamics
MD	Molecular Dynamics
MM	Molecular Mechanics
MMC	Metropolis Monte Carlo
MCMM	Monte Carlo Multi Minimum
PFOS	Potential Function for Oligosaccharides
QM	Quantum Mechanics
RAMM	RAndom Molecular Mechanics
SD	Stochastic Dynamics

References

1. Bohne A, Lang E, von der Lieth C-W: SWEET – WWW-based rapid 3D construction of oligo- and polysaccharides. *Bioinformatics* 1999, **15**:767–768.
2. Frank M, Gutbrod P, Hassayoun C, von der Lieth C-W: Dynamic Molecules: molecular dynamics for everyone. An Internet-based access to molecular dynamic simulations: basic concepts. *J Mol Model* 2003, **9**:308–315.
3. Case DA, Cheatham TE III, Darden T, *et al.*: The Amber biomolecular simulation programs. *J Comput Chem* 2005, **26**:1668–1688.
4. Van der Spoel D, Lindahl E, Hess B, *et al.*: Gromacs: fast, flexible, and free. *J Comput Chem* 2005, **26**:1701–1718.
5. Phillips JC, Braun R, Wang W, *et al.*: Scalable molecular dynamics with NAMD. *J Comput Chem* 2005, **26**:1781–1802.
6. MacKerell AD, Bashford D, Bellott M, *et al.*: All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 1998, **102**:3586–3616.
7. Eklund R, Widmalm G: Molecular dynamics simulations of an oligosaccharide using a force field modified for carbohydrates. *Carbohydr Res* 2003, **338**:393–398.
8. Mackerell AD Jr: Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem* 2004, **25**:1584–1604.
9. Tschampel SM, Kirschner KN, Woods RJ: Incorporation of carbohydrates into macromolecular force fields: development and validation. In *NMR Spectroscopy and Computer Modeling of*

- Carbohydrates* (eds Vliegthart JFG, Woods RJ), ACS Symposium Series, Vol. 930. Washington, DC: American Chemical Society; 2006, pp. 235–257.
- Lii JH, Ma BY, Allinger NL: Importance of selecting proper basis set in quantum mechanical studies of potential energy surfaces of carbohydrates. *J Comput Chem* 1999, **20**:1593–1603.
 - Tvaroska I: Structural insights into the catalytic mechanism and transition state of glycosyltransferases using *ab initio* molecular modeling. *TIGG* 2005, **17**:177–190.
 - Woods RJ, Chappelle R: Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates. *THEOCHEM* 2000, **527**:149–156.
 - Kirschner KN, Yongye AB, Tschampel SM, *et al.*: GLYCAM06: a generalizable biomolecular force field. *Carbohydrates. J Comput Chem* 2008, **29**:622–655.
 - Hemmingsen L, Madsen DE, Esbensen AL, *et al.*: Evaluation of carbohydrate molecular mechanical force fields by quantum mechanical calculations. *Carbohydr Res* 2004, **339**:937–948.
 - Kuttel M, Brady JW, Naidoo KJ: Carbohydrate solution simulations: producing a force field with experimentally consistent primary alcohol rotational frequencies and populations. *J Comput Chem* 2002, **23**:1236–1243.
 - Kony D, Damm W, Stoll S, Van Gunsteren WF: An improved OPLS-AA force field for carbohydrates. *J Comput Chem* 2002, **23**:1416–1429.
 - Senderowitz H, Parish C, Still WC: Carbohydrates: united atom AMBER* parameterization of pyranoses and simulations yielding anomeric free energies. *J Am Chem Soc* 1996, **118**:2078–2086.
 - Stenutz R, Carmichael I, Widmalm G, Serianni AS: Hydroxymethyl group conformation in saccharides: structural dependencies of $^2J(\text{HH})$, $^3J(\text{HH})$, and $^1J(\text{CH})$ spin–spin coupling constants. *J Org Chem* 2002, **67**:949–958.
 - Gonzalez-Outeirino J, Kirschner KN, Thobhani S, Woods RJ: Reconciling solvent effects on rotamer populations in carbohydrates – a joint MD and NMR analysis. *Can J Chem* 2006, **84**:569–579.
 - Rohfritsch PF, Frank M, Sandstrom C, *et al.*: Comparative ^1H NMR and molecular modeling study of hydroxy protons of beta-D-Galp-(1–4)-beta-D-GlcpNAc-(1–2)-alpha-D-Manp-(1-O)(CH(2))(7)C H(3) analogues in aqueous solution. *Carbohydr Res* 2007, **342**:597–609.
 - Zhang QM, Jaroniec J, Lee G, Marszalek PE: Direct detection of inter-residue hydrogen bonds in polysaccharides by single-molecule force spectroscopy. *Angew Chem Int Ed* 2005, **44**:2723–2727.
 - Almond A, Sheehan JK: Predicting the molecular shape of polysaccharides from dynamic interactions with water. *Glycobiology* 2003, **13**:255–264.
 - Kirschner KN, Woods RJ: Solvent interactions determine carbohydrate conformation. *Proc Natl Acad Sci USA* 2001, **98**:10541–10545.
 - Barrows SE, Dulles FJ, Cramer CJ, *et al.*: Relative stability of alternative chair forms and hydroxymethyl conformations of beta-D-glucopyranose. *Carbohydr Res* 1995, **276**:219–251.
 - Levitt M, Warshel A: Extreme conformational flexibility of the furanose ring in DNA and RNA. *J Am Chem Soc* 1978, **100**:2607–2613.
 - French AD, Tran V: Analysis of fructofuranose conformations by molecular mechanics. *Biopolymers* 1990, **29**:1599–1611.
 - Cremer D, Pople JA: General definition of ring puckering coordinates. *J Am Chem Soc* 1975, **97**:1354–1358.
 - Altona C, Sundaralingam M: Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J Am Chem Soc* 1972, **94**:8205–8212.
 - Hassel O, Ottar B: The structure of molecules containing cyclohexane or pyranose rings. *Acta Chem Scand* 1947, **1**:929–942.
 - Mulloy B, Forster MJ: Conformation and dynamics of heparin and heparan sulfate. *Glycobiology* 2000, **10**:1147–1156.

31. Ferro DR, Provasoli A, Ragazzi M, *et al.*: Conformer populations of L-iduronic acid residues in glycosaminoglycan sequences. *Carbohydr Res* 1990, **195**:157–167.
32. Remko M, von der Lieth C-W: Conformational structure of some trimeric and pentameric structural units of heparin. *J Phys Chem A* 2007, **111**:13484–13491.
33. Juaristi E, Cuevas G: Recent studies of the anomeric effect. *Tetrahedron* 1992, **48**:5019–5087.
34. Cramer CJ, Truhlar DG, French AD: *Exo*-anomeric effects on energies and geometries of different conformations of glucose and related systems in the gas phase and aqueous solution. *Carbohydr Res* 1997, **298**:1–14.
35. Lii JH, Chen KH, Durkin KA, Allinger NL: Alcohols, ethers, carbohydrates, and related compounds. II. The anomeric effect. *J Comput Chem* 2003, **24**:1473–1489.
36. Tvaroska I, Carver JP: The anomeric and *exo*-anomeric effects of a hydroxyl group and the stereochemistry of the hemiacetal linkage. *Carbohydr Res* 1998, **309**:1–9.
37. Asensio JL, Canada FJ, Cheng X, *et al.*: Conformational differences between *O*- and *C*-glycosides: the alpha-O-man-(1-1)-beta-Gal/alpha-C-Man-(1-1)-beta-Gal case – a decisive demonstration of the importance of the *exo*-anomeric effect on the conformation of glycosides. *Chem Eur J* 2000, **6**:1035–1041.
38. Dabrowski J, Kozar T, Grosskurth H, Nifantev NE: Conformational mobility of oligosaccharides – experimental evidence for the existence of an *anti* conformer of the Gal1-Beta-1-3Glc-Beta-1-OME disaccharide. *J Am Chem Soc* 1995, **117**:5534–5539.
39. Landersjo C, Stenutz R, Widmalm G: Conformational flexibility of carbohydrates: a folded conformer at the phi dihedral angle of a glycosidic linkage. *J Am Chem Soc* 1997, **119**:8695–8698.
40. Hoog C, Landersjo C, Widmalm G: Oligosaccharides display both rigidity and high flexibility in water as determined by C-13 NMR relaxation and H-1,H-1 NOE spectroscopy: evidence of *anti*-phi and *anti*-psi torsions in the same glycosidic linkage. *Chem Eur J* 2001, **7**:3069–3077.
41. Muhlenhoff M, Eckhardt M, Gerardy-Schahn R: Polysialic acid: three-dimensional structure, biosynthesis and function. *Curr Opin Struct Biol* 1998, **8**:558–564.
42. Brisson JR, Baumann H, Imberty A, *et al.*: Helical epitope of the group-B meningococcal alpha(2-8)-linked sialic acid polysaccharide. *Biochemistry* 1992, **31**:4996–5004.
43. Brant DA, Christ MD: Realistic conformational modeling of carbohydrates. In *Computer Modeling of Carbohydrate Molecules* (eds French AD, Brady JW), ACS Symposium Series, Vol **430**. Washington, DC: American Chemical Society; 1990, pp. 42–68.
44. Rodriguez-Carvajal MA, Imberty A, Perez S: Conformational behavior of chondroitin and chondroitin sulfate in relation to their physical properties as inferred by molecular modeling. *Biopolymers* 2003, **69**:15–28.
45. Naidoo KJ, Brady JW: Calculation of the Ramachandran potential of mean force for a disaccharide in aqueous solution. *J Am Chem Soc* 1999, **121**:2244–2252.
46. Naidoo KJ, Kuttel M: Water structure about the dimer and hexamer repeat units of amylose from molecular dynamics computer simulations. *J Comput Chem* 2001, **22**:445–456.
47. Naidoo KJ, Chen JYJ: The role of water in the design of glycosidic linkage flexibility. *Mol Phys* 2003, **101**:2687–2694.
48. Almond A: Towards understanding the interaction between oligosaccharides and water molecules. *Carbohydr Res* 2005, **340**:907–920.
49. Stortz CA: Additive effects in the modeling of oligosaccharides with MM3 at high dielectric constants: an approach to the ‘multiple minimum problem’. *Carbohydr Res* 2006, **341**:663–671.
50. Chen J, Brooks CL III, Khandogin J: Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr Opin Struct Biol* 2008, **18**:140–148.
51. Stortz CA: Disaccharide conformational maps: how adiabatic is an adiabatic map? *Carbohydr Res* 1999, **322**:77–86.
52. Stortz CA, Cerezo AS: Potential energy surfaces of alpha-(1-3)-linked disaccharides calculated with the MM3 force-field. *J Carbohydr Chem* 2002, **21**:355–371.

53. French AD, Kelterer AM, Johnson GP, *et al.*: Constructing and evaluating energy surfaces of crystalline disaccharides. *J Mol Graph Model* 2000, **18**:95–107.
54. Thibaudeau C, Stenutz R, Hertz B, *et al.*: Correlated C–C and C–O bond conformations in saccharide hydroxymethyl groups: parametrization and application of redundant ^1H – ^1H , ^{13}C – ^1H , and ^{13}C – ^{13}C NMR *J*-couplings. *J Am Chem Soc* 2004, **126**:15668–15685.
55. French AD, Johnson GP: Advanced conformational energy surfaces for cellobiose. *Cellulose* 2004, **11**:449–462.
56. French AD: Rigid- and relaxed-residue conformational analysis of cellobiose using the computer program MM2. *Biopolymers* 1988, **27**:1519–1525.
57. Peters T, Meyer B, Stuikeprill R, *et al.*: A Monte-Carlo method for conformational-analysis of saccharides. *Carbohydr Res* 1993, **238**:49–73.
58. von der Lieth C-W, Kozar T, Hull WE: A (critical) survey of modelling protocols used to explore the conformational space of oligosaccharides. *THEOCHEM* 1997, **395**:225–244.
59. French AD, Kelterer AM, Johnson GP, *et al.*: HF/6–31G*energy surfaces for disaccharide analogs. *J Comput Chem* 2001, **22**:65–78.
60. Frank M, Bohne-Lang A, Wetter T, von der Lieth C-W: Rapid generation of a representative ensemble of *N*-glycan conformations. *In Silico Biol* 2002, **2**:427–439.
61. Lemieux RU, Bock K: The conformational analysis of oligosaccharides by ^1H -NMR and HSEA calculation. *Arch Biochem Biophys* 1983, **221**:125–134.
62. Tvaroska I, Perez S: Conformational-energy calculations for oligosaccharides: a comparison of methods and a strategy of calculation. *Carbohydr Res* 1986, **149**:389–410.
63. Bohne A, Lang E, von der Lieth C-W: W3-Sweet: carbohydrate modeling by Internet. *J Mol Model* 1998, **4**:33–43.
64. French AD: Comparisons of rigid and relaxed conformational maps for cellobiose and maltose. *Carbohydr Res* 1989, **188**:206–211.
65. French AD, Dowd MK: Exploration of disaccharide conformations by molecular mechanics. *THEOCHEM* 1993, **286**:183–201.
66. Gelin BR, Karplus M: Role of structural flexibility in conformational calculations. Application to acetylcholine and beta-methylacetylcholine. *J Am Chem Soc* 1975, **97**:6996–7006.
67. Allinger NL, Rahman M, Lii JH: A molecular mechanics force-field (MM3) for alcohols and ethers. *J Am Chem Soc* 1990, **112**:8293–8307.
68. Kozar T, Petrak F, Galova Z, Tvaroska I: RAMM – a new procedure for theoretical conformational-analysis of carbohydrates. *Carbohydr Res* 1990, **204**:27–36.
69. Ponder J: Tinker 4; <http://dasher.wustl.edu/tinker/>.
70. Stortz CA: Comparative performance of MM3(92) and two TINKER MM3 versions for the modeling of carbohydrates. *J Comput Chem* 2005, **26**:471–483.
71. Frank M: Conformational Analysis Tools (CAT); <http://www.md-simulations.de/CAT/>.
72. Imberty A, Mikros E, Koca J, *et al.*: Computer-simulation of histo-blood group oligosaccharides – energy maps of all constituting disaccharides and potential-energy surfaces of 14 Abh and Lewis carbohydrate antigens. *Glycoconj J* 1995, **12**:331–349.
73. Frank M, Lutteke T, von der Lieth C-W: GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res* 2007, **35**:287–290.
74. Kozar T, von der Lieth C-W: Efficient modelling protocols for oligosaccharides: from vacuum to solvent. *Glycoconj J* 1997, **14**:925–933.
75. Li Z, Scheraga HA: Monte Carlo minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci USA* 1987, **84**:6611–6615.
76. Chang G, Guida WC, Still WC: An internal-coordinate Monte Carlo method for searching conformational space. *J Am Chem Soc* 1989, **111**:4379–4386.
77. Koca J: Computer-program Cicada – traveling along conformational potential-energy hypersurface. *THEOCHEM* 1994, **308**:13–24.

78. Koca J, Perez S, Imberty A: Conformational analysis and flexibility of carbohydrates using the Cicada approach with MM3. *J Comput Chem* 1995, **16**:296–310.
79. Nahmany A, Strino F, Rosen J, *et al.*: The use of a genetic algorithm search for molecular mechanics (MM3)-based conformational analysis of oligosaccharides. *Carbohydr Res* 2005, **340**:1059–1064.
80. Parish C, Lombardi R, Sinclair K, *et al.*: A comparison of the Low Mode and Monte Carlo conformational search methods. *J Mol Graph Model* 2002, **21**:129–150.
81. Casset F, Imberty A, duPenhoat CH, *et al.*: Validation of two conformational searching methods applied to sucrose: simulation of NMR and chiro-optical data. *THEOCHEM* 1997, **395**:211–224.
82. York WS, Thomsen JU, Meyer B: The conformations of cyclic (1→2)-beta-D-glucans: application of multidimensional clustering analysis to conformational data sets obtained by Metropolis Monte Carlo calculations. *Carbohydr Res* 1993, **248**:55–80.
83. Adeyeye J, Azurmendi HF, Stroop CJM, *et al.*: Conformation of the hexasaccharide repeating subunit from the *Vibrio cholerae* O139 capsular polysaccharide. *Biochemistry* 2003, **42**:3979–3988.
84. Furlan S, La Penna G, Perico A, Cesaro A: Conformational dynamics of hyaluronan oligomers in solution. 3. Molecular dynamics from Monte Carlo replica-exchange simulations and mode-coupling diffusion theory. *Macromolecules* 2004, **37**:6197–6209.
85. Rundlof T, Eriksson L, Widmalm G: A conformational study of the trisaccharide beta-D-Glcp-(1→2)[beta-D-Glcp-(1→3)]alpha-D-Glcp-OMe by NMR NOESY and TROESY experiments, computer simulations, and X-ray crystal structure analysis. *Chem Eur J* 2001, **7**:1750–1758.
86. Metropolis N, Rosenbluth AW, Rosenbluth MN, *et al.*: Equation of state calculations by fast computing machines. *J Chem Phys* 1953, **21**:1087–1092.
87. Stuikeprill R, Meyer B: A new force-field program for the calculation of glycopeptides and its application to a heptacosapeptide-decasaccharide of immunoglobulin-G1 – importance of 1–6-glycosidic linkages in carbohydrate–peptide interactions. *Eur J Biochem* 1990, **194**:903–919.
88. Bouzida D, Kumar S, Swendsen RH: Efficient Monte-Carlo methods for the computer simulation of biological molecules. *Phys Rev A* 1992, **45**:8894–8901.
89. Karplus M, McCammon JA: Protein structural fluctuations during a period of 100 ps. *Nature* 1979, **277**:578.
90. Berendsen HJ: Molecular dynamics studies of proteins and nucleic acids. *Curr Opin Struct Biol* 1991, **1**:191–195.
91. Vangunsteren WF: Molecular-dynamics studies of proteins. *Curr Opin Struct Biol* 1993, **3**:277–281.
92. Brady JW: Theoretical studies of oligosaccharide structure and conformational dynamics. *Curr Opin Struct Biol* 1991, **1**:711–715.
93. Karplus M, McCammon JA: Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 2002, **9**:646–652.
94. Freddolino PL, Arkhipov AS, Larson SB, *et al.*: Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* 2006, **14**:437–449.
95. Sanbonmatsu KY, Tung CS: High performance computing in biology: multimillion atom simulations of nanoscale systems. *J Struct Biol* 2007, **157**:470–480.
96. Brady JW: Molecular Dynamics Simulations of α -D-glucose. *J Am Chem Soc* 1986, **108**:8153–8160.
97. Brady JW: Molecular-dynamics simulations of alpha-D-glucose in aqueous solution. *J Am Chem Soc* 1989, **111**:5155–5165.
98. Yan ZY, Bush CA: Molecular dynamics simulations and the conformational mobility of blood group oligosaccharides. *Biopolymers* 1990, **29**:799–811.
99. Siebert HC, Reuter G, Schauer R, *et al.*: Solution conformations of GM3 gangliosides containing different sialic acid residues as revealed by NOE-based distance mapping, molecular mechanics, and molecular dynamics calculations. *Biochemistry* 1992, **31**:6962–6971.

100. Woods RJ, Edge CJ, Dwek RA: Protein surface oligosaccharides and protein function. *Nat Struct Biol* 1994, **1**:499–501.
101. Balaji PV, Qasba PK, Rao VS: Molecular dynamics simulations of high-mannose oligosaccharides. *Glycobiology* 1994, **4**:497–515.
102. Poveda A, Asensio JL, MartinPastor M, Jimenez Barbero J: Exploration of the conformational flexibility of the Le(X) related oligosaccharide GalNAc alpha(1–3)Gal beta(1–4)[Fuc alpha 1–3]Glc by H-1 NMR relaxation measurements and molecular dynamics simulations. *Chem Commun* 1996: 421–422.
103. von der Lieth C-W, Frank M, Lindhorst TK: Molecular dynamics simulations of glycoclusters and glycodendrimers. *Rev Mol Biotechnol* 2002, **90**:311–337.
104. Corzana F, Motawia MS, Du Penhoat CH, *et al.*: A hydration study of (1–4) and (1–6) linked alpha-glucans by comparative 10 ns molecular dynamics simulations and 500-MHz NMR. *J Comput Chem* 2004, **25**:573–586.
105. Lee G, Nowak W, Jaroniec J, *et al.*: Molecular dynamics simulations of forced conformational transitions in 1,6-linked polysaccharides. *Biophys J* 2004, **87**:1456–1465.
106. Gonzalez-Outeirino J, Kadirvelraj R, Woods RJ: Structural elucidation of type III group B *Streptococcus* capsular polysaccharide using molecular dynamics simulations: the role of sialic acid. *Carbohydr Res* 2005, **340**:1007–1018.
107. Landersjo C, Jansson JLM, Maliniak A, Widmalm G: Conformational analysis of a tetrasaccharide based on NMR spectroscopy and molecular dynamics simulations. *J Phys Chem B* 2005, **109**:17320–17326.
108. Lerbret A, Bordat P, Affouard F, *et al.*: How homogeneous are the trehalose, maltose, and sucrose water solutions? An insight from molecular dynamics Simulations. *J Phys Chem B* 2005, **109**:11046–11057.
109. Qian X, Nimlos MR, Davis M, *et al.*: *Ab initio* molecular dynamics simulations of beta-D-glucose and beta-D-xylose degradation mechanisms in acidic aqueous solution. *Carbohydr Res* 2005, **340**:2319–2327.
110. Neelov IM, Adolf DB, McLeish TC, Paci E: Molecular dynamics simulation of dextran extension by constant force in single molecule AFM. *Biophys J* 2006, **91**:3579–3588.
111. Almond A: Biomolecular dynamics: testing microscopic predictions against macroscopic experiments. In *NMR Spectroscopy and Computer Modeling of Carbohydrates* (eds Vliegthart JFG, Woods RJ), ACS Symposium Series, Vol. 930. Washington, DC: American Chemical Society; 2006, pp. 156–169.
112. Widmalm G: General NMR Spectroscopy of Carbohydrates and Conformational Analysis in Solution. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP): Oxford: Elsevier; 2007, pp. 101–132.
113. Jensen MO, Yin Y, Tajkhorshid E, Schulten K: Sugar transport across lactose permease probed by steered molecular dynamics. *Biophys J* 2007, **93**:92–102.
114. Kony DB, Damm W, Stoll S, *et al.*: Explicit-solvent molecular dynamics simulations of the polysaccharide schizophyllan in water. *Biophys J* 2007, **93**:442–455.
115. Jorgensen WL: Monte Carlo simulation of n-butane in water. Conformational evidence for the hydrophobic effect. *J Chem Phys* 1982, **77**:5757–5765.
116. Pastor RW: Molecular dynamics and Monte Carlo simulations of lipid bilayers. *Curr Opin Struct Biol* 1994, **4**:486–492.
117. Lins RD, Hunenberger PH: A new GROMOS force field for hexopyranose-based carbohydrates. *J Comput Chem* 2005, **26**:1400–1412.
118. Woods RJ, Dwek RA, Edge CJ, Fraserreid B: Molecular mechanical and molecular dynamical simulations of glycoproteins and oligosaccharides. 1. Glycam-93 parameter development. *J Phys Chem* 1995, **99**:3832–3846.
119. Basma M, Sundara S, Calgan D, *et al.*: Solvated ensemble averaging in the calculation of partial atomic charges. *J Comput Chem* 2001, **22**:1125–1137.

120. Rutherford TJ, Spackman DG, Simpson PJ, Homans SW: 5 nanosecond molecular-dynamics and nmr-study of conformational transitions in the sialyl-Lewis-X antigen. *Glycobiology* 1994, **4**:59–68.
121. Engelsen SB, Dupenhoat CH, Perez S: Molecular relaxation of sucrose in aqueous-solution – how a nanosecond molecular-dynamics simulation helps to reconcile NMR data. *J Phys Chem* 1995, **99**:13334–13351.
122. Hardy BJ, Egan W, Widmalm G: Conformational analysis of the disaccharide alpha-L-Rhap-(1–2)-alpha-L-Rhap-OMe – comparison of dynamics simulations with nmr experiments. *Int J Biol Macromol* 1995, **17**:149–160.
123. Ott KH, Meyer B: Molecular dynamics simulations of maltose in water. *Carbohydr Res* 1996, **281**:11–34.
124. Almond A, Sheehan JK, Brass A: Molecular dynamics simulations of the two disaccharides of hyaluronan in aqueous solution. *Glycobiology* 1997, **7**:597–604.
125. Venable RM, Bizik F, Henderson TJ, Egan W: Molecular dynamics simulations of an alpha-(2→8)-linked sialic acid tetramer in vacuum and solvent. *THEOCHEM* 1997, **395**:375–388.
126. Ueda K, Brady JW: Molecular dynamics simulations of carrabiose. *Biopolymers* 1997, **41**:323–330.
127. Daggett V: Long timescale simulations. *Curr Opin Struct Biol* 2000, **10**:160–164.
128. Frank M: Conformational analysis of oligosaccharides in the free and the bound state. Dissertation, Heidelberg University; 2000.
129. Apweiler R, Hermjakob H, Sharon N: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999, **1473**:4–8.
130. Berman H, Henrick K, Nakamura H, Markley JL: The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 2007, **35**:D301–D303.
131. Bohne-Lang A, von der Lieth C-W: GlyProt: *in silico* glycosylation of proteins. *Nucleic Acids Res* 2005, **33**:W214–W219.
132. Lutteke T, Frank M, von der Lieth C-W: Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res* 2005, **33**:D242–D246.
133. Naidoo KJ, Brady JW: Molecular dynamics simulations of a glycoprotein: the lectin from *Erythrina corallodendron*. *THEOCHEM* 1997, **395**:469–475.
134. Blanchard V, Frank M, Leeftang BR, *et al.*: The structural basis of the difference in sensitivity for PNGase F in the de-*N*-glycosylation of the native bovine pancreatic ribonucleases B and BS. *Biochemistry* 2008, **47**:3435–3446.
135. Molinero V, Goddard WA: M3B: A coarse grain force field for molecular simulations of malto-oligosaccharides and their water mixtures. *J Phys Chem B* 2004, **108**:1414–1427.
136. Loris R, Imberty A, Beeckmans S, *et al.*: Crystal structure of *Pterocarpus angolensis* lectin in complex with glucose, sucrose, and turanose. *J Biol Chem* 2003, **278**:16297–16303.
137. Pellegrini L, Burke DF, von Delft F, *et al.*: Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin. *Nature* 2000, **407**:1029–1034.
138. Liu Q, Brady JW: Anisotropic solvent structuring in aqueous sugar solutions. *J Am Chem Soc* 1996, **118**:12276–12286.
139. Humphrey W, Dalke A, Schulten K: VMD: Visual molecular dynamics. *J Mol Graphics* 1996, **14**:33–38.

Synergy of Computational and Experimental Methods in Carbohydrate 3D Structure Determination and Validation

Thomas Lütteke¹ and Martin Frank²

¹*Faculty of Veterinary Medicine, Institute of Biochemistry and Endocrinology, Justus-Liebig University Gießen, 35392 Gießen, Germany*

²*Deutsches Krebsforschungszentrum (German Cancer Research Centre), Molecular Structure Analysis Core Facility – W160, 69120 Heidelberg, Germany*

20.1 Introduction

For a complete understanding of the molecular processes in which carbohydrates are involved, such as protein–carbohydrate interactions and the impact of glycosylation on protein function, a knowledge of the 3D structure of the carbohydrate, the protein–carbohydrate complex, or the glycoprotein is often indispensable [1, 2]. The foremost methods of carbohydrate 3D structure determination are nuclear magnetic resonance (NMR) spectroscopy [3–5] and X-ray crystallography [6–8]. In the experimental determination of glycoprotein 3D structures, a variety of problems can arise. Due to the methods of protein expression and crystallization, the determination of the glycan part of a natural glycoprotein is often not possible. The glycan chains, for example, often hamper crystal growth and therefore in many cases are removed from the proteins before growing crystals. If proteins are expressed in bacterial cells, where the mammalian glycosylation machinery does not exist (see Section 8.1), there are no glycan chains attached to the protein, even if the protein is glycosylated in nature. Also, in those cases where glycans are present in a crystal, they are often fairly flexible and therefore do not yield enough electron density to allow determination of their structure [9, 10]. Consequently, the experimental determination of a 3D structure of a carbohydrate frequently needs to be supported by the application of theoretical methods (see also Chapter 19) [11–13]. This is especially true for the interpretation of NMR data because the available experimental restraints [e.g. atom distances derived from nuclear Overhauser effect (NOE) analysis] are frequently not sufficient to determine the 3D structure of a given carbohydrate unambiguously. The fact that carbohydrates are generally flexible molecules adds further complication to the situation: if two or more interconverting conformations of the carbohydrate coexist in an NMR sample, this will result, for example, in an average value of the measured NOE intensities that may not represent an existing conformation, but rather a “virtual conformation” [14]. The only way to judge whether the NOE data are in agreement with a single “low-energy conformation” is to perform a conformational analysis of the

carbohydrate structure using molecular modeling methods [13, 15–19]. A frequently used approach is to calculate conformational energy maps for the glycosidic linkages [20–23] and check whether the experimentally derived inter-glycosidic atom distances are in agreement with a structure representing a local minimum on the conformational map [24]. Theoretical methods can also be used to predict and compare experimentally observable features such as vicinal coupling constants, ensemble average NOE build-up curves [25], and residual dipolar couplings [26, 27]. If the predicted and experimental data are in satisfactory agreement, it can be assumed that the theoretical model – which may be a mixture of conformations – is a good approximation of the conformational ensemble present in the real sample.

The synergy between experiment and theory is also reflected in the fact that high-quality experimental data usually form the basis to determine the correct parameters for force field development [28–31], and later on force field-based methods can be applied in validation tools for crystal structures [23, 32] (see Section 20.2.4). In addition, experimentally resolved 3D structures are used as starting structures for docking calculations or molecular dynamics simulations. It is therefore important to have easy access to experimentally determined 3D structures of carbohydrates and their complexes with proteins.

This chapter gives an overview of how computational methods can be used to find and analyze carbohydrate 3D structures in the Protein Data Bank (PDB). Because carbohydrate sequences are not directly encoded in the PDB entries, the identity of a carbohydrate has to be determined by analyzing the atom types and coordinates [33]. Once the carbohydrate sequences have been assigned, this information can be used together with the atom coordinates, for example, to derive statistics about preferred torsion values of specific glycosidic linkages. The comparison of experimental and calculated data is also illustrated. Finally, the application of *Karplus equations* and *distance mapping* methods to support the determination of carbohydrate 3D structure in solution by NMR methods is briefly described (Section 20.3).

20.2 Data Mining 3D Structures of Carbohydrates in the Protein Data Bank

20.2.1 Databases as Information Resources for Structural Glycobiology

The two major databases where experimentally determined carbohydrate structures are stored are the Cambridge Structural Database (CSD) (<http://www.ccdc.cam.ac.uk/products/csd/>) and the Protein Data Bank (PDB) (<http://www.rcsb.org/pdb/>). Many crystal structures of small oligosaccharides are also accessible through the Glyco3D web interface (<http://www.cermav.cnrs.fr/glyco3d/>). Of the 4000 CSD entries classified as carbohydrates, many are monosaccharides [34]. Since carbohydrate entries in the CSD are not directly accessible over the Internet, we focus in this chapter on carbohydrate entries in the PDB. The PDB [35, 36] is the largest publicly available collection of biomolecular 3D structures. Most of the entries are proteins, with some DNA structures and protein–DNA complexes also present. Purely carbohydrate structures are rarely found in the PDB (some examples are shown in Chapter 18, Figure 18.1), but a considerable number [estimated to be approximately 7% (see below)] of the protein structures also contain carbohydrate moieties – covalently bound glycans and non-covalently bound as ligands (see Figures 20.1 and 20.2).

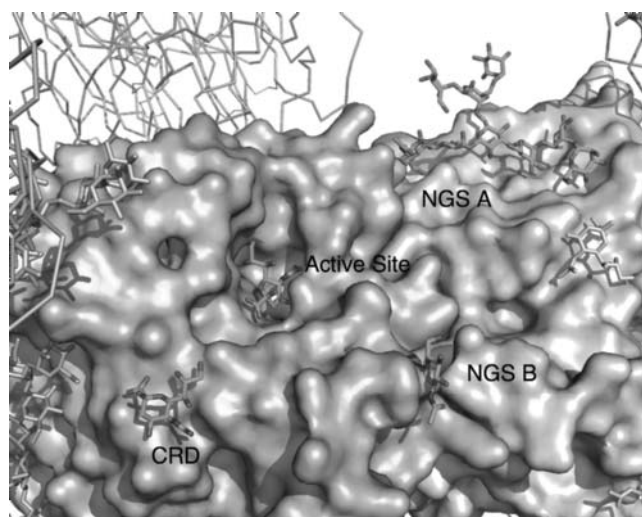


Figure 20.1 X-ray structure of avian influenza virus neuraminidase (PDB code 1MWE) [90]. Neuraminidase is an enzyme that cleaves terminal sialic acids (Neu5Ac) from glycoconjugates found on cell surfaces. The X-ray structure shows a Neu5Ac molecule bound in the active side of the enzyme, and a second Neu5Ac molecule bound to a lectin-like carbohydrate recognition domain (CRD). All three *N*-glycosylation sites (NGS) are glycosylated. A GlcNAc₂Man₇ glycan (named NGS A in the figure) is well resolved due to the interaction with a neighboring protein molecule in the crystal. However, for the remaining two glycosylation sites only the GlcNAc fragments could be resolved because of the lack of electron density, probably because the outer parts of the *N*-glycans are still too flexible in the crystal (see NGS B; the third NGS with a GlcNAc₂ fragment is not visible in the figure). A full-color version of this figure is included in the Plate section of this book.

There have been several attempts to gain information on properties of carbohydrates from statistical examinations of PDB entries. The main target of these studies have been *N*-glycans [9–11, 37]. In 1995, Imberty and Pérez [9] analyzed torsion angles of 44 *N*-glycan chains taken from 29 PDB entries. The main focus of this study was on the linkages between Asn and the proximal β -D-GlcpNAc residue, the Asn side-chain torsions, the orientation of the acetamido and the hydroxymethyl groups of β -D-GlcpNAc, and the backbone conformations of the glycoproteins. Almost a decade later, Petrescu *et al.* [37] performed a similar analysis, the number of occupied *N*-glycosylation sites in the PDB having grown to 1683 by then. Both studies revealed similar results regarding the torsion angles of the β -D-GlcpNAc-(1-*N*)-Asn linkage. The observed torsions of about -90° for ϕ_N ($O_5-C_1-N_8-C_\gamma$) and about 180° for ψ_N ($C_1-N_8-C_\gamma-C_\beta$), with ψ_N occupying a broader range of conformations than ϕ_N , correspond well with the values measured from small-molecule crystal structures of analogs of this linkage [38]. The comparison of the Asn side-chain torsions of occupied and unoccupied *N*-glycosylation sites revealed noticeable differences in the latter study. Both occupied and unoccupied Asn side-chains exhibit χ_1 torsions ($N-C_\alpha-C_\beta-C_\gamma$) of -60 , 60 or 180° , corresponding to the g^- , the g^+ , and the t conformer [39], respectively (Figure 20.3). As the Asn C_γ atom is not a tetrahedral carbon, the χ_2 torsion ($C_\alpha-C_\beta-C_\gamma-O_\delta$) does not display the three-fold staggered conformations, but shows a wide distribution centered at about 0° (180° when defined as $C_\alpha-C_\beta-C_\gamma-N_\delta$).

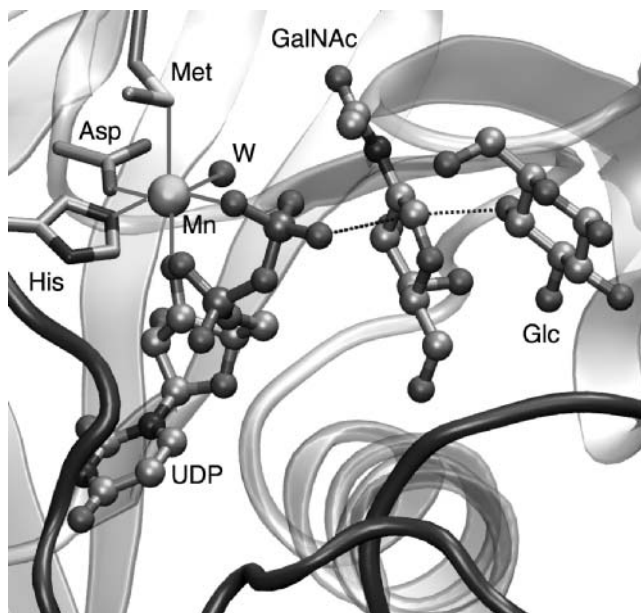


Figure 20.2 X-ray structure of the catalytic domain of bovine β -1,4-galactosyltransferase-I (Gal-T1) (PDB code 2FYD) [91]. Gal-T1 catalyzes the transfer of galactose from UDP-Gal to *N*-acetylglucosamine (GlcNAc), for example as part of the pathway to build complex-type biantennary *N*-glycans on glycoproteins [92]. In milk, Gal-T1 interacts with α -lactalbumin to form the lactose synthase complex (LS), which transfers the galactose moiety from UDP-Gal to free glucose to produce lactose (“milk sugar”). The X-ray structure shows a snapshot along the enzyme catalytic pathway of the Gal transfer. By replacing UDP-Gal by UDP-GalNAc, the chemical reaction could be trapped in the S_N2 -type transition state [91].

This distribution is much smaller for glycosylated than for non-glycosylated Asn residues. Furthermore, the relative populations of the three conformers of the χ_1 torsion change upon glycosylation (Figure 20.3). In unoccupied Asn side-chains, the g^- conformer is preferred over the t conformer, whereas in occupied Asn the t conformer is found more frequently than the g^- conformer. The g^+ conformer is the rarest one in both glycosylated and non-glycosylated Asn residues [37]. From the small dataset that was available in 1995, these differences could not be seen, so Imberty and Pérez assumed at that time that *N*-glycosylation does not have a significant effect on Asn side-chain conformation. The examples of the β -D-GlcpNAc-(1-*N*)-Asn linkage torsions and the Asn side-chain torsions show that even rather small datasets can yield information on preferred conformations of glycosidic linkages, but that some specific properties may only be seen in larger datasets.

Analysis of torsion angles of various kinds of glycosidic linkages reveals that both the preferred torsions and the degree of conformational dispersion depend on the linkage position and the participating monosaccharide residues [10, 11] (see Section 19.2). As a result of the additional rotatable bond, most scatter is seen with 1–6 linkages. For ω torsions ($O_5-C_5-C_6-O_6$), three staggered conformations are possible, which are named *gg*, *gt*, and *tg* (see Figure 20.4a). In monosaccharides with an axial OH group at position 4, such as D-Galp, mostly the *gt* conformation is seen, whereas monosaccharides with an equatorial

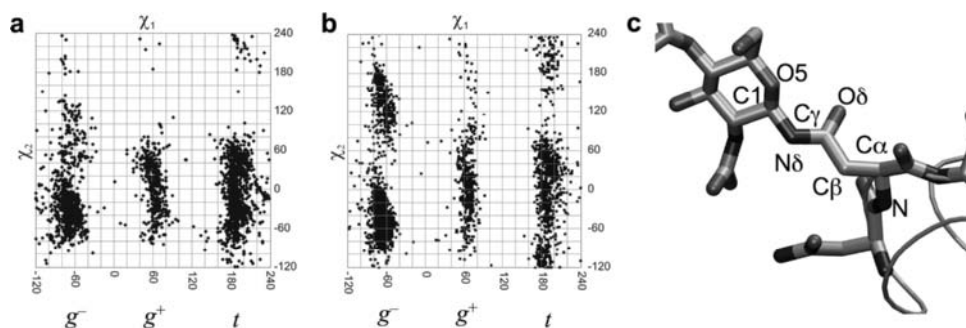


Figure 20.3 Comparison of Asn side-chain torsions of occupied *N*-glycosylation sites (a), and all Asn residues (b), in the PDB. Atom names are defined in (c). Glycosylated Asn side-chains occupy a narrower range of χ_2 (C_α - C_β - C_γ - O_δ) torsion angles, and the relative frequencies of the three staggered conformations of the χ_1 (N - C_α - C_β - C_γ) torsion angle differ between occupied *N*-glycosylation sites and all Asn residues.

4-OH group, such as D-Glcp or D-Manp, are preferably present in *gg* and *gt* conformations [10] (Figure 20.4b–d; for more details, see Section 19.2).

20.2.2 Searching 3D Structures of Carbohydrates in the Protein Data Bank

After the 3D structures of carbohydrate moieties have been experimentally resolved, researchers are often confronted with another problem: the large number of different monosaccharides and the fact that these are frequently modified require relatively long residue names for the carbohydrate residues (see Chapter 2). This causes some problems with respect to storage of the experimental result in an established computer-readable format. Common monosaccharide names often exceed the space that is allotted to residue names in file formats used to store 3D structural information. For example, the file format of the PDB (see below) uses residue names consisting of three characters. This is sufficient to encode amino acids, for which a commonly accepted three-letter code exists, but not to describe carbohydrates using common nomenclature. Therefore, it is often difficult to see readily if a PDB residue name encodes a carbohydrate. Furthermore, there are some redundancies within these residue names and, in addition, until recently many PDB carbohydrate residues were ambiguously defined and encoded; for example, for both the α - and the β -anomers of a monosaccharide moiety the same residue code was used. During the PDB remediation [40], many residue names have recently been redefined. Thereby, most of the ambiguities and redundancies have been removed, but this does not solve the problem of the “unusual” residue names. For these reasons, it is difficult for glycobiologists to find 3D structures of the carbohydrates of their interest.

The following sections describe an algorithm that overcomes this problem by locating and assigning carbohydrate information in 3D structural data without making use of the residue notation present in the structure file. The description of the algorithm is followed by an overview of carbohydrate structures that were detected in the PDB using that algorithm.

20.2.2.1 Detection of Carbohydrate Rings. The detection of carbohydrate residues starts with searching the 3D structural data for rings. Among them, potential carbohydrate rings are marked. These must consist of five or six atoms, one of which must be an oxygen

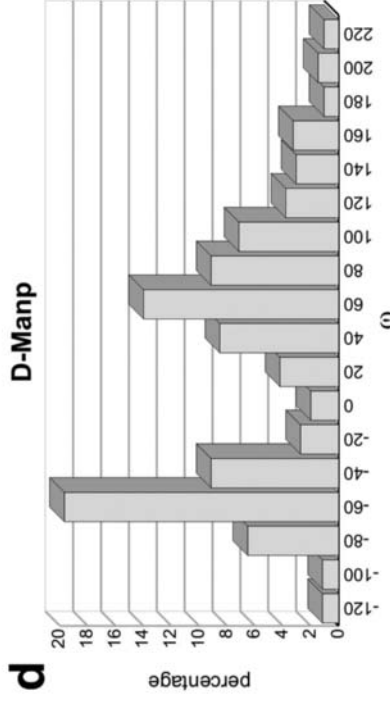
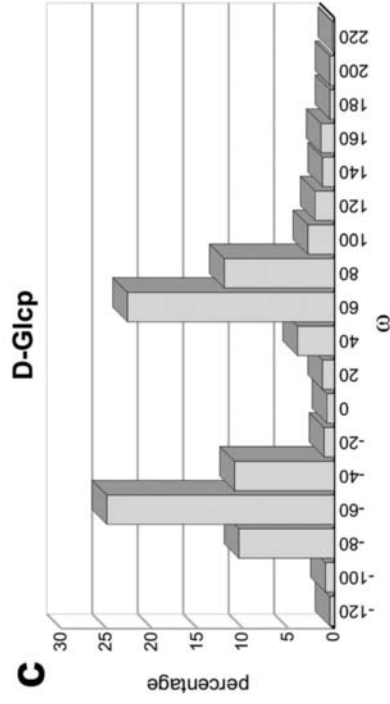
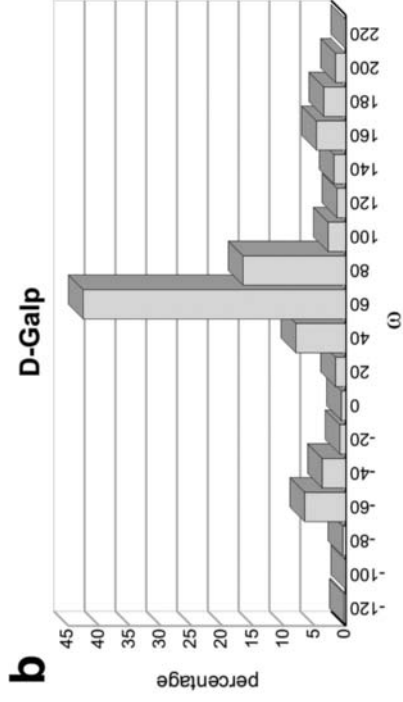
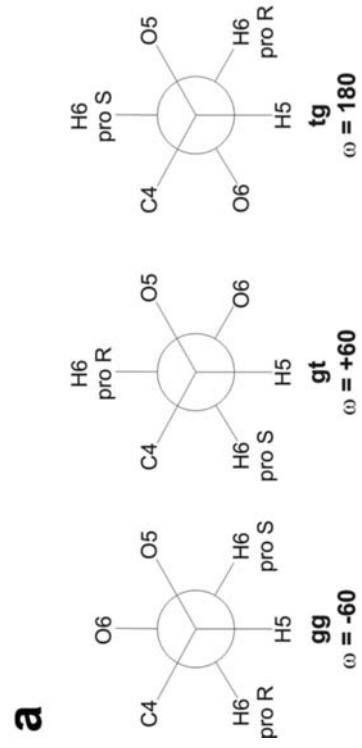


Figure 20.4 The ω ($O_5-C_5-C_6-O_6$) torsions usually adopt one of the three staggered conformations *gg*, *gt*, and *tg* (a). In residues with an axial hydroxyl group at position 4, such as D-Galp, mostly the *gt* conformation is present (b), while both the *gg* and the *gt* conformations are observed in residues with an equatorial hydroxyl group at position 4, such as D-Glcp (c) or D-Manp (d).

atom, whereas all the other atoms in the ring must be carbon. Furthermore, the ring must not be planar, and one of the carbons that are attached to the ring oxygen, the anomeric center, must be connected to an exocyclic oxygen, nitrogen, or sulfur atom. However, there are several carbohydrate rings within the PDB where the exocyclic oxygen at the anomeric center is missing. Therefore, in the case where no such exocyclic oxygen atom is found, the two carbon atoms attached to the ring oxygen atom have to be checked again. If one of them is bonded to a non-ring carbon, this must be the C5 atom (or, if the residue is not an aldopyranose, the respective atom of the present residue type). Consequently, the other carbon that is connected to the ring oxygen must be the anomeric center. A visual inspection of such structures revealed that in many cases the anomeric center of such a ring is in the proximity of the N_δ atom of an Asn side-chain or of a hydroxyl oxygen atom of another carbohydrate ring, but the distance between the anomeric center and the potential glycosidic atom is slightly too large to form a bond (see Figure 20.7d). One can assume that in those cases the bond actually exists, but the relative positions of the residues are not well resolved in the 3D structure, due for example to a low resolution of the structure. Therefore, such bonds can be assigned by looking for potential glycosidic atoms in the proximity of the anomeric center. Amino acid atoms that are not known to form glycosidic bonds are excluded, and the same applies to atoms that belong to the same residue as the anomeric center that misses the glycosidic atom. The latter check is necessary to prevent the introduction of wrong linkages within one monosaccharide unit. For each potential glycosidic atom a penalty score is calculated from the relative deviations of bond length and bond angle from the expected values. Only atoms with a penalty score below a given threshold are considered. In case more than one potential glycosidic atom is found, the one with the lowest penalty score is assigned as the correct glycosidic atom.

20.2.2.2 Detection of Carbohydrate Chains. After the monosaccharide units have been determined, the rings, which in the first step were assigned as carbohydrate rings, are searched for reducing ends, that is, for rings the anomeric center of which is not glycosidically linked to another carbohydrate ring. The reducing end was chosen as a starting point for the chain detection because each carbohydrate chain has only one reducing end. An exception is formed by chains in which two anomeric centers are glycosidically connected and therefore no reducing end exists, as in sucrose. In such a case, both sub-chains are analyzed independently, and the data are merged afterwards.

If a reducing end is found, the monosaccharide type of that residue is assigned (see below). Subsequently, the non-anomeric carbon atoms, for example C₂ to C₆ in the case of aldohexoses, are checked for further glycosidically bonded carbohydrate rings, which are recursively handled in the same way. In this manner, an entire carbohydrate chain is analyzed. Finally, the ring at the reducing end is reinvestigated, by checking to see if it is glycosidically linked to a non-carbohydrate residue, for example, an amino acid in case of *N*- or *O*-glycan chains. From the data collected in these steps, a unique description of the carbohydrate chain can be generated in LINUCS [41] notation, a linear, unique notation for carbohydrate structures (see Chapter 3).

20.2.2.3 Assignment of Monosaccharide Types. To distinguish between the different types of monosaccharide residues, a code containing the ring size and the stereochemistry of the chiral centers is generated for each carbohydrate ring. The ring size is determined by a simple count of atoms in the ring. To identify the stereochemistry of an atom C_{*n*}, a virtual torsion angle C_{*n*-1}-O_{*n*}-H_{*n*}-C_{*n*+1} is calculated (Figure 20.5). If the hydroxyl group

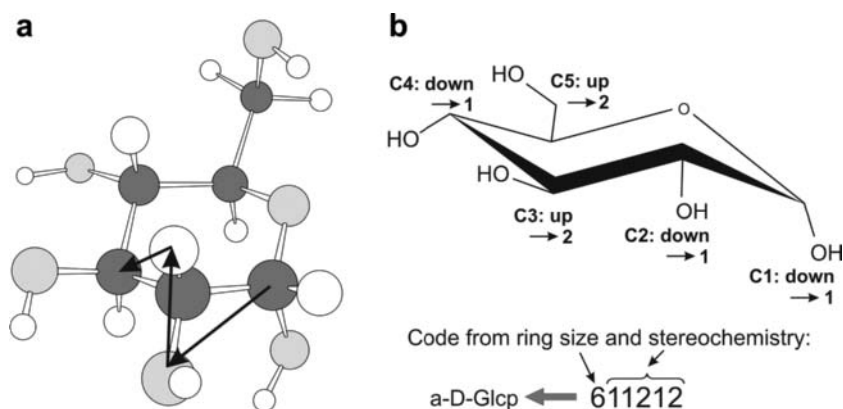


Figure 20.5 Assignment of monosaccharide residue types from their 3D structure. To detect the stereochemistry of a ring carbon C_n , a virtual torsion angle ($C_{n-1}-O_n-H_n-C_{n+1}$) is defined (a). A code is generated from the ring size of the monosaccharide and the stereochemistry of the ring carbons (b). The torsion angle as defined in (a) is either positive or negative, which is represented by a 1 or 2, respectively, in the code. Using a lookup table, the code is matched to a residue name. Reprinted from [33] with permission from Elsevier.

points down relative to the plane spanned by the atoms C_{n-1} , C_n , and C_{n+1} , this angle is positive, which is represented by a “1” in the code. If the hydroxyl group points up, the angle is negative, and a “2” is added to the code. A “0” is assigned to achiral positions. After the code has been completed, it is compared with the values stored in a lookup table, in which the assignments of the codes to monosaccharide residue types are stored. After the basic type of a carbohydrate residue, for example “b-D-Glcp”, has been assigned in this way, modifications such as acetylation or methylation are determined to reveal the full CarbBank style (see Section 3.2.2) residue name, such as “b-D-GlcpNAc”.

20.2.2.4 Implementation of the Algorithm. The approach described above is implemented in the `pdb2linucs` software, which is available online through a web interface at the GLYCOSCIENCES.de web portal (<http://www.glycosciences.de/tools/pdb2linucs/>). This program searches PDB structures for carbohydrates and displays the detected carbohydrate chains using the LINUCS notation. On the results page, the 3D structure can be viewed with Jmol or Chime. The detected carbohydrate residues are highlighted within the 3D structure view.

Since the LINUCS nomenclature is optimized for use in computer programs and is not easy to read for humans, the web interface can also display the results in IUPAC notation. The translation of LINUCS into IUPAC nomenclature is done utilizing the LiGraph tool (<http://www.glycosciences.de/tools/LiGraph/>). The same program is used to generate pictograms of the carbohydrate chains, which are accessed through a link on the results page. If there is an entry in the GLYCOSCIENCES.de database (the former SweetDB) corresponding to a carbohydrate chain found in a PDB structure file, a direct link to that entry is established. Thereby, `pdb2linucs` allows a cross-linking between a proteomics and a glycomics database. In case the 3D structure contains covalently bound glycans, the amino acid sequence of the protein is also displayed. Glycosylation sites are marked in the sequence.

Table 20.1 Carbohydrates in the Protein Data Bank^a. The majority of carbohydrate compounds in the PDB (March 2008) are either *N*-glycosidically linked or non-covalently bound. *O*-Glycosidically linked chains are in the minority.

Chain type	PDB entries ^b			Carbohydrate chains			Carbohydrate residues		
	Count	Ratio (%) ^c		Count	Ratio (%) ^c		Count	Ratio (%) ^c	
Glycans total	1716	48.2	–	7181	57.6	–	13311	58.6	–
Asn	1595	44.8	92.9	6398	51.4	89.1	12399	54.6	93.1
Ser	101	2.8	5.9	331	2.7	4.6	387	1.7	2.9
Thr	58	1.6	3.4	346	2.8	4.8	382	1.7	2.9
Glu	43	1.2	2.5	88	0.7	1.2	116	0.5	0.9
Asp	17	0.5	1.0	17	0.1	0.2	26	0.1	0.2
Tyr	1	0.03	0.1	1	0.01	0.01	1	0.004	0.01
Ligands	2142	60.2	–	5277	42.4	–	9400	41.4	–
Total	3561	–	–	12458	100	–	22711	100	–

^a To exclude wrong positive hits, only residues whose PDB residue names were known to be carbohydrate residues were counted.

^b The total number of PDB entries with carbohydrates is lower than the sum of those with *N*- or *O*-glycans and those with non-covalently bound ligands, since some entries contain both glycans and ligands. The same applies to the number of entries with glycans and the numbers of entries with the different types of amino acids, to which the glycans are linked.

^c Ratios for “Glycans total” and “Ligands” are calculated with reference to the total numbers of entries, chains, or residues, respectively. For the different types of glycans, both the portions of the total numbers (left column) and of the total number of glycans (right column) are given.

20.2.3 Overview of Carbohydrate Structures Detected in the PDB

In the PDB release of March 2008, which consisted of about 50 000 entries, a total of 3561 entries that contain carbohydrate residues were found. The majority of these carbohydrate chains are *N*-glycans or non-covalently bound ligands, whereas *O*-glycosidically linked glycans form a minority (Table 20.1). In total, the 12 458 carbohydrate chains that were found in the PDB consist of 22 711 monosaccharide residues. This means that the average chain length is about 1.8 residues, which shows that the majority of carbohydrate chains in the PDB are rather short. More than 80% of all chains are mono- or disaccharides, but some longer chains of up to 16 residues are also present (Table 20.2).

The most frequent carbohydrate residue present in the PDB is β -D-GlcpNAc (Figure 20.6d). This arises from the fact that more than 50% of all carbohydrate residues in the PDB are found in *N*-glycosidically bound chains. Of these, in most cases only the first one to three residues are resolved in the 3D structures (see Table 20.2), because the residues further away from the reducing end are often too flexible to provide enough electron density for a determination of their 3D structures, or have been clipped prior to crystallization. Since the first two residues in an *N*-glycan chain are β -D-GlcpNAc, more than two-thirds of the *N*-glycosidically bound carbohydrate residues in the PDB are of this type. The residues that are present in the *N*-glycan core (β -D-GlcpNAc and D-Manp) together form 88.5% of all monosaccharide units in the *N*-glycan chains in the PDB (Figure 20.6a). The remaining 11.5% mainly consist of α -D-GlcpNAc and L-Fucp residues. Beyond that, D-Galp, D-Xylp, D-Glcp and D-Neup5Ac are found in *N*-glycans in the PDB. As there is no biological pathway known for α -D-GlcpNAc in *N*-glycan chains [11], it is very likely that those are incorrectly resolved structures that in nature are also present as β -D-GlcpNAc residues.

Table 20.2 Chain length (i.e. number of monosaccharide units per chain) of carbohydrate chains found in the Protein Data Bank.

Length	N-Glycan	O-Glycan	Ligand	Total
1	3575	707	3044	7326
2	1456	50	1240	2746
3	672	13	504	1189
4	203	5	205	413
5	184	4	176	364
6	143	2	53	198
7	74	2	24	100
8	49	–	17	66
9	35	–	5	40
10	6	–	3	9
11	1	–	3	4
12	–	–	1	1
16	–	–	2	2

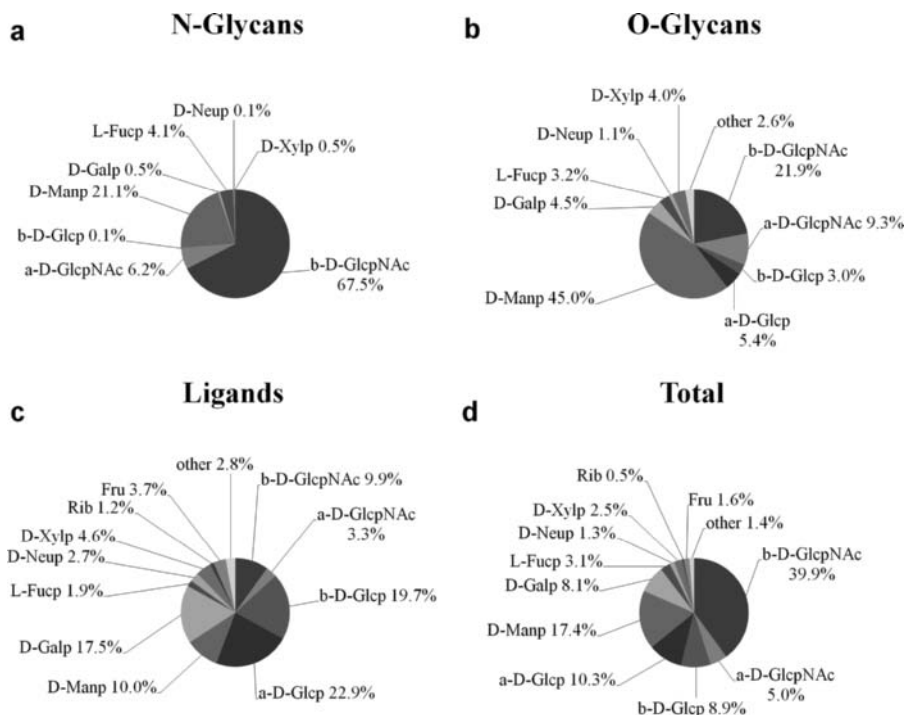


Figure 20.6 Distribution of monosaccharide residue types that were detected in the PDB. In *N*-glycosidically linked chains, only D-GlcpNAc, D-Manp, D-Glcp, D-Galp, L-Fucp, D-Neup5Ac, and D-Xylp residues were found. Approximately two-thirds of all residues in *N*-glycan chains are β -D-GlcpNAc (a). Although almost 10 times more *N*-glycan than *O*-glycan chains are present, the variation of residue types is larger in the latter class (b). Non-covalently bound ligands, which are not as dependent on biological pathways as covalently linked glycan chains, show an even larger variability of residue types, with glucose residues predominating (c). In total, almost 90% of all monosaccharide units that are present in the PDB are of a glucose, mannose, or galactose base type (d).

The relatively low ratio of galactose and sialic acid residues that were detected in *N*-glycan chains can be explained by the fact that these residues mainly appear in the terminal regions, beyond the GlcNAc₂Man₃ core, but about 90% of the *N*-glycosidically linked glycans in the PDB consist of no more than three residues (see Table 20.2). For the same reason, L-Fucp is found in *N*-glycan chains in the PDB almost exclusively as core-fucosylation. Only one PDB entry (1LGC) features L-Fucp as a terminal modification of an *N*-glycan antenna.

In *O*-glycan chains, D-Manp (45.0%) is the most frequent residue, but the different D-Glcp-based residues together (39.6%) form only a slightly smaller fraction (Figure 20.6b). Although there are significantly fewer residues present in *O*-glycan than in *N*-glycan chains (see Table 20.1), the variety of different monosaccharides is larger for *O*-glycosidically linked glycans. This arises from the fact that unlike *N*-glycans, there is no single core structure common to all *O*-glycans (see Chapter 8). In non-covalently bound ligands, which are not dependent on biological pathways, the variety of different carbohydrate residues is even larger (Figure 20.6c).

20.2.3.1 Erroneous Carbohydrate Entries in the PDB. Within the carbohydrate parts of PDB entries, a relatively high error frequency is found. About 30% of all PDB entries that contain carbohydrates are erroneous [33]. The main sources of these errors are inconsistencies between the residue names used in the PDB files and the carbohydrate residues found in the 3D structural data. This problem has two obvious reasons. On the one hand many experimentalists who deposit biomolecular 3D structures in the PDB are mainly interested in the protein part and are less concerned about any associated carbohydrates. On the other hand, as mentioned above, the PDB file format allows only three letters for the residue names, but carbohydrate residue names usually require more space than this. Therefore, the three-letter codes used in the PDB are usually very cryptic, so that even for carbohydrate experts it is often difficult to assign the correct names. The recent remediation of the PDB entries [40] has solved most of the notation inconsistencies by assigning the correct names to the non-protein residues. The mismatches between residue names and actual 3D structure, however, are not always due to the selection of wrong residue names but can also indicate errors in the 3D structure. Therefore, the renaming of residues solves those problems that are based on wrong notation but hides the problems that result from erroneous coordinate data.

Other major sources of errors aside from residue notation are connected with *N*-glycan structures without any known biosynthetic pathway and with the information stored on atom connections. Surprisingly, surplus connections are almost as frequently found as missing connections, which can lead to awkward structures (Figure 20.7a and b). In some structures, up to hexavalent carbons can be found, with some “bond lengths” of more than 60 Å. Surplus atoms seldom occur, and are only found within the glycosidic linkages. A problem that occurs mainly at the linkage between *N*-glycan chains and the protein, but sometimes also between individual residues of a carbohydrate chain, concerns the relative positions of the residues. If the distance between the anomeric carbon of a carbohydrate ring and the atom to which it is linked is too large, no link is detected in the structure (Figure 20.7d). The proximal D-GlcpNAc ring in Figure 20.7d illustrates another issue. This ring is not in a chair conformation but in a rather distorted conformation. As a result, linking the anomeric center of this ring to the N_δ2 atom of the Asn would form an α-D-GlcpNAc instead of a β-D-GlcpNAc residue.

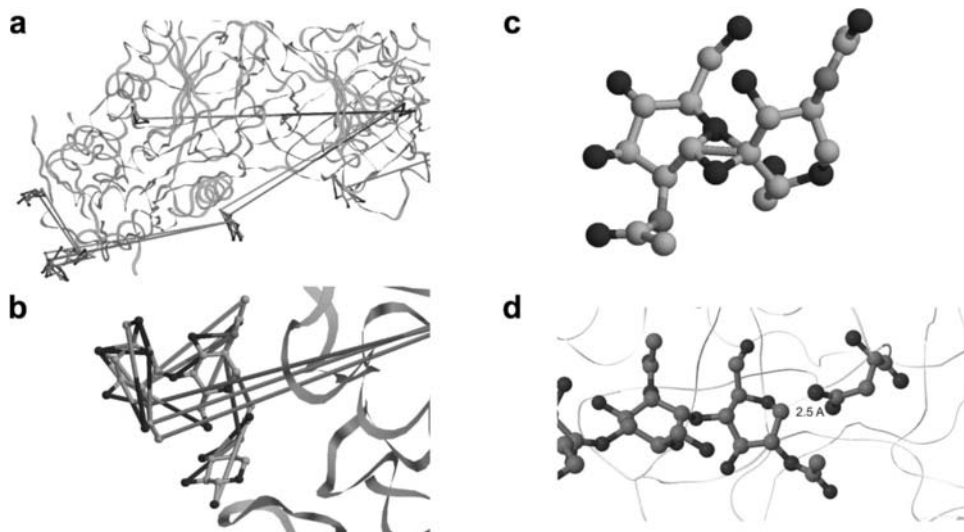


Figure 20.7 Errors within the connection information form the second largest class of errors found in the carbohydrate parts of PDB entries. Surplus connections are almost as frequent as missing ones. In cases where wrongly connected atoms are far distant from each other, such errors can be observed on the first view (a). If there are many surplus connections ranging over short or medium distances present, they can still be fairly easily detected visually when looking at the carbohydrate residues (b). However, if only a few short, extra connections are present, without a check program they can only be observed by zooming in and looking carefully at the affected residues (c). If residues that are involved in a glycosidic linkage are positioned too far away from each other in the 3D structural data, the linkage cannot be detected correctly any longer (d). (a, b, PDB entry 1DZQ; c, PDB entry 1BCS; d, PDB entry 1ABR).

Another reason for the relatively high error rate within carbohydrate structures is the lack of check programs. For the protein parts of PDB entries, several check programs are available [42]. For carbohydrate components, however, it is only recently that two check programs were developed. The first one, *pdb-care* (PDB CARbohydrate RESidue check; <http://www.glycosciences.de/tools/pdb-care/>) [43], is based on *pdb2linucs* (see above). For each carbohydrate residue detected, the assigned residue name is compared with that expected from the three-letter residue name used in the PDB file. Discrepancies are reported to the user. If available, a three-letter code for the detected residue type is also given in case of residue mismatches, which aids the user in choosing a correct notation. In addition to the notation check, bond lengths and atom valences are also inspected to detect errors within the connectivity information. Currently, *pdb-care* only checks the notation of single residues. No check is done for the biological likeliness of carbohydrate chains. For *N*-glycan structures, this can be analyzed using *getCarbo* (<http://www.glycostructures.jp>), the second check program that is currently available [44]. This software detects carbohydrates in PDB files in a way similar to *pdb2linucs*. Subsequently, it compares *N*-glycosidically bound glycans with those present in the KEGG Glycan (<http://www.genome.jp/kegg/glycan/>) database [45]. If no corresponding entry is found in that database, no biological pathway is known so far for the structure, which means that the 3D coordinates may be erroneous. The use of these two programs to validate 3D structure files which contain carbohydrates, before sending them to the PDB, would certainly reduce the amount of errors in carbohydrate 3D structures.

20.2.4 Comparison of Carbohydrate Conformation Data Derived from the PDB with Theoretical Data

Experimental data on carbohydrate conformation, such as the preferred orientation(s) of specific glycosidic linkages, can be derived directly from databases containing 3D structure data, for example from carbohydrates found in entries of the PDB (see Section 20.2.1). The GlyTorsion software (<http://www.glycosciences.de/tools/glytorsion/>) [46] analyzes the PDB and provides direct access to the individual values of the glycosidic torsion angles ϕ and ψ (and ω). Each single entry represents only one snapshot of the possible orientations of a glycosidic linkage, but the more data that are available for a disaccharide fragment representing the linkage, the better they reflect the conformational space that can be occupied by the linkage [10].

Calculated conformational energy maps (see Section 19.3.1) derived from Glyco-MapsDB [23] and experimentally resolved structures can readily be compared by plotting the torsion angles found in PDB structures onto the respective maps. Here, two cases have to be distinguished: (a) the carbohydrate chain in the PDB structure and the position of the linkage within that chain exactly match the carbohydrate chain that was used to calculate the map and the location of the linkage within that chain, respectively; or (b) the disaccharide fragment in the PDB structure is the same as that in the map, but is present in a different carbohydrate chain or at a different location within the chain (fragment match). This distinction is necessary because the conformational space that can be occupied by a glycosidic linkage is governed not only by the linkage type and the two residues that are linked via this linkage, but also by neighboring residues. For example, if a glycosidic linkage is present in a PDB entry in a disaccharide, and the linkage in the map is part of a branched oligosaccharide, then the experimental structure might be present in a conformation that cannot be accessed by the linkage in the map due to steric hindrance. Thus, the proportion of experimental structures with glycosidic linkages located within energetically favorable regions of the appropriate map in general is higher among the exactly matched structures than among the fragment matches. Figure 20.8 illustrates this for the conformational map of one of the glycosidic linkages in the trisaccharide β -D-GlcpNAc-(1-4)- β -D-GlcpNAc-(1-4)- β -D-GlcpNAc. More than 70% of the β -D-GlcpNAc-(1-4)- β -D-GlcpNAc disaccharide fragments of experimentally resolved structures are located within energy areas of 0–1.5 kcal mol⁻¹, and energy areas between 0 and 3.0 kcal mol⁻¹ include approximately 80% of the structures. Within the exact structure matches, more than 90% of the structures are found in the most favorable areas with calculated energy between 0 and 1.5 kcal mol⁻¹. However, the total number of exact matches (29) is considerably smaller than that of the fragment matches (2614), which lowers the statistical significance of the latter results. In general, the data derived from the carbohydrates in the PDB are in good accord with the theoretical data, which supports the correctness of the computational results.

A second application that makes use of a comparison of experimental and computational data is the *carbohydrate Ramachandran plot* or “*carp*” tool (www.glycosciences.de/tools/carp/) [46]. The “classical” Ramachandran plot of protein backbone torsions [47] is frequently used to evaluate the quality of protein 3D structures [42, 48]. Combinations of ϕ/ψ torsion angles that are outside the “allowed” areas indicate potential errors in the protein structure. In principle, this approach can also be applied to carbohydrate structures. However, as the possible conformations of a glycosidic linkage depend on several factors such as the residues involved, the linkage type, or neighboring residues, a linkage-specific evaluation is required. This can be done in two ways: by comparing the torsions that are

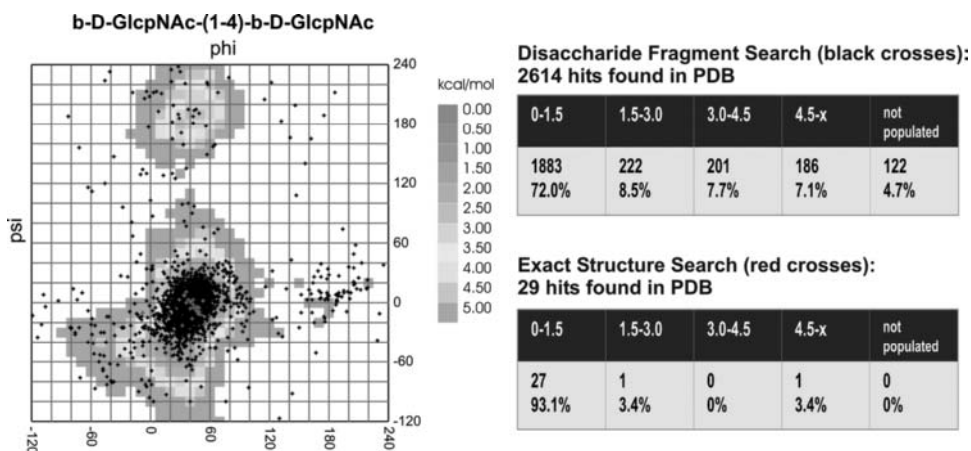


Figure 20.8 Comparison of experimental and theoretical torsion angle data for the β -D-GlcpNAc-(1-4)- β -D-GlcpNAc linkage. Comparison of experimental data with calculated conformational energy maps can be done by plotting the experimentally determined torsions onto the energy map. In those cases where the experimental torsions were obtained from carbohydrate chains that exactly match the one that was used to calculate the energy map (grey crosses), more than 90% of all torsions are located in low-energy areas of 0–1.5 kcal mol⁻¹. Of those torsions that were derived from disaccharide fragments that are identical with the one from which the map was computed, but are part of a different carbohydrate chain (black crosses), still more than 70% are found in low-energy areas of 0–1.5 kcal mol⁻¹, and energy areas of 0–3 kcal mol⁻¹ cover more than 80% of the β -D-GlcpNAc-(1-4)- β -D-GlcpNAc linkage torsions that are present in the PDB. A full-color version of this figure is included in the Plate section of this book.

present in the candidate structure with that present in other PDB entries or with calculated conformational maps. When using conformational maps for the comparison, one should keep in mind that the maps are calculated from carbohydrate chains that are free in solution or the gas phase, whereas the majority of carbohydrates in the PDB are either covalently or non-covalently bound to proteins, which can influence the conformations of the glycosidic linkages. The high conformance of experimentally observed torsions and computed energy maps suggests that the carbohydrate chains are usually bound to the protein in a low-energy conformation. However, in some cases the binding to a protein results in conformations that differ from those which the carbohydrate preferably adopts in its free state [49]. Therefore, “unusual” conformations are not necessarily erroneous. Nevertheless, this technique can aid researchers in finding potential problems in 3D structural data.

20.3 Supporting Interpretation of NMR Experiments by Using Theoretical Methods

Recognition of the relationship between molecular conformations and the biological function of oligosaccharides has stimulated interest in their structure and conformational properties in solution. For the determination of the 3D structure of complex carbohydrates in solution, NMR is the preferred experimental technique [5]. ¹H, ¹³C, ¹⁵N, and ³¹P are the most commonly used magnetically active nuclear sensors present in carbohydrates, that signal information about their local environment in an NMR experiment. Typical parameters that

are measured are the chemical shift [50], homo- and heteronuclear scalar coupling constants [51], residual dipolar couplings [26, 52, 53], relaxation times [54], and NOE intensities [55, 56]. However, since oligosaccharides have several degrees of internal motional freedom, the observed parameters represent averaged values and the “average” conformation deduced from NMR measurements may not have any direct physical significance, and may not correspond to any of the actual molecular conformations. Therefore, it is necessary to “decompose” the data into contributions from individual conformations and theoretical conformational analysis (see Chapter 19) is required to generate a set of reasonable 3D models of carbohydrates that can be used as a starting point for the interpretation of the NMR data.

It is beyond the scope of this chapter to review comprehensively the NMR methods used to derive information on carbohydrate 3D structure and dynamics. Excellent reviews and books already exist [3–5, 11, 12, 57, 58] and the interested reader will find the experimental details in the original literature.

A key objective of conformational studies of oligosaccharides in solution is the assessment of the orientation of the glycosidic linkage torsions ϕ and ψ (Figure 20.9a), since these torsions determine the overall shape of a carbohydrate. In this section, we will therefore highlight two important properties observable by NMR that can be readily compared with computationally calculated values: scalar 3J coupling constants and NOE intensities.

20.3.1 Determination of Torsion Angles Based on 3J Coupling Constants

Scalar coupling constants that are sensitive to the relative orientation of two chemical bonds are of particular interest for the determination of carbohydrate conformation [51, 59–63]. The dependence of the three bond coupling constants 3J on the torsion angle φ can be

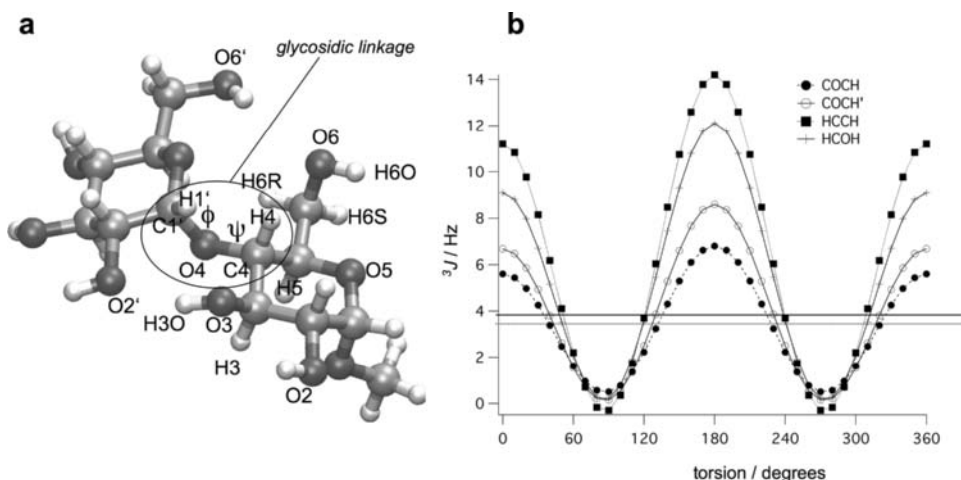


Figure 20.9 Determination of torsion values from scalar vicinal coupling constants using Karplus equations. (a) Low-energy structure of maltose with atom labels. (b) Graphical plots of various Karplus equations: heteronuclear: COCH (Equation 20.3), COCH' (Equation 20.4); and homonuclear: HCCH (Equation 20.2), HCOH (Equation 20.5). The horizontal lines indicate the experimentally determined [75] coupling constants for maltose: $^3J(\text{H1}'\text{-C1}'\text{-O4-C4})$ (ϕ) = 3.5 ± 0.3 Hz and $^3J(\text{C1}'\text{-O4-C4-H4})$ (ψ) = 3.9 ± 0.2 Hz.

described by a Karplus [64] relationship of the general form

$${}^3J = A \cos^2 \varphi + B \cos \varphi + C \quad (20.1)$$

The parameters A , B , and C have to be derived for each set of connected atom types individually, either by fitting experimental values [65, 66] or by fitting coupling constants calculated using quantum mechanics (DFT) methods [67].

One of the first proposed Karplus equations for the H–C–C–H segment was [68]

$${}^3J_{\text{HCCH}} = 13.0 \cos^2 \varphi - 1.5 \cos \varphi - 0.3 \quad (20.2)$$

For C–O–C–H segments, present in glycosidic linkages, commonly used Karplus equations are [65, 67]

$${}^3J_{\text{COCH}} = 5.7 \cos^2 \varphi - 0.6 \cos \varphi + 0.5 \quad (20.3)$$

$${}^3J_{\text{COCH}} = 7.49 \cos^2 \varphi - 0.96 \cos \varphi + 0.15 \quad (20.4)$$

Using one of these equations, the orientation of the glycosidic linkage torsions can be determined based on measurement of heteronuclear ${}^1\text{H}$ – ${}^{13}\text{C}$ 3J coupling constants. However, complex proton multiplicity and similar magnitude of homo- and heteronuclear couplings render a direct estimation of J_{COCH} difficult [5].

Hydroxyl protons exchange in D_2O solution and are therefore not detectable by NMR spectroscopy under “physiological conditions”. However, coupling constants of the H–C–O–H segment can be determined when using $\text{DMSO-}d_6$ as solvent [69] or in aqueous solutions under *supercooled* conditions [70]. The angular dependence of the OH group rotation can be described by the Karplus relationship [68]:

$${}^3J_{\text{HCOH}} = 10.4 \cos^2 \varphi - 1.5 \cos \varphi + 0.2 \quad (20.5)$$

More Karplus equations can be found in the literature. Of particular interest are equations to determine the orientation of the exocyclic hydroxy methyl group [51, 59, 71].

Average 3J coupling constants can easily be calculated from an ensemble of structures generated by molecular dynamics (MD) or Metropolis Monte Carlo (MMC) simulation and compared with experimental values [72–74]. Graphical representations of the Karplus equations given above are shown in Figure 20.9b. The experimental values for maltose are ${}^3J(\text{H1}'\text{--C1}'\text{--O4--C4}) (\phi) = 3.5 \pm 0.3 \text{ Hz}$ and ${}^3J(\text{C1}'\text{--O4--C4--H4}) (\psi) = 3.9 \pm 0.2 \text{ Hz}$ [75]. This gives clear evidence that ϕ and ψ do not adopt an *anti* (180°) or *syn* (0°) orientation, which would result in a higher coupling constant. Since four torsion values are theoretically in agreement with the measured 3J coupling constants, further information is required to determine unambiguously the orientation of the glycosidic torsions (see next section).

20.3.2 Determination of H–H Distances from ${}^1\text{H}$, ${}^1\text{H}$ Cross-relaxation Experiments

In addition to heteronuclear ${}^{13}\text{C}$ – ${}^1\text{H}$ 3J coupling constants, inter-residue NOEs are of particular importance in determining the orientation of the glycosidic torsions ϕ and ψ .

The nuclear Overhauser effect (NOE) is based on the fact that two magnetic nuclei with a spin experience each other's magnetic dipole moment through space, and magnetization can be transferred from one spin to the other (cross-relaxation). Consequently, if the spin of one interacting proton H_A is selectively disturbed by saturation or inversion with radiofrequency pulses, the NOE on the spin of the second proton H_B can be measured. The NOE enhancement depends on the overall correlation time of the molecule τ_c and is approximately proportional to r^{-6} , where r is the distance between the two protons H_A and H_B . Using a very basic approach, the isolated spin-pair approximation (ISPA) [76, 77], the distance between two protons can be calculated by using the simple relationship

$$\text{NOE}/\text{NOE}_{\text{ref}} = r^{-6}/r_{\text{ref}}^{-6} \quad (20.6)$$

where NOE_{ref} is the NOE intensity measured between two protons separated by a known distance r_{ref} . Usually distances between H atoms within a pyranose ring can serve as a good reference because it can be assumed that the ring is relatively rigid in most cases and therefore the H–H distance does not change significantly. Generally, NOEs in oligosaccharides are only detectable for H–H distances smaller than 4 Å and are qualitatively grouped into strong, medium, and weak NOEs.

Once the measured NOEs have been translated into H–H distances, these values can be used as distance restraints in a molecular dynamics simulation in order to generate a structure that is in agreement with all NOE data. This approach is used routinely to determine 3D structures of proteins by NMR spectroscopy, and works fairly well for molecules that exist in a single, well-defined conformation. However, in the case of flexible carbohydrates, the measured NOEs may result from a mixture of conformations each of which is in agreement with a subset of the NOEs only. In this case, if the weights of the NMR restraints are too high during the MD simulation the outcome of the modeling, that is very strongly biased towards generating a single average structure that satisfies all restraints, may be a “virtual conformation” that may not represent a real conformation.

An alternative approach that can be used to evaluate experimental and theoretical NOE data is to generate an ensemble of structures using theoretical methods such as MD simulation in explicit solvent without any restraints. From the generated data, NOE intensities and build-up curves can be calculated using the full relaxation matrix approach [78–83]. If the calculated NOEs and build-up curves are in good agreement with the experimental values, it can be assumed that the theoretical conformational ensemble is a good approximation of the real ensemble present in the NMR sample. Additionally, the internal flexibility of oligosaccharides can be estimated based on calculation of generalized order parameters S^2 that can be obtained from normalized rotation correlation plots calculated from MD data. The generalized order parameter of a pair of nuclei is a measure for the spatial restriction of the internal reorientational motion with correlation time τ and can be determined from NMR experiments using ^{15}N or ^{13}C relaxation rates [84].

In order to check which ϕ/ψ values would be in agreement with an experimentally observed NOE a systematic rigid rotation around the glycosidic torsions of a disaccharide fragment can be performed and for each ϕ/ψ orientation the theoretical interglycosidic NOE intensity of an H–H pair of interest can be calculated and those values that are in agreement with the experimental restraints are marked on the ϕ/ψ map [85]. The calculation of accurate NOE intensities using the relaxation matrix approach is a relatively complex task and normally it is sufficient to plot directly the calculated H–H distances as a function

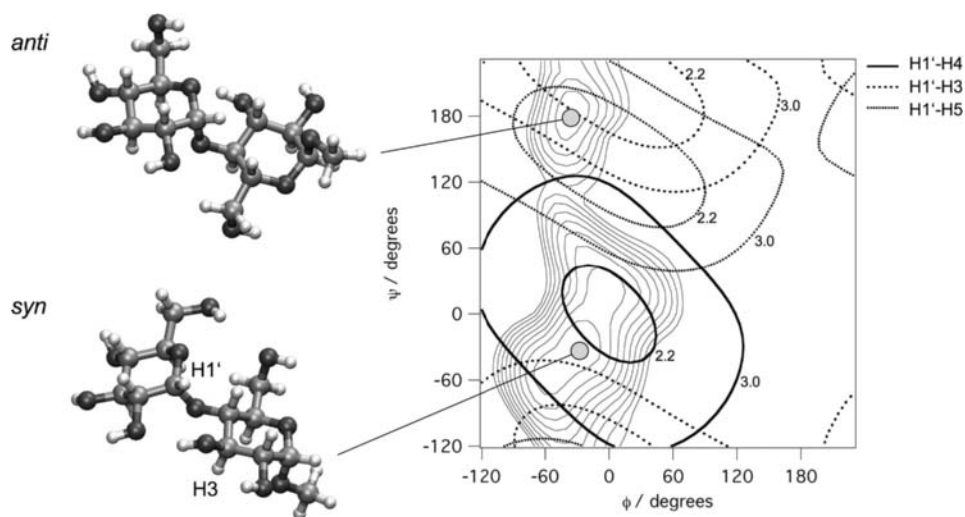


Figure 20.10 Distance mapping plot for maltose: distances between H1' and protons on the adjacent glucose residue are overlaid on an energy map of the glycosidic ϕ/ψ torsions. The lower and upper limits for the distance contours are 2.2 and 3.0 Å, respectively. In the background an adiabatic map calculated with TINKER/MM3 ($\epsilon = 4$) is displayed. Contour lines of the adiabatic map are plotted in steps of 1 kcal mol⁻¹ up to 10 kcal mol⁻¹ (outer contour). Two minimized conformations of maltose representing the *syn* and *anti* local minima are shown (gray circles). A short H1'–H4 distance (~ 2.3 Å, strong NOE) and a longer H1'–H3 distance (~ 3.1 Å, medium NOE) would be characteristic for the *syn* conformer. An NOE for H1'–H5 would give evidence for the *anti* conformer.

of the glycosidic torsions (Figure 20.10). This method has been named “distance mapping” [17, 86] and is available as a free service at www.glycosciences.de. Using a web interface, a disaccharide representing the glycosidic linkage can be built using SWEET II [87]. After selecting all the atom pairs for which an NOE has been detected experimentally, a rigid-residue systematic search is performed and the contours representing the experimental upper and lower distance limits for each atom pair are output as SVG graphics, overlaid on a ϕ/ψ energy map. Intersecting distance contours indicate ϕ/ψ regions where more than one NOE restraint would be satisfied and which therefore represent a likely conformation of the linkage. Conformational energy maps (see Chapter 19.3) are used as background information for the distance maps in order to exclude high-energy orientations. Some examples where distance mapping has been applied successfully can be found in the literature [24, 88, 89]. An example of “distance mapping” for maltose [α -D-Glcp-(1,4)-D-Glcp] is shown in Figure 20.10. The experimental relative NOE intensities for maltose are H1'–H3 = 1.0, H1'–H4 = 5.9, and H1'–H2' = 7.9 [75]. The distance between H1' and H2' is relatively fixed in a chair conformation of a pyranose ring (see Chapter 19.2) and is about 2.4 Å. The weaker NOEs for H1'–H4 and H1'–H3 indicate, respectively, that the H1'–H4 distance is slightly larger than 2.4 Å and that there are not many structures in the ensemble with short H1'–H3 distances of about 2.2–3.0 Å. This makes the existence of an *anti* ψ conformation very unlikely. This is also in agreement with the heteronuclear 3J coupling constants (see the previous section). It is therefore very likely that the solution conformation of maltose is dominated by conformations that are located in the so-called *syn*-minimum (ϕ/ψ is in the range -40° to -20°).

20.4 Conclusion and Future Prospects

The foremost experimental techniques to resolve the 3D structures of carbohydrates are NMR spectroscopy and X-ray crystallography. The Protein Data Bank (PDB) is the largest publicly available resource of such 3D structures. About 7% of its entries contain carbohydrate residues, most of which are covalently linked to glycoproteins or non-covalently bound in protein–carbohydrate complexes. From these data, information about preferred torsion angles or similar properties of the carbohydrates can be determined. This information forms the basis of several computational methods such as modeling studies, or can be used to validate calculated data. In NMR studies, the application of theoretical methods aids researchers in interpreting the primary experimental data. Unfortunately, a rather high rate of errors is present within the carbohydrate residues in the PDB. Recently software tools have become available to identify such errors automatically, which can help experimentalists to find problems before submission of their data to public databases and also users of the data to filter out erroneous structures. However, the standard software used by experimentalists to interpret the primary data is tailored mainly for protein NMR or X-ray crystallography. Carbohydrates are not very well supported at present. Therefore, a better implementation of routines dedicated to carbohydrate structures into the standard software used in X-ray crystallography would be an important step to improving the quality of carbohydrate 3D structural data in the PDB.

Abbreviations

3D	three-dimensional
DFT	density functional theory
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
MD	molecular dynamics
MMC	Metropolis Monte Carlo
PDB	Protein Data Bank

References

1. Imberty A, Perez S: Structure, conformation, and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations. *Chem Rev* 2000, **100**: 4567–4588.
2. Woods RJ: Three-dimensional structures of oligosaccharides. *Curr Opin Struct Biol* 1995, **5**:591–598.
3. Jimenez-Barbero J, Peters T (eds): *NMR Spectroscopy of Glycoconjugates*. Weinheim: Wiley-VCH Verlag GmbH; 2003.
4. Vliegthart JFG, Woods RJ (eds): *NMR Spectroscopy and Computer Modeling of Carbohydrates*, ACS Symposium Series, Vol. **930**. Washington, DC: American Chemical Society; 2006.
5. Widmalm G: General NMR spectroscopy of carbohydrates and conformational analysis in solution. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Oxford: Elsevier; 2007, pp. 101–132.
6. Perez S: Oligosaccharide and polysaccharide conformations by diffraction methods. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Oxford: Elsevier; 2007, pp. 193–219.

7. Buts L, Loris R, Wyns L: X-ray crystallography of lectins. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Oxford: Elsevier; 2007, pp. 221–249.
8. Qasba PK, Ramakrishnan B: X-ray crystal structures of glycosyltransferases. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Oxford: Elsevier; 2007, pp. 251–281.
9. Imberty A, Pérez S: Stereochemistry of the *N*-glycosylation sites in glycoproteins. *Protein Eng* 1995, **8**:699–709.
10. Petrescu AJ, Petrescu SM, Dwek RA, Wormald MR: A statistical analysis of *N*- and *O*-glycan linkage conformations from crystallographic data. *Glycobiology* 1999, **9**:343–352.
11. Wormald MR, Petrescu AJ, Pao YL, *et al.*: Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modelling. *Chem Rev* 2002, **102**:371–386.
12. Peters T, Pinto BM: Structure and dynamics of oligosaccharides: NMR and modeling studies. *Curr Opin Struct Biol* 1996, **6**:710–720.
13. Weimar T, Woods RJ: Combining NMR and simulation methods in oligosaccharide conformational analysis. In *NMR Spectroscopy of Glycoconjugates* (eds Jimenez-Barbero J, Peters T). Weinheim: Wiley-VCH Verlag GmbH; 2003, pp. 111–144.
14. Cumming DA, Carver JP: Virtual and solution conformations of oligosaccharides. *Biochemistry* 1987, **26**:6664–6676.
15. French AD, Brady JW (eds): *Computer Modeling of Carbohydrate Molecules*. Washington, DC: American Chemical Society; 1990.
16. Kozar T, von der Lieth C-W: Efficient modelling protocols for oligosaccharides: from vacuum to solvent. *Glycoconj J* 1997, **14**:925–933.
17. von der Lieth C-W, Kozar T, Hull WE: A (critical) survey of modelling protocols used to explore the conformational space of oligosaccharides. *THEOCHEM* 1997, **395**:225–244.
18. Woods RJ: Computational carbohydrate chemistry: what theoretical methods can tell us. *Glycoconj J* 1998, **15**:209–216.
19. Perez S: Molecular modeling in glycoscience. In *Comprehensive Glycoscience – from Chemistry to Systems Biology*, Vol. 2 (ed. Kamerling JP). Oxford: Elsevier; 2007, pp. 347–388.
20. French AD, Kelterer AM, Johnson GP, *et al.*: Constructing and evaluating energy surfaces of crystalline disaccharides. *J Mol Graphics Modell* 2000, **18**:95–107.
21. Imberty A, Mikros E, Koca J, *et al.*: Computer-simulation of histo-blood group oligosaccharides – energy maps of all constituting disaccharides and potential-energy surfaces of 14 Abh and Lewis carbohydrate antigens. *Glycoconj J* 1995, **12**:331–349.
22. Stortz CA: Disaccharide conformational maps: how adiabatic is an adiabatic map? *Carbohydr Res* 1999, **322**:77–86.
23. Frank M, Lütke T, von der Lieth C-W: GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res* 2007, **35**:287–290.
24. Siebert HC, Reuter G, Schauer R, *et al.*: Solution conformations of GM3 gangliosides containing different sialic acid residues as revealed by NOE-based distance mapping, molecular mechanics, and molecular dynamics calculations. *Biochemistry* 1992, **31**:6962–6971.
25. Weimar T, Peters T, Perez S, Imberty A: Combined NMR, grid search MM3 and Metropolis Monte Carlo GEGOP studies of two L-fucose containing disaccharides: alpha-L-Fuc-(1,4)-beta-D-GlcNAc-OME and alpha-L-Fuc-(1,6)-beta-D-GlcNAc-OME. *THEOCHEM* 1997, **395**:297–311.
26. Prestegard JH, Yi X: Structure and dynamics of carbohydrates using residual dipolar couplings. In *NMR Spectroscopy and Computer Modeling of Carbohydrates* (eds Vliegthart JFG, Woods RJ), ACS Symposium Series, Vol. **930**. Washington, DC: American Chemical Society; 2006, pp. 40–59.
27. Almond A: Biomolecular dynamics: testing microscopic predictions against macroscopic experiments. In *NMR Spectroscopy and Computer Modeling of Carbohydrates* (eds Vliegthart JFG,

- Woods RJ), ACS Symposium Series, Vol. **930**. Washington, DC: American Chemical Society; 2006, pp. 156–169.
28. Weiner SJ, Kollman PA, Case DA, *et al.*: A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* 1984, **106**:765–776.
 29. Allinger NL, Yuh YH, Lii JH: Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J Am Chem Soc* 1989, **111**:8551–8566.
 30. Rasmussen K: How to develop force fields: an account of the emergence of potential energy functions for saccharides. *THEOCHEM* 1997, **395**:91–106.
 31. Kirschner KN, Yongye AB, Tschampel SM, *et al.*: GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J Comput Chem* 2008, **29**:622–655.
 32. Crispin M, Stuart DI, Jones EY: Building meaningful models of glycoproteins. *Nat Struct Mol Biol* 2007, **14**:354; discussion 354–355.
 33. Lutteke T, Frank M, von der Lieth C-W: Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr Res* 2004, **339**:1015–1020.
 34. Perez S, Mulloy B: Prospects for glycoinformatics. *Curr Opin Struct Biol* 2005, **15**:517–524.
 35. Berman HM, Westbrook J, Feng Z, *et al.*: The Protein Data Bank. *Nucleic Acids Res* 2000, **28**:235–242.
 36. Berman H, Henrick K, Nakamura H, Markley JL: The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 2007, **35**:D301–D303.
 37. Petrescu AJ, Milac AL, Petrescu SM, *et al.*: Statistical analysis of the protein environment of *N*-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology* 2004, **14**:103–114.
 38. Lakshmanan T, Sriram D, Priya K, Loganathan D: On the structural significance of the linkage region constituents of *N*-glycoproteins: an X-ray crystallographic investigation using models and analogs. *Biochem Biophys Res Commun* 2003, **312**:405–413.
 39. Janin J, Wodak S: Conformation of amino acid side-chains in proteins. *J Mol Biol* 1978, **125**:357–386.
 40. Henrick K, Feng Z, Bluhm WF, *et al.*: Remediation of the protein data bank archive. *Nucleic Acids Res* 2008, **36**:D426–D433.
 41. Bohne-Lang A, Lang E, Forster T, von der Lieth C-W: LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res* 2001, **336**:1–11.
 42. Laskowski RA, MacArthur MW, Thornton JM: Validation of protein models derived from experiment. *Curr Opin Struct Biol* 1998, **8**:631–639.
 43. Lutteke T, von der Lieth C-W: pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics* 2004, **5**:69.
 44. Nakahara T, Hashimoto R, Nakagawa H, *et al.*: Glycoconjugate Data Bank: Structures—an annotated glycan structure database and *N*-glycan primary structure verification service. *Nucleic Acids Res* 2008, **36**:D368–D371.
 45. Hashimoto K, Goto S, Kawano S, *et al.*: KEGG as a glycome informatics resource. *Glycobiology* 2006, **16**:63R–70R.
 46. Lutteke T, Frank M, von der Lieth C-W: Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res* 2005, **33**:D242–246.
 47. Ramachandran GN, Ramakrishnan C, Sasisekharan V: Stereochemistry of polypeptide chain configurations. *J Mol Biol* 1963, **7**:95–99.
 48. Abola EE, Bairoch A, Barker WC, *et al.*: Quality control in databanks for molecular biology. *BioEssays* 2000, **22**:1024–1034.
 49. Raman R, Venkataraman G, Ernst S, *et al.*: Structural specificity of heparin binding in the fibroblast growth factor family of proteins. *Proc Natl Acad Sci USA* 2003, **100**:2357–2362.
 50. Roslund MU, Tahtinen P, Niemitz M, Sjöholm R: Complete assignments of the ^1H and ^{13}C chemical shifts and $J(\text{H,H})$ coupling constants in NMR spectra of D-glucopyranose and all D-glucopyranosyl-D-glucopyranosides. *Carbohydr Res* 2008, **343**:101–112.

51. Thibaudeau C, Stenutz R, Hertz B, *et al.*: Correlated C–C and C–O bond conformations in saccharide hydroxymethyl groups: parametrization and application of redundant ^1H – ^1H , ^{13}C – ^1H , and ^{13}C – ^{13}C NMR *J*-couplings. *J Am Chem Soc* 2004, **126**:15668–15685.
52. Kiddle GR, Homans SW: Residual dipolar couplings as new conformational restraints in isotopically C-13-enriched oligosaccharides. *FEBS Lett* 1998, **436**:128–130.
53. Tolman JR: Dipolar couplings as a probe of molecular dynamics and structure in solution. *Curr Opin Struct Biol* 2001, **11**:532–539.
54. Mobli M, Nilsson M, Almond A: The structural plasticity of heparan sulfate NA-domains and hence their role in mediating multivalent interactions is confirmed by high-accuracy ^{15}N -NMR relaxation studies. *Glycoconj J* 2008, **25**:401–414.
55. Lemieux RU, Bock K: The conformational analysis of oligosaccharides by ^1H -NMR and HSEA calculation. *Arch Biochem Biophys* 1983, **221**:125–134.
56. Zagrovic B, van Gunsteren WF: Comparing atomistic simulation data with the NMR experiment: how much can NOEs actually tell us? *Proteins: Struct Funct Bioinf* 2006, **63**:210–218.
57. Vanhalbeek H: NMR developments in structural studies of carbohydrates and their complexes. *Curr Opin Struct Biol* 1994, **4**:697–709.
58. Duus JO, Gottfredsen CH, Bock K: Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations. *Chem Rev* 2000, **100**:4589–4614.
59. Haasnoot CAG, de Leeuw FAAM, Altona C: The relationship between proton–proton NMR coupling constants and substituent electronegativities – I: an empirical generalization of the Karplus equation. *Tetrahedron* 1980, **36**:2783.
60. Tvaroska I, Gajdos J: Angular dependence of vicinal carbon–proton coupling-constants for conformational studies of the hydroxymethyl group in carbohydrates. *Carbohydr Res* 1995, **271**:151–162.
61. Cheetham NW, Dasgupta P, Ball GE: NMR and modelling studies of disaccharide conformation. *Carbohydr Res* 2003, **338**:955–962.
62. Rundlof T, Widmalm G: NMR analysis of the trisaccharide 2'-fucosyllactose by heteronuclear *trans*-glycosidic coupling constants and molecular simulations. *Magn Reson Chem* 2001, **39**:381–385.
63. Gonzalez-Outeirino J, Kadirvelraj R, Woods RJ: Structural elucidation of type III group B *Streptococcus* capsular polysaccharide using molecular dynamics simulations: the role of sialic acid. *Carbohydr Res* 2005, **340**:1007–1018.
64. Karplus M: Contact electron-spin coupling of nuclear magnetic moments. *J Chem Phys* 1959, **30**:11–15.
65. Tvaroska I, Hricovini M, Petrakova E: An attempt to derive a new Karplus-type equation of vicinal proton carbon coupling-constants for C–O–C–H segments of bonded atoms. *Carbohydr Res* 1989, **189**:359–362.
66. Bose B, Zhao S, Stenutz R, *et al.*: Three-bond C–O–C–C spin-coupling constants in carbohydrates: development of a Karplus relationship. *J Am Chem Soc* 1998, **120**:11158–11173.
67. Cloran F, Carmichael I, Serianni AS: Density functional calculations on disaccharide mimics: studies of molecular geometries and *trans*-O-glycosidic $^3J_{\text{COCH}}$ and $^3J_{\text{COCC}}$ spin-couplings. *J Am Chem Soc* 1999, **121**:9843–9851.
68. Fraser R, Kaufman M, Morand P, Govil G: Stereochemical dependence of vicinal H–C–O–H coupling constants. *Can J Chem* 1969, **47**:403–409.
69. Dabrowski J, Poppe L: Hydroxyl and amido groups as long-range sensors in conformational-analysis by nuclear Overhauser enhancement - a source of experimental-evidence for conformational flexibility of oligosaccharides. *J Am Chem Soc* 1989, **111**:1510–1511.
70. Poppe L, van Halbeek H: NMR spectroscopy of hydroxyl protons in supercooled carbohydrates. *Nat Struct Biol* 1994, **1**:215–216.

71. Stenutz R, Carmichael I, Widmalm G, Serianni AS: Hydroxymethyl group conformation in saccharides: structural dependencies of $^2J(\text{HH})$, $^3J(\text{HH})$, and $^1J(\text{CH})$ spin–pin coupling constants. *J Org Chem* 2002, **67**:949–958.
72. Brisson JR, Uhrinova S, Woods RJ, *et al.*: NMR and molecular dynamics studies of the conformational epitope of the type III group B *Streptococcus* capsular polysaccharide and derivatives. *Biochemistry* 1997, **36**:3278–3292.
73. Corzana F, Motawia MS, Du Penhoat CH, *et al.*: A hydration study of (1–4) and (1–6) linked alpha-glucans by comparative 10 ns molecular dynamics simulations and 500-MHz NMR. *J Comput Chem* 2004, **25**:573–586.
74. Landersjo C, Jansson JLM, Maliniak A, Widmalm G: Conformational analysis of a tetrasaccharide based on NMR spectroscopy and molecular dynamics simulations. *J Phys Chem B* 2005, **109**:17320–17326.
75. Shashkov AS, Lipkind GM, Kachetkov NK: Nuclear overhauser effects for methyl β -maltoside and the conformational states of maltose in aqueous solution. *Carbohydr Res* 1986, **147**:175–182.
76. Homans SW, Dwek RA, Rademacher TW: Tertiary structure in *N*-linked oligosaccharides. *Biochemistry* 1987, **26**:6553–6560.
77. Woods RJ, Pathiaseril A, Wormald MR, *et al.*: The high degree of internal flexibility observed for an oligomannose oligosaccharide does not alter the overall topology of the molecule. *Eur J Biochem* 1998, **258**:372–386.
78. Leeflang BR, Kroonbatenburg LMJ: Crosrel – full relaxation matrix analysis for NOESY and ROESY NMR-Spectroscopy. *J Biomol NMR* 1992, **2**:495–518.
79. Widmalm G, Byrd RA, Egan W: A conformational study of α -L-Rhap-(1 \rightarrow 2)- α -L-Rhap-(1 \rightarrow OMe) by NMR nuclear Overhauser effect spectroscopy (NOESY) and molecular-dynamics calculations. *Carbohydr Res* 1992, **229**:195–211.
80. Kroonbatenburg LMJ, Kroon J, Leeflang BR, Vliegenthart JFG: Conformational analysis of methyl beta-cellobioside by ROESY NMR spectroscopy and MD simulations in combination with the Crosrel method. *Carbohydr Res* 1993, **245**:21–42.
81. Weimar T, Meyer B, Peters T: Conformational analysis of α -D-Fuc-(1 \rightarrow 4)- β -D-GlcNAc-OMe – one-dimensional transient NOE experiments and Metropolis Monte-Carlo simulations. *J Biomol NMR* 1993, **3**:399–414.
82. Peters T, Weimar T: Assessing glycosidic linkage flexibility – conformational-analysis of the repeating trisaccharide unit of *Aeromonas salmonicida*. *J Biomol NMR* 1994, **4**:97–116.
83. Asensio JL, Jimenez-Barbero J: The use of the AMBER force field in conformational analysis of carbohydrate molecules: determination of the solution conformation of methyl alpha-lactoside by NMR spectroscopy, assisted by molecular mechanics and dynamics calculations. *Biopolymers* 1995, **35**:55–73.
84. Lommerse JPM, Kroonbatenburg LMJ, Kroon J, *et al.*: Conformations and internal mobility of a glycopeptide derived from bromelain using molecular-dynamics simulations and NOESY analysis. *J Biomol NMR* 1995, **6**:79–94.
85. Brisson JR, Carver JP: Solution conformation of α -D-(1–3)- and A-D-(1–6)-linked oligomannosides using proton nuclear magnetic resonance. *Biochemistry* 1983, **22**:1362–1368.
86. Poppe L, von der Lieth C-W, Dabrowski J: Conformation of the glycolipid globoside head group in various solvents and in the micelle-bound state. *J Am Chem Soc* 1990, **112**:7762–7771.
87. Bohne A, Lang E, von der Lieth C-W: SWEET – WWW-based rapid 3D construction of oligo- and polysaccharides. *Bioinformatics* 1999, **15**:767–768.
88. Gilleron M, Siebert HC, Kaltner H, *et al.*: Conformer selection and differential restriction of ligand mobility by a plant lectin. Conformational behaviour of Gal β 1–3GlcNAc β 1-R, Gal β 1–3GalNAc β 1-R and Gal β 1–2Gal β 1-R' in the free state and complexed with galactoside-specific mistletoe lectin as revealed by random-walk and conformational-clustering molecular-mechanics calculations, molecular-dynamics simulations and nuclear Overhauser experiments. *Eur J Biochem* 1998, **252**:416–427.

89. Kozar T, Nifant'ev NE, Grosskurth H, *et al.*: Conformational changes due to vicinal glycosylation: the branched α -L-Rhap(1-2)[β -D-Galp(1-3)]- β -D-Glc1-OMe trisaccharide compared with its parent disaccharides. *Biopolymers* 1998, **46**:417-432.
90. Varghese JN, Colman PM, van Donkelaar A, *et al.*: Structural evidence for a second sialic acid binding site in avian influenza virus neuraminidases. *Proc Natl Acad Sci USA* 1997, **94**:11808-11812.
91. Ramakrishnan B, Ramasamy V, Qasba PK: Structural snapshots of β -1,4-galactosyltransferase-I along the kinetic pathway. *J Mol Biol* 2006, **357**:1619-1633.
92. Ramasamy V, Ramakrishnan B, Boeggeman E, *et al.*: Oligosaccharide preferences of β 1,4-galactosyltransferase-I: crystal structures of Met340His mutant of human β 1,4-galactosyltransferase-I with a pentasaccharide and trisaccharides of the *N*-glycan moiety. *J Mol Biol* 2005, **353**:53-67.

**Section 6:
Protein–Carbohydrate
Interaction**

21 Structural Features of Lectins and Their Binding Sites

Remy Loris

Structural Biology Brussels, Vrije Universiteit Brussel and Department of Molecular and Cellular Interactions, VIB, Pleinlaan 2, B-1050 Brussels, Belgium

21.1 Introduction

Complex glycans, displayed on proteins or on cell surfaces, are often used as an identity tag. They may signal the type of cell, the location of a protein in (or out of) the cell, or if a protein should be taken out of circulation. They are also useful recognition tags for pathogenic microorganisms that intend to invade tissues or cells of a potential host.

The molecules that read this glycode are called lectins. Because there are so many different situations in which deciphering of a glyco-message is necessary, there is an equally large number of different lectin families. Crystal structures of members of the different animal and plant lectin families have revealed a wide variety of lectin folds and carbohydrate binding site architectures. Despite this large variability, a number of interesting cases of both convergent and divergent evolution among plant, animal, and bacterial lectins are noted. These similarities exist at the level of the protein fold, the architecture of the binding site, and quaternary association, and may be derived from similar functional needs. In this chapter, the most important families of lectins will be visited and common principles for carbohydrate recognition will be discussed.

21.2 The Lectin Family Album

21.2.1 *Lectins Targeting Glycoproteins to Specific Subcellular Locations in Eukaryotes*

Lectins play a major role in the cellular physiology of eukaryotes. They are crucial for the quality control of folding and correct intracellular trafficking of glycoproteins [1]. Calnexin, an integral membrane protein, and its soluble relative calreticulin, bind to a monoglucosylated high-mannose intermediate of *N*-linked glycans. They act as molecular chaperones that assist in the productive folding of newly synthesized *N*-linked glycoproteins

in the endoplasmatic reticulum. ERGIC53 mediates transport of glycoproteins from the rough endoplasmatic reticulum to the Golgi apparatus and VIP36 does a similar job cycling between the Golgi apparatus and the plasma membrane. The cation-dependent mannose-6-phosphate receptor delivers newly synthesized soluble acid hydrolases that carry mannose-6-phosphate on their *N*-linked glycans to the lysosomes.

21.2.2 *Plant-specific Lectins*

It is often believed that the very first lectins ever discovered were lectins from plants. Indeed, many plant species are valuable sources of lectins, which are often used as tools in glycoresearch. The best known lectin family from plants is the legume lectin family [2]. These lectins are usually, but not exclusively, found in the seeds of leguminous plants. Their carbohydrate specificities are the most variable among all lectin families and cover Man/Glc, Gal, GalNAc, GlcNAc, Fuc, and complex specificities.

Although by far the largest and best studied group of plant lectins, the legume lectins were not the first plant lectins to be discovered. That honor goes to ricin, a highly toxic protein from the seeds of castor beans. In 1888, Stillmark discovered that extracts from castor bean seeds were capable of agglutinating specific types of erythrocytes, and that this agglutination was inhibited by galactose, but not by other monosaccharides [3]. The active substance was thus found to be highly specific and termed an agglutinin or lectin (from the Latin *legere* – to choose). Ricin is a ribosome-inactivating protein (RIP) consisting of a lectin subunit, usually galactose specific, covalently coupled to a toxin subunit.

Several other families of plant lectins exist. Lectins consisting of repeats of hevein domains (which bind GlcNAc oligomers) were first discovered in grasses, but are found in a large variety of plants such as the stinging nettle, pokeweed, the rubber tree, potato, and tomato. Hevein lectin domains are also commonly found as N-terminal domains of plant chitinases. The bulbs and sometimes vegetative parts of monocotyledonous plants contain yet another type of closely related lectins, all mannose specific [4, 5]. Finally, a smaller family of lectins are those related to jacalin, which contains both galactose- and mannose-specific members [6, 7].

The major function of lectins from higher plants seems to be defense against pathogens and predators, at least for those lectins that are present in sufficiently large amounts. Specific seed lectins were found to inhibit the growth of phytopathogenic fungi and to be toxic when fed to insects [8, 9]. For the lectins that are sometimes present in low concentrations in vegetative parts of the plants, the function is less clear. In one specific case, DB46 from the leguminous plant *Dolichos biflorus*, a role in the symbiosis with the nitrogen-fixating *Rhizobium* bacteria has been established [10].

Fungal lectins are distinct from the lectins of higher plants. Different families exist, usually unrelated to those of higher plants. Although some of them may be involved in defense against pathogens and insects, or in storage of nutrients, a major role seems to involve host association in the contexts of parasitism and symbiosis. Fungal lectins are understudied with respect to those of higher plants, animals, and bacteria. To date, the crystal structures of four fungal lectins have been determined: one member of the galectin family [11] and three showing unique folds [12–14].

21.2.3 *Animal-specific Lectins*

In contrast to plants, which usually contain one or two lectins that are highly abundant in storage organs such as seeds, bulbs, or rhizomes, animals contain a much larger variety of lectins that carry out many different functions. Although not well known, the first published lectin hemagglutinating activity by animal proteins dates from 1886 and lectin activity from rattlesnake venom may have been observed as early as 1860 [15].

There are two large families of animal lectins that are found in both vertebrates and invertebrates: the C-type lectins [16] and the galectins [17, 18].

The C-type lectins constitute by far the largest family of animal lectins. They are found both as part of membrane proteins and in soluble forms. They play a major role in the innate immune system of both vertebrates and invertebrates and the adaptive immune system of vertebrates [19, 20]. For example, the mannose receptor is involved in the uptake of pathogens. Selectins mediate cell–cell adhesion and migration [21].

Galectins form a highly conserved family of soluble galactose-binding lectins. The widespread occurrence of multiple members in most vertebrates suggests that they are involved in a variety of crucial cellular functions. Indeed, roles in cell–cell adhesion, cell migration, cell proliferation, chemotaxis, apoptosis, and neurite elongation have been demonstrated. Together with the C-type lectins they have a role in complement activation, and function in both the innate and adaptive immune system [19].

In both vertebrates and invertebrates, many other lectin families have been discovered [22]. Invertebrate lectins are usually involved in innate immunity. Binding of these lectins to foreign carbohydrate-bearing substances leads to agglutination, endocytosis by phagocytic cells, and the activation of protease cascades, ultimately resulting in clotting, melanization, or killing of the pathogen by the complement system. Such lectins with known structures include the tachylectins from the horseshoe crab [23, 24], a fucoslectin found in the serum of the European eel [25], and CEL-III from the sea cucumber [26].

In vertebrates, one can discern at least seven additional lectin families, and some may still remain to be discovered. The siglecs or I-type lectins are sialic acid-binding lectins that are expressed in the hematopoietic and immune systems [27]. One group of siglecs seems to be involved in the regulation of leukocyte function. In addition to its eight tandem C-type lectin domain repeats, the mannose receptor on macrophage and epithelial cells contains an additional cysteine-rich lectin domain (Cys-MR) that recognizes sulfated carbohydrates [28]. Fibroblast growth factors (FGFs) contain heparin-binding domains that modulate FGF dimerization and activation. The hepatocyte growth factor/scatter factor (HGF/SF) contains a heparin-binding domain and is involved in the development of placenta and liver [29, 30]. The pentraxins are oligomeric plasma proteins conserved through vertebrate evolution that participate in inflammation and host defense, although additional functions are likely [31]. The spermadhesin family includes lectins that mediate sperm-egg binding [32]. Ym1 is a lectin that is secreted by activated peritoneal macrophages from mouse [33]. This lectin with a TIM barrel fold is homologous to certain chitinases, but its function is as yet unknown. Similarly, the normal function of human cartilage glycoprotein-39 (HCGP-39), a lectin that might play a role in rheumatoid arthritis and belonging to the same family as Ym1, is also unknown [34, 35].

21.2.4 Lectins from Prokaryotes and Viruses

Prokaryotes lack the complex carbohydrate chemistry found in eukaryotes. Carbohydrate chemistry in prokaryotes is largely confined to (i) basic metabolism, (ii) cell wall synthesis and (iii) the synthesis of osmolytes. Nevertheless, several kinds of lectins are produced by prokaryotic species. Their role is not to recognize glycans produced by their own species, but to recognize those from other organisms while interacting in a parasite–host or symbiont–host relationship.

The best studied bacterial lectins are those associated with fimbriae [36, 37]. Usually they are found at the tip of fimbriae, but sometimes the major fimbrial subunit itself displays carbohydrate binding activity. Alternatively, integral membrane adhesins such as OpcA from *Neisseria meningitidis* [38] and OmpA from *Escherichia coli* [39] are used for attachment prior to invasion. Other bacterial lectins such as LecB from *Pseudomonas aeruginosa* play a role in the formation of biofilms [40]. In addition, lectin domains are encountered as part of several bacterial proteins such as pertussis toxin from *Bordetella pertussis* [41] and non-catalytic carbohydrate binding modules of polysaccharide-degrading enzymes.

As in prokaryotes, viruses display lectins to permit attachment to host cells. These may simply be the major capsid proteins as in foot-and-mouth disease virus [42] or mouse polyomavirus [43], dedicated proteins that stick out of the surface of the viral capsid as in *Rheoviridae* [44] and adenoviruses [45], membrane glycoproteins present in the envelope as in influenza virus [46, 47], or tailspike proteins on the attachment apparatus of bacteriophages such as phage P22 [48].

21.3 A Pantheon of Lectin Folds

21.3.1 Folds in Common Between Plant and Animal Lectins

With so many different families of lectins involved in distinct functions, it should come as no surprise that there is also a plethora of lectin folds. One common denominator is nevertheless omnipresent: the overwhelming presence of β -sheet. Most common are the β -sandwich type of folds such as the legume lectin fold and related folds [49], the CUB fold of the spermadhesins [50], the immunoglobulin folds, and the folds of calnexin [51] and of certain bacterial lectins such as LecA [52] and LecB [53, 54] from *P. aeruginosa* (Figure 21.1).

It should be noted that the frequently used names “jelly roll” and “ β -barrel” folds are not appropriate in this context. The jelly roll topology forms a distinct subset of the possible topologies of β -sandwich architectures. True jelly roll topologies have been identified in certain viral coat proteins, but never in lectins. A β -barrel implies a closed β -sheet structure and not two sheets sandwiched upon each other.

A second class of all- β -folds commonly found in lectins are those that contain internal symmetry (Figure 21.2). Most frequently observed are three-fold internal symmetries such as in the β -trefoil domains and the type I and II β -prisms [6, 55, 56]. Other folds that can be included in this class are the β -propeller folds. Tachylectin-2, a lectin involved in the innate immune system of the Japanese horseshoe crab, consists of a five-bladed β -propeller [23], while the lectin from the mushroom *Aleuria aurantia* forms a six-bladed β -propeller [14].

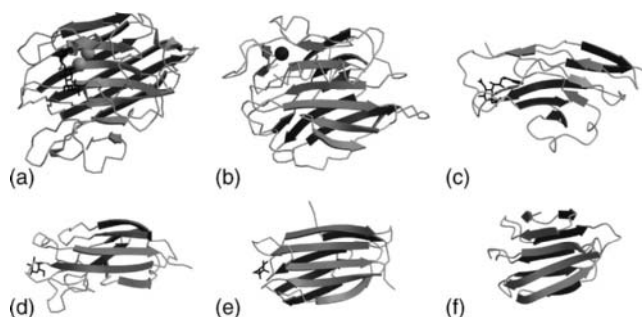


Figure 21.1 Lectins with β -sandwich folds. The main architecture is two (usually entirely antiparallel) β -sheets with variable topology sandwiched upon each other. (a) Concanavalin A; (b) calnexin; (c) siglec-1; (d) LecA; (e) LecB; (f) the spermadhesin PSP-1.

A third but smaller group of lectins adopt mixed α/β -folds (Figure 21.3). The best known such family are the C-type lectins [57], but α/β -folds are also found in heparan sulfate-binding chemokines [58] and the heparin-binding NK1 fragment of hepatocyte growth factor [59]. Ym1, a secretory lectin synthesized by activated murine peritoneal macrophages [33], and HCGP39, the 39 kDa glycoprotein from human articular chondrocytes [34, 35], adopt the TIM barrel fold consisting of a parallel eight-stranded β -barrel flanked by eight

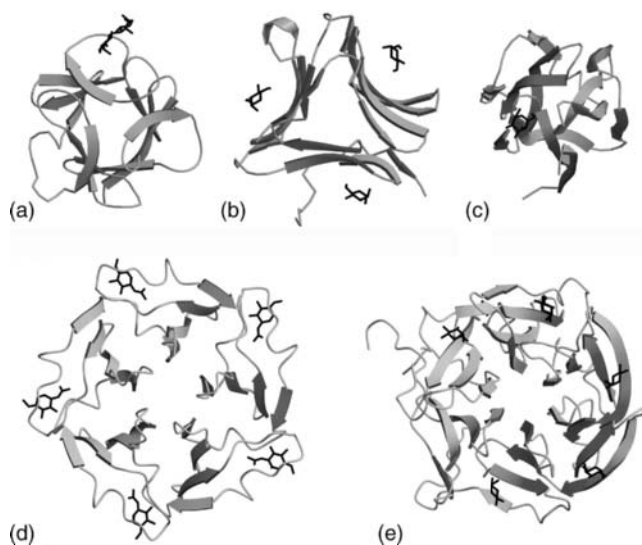


Figure 21.2 All- β lectin folds with internal symmetry viewed down their molecular pseudosymmetry axes. (a) Ricin; (b) snowdrop lectin; (c) jacalin; (d) tachylectin-2; (e) *Aleuria aurantia* lectin. They can be divided in the three-fold symmetric β -prism folds (ricin, jacalin, and snowdrop lectin) and the β -propellers (tachylectin-2 and *A. aurantia* lectin). Ricin and jacalin do not exploit the three-fold internal pseudosymmetry to generate multivalency and have only one carbohydrate binding site on each domain. Tachylectin-2 fully exploits its five-fold internal pseudosymmetry. The *A. aurantia* lectin, on the other hand, has only five active binding sites despite its six-fold internal pseudosymmetry.

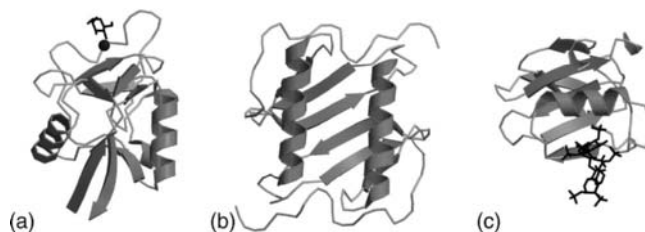


Figure 21.3 Lectins with α/β -folds. (a) C-type lectin domain of the rat mannose-binding protein; (b) IL-8, a chemokine; (c) NK1 fragment of HGF/SF.

α -helices. Interestingly, to date no single lectin has been described with an all- α -fold. The reason for this is unclear as there is no reason why all- α would be less suitable for carbohydrate recognition. In fact, the architecture of many carbohydrate processing enzymes such as lysozyme consists mainly of α -helices. Indeed, one can easily imagine an all- α carbohydrate-active enzyme to evolve into a lectin by loss of one or more catalytic residues followed by fine-tuning of its carbohydrate specificity by further mutations (as has apparently happened for HCGP39, which shows clear homology to certain chitinases).

21.3.2 Divergent and Convergent Evolution

Although there are many different lectin folds, certain folds are used by different families of lectins [60]. When in the early 1990s the first crystal structures of members of the galectin and pentraxin family became available, a striking resemblance with the legume lectin fold was immediately apparent (Figure 21.4). A similar fold has since then also been observed for the rhesus rotavirus VP4 sialic acid-binding domain [44]. These resemblances are most likely the consequence of convergent evolution, the reinvention of the same fold at different times during evolution. Not only is there no detectable sequence identity between the three families of proteins, their folds are circularly permuted, putting the N- and C-termini at different relative positions. Furthermore, their carbohydrate binding sites are distinctly located and show no topological relationships. ERGIC-53, a mannose-specific lectin involved in glycoprotein export from the endoplasmic reticulum, on the other hand, seems to be a genuine member of the legume lectin family despite a sequence identity of less than 20% (Figure 21.4). Not only is its topology identical with that of the legume lectins, but also the carbohydrate binding site shows all the key characteristics typical for the legume lectin family: a bound calcium ion stabilizes a non-proline *cis*-peptide bond, thus positioning two crucial aspartate and asparagine side-chains [61]. In addition, a loop

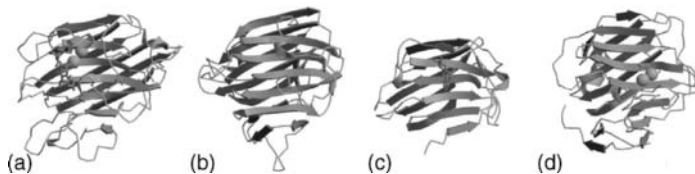


Figure 21.4 Legume lectin folds in plant and animal lectins. (a) Concanavalin A; (b) ERGIC-53; (c) human galectin-7; (d) serum amyloid protein.

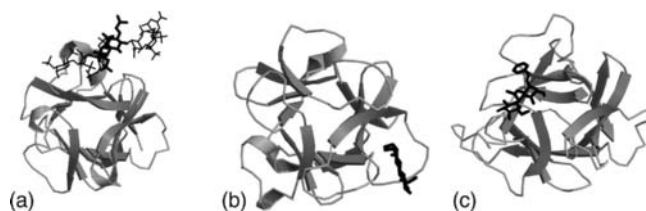


Figure 21.5 β -Trefoil folds in plant and animal lectins. (a) Fibroblast growth factor-2; (b) ricin; (c) *Amaranthus* lectin.

that in the legume lectin family is thought to be a key determinant of monosaccharide specificity is seen in ERGIC-53 with a conformation identical with what is seen only for the Man-specific legume lectins.

Convergent evolution has been observed for several other animal and plant lectin families. Several lectins exhibit the β -trefoil fold, a fairly common fold, that was first identified in soybean trypsin inhibitor. Examples of this fold have been discovered twice in plant lectins [55, 62], for ricin and amarantin, and also three times in animal lectins [26, 63, 64], for the fibroblast growth factors (FGF), for the cysteine-rich domain of the mannose receptor (cys-MR), and for the hemolytic lectin CEL-III from sea cucumber (Figure 21.5). Although the fold is three-fold symmetric, only a single carbohydrate binding site is observed on each β -trefoil domain. With the possible exception of CEL-III (of which the carbohydrate binding site is not known), the different locations of the carbohydrate binding sites indicates that the plant and animal lectins are probably not evolutionarily related to each other.

A third example is the hevein fold, named after a small carbohydrate binding protein found in the latex of rubber trees [65]. Hevein domains are small (30–45 amino acids) carbohydrate binding domains found in a variety of lectins or as N-terminal domains of chitinases. This fold is also found in certain snake toxins [66] such as the heparin-binding cobra venom cardiotoxin [67].

Finally, the immunoglobulin fold, a very common fold that is predominant in recognition contexts, is observed in different unrelated lectin families (Figure 21.6). These are the I-type

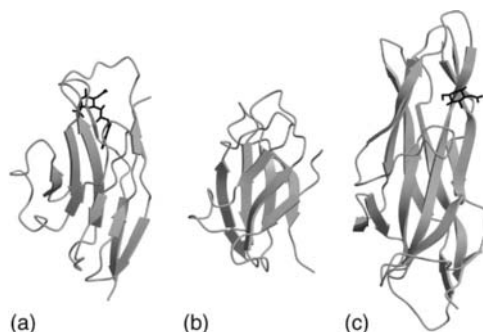


Figure 21.6 Lectins with an immunoglobulin fold. (a) The N-terminal domain of siglec-1 in complex with a benzylated sialic acid; (b) the carbohydrate binding domain of Fve, an immunomodulatory lectin from the fruiting body of the mushroom *Flammulina velutipes*; (c) the F17G lectin domain from the adhesin at the tip of the F17 pili in complex with GlcNAc.

lectins or siglecs involved in cell–cell interactions and signaling functions in the immune, nervous, and hematopoietic systems, the fungal immunomodulatory lectin Fve [13], and also all bacterial adhesins located on the tips of various pili. The crystal structures of three such adhesins have been determined [68–70]. The structures of the three lectin domains from these adhesins are nevertheless structurally fairly diverse and their carbohydrate binding sites are at different locations despite their common context and mode of assembly into the pilus.

21.4 Structural Basis of Carbohydrate Recognition

If we want to look for common recognition principles in a family of proteins as diverse as lectins, we should look at the ligands and ask what is special about them. It turns out that specifically recognizing a carbohydrate is far more challenging than recognizing a protein or nucleic acid. Carbohydrates are dominated by a single functional group: hydroxyl. Discriminating between different carbohydrates is therefore a difficult job.

A good example is UEA-II, a lectin from the seeds of *Ulex europaeus* that belongs to the chitobiose-specific [GlcNAc(β 1–4)GlcNAc] group of legume lectins [71]. Interestingly, the lectin binds with a similar affinity the seemingly unrelated trisaccharide fucosyllactose using a set of equivalent hydrogen bonds and van der Waals interactions (Figure 21.7).

21.4.1 Carbohydrate Binding Sites

Generally, lectins bind their ligands with low affinity. A typical association constant for a lectin–monosaccharide interaction resides in the millimolar range. Longer oligosaccharides are sometimes bound with slightly higher affinity, but it is rare to see an association constant reaching the micromolar range. This contrasts sharply with other typical molecular associations that are relevant in biological systems. For example, antigen–antibody interactions typically occur in the subnanomolar range and protein–DNA interactions are often still

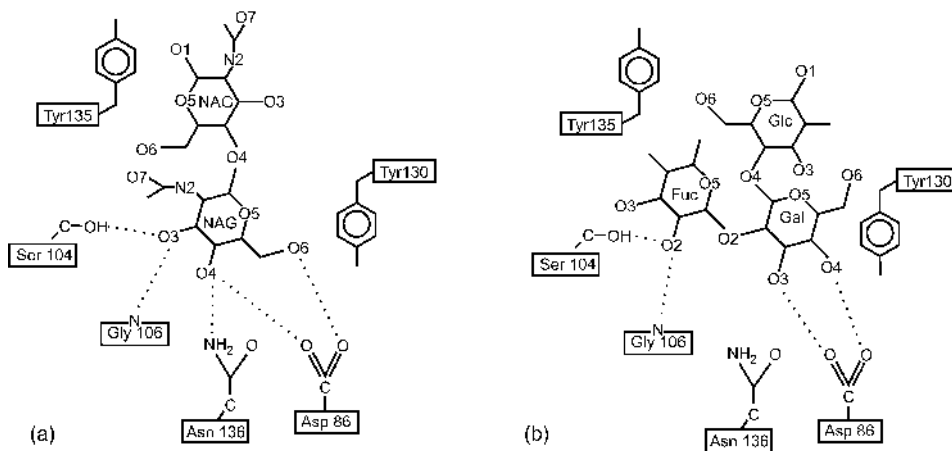


Figure 21.7 Recognition of (a) chitobiose and (b) fucosyllactose by UEA-II. The same residues make a set of equivalent hydrogen bonds and van der Waals interactions with the two saccharides.

tighter. In fact, for many biologically relevant recognition events, millimolar affinities are considered aspecific.

In most lectins, the carbohydrate binding site is dominated by a monosaccharide binding site that fits one or two types of monosaccharide. Hence lectins have historically been classified based on their monosaccharide specificity: galactose, mannose/glucose, fucose, and so on. The large majority of lectins that have been characterized recognize mannose or galactose. Fucose- and sialic acid-specific lectins are rarer.

The monosaccharide binding site usually interacts with the carbohydrate ligand via three or four key hydrogen bonds (with amino acid residues and/or with bound calcium ions), determining the relative orientations of two or three of the sugar hydroxyl groups, and via hydrophobic interactions with one or more aromatic side-chains that often are oriented roughly parallel to the sugar ring.

Most lectin families specialize in one type of carbohydrate. For example, all galectins bind to terminal galactose residues, all hevein domains bind oligomers of GlcNAc, all monocot bulb lectins are mannose-specific, and all siglecs require sialic acids. There are, however, families for which the carbohydrate specificities of their members vary. The legume lectins and the C-type lectins both have a large repertoire of carbohydrate ligands. In both cases, this is established by a core of conserved residues that provide several key hydrogen bonds and hydrophobic contacts to the sugar residue in the primary binding site. This conserved core is flanked by two variable loops that determine both monosaccharide and subsite specificity. In the case of the legume lectins, the large number of available crystal structures suggests that the variable loops adopt canonical loop structures [72] similar to what has been observed for antibodies (Figure 21.8). Although evolutionarily unrelated, the legume lectin family and the C-type lectin family seem to have adapted in a similar way to a similar functional challenge [60].

The distinction between lectins and non-lectin family members is not always very clear. For example, L-ficolin, a plasma protein capable of activating the complement system, was classified as a lectin because it binds *N*-acetylglucosamine and *N*-acetylgalactosamine. Recent results, however, indicate that only the *N*-acetyl group is recognized, inside or outside a carbohydrate context, and that the protein is thus a pattern recognition molecule specific for acetyl groups. [73]. Having said this, it can be noted that many – if not most – large lectin families contain members that have acquired different recognition functions that do not involve carbohydrate recognition. The monocot mannose-binding lectins contain apparent non-lectin members such as the sweet-tasting protein curculin [74], and the Charcot–Leyden crystal protein belongs to the galectin family [75]. The legume lectin family contains a series of α -amylase as well as the arcelins, defense proteins for which the (non-carbohydrate) target is unknown [76]. In both cases, carbohydrate binding is structurally impaired [77, 78]. Nevertheless, for the α -amylase inhibitors, the recognition site is at the same position as the carbohydrate binding site of the lectin members of the family, although its typical architecture is not retained [78]. Some C-type lectin-like antifreeze proteins are even more surprising as in this case the key features of the carbohydrate binding sites are left intact, but recruited for their novel function [79].

21.4.2 Quaternary Structure and Avidity

All the structural data that have been discussed until now do not solve a fundamental question: how is a biologically relevant activity possible by proteins that are characterized by such weak affinities and selectivities for their ligands? The answer lies in numbers. Most

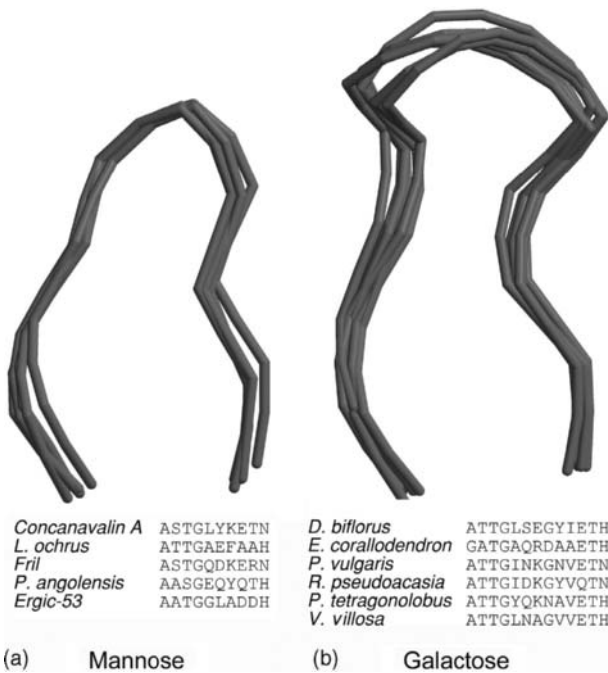


Figure 21.8 Canonical loop structures that determine carbohydrate specificity in the legume lectins. (a) Superposition of the “specificity-determining loops” of a series of mannose-specific legume lectins including ERGIC-53; (b) a series of similar loops, but from galactose-specific legume lectin members. The amino acid sequences of the loops are shown below each superposition.

lectins are either multimeric proteins with two or more carbohydrate binding subunits, or they are multidomain proteins with multiple binding sites. Thus, multivalence is a fundamental property of lectins that is used to increase avidity and specificity. By having multiple binding sites for the same low-affinity ligand, lectins will achieve an apparent higher affinity or avidity. This can best be compared with the way in which Velcro gains its strength.

Perhaps even more important is that multiple binding sites can also result in an additional layer of specificity. When the spacings of the low-affinity ligands agree with the spacings (and relative orientations) of the multivalent lectin, a high avidity is observed. When the same low-affinity ligands are spaced in a different way to the lectin binding sites, the avidity effect will not occur.

This property leads to an interesting phenomenon, the formation of homogeneous crosslinked lattices of a crystalline nature [80], and in some cases leads to macroscopic crystals when multivalent lectins are used to precipitate glycoproteins and branched multivalent oligosaccharides. Originally discovered with the legume lectins, this phenomenon appears widespread and examples of crystal structures where the lattice is determined by the crosslinking carbohydrate have been determined for lectins belonging to several families [81–83].

The diversity by which multivalency is generated is as large as the different types of lectin folds, but again some interesting examples of convergent evolution are noted. One way to generate a multivalent lectin is by domain duplication. The oldest known examples

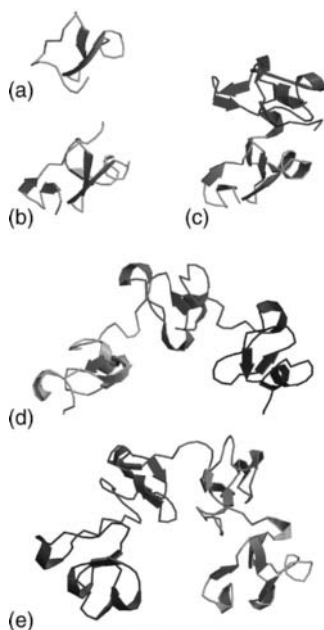


Figure 21.9 Generating multivalency by domain duplication as exemplified by the hevein family. (a) The antimicrobial peptide AcAmp constitutes three-quarters of a hevein domain; (b) single domain hevein from the rubber tree; (c) double domain lectin from stinging nettle; (d) the pokeweed lectin C contains three hevein domains; (e) the wheat germ agglutinin monomer consists of four consecutive hevein domains. In each protein, the relative orientations of the domains is unique.

are wheat germ agglutinin (WGA) and related hevein-domain lectins (Figure 21.9). WGA consists of four consecutive hevein domains [84]. Similarly, galectins 6, 4, 8 and 9 consist of a tandem repeat of two homologous carbohydrate binding domains. Also, C-type lectin domains are often found in multiple copies on a single polypeptide.

Multivalency can also be generated by oligomerization. Lectin oligomers can be generated either by the oligomerization potential of the lectin domains themselves or by additional domains that mediate oligomerization. The former strategy is adopted by most plant lectins, several galectins, the spermadhesins, LecA and LecB from *P. aeruginosa*, and the cation-dependent mannose-6-phosphate receptor. The latter strategy is observed for some C-type lectins, galectin-3, the fungal lectin Fve, and *Amaranthus* lectin.

Also, by varying their mode of oligomerization, it is possible to change the specificity by selecting for multivalent carbohydrate ligands with different relative positions of their monovalent epitopes. This is indeed what is observed in several lectin families [49, 60], the most variable being the legume lectins. Nine different quaternary associations (one monomer, four types of dimers, and four types of tetramers) have been described. Interestingly, several of the dimers and tetramers observed in the legume lectin family are also found in animal lectins with a related β -sandwich fold such as the galectins and the spermadhesins (Figure 21.10). It is therefore tempting to suggest that the legume lectin fold and related β -sandwich folds as found in the galectins, pentraxins, and spermadhesins were chosen during evolution just because they allow an easy evolution of quaternary structures. This then allows them to adapt easily to the different spacings of carbohydrate ligands

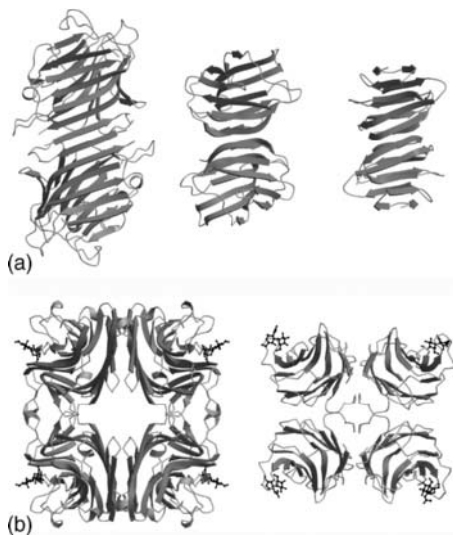


Figure 21.10 Convergent evolution of quaternary structures in lectins with β -sandwich folds. (a) The canonical legume lectin dimer as observed in lentil lectin (left) is also frequently found in the galectin family (middle) and the spermadhesins (right); (b) a very similar tetrameric association observed in *Griffonia simplicifolia* lectin I-B4 (left) and a fungal galectin (right).

presented on, for example, cell surfaces. It will nevertheless be very difficult to prove such a concept experimentally.

A third way to generate multivalence is by using a domain fold with internal (pseudo)symmetry. A good example is tachylectin-2 (Figure 21.2d). This lectin binds GlcNAc and GalNAc as part of the innate immune system of horseshoe crab. Its crystal structure resembles a five-bladed propeller [23], each blade consisting of a four-stranded antiparallel β -sheet of identical topology. Five virtually identical binding sites are generated at each interface between two adjacent blades. Similarly, the monocot mannose-binding lectins have an internal three-fold symmetry [56], and each of the three subdomains has mannose-binding potential (Figure 21.2b).

In many cases, several strategies to obtain multivalence are combined (Figure 21.11). A notorious example is wheat germ agglutinin, which is a dimer of a polypeptide chain containing four hevein domains, resulting in eight carbohydrate binding sites. Equally, the trivalent monomer of snowdrop lectin associates into a tetramer, generating a total of 12 carbohydrate binding sites.

21.5 Conclusion

Structural bioinformatics and the successful prediction of protein–carbohydrate interactions depend crucially on the availability of a sufficiently large database of protein–carbohydrate complexes, and also a correspondingly large set of thermodynamic binding parameters. Whereas for many years the number of protein–carbohydrate complexes in the Protein Data Bank was limited to a small number of lectins belonging to only a few families, mostly of plant origin, the last decade has witnessed an explosion of crystal structures.

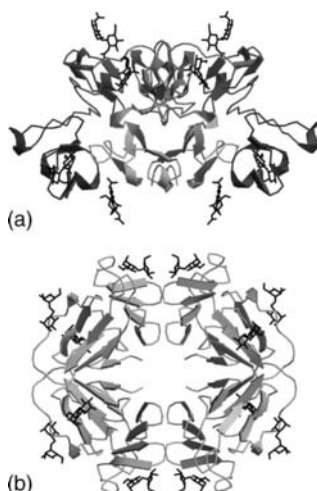


Figure 21.11 Highly multivalent lectins. (a) Wheat germ agglutinin combines domain duplication with dimer formation; (b) snowdrop lectin combines internal three-fold symmetry with tetramer formation.

In this chapter, an overview of the different lectins folds and the strategies that they use to select and bind carbohydrates has been given. A number of general principles emerge that create highly specific high-avidity binders out of low-selectivity, low-affinity building blocks.

References

1. Schrag JD, Procopio DO, Cygler M, *et al.*: Lectin control of protein folding and sorting in the secretory pathway. *Trends Biochem Sci* 2003, **28**:49–57.
2. Sharon N, Lis H: Legume lectins – a large family of homologous proteins. *FASEB J* 1990, **4**:3198–3208.
3. Stillmark H: Über Ricin, ein giftiges Ferment aus den Samen von *Ricinus communis L.* und einigen anderen Euphorbiaceen. PhD Thesis, Dorpat University; 1888.
4. Van Damme EJM, Smeets K, Peumans WJ: The mannose-binding monocot lectins and their genes. In *Lectins, Biomedical Perspectives* (eds Pusztai A, Bardocz S). London: Taylor and Francis; 1995, pp. 59–80.
5. Barre A, Bourne Y, Van Damme EJM, *et al.*: Mannose-binding plant lectins: different structural scaffolds for a common sugar-recognition process. *Biochimie* 2001, **83**:645–651.
6. Sankaranarayanan R, Sekar K, Banerjee R, *et al.*: A novel mode of carbohydrate recognition in jacalin, a Moraceae plant lectin with a β -prism fold. *Nat Struct Biol* 1996, **3**:596–603.
7. Pratap JV, Jeyaprakash AA, Rani PG, *et al.*: Crystal structures of artocarpin, a Moraceae lectin with mannose specificity, and its complex with methyl- α -D-mannose: implications to the generation of carbohydrate specificity. *J Mol Biol* 2002, **317**:237–247.
8. Chrispeels MJ, Raikhel NV: Lectins, lectin genes, and their role in plant defense. *Plant Cell* 1991, **3**:1–9.
9. Gatehouse AMR, Powell KS, Van Damme EJM, Gatehouse JA: Insecticidal properties of plant lectins. In *Lectins, Biomedical Perspectives* (eds Pusztai A, Bardocz S). London: Taylor and Francis; 1995, pp. 35–57.

10. Etzler ME, Kalsi G, Ewing NN, *et al.*: A nod factor binding lectin with apyrase activity from legume roots. *Pro. Natl Acad Sci USA* 1999, **96**:5856–5861.
11. Walser PJ, Haebel PW, Kunzler M, *et al.*: Structure and functional analysis of the fungal galectin CGL2. *Structure* 2004, **12**:689–702.
12. Birck C, Damian L, Marty-Detraves C, *et al.*: A new lectin family with structure similarity to actinoporins revealed by the crystal structure of *Xerocomus chrysenteron* lectin XCL. *J Mol Biol* 2004, **344**:1409–1420.
13. Paaventhan P, Joseph JS, Seow SV, *et al.*: A 1.7 Å structure of Fve, a member of the new fungal immunomodulatory protein family. *J Mol Biol* 2003, **332**:461–470.
14. Wimmerova M, Mitchell E, Sanchez J-F, *et al.*: Crystal structure of fungal lectin: six-bladed β-propeller fold and novel fucose recognition mode for *Aleuria aurantia* lectin. *J Biol Chem* 2003, **278**:27059–27067.
15. Kilpatrick DC: Animal lectins: a historical introduction and overview. *Biochim Biophys Acta* 2002, **1572**:187–197.
16. Weis WI, Taylor ME, Drickamer K: The C-type lectin superfamily in the immune system. *Immunol Rev* 1998, **163**:19–34.
17. Barondes SH, Cooper DNW, Gitt MA, Leffler H: Galectins. Structure and function of a large family of animal lectins. *J Biol Chem* 1994, **269**:20807–20810.
18. Cooper DNW, Barondes SH: God must love galectins; he made so many of them. *Glycobiology* 1999, **9**:979–984.
19. Vasta GR, Quesenberry M, Ahmed H, O’Leary N: C-type lectins and galectins mediate innate and adaptive immune functions: their roles in the complement activation pathway. *Dev Comp Immunol* 1999, **23**:401–420.
20. McGreal EP, Martinez-Pomares L, Gordon S: Divergent roles for C-type lectins expressed by cells of the innate immune system. *Mol Immunol* 2004, **41**:1109–1121.
21. Cambi A, Figdor CG: Dual function of C-type lectin-like receptors in the immune system. *Curr Opin Cell Biol* 2003, **15**:539–546.
22. Gabius H-J: Animal lectins. *Eur J Biochem* 1997, **243**:543–576.
23. Beisel HG, Kawabata S, Iwanaga S, *et al.*: Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab *Tachypleus tridentatus*. *EMBO J* 1999, **18**:2313–2322.
24. Kairies N, Beisel H-G, Fuentes-Prior P, *et al.*: The 2.0-Å crystal structure of tachylectin 5A provides evidence for the common origin of the innate immunity and the blood coagulation systems. *Proc Natl Acad Sci USA* 2001, **98**:13519–13524.
25. Bianchet MA, Odom EW, Vasta GR, Amzel LM: A novel fucose recognition fold involved in innate immunity. *Nat Struct Biol* 2002, **9**:628–634.
26. Uchida T, Yamasaki T, Eto S, *et al.*: Crystal structure of the hemolytic lectin CEL-III isolated from the marine invertebrate *Cucumaria echinata*: implications of domain structure for its membrane pore-formation mechanism. *J Biol Chem* 2004, **279**:37133–37141.
27. Crocker PR: Siglecs: sialic-acid-binding immunoglobulin-like lectins in cell–cell interactions and signalling. *Curr Opin Struct Biol* 2002, **12**:609–615.
28. Stahl PD, Ezekowitz RA: The mannose receptor is a pattern recognition receptor involved in host defense. *Curr Opin Immunol* 1998, **10**:50–55.
29. Nakamura T, Nishizawa T, Hagiya M, *et al.*: Molecular cloning and expression of human hepatocyte growth factor. *Nature* 1989, **342**:440–443.
30. Schmidt C, Bladt F, Goedecke S, *et al.*: Scatter factor/hepatocyte growth factor is essential for liver development. *Nature* 1995, **373**:699–702.
31. Gewurz H, Zhang XH, Lint TF: Structure and function of the pentraxins. *Curr Opin Immunol* 1995, **7**:54–64.
32. Calvete JJ, Sanz L, Dostalova Z, Töpfer-Petersen E: Spermadhesins: spermcoating proteins involved in capacitation and zona-pellucida binding. *Fertilität* 1995, **11**:35–40.

33. Sun Y-J, Chang N-CA, Hung S-I, *et al.*: The crystal structure of a novel mammalian lectin, Ym1, suggests a saccharide binding site. *J Biol Chem* 2001, **276**:17507–17514.
34. Houston DR, Recklies AD, Krupa JC, van Aalten DMF: Structure and ligand-induced conformational change of the 39-kDa glycoprotein from human articular chondrocytes. *J Biol Chem* 2003, **278**:30206–30212.
35. Fusetti F, Pijning T, Kalk KH, *et al.*: Crystal structure and carbohydrate-binding properties of the human cartilage glycoprotein-39. *J Biol Chem* 2003, **278**:37753–37760.
36. Soto GE, Hultgren SJ: Bacterial adhesins: common themes and variations in architecture and assembly. *J Bacteriol* 1999, **181**:1059–1071.
37. Knight SD, Berglund J, Choudhury D: Bacterial adhesins: structural studies reveal chaperone function and pilus biogenesis. *Curr Opin Chem Biol* 2000, **4**:653–660.
38. Prince SM, Achtman M, Derrick JP: Crystal structure of the OpcA integral membrane adhesin from *Neisseria meningitidis*. *Proc Natl Acad Sci USA* 2002, **99**:3417–3421.
39. Datta D, Vaidehi N, Floriano WB, *et al.*: Interaction of *E. coli* outer-membrane protein A with sugars on the receptors of the brain microvascular endothelial cells. *Proteins* 2003, **50**:213–221.
40. Tielker D, Hacker S, Loris R, *et al.*: *Pseudomonas aeruginosa* lectin LecB is located in the outer membrane and is involved in biofilm formation. *Microbiology* 2005, **151**:1313–1323.
41. Heerze LD, Armstrong GD: Comparison of the lectin-like activity of pertussis toxin with two plant lectins that have differential specificities for $\alpha(2-6)$ - and $\alpha(2-3)$ -linked sialic acid. *Biochem Biophys Res Commun* 1990, **172**:1224–1229.
42. Fry EE, Lea SM, Jackson T, *et al.*: The structure and function of a foot-and-mouth disease virus–oligosaccharide receptor complex. *EMBO J* 1999, **18**:543–554.
43. Stehle T, Harrison SC: High-resolution structure of a polyomavirus VP1–oligosaccharide complex: implications for assembly and receptor binding. *EMBO J* 1997, **16**:5139–5148.
44. Dormitzer PR, Sun Z-YJ, Wagner G, Harrison SC: The rhesus rotavirus VP4 sialic acid binding domain has a galectin fold with a novel carbohydrate binding site. *EMBO J* 2002, **21**:885–897.
45. Burmeister WP, Guilligay D, Cusack S, *et al.*: Crystal structure of species D adenovirus fiber knobs and their sialic acid binding sites. *J Virol* 2004, **78**:7727–7736.
46. Weis WI, Cusack SC, Brown JH, *et al.*: The structure of a membrane fusion mutant of the influenza virus haemagglutinin. *EMBO J* 1990, **9**:17–24.
47. Sauter NK, Glick GD, Crowther RL, *et al.*: Crystallographic detection of a second ligand binding site in influenza virus hemagglutinin. *Proc Natl Acad Sci USA* 1992, **89**:324–328.
48. Steinbacher S, Baxa U, Miller S, *et al.*: Crystal structure of phage P22 tailspike protein complexed with *Salmonella* sp. O-antigen receptors. *Proc Natl Acad Sci USA* 1996, **93**:10584–10588.
49. Loris R, Hamelryck T, Bouckaert J, Wyns L: Legume lectin structure. *Biochim Biophys Acta* 1998, **1383**:9–36.
50. Romero A, Romão MJ, Varela PF, *et al.*: The crystal structures of two spermadhesins reveal the CUB domain fold. *Nat Struct Biol* 1997, **4**:783–788.
51. Schrag JD, Bergeron JJM, Li Y, *et al.*: The structure of calnexin, an ER chaperone involved in quality control of protein folding. *Mol Cell* 2001, **8**:633–644.
52. Cioci G, Mitchell EP, Gautier C, *et al.*: Structural basis of calcium and galactose recognition by the lectin PA-IL of *Pseudomonas aeruginosa*. *FEBS Lett* 2003, **555**:297–301.
53. Mitchell E, Houles C, Sudakevitz D, *et al.*: Structural basis for oligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients. *Nat Struct Biol* 2002, **9**:918–921.
54. Loris R, Tielker D, Jaeger K-E, Wyns L: Structural basis of carbohydrate recognition by the lectin LecB from *Pseudomonas aeruginosa*. *J Mol Biol* 2003, **331**:861–870.
55. Montfort W, Villafranca JE, Monzingo AF, *et al.*: The three-dimensional structure of ricin at 2.8 Å. *J Biol Chem* 1987, **262**:5398–5403.
56. Hester G, Kaku H, Goldstein IJ, Wright CS: Structure of mannose-specific snowdrop (*Galanthus nivalis*) lectin is representative of a new plant lectin family. *Nat Struct Biol* 1995, **2**:472–479.

57. Drickamer K: C-Type lectin-like domains. *Curr Opin Struct Biol* 1999, **9**:585–590.
58. Lortat-Jacob H, Grosdidier A, Imberty A: Structural diversity of heparan sulfate binding domains in chemokines. *Proc Natl Acad Sci USA* 2002, **99**:1229–1234.
59. Lietha D, Chirgadze DY, Mulloy B, *et al.*: Crystal structures of NK1-heparin complexes reveal the basis for NK1 activity and enable engineering of potent agonists of the MET receptor. *EMBO J* 2001, **20**:5543–5555.
60. Loris R: Principles of structures of animal and plant lectins. *Biochim Biophys Acta* 2002, **1572**:198–208.
61. Velloso LM, Svensson K, Schneider G, *et al.*: Crystal structure of the carbohydrate recognition domain of p58/ERGIC-53, a protein involved in glycoprotein export from the endoplasmic reticulum. *J Biol Chem* 2002, **277**:15979–15984.
62. Transue TR, Smith AK, Mo H, *et al.*: Structure of benzyl T-antigen disaccharide bound to *Amaranthus caudatus* agglutinin. *Nat Struct Biol* 1997, **4**:779–783.
63. DiGabriele AD, Lax I, Chen DL, *et al.*: Structure of a heparin-linked biologically active dimer of fibroblast growth factor. *Nature* 1998, **393**:812–817.
64. Liu Y, Chirino AJ, Misulovin Z, *et al.*: Crystal structure of the cysteine-rich domain of mannose receptor complexed with a sulfated carbohydrate ligand. *J Exp Med* 2000, **191**:1105–1116.
65. Van Parijs J, Broekaert WF, Goldstein IJ, Peumans WJ: Hevein: an antifungal protein from rubber-tree (*Hevea brasiliensis*) latex. *Planta* 1991, **183**:258–264.
66. Drenth J, Low BW, Richardson JS, Wright CS: The toxin-agglutinin fold. A new group of small protein structures organized around a four-disulfide core. *J Biol Chem* 1980, **255**:2652–2655.
67. Sue S-C, Brisson J-R, Chang S-C, *et al.*: Structures of heparin-derived disaccharide bound to cobra cardiotoxins: context-dependent conformational change of heparin upon binding to the rigid core of the three-fingered toxin. *Biochemistry* 2001, **40**:10436–10446.
68. Buts L, Bouckaert J, De Genst E, *et al.*: The fimbrial adhesin F17-G of enterotoxigenic *Escherichia coli* has an immunoglobulin-like lectin domain that binds *N*-acetylglucosamine. *Mol Microbiol* 2003, **49**:705–715.
69. Choudhury D, Thompson A, Stojanoff V, *et al.*: X-ray structure of the FimC–FimH chaperone–adhesin complex from uropathogenic *Escherichia coli*. *Science* 1999, **285**:1061–1066.
70. Dodson KW, Pinkner JS, Rose T, *et al.*: Structural basis of the interaction of the pyelonephritic *E. coli* adhesin to its human kidney receptor. *Cell* 2001, **105**:733–743.
71. Loris R, De Greve H, Dao-Thi M-H, *et al.*: Structural basis of carbohydrate recognition by lectin II from *Ulex europaeus*, a protein with a promiscuous carbohydrate-binding site. *J Mol Biol* 2000, **301**:987–1002.
72. Loris R, Van Walle I, De Greve H, *et al.*: Structural basis of oligomannose recognition by the *Pterocarpus angolensis* seed lectin. *J Mol Biol* 2004, **335**:1227–1240.
73. Krarup A, Thiel S, Hansen A, *et al.*: L-Ficolin is a pattern recognition molecule specific for acetyl groups. *J Biol Chem* 2004, **279**:47513–47519.
74. Yamashita H, Theerasilp S, Aiuchi T, *et al.*: Purification and complete amino acid sequence of a new type of sweet protein with taste-modifying activity, curculin. *J Biol Chem* 1990, **265**:15770–15775.
75. Leonidas DD, Elbert BL, Zhou Z, *et al.*: Crystal structure of human Charcot–Leyden crystal protein, an eosinophil lysophospholipase, identifies it as a new member of the carbohydrate-binding family of galectins. *Structure* 1995, **3**:1379–1393.
76. Mirkov TE, Wahlstrom JM, Hagiwara K, *et al.*: Evolutionary relationships among proteins in the phytohemagglutinin–arcelin– α -amylase inhibitor family of the common bean and its relatives. *Plant Mol Biol* 1994, **26**:1103–1113.
77. Hamelryck TW, Poortmans F, Goossens A, *et al.*: Crystal structure of arcelin-5, a lectin-like defense protein from *Phaseolus vulgaris*. *J Biol Chem* 1996, **271**:32796–32802.

78. Bompard-Gilles C, Rousseau P, Rouge P, Payan F: Substrate mimicry in the active center of a mammalian α -amylase: structural analysis of an enzyme–inhibitor complex. *Structure* 1996, **4**:1441–1452.
79. Li Z, Lin Q, Yang DSC, *et al.*: The role of Ca^{2+} -coordinating residues of herring antifreeze protein in antifreeze activity. *Biochemistry* 2004, **43**:14547–14554.
80. Sacchettini JC, Baum LG, Brewer CF: Multivalent protein–carbohydrate interactions. A new paradigm for supermolecular assembly and signal transduction. *Biochemistry* 2001, **40**:3009–3015.
81. Olsen LR, Dessen A, Gupta D, *et al.*: X-ray crystallographic studies of unique cross-linked lattices between four isomeric biantennary oligosaccharides and soybean agglutinin. *Biochemistry* 1997, **36**:15073–15080.
82. Bourne Y, Bolgiano B, Liao DI, *et al.*: Crosslinking of mammalian lectin (galectin-1) by complex biantennary saccharides. *Nat Struct Biol* 1994, **1**:863–870.
83. Saul FA, Rovira P, Boulot G, *et al.*: Crystal structure of *Urtica dioica* agglutinin, a superantigen presented by MHC molecules of class I and class II. *Struct Fold Des* 2000, **8**:593–603.
84. Wright CS: The crystal structure of wheat germ agglutinin at 2.2 Å resolution. *J Mol Biol* 1977, **111**:439–457.

Statistical Analysis of Protein–Carbohydrate Complexes Contained in the PDB

Thomas Lütteke¹ and Claus-Wilhelm von der Lieth²

¹*Faculty of Veterinary Medicine, Institute of Biochemistry and Endocrinology, Justus-Liebig University Gießen, 35392 Gießen, Germany*

²*Formerly at the Central Spectroscopic Unit, Deutsches Krebsforschungszentrum (German Cancer Research Center), 69120 Heidelberg, Germany*

22.1 Introduction

For many of the biological functions of carbohydrates, especially their roles in cell–cell and cell–matrix recognition events, a specific recognition of the carbohydrate structures is required. Recognition events that involve carbohydrates on cell surfaces range from fertilization and cellular differentiation to maturation and apoptosis. Glycan structures on the surfaces of glycoproteins serve as a cellular address code that is used for targeted protein transport and also clearance from the circulatory system [1–3]. In addition, carbohydrates play an important role in infections, where pathogens such as viruses or bacteria identify their target host cells by the carbohydrates on the cell surfaces, and in the immune response [3–5]. The majority of these recognition events are mediated by carbohydrate binding proteins, the lectins, most of which bind specifically to certain carbohydrate epitopes [6] (see Chapter 21). For carbohydrate active enzymes (see Chapter 5), a specific distinction between different carbohydrates is also indispensable [7]. It is not fully understood yet how these proteins are able to distinguish precisely between very similar carbohydrate structures.

Even proteins that share a common recognition specificity, for example, for galactose residues, do not necessarily have strong similarities in terms of primary sequence or tertiary structure [8, 9]. On the other hand, proteins with common structural features can be selective for different types of carbohydrates [10]. Therefore, the specificity for certain carbohydrates obviously lies in the local structure of the carbohydrate binding site rather than the overall structure of the protein. Insights into the structural requirements for carbohydrate binding and distinction between different carbohydrates can be gained from the 3D structures of carbohydrate–protein complexes [10–16]. The major methods for resolving such 3D structural data are X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. The largest publicly available collection of such 3D structures is the Protein Data Bank (PDB) [17] (see Chapter 20), which contains about 2500 entries on

carbohydrate–protein complexes (May 2007). These complexes comprise approximately 5900 carbohydrate chains consisting of more than 10 000 carbohydrate residues.

A statistical analysis of the carbohydrate–protein complexes in the PDB yields insights into the amino acids that are necessary for the selective binding of carbohydrates.

22.2 Generation of Datasets

For the analysis of carbohydrate–protein interactions, carbohydrate residues in the PDB entries are detected using the `pdb2linucs` software (see Chapter 20, Section 20.2.2.4). For each monosaccharide unit, all amino acid atoms within a distance of 4 Å are determined. To avoid calculating the distances between all pairs of carbohydrate–amino acid atoms, distances between the geometric centers of residues are evaluated in a first step. Only if such a distance is below a cut-off value are distances between single atoms calculated. Data about PDB entries, carbohydrate chains, carbohydrate and amino acid residues and the corresponding pairs of carbohydrate and amino acid atoms that are distant from each other by not more than 4 Å are stored in an XML file.

22.3 Analysis of Datasets: GlyVicinity

To analyze statistically the data stored in the XML file described above, the GlyVicinity software was developed. A web interface to GlyVicinity is available at the glycosciences.de portal [www.glycosciences.de/tools/glyvicinity/].

Analysis is performed in several steps. In the first step, the residue of interest, for example “b-D-Galp” or “a-D-Neup5Ac”, is entered (in CarbBank notation; see Chapter 3). In addition to this, the type of carbohydrate chains to be analyzed (covalently bound glycans, ligands or all chains) is selected (Figure 22.1a). To investigate the amino acids in the neighborhood of non-covalently bound carbohydrates, the “ligands only” option has to be selected here. In this step, the amino acids within a radius of up to 4 Å around carbohydrate rings of the selected residue type are counted. Absolute counts, relative portions, and deviations from natural abundances can be displayed graphically in diagrams or numerically in tables (Figure 22.1b and c).

In the second step, a refined analysis that includes not only residues but also single atoms can be performed. In the web interface, the user gets to this analysis for a specific amino acid by clicking on the amino acid name in the text output. In this step, it is possible to examine which atoms of the amino acid and of the carbohydrate interact frequently (Figure 22.1d). This can yield insights into the kind of interactions that are predominantly formed between the carbohydrate and specific kinds of amino acids (see below).

To obtain more information about the protein–carbohydrate complexes from which the displayed data were derived, the user can access a list of the PDB entries and carbohydrate chains and also the interacting residues and atoms in a third step. This list is reached by clicking on an atom name or count value in the results of step two. A click on an atom name lists all interactions in which the respective atom is involved, independent of the kind of interacting atom. When clicking on a count value, both related atoms (carbohydrate and amino acid) must interact with each other to be included in the list (Figure 22.1e). From this step, direct links to a 3D structure view of the respective PDB entry, in which the

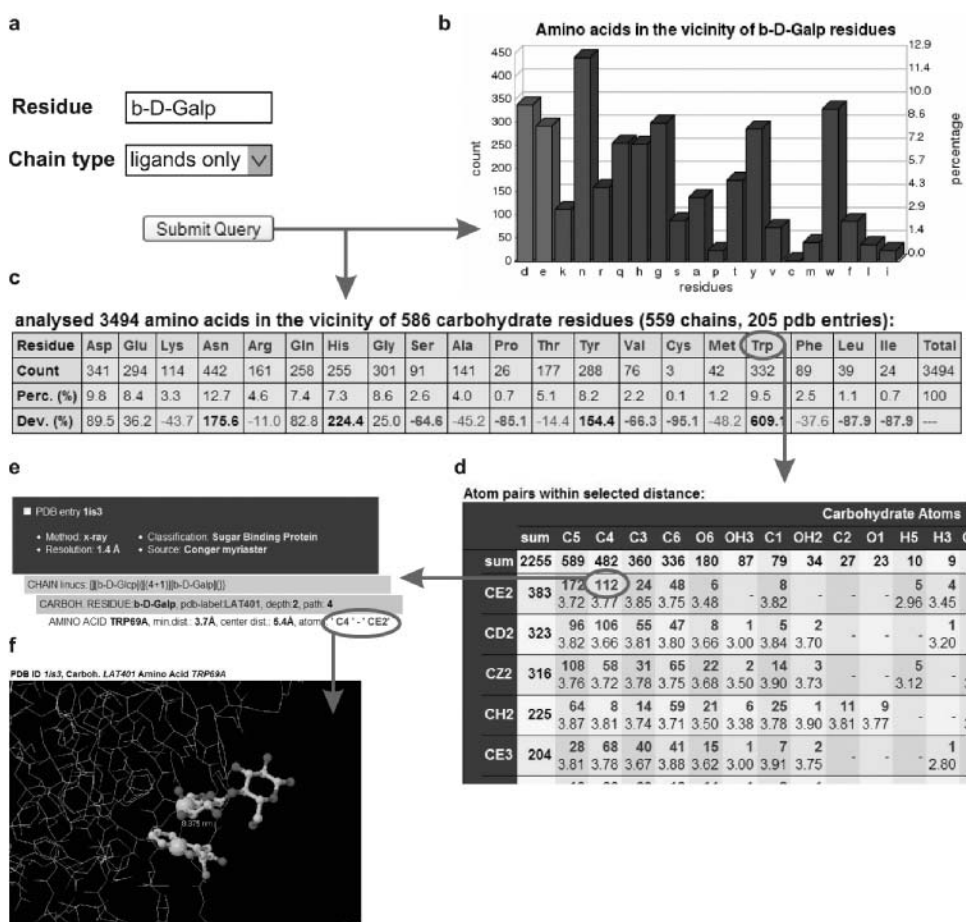


Figure 22.1 Workflow in the GlyVicinity web interface. After selecting a carbohydrate residue and the type of chains (glycan, ligand, or all) to be analyzed (a), data about the amino acids within a 4 Å radius around the selected carbohydrate residues are displayed graphically as diagrams (b) and numerically as tables (c). Statistics about the atoms involved in interactions can be displayed separately for each amino acid (d). Detailed data on the PDB entries, the carbohydrate chains and the residues from which the data were obtained (e) can be accessed in addition to the 3D structures with the selected atoms and residues highlighted (f). A full-color version of this figure is included in the Plate section of this book.

selected interacting atoms are highlighted, is provided (Figure 22.1f). For the display of the 3D structure, the Jmol applet or the Chime plugin can be used.

22.3.1 Amino Acids Found in the Vicinity of Carbohydrate Residues

Within a distance of 4 Å around non-covalently bound carbohydrate residues in the PDB, polar amino acids are more frequently found than non-polar amino acids. The aromatic amino acids Tyr and especially Trp form a remarkable exception. Together, their portion of the total interactions approaches 20% (Figure 22.2a). Since the different types of amino

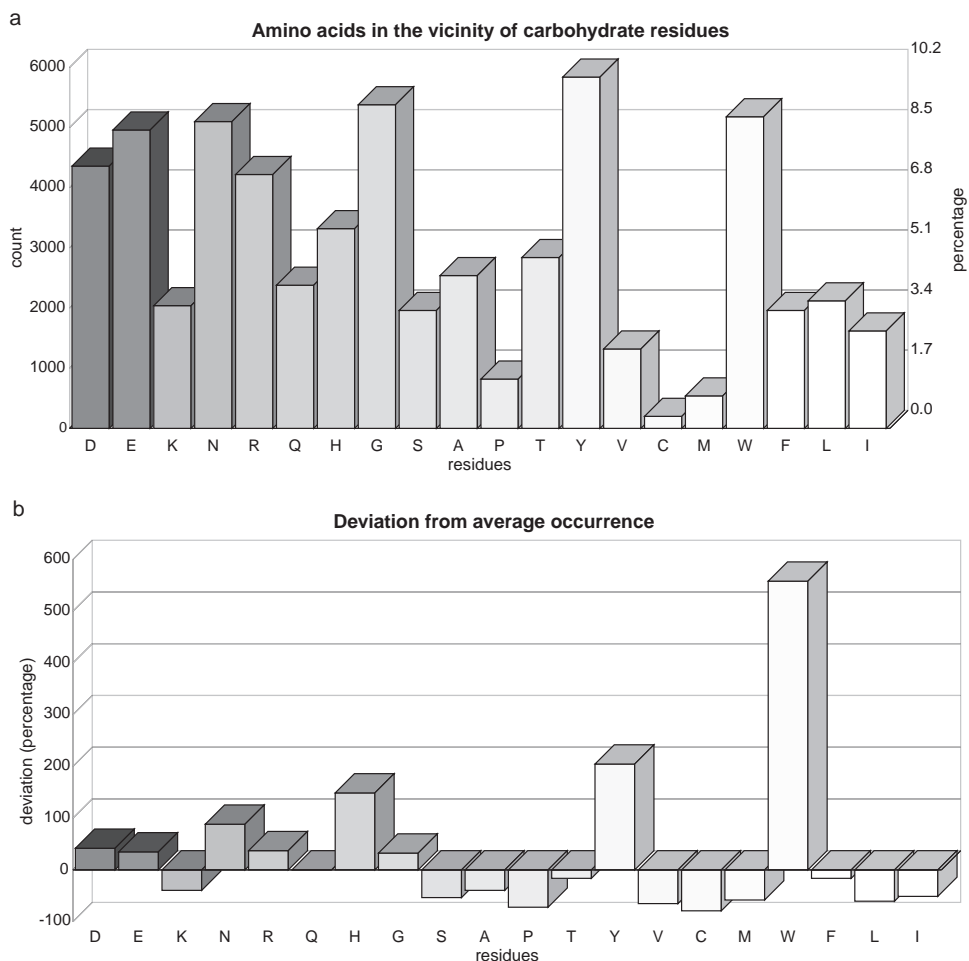


Figure 22.2 Amino acids found within a 4 Å radius around carbohydrate residues. (a) Absolute count; (b) deviations from natural abundances. Residues are sorted and color-coded by polarity from polar residues (left, dark gray) to non-polar residues (right, light gray). With the exception of Lys (K), Tyr (Y), and Trp (W), polar residues are over-represented, whereas the occurrence of non-polar amino acids is below the natural abundance.

acids occur with different frequencies in nature – for example, Leu is the most frequently found, whereas Trp is the rarest amino acid [18] – absolute counts are not necessarily sufficient for evaluating the importance of the different amino acids in glycan binding. Therefore, the deviations of the occurrences of the amino acids from their natural abundances are also computed. Deviations are calculated by the equation

$$Dev_{aa} = \frac{\frac{num_{aa}}{num_{total}} - nat_{aa}}{nat_{aa}} \times 100 \quad (22.1)$$

The relative frequency of an amino acid is determined by dividing its absolute count (num_{aa}) by the total number of residues that were analyzed (num_{total}). From this, the natural abundance of that amino acid (nat_{aa}) is subtracted (abundances values are taken from [18]). For normalization purposes, the result is divided afterwards by the natural abundance. Multiplication by 100 then yields a percentage value.

In the vicinity of non-covalently bound carbohydrate residues in the PDB, all polar amino acids except for Lys are over-represented, whereas the non-polar amino acids apart from the aromatic residues Tyr, Trp, and Phe are under-represented (Figure 22.2b). The strikingly large deviation for Trp results from the fact that this is the amino acid with the lowest natural abundance, but is among the most frequently found residues in protein–carbohydrate interactions in the PDB.

22.3.1.1 Differences Between Various Carbohydrate Residues The GlyVicinity approach can be used to investigate which amino acids are responsible for the specific binding of certain carbohydrate residues. In the following, this is demonstrated for β -D-Galp, D-GlcpNAc, α -D-Neup5Ac, and sulfated residues.

The occurrences of amino acids in the spatial vicinity of β -D-Galp residues are similar to those of the entire dataset (Figure 22.3a and b). Trp shows the highest deviation from its natural abundance, although in absolute counts Asn is more frequently found around β -D-Galp. His is also found to be relatively clearly over-represented, while the deviation for Tyr is less pronounced than in the complete dataset. Among the polar amino acids, Arg is slightly under-represented in addition to Lys. Around D-GlcpNAc residues, the aromatic amino acids Trp and Tyr are predominant. Surprisingly, apart from Asn, polar residues are under-represented here (Figure 22.3c and d).

In the vicinity of α -D-Neup5Ac residues, which are negatively charged at physiological pH, one would expect an increase in positively charged and a decrease in negatively charged

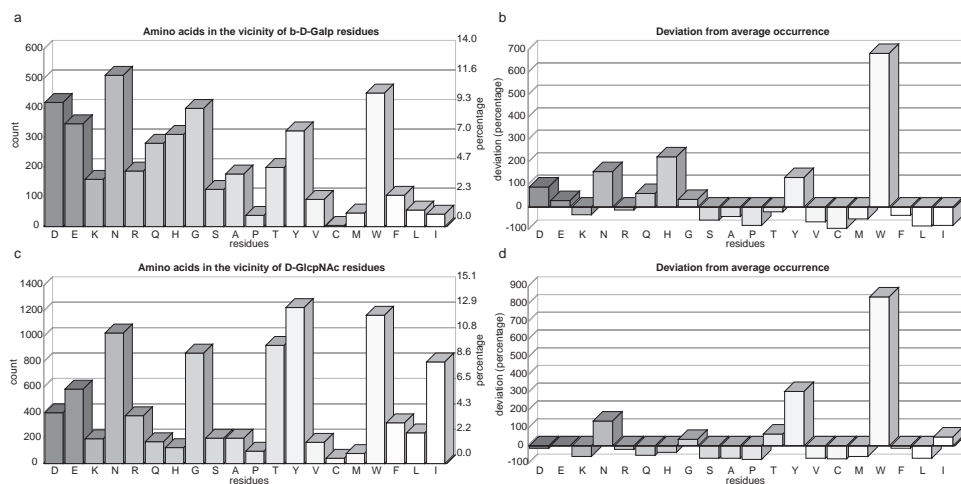


Figure 22.3 Within a 4 Å radius around β -D-Galp (a, b) and D-GlcpNAc (c, d), similar amino acids as in the entire dataset (see Figure 22.2) are found. The major differences between Gal and GlcNAc are observed in the pattern of the deviations of polar amino acids from their average occurrence and also the occurrences of Trp (W) and Cys (C) (b, d).

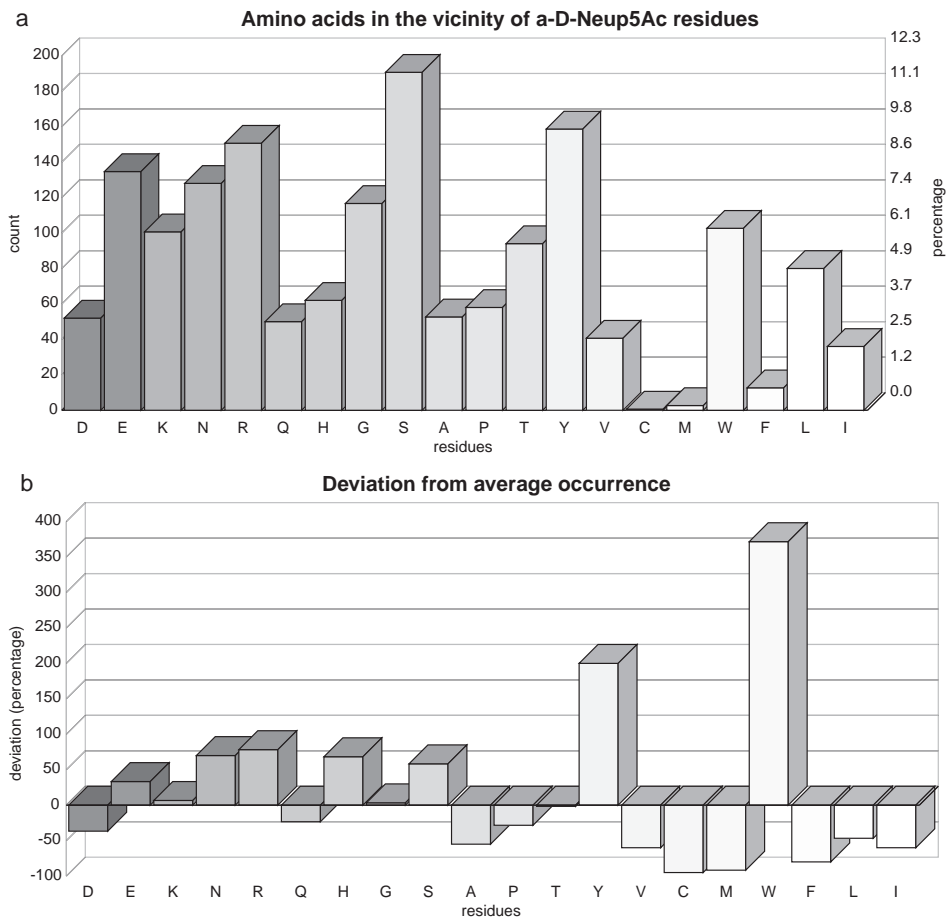


Figure 22.4 Despite the negative charge of α -D-Neup5Ac, the also negatively charged amino acid Glu (E) is over-represented in the vicinity of this carbohydrate. Asp (D), the second negatively charged amino acid, is under-represented, but in absolute counts is about as frequent as Gln (Q), which bears a partial positive charge. In comparison with the uncharged carbohydrates β -D-Galp and D-GlcpNAc (see Figure 22.3), among the positively charged amino acids only Lys (K) and Arg (R) are more frequent in the vicinity of α -D-Neup5Ac, while Asn (N) and Glu (Q) are found less frequently here than around the uncharged carbohydrates. The occurrence of His (H) in the proximity of α -D-Neup5Ac is in between the occurrences of this amino acid in the vicinities of β -D-Galp and D-GlcpNAc.

amino acids compared with the uncharged carbohydrates. However, among the positively charged residues Lys is found to be present at about its natural abundance, whereas Gln, which bears a partial positive charge, is even under-represented (Figure 22.4). Asn, Arg, and His are the only positively charged amino acids that are more than 50% over-represented here. Of the negatively charged amino acids, Asp is under-represented, but Glu is over-represented. In total counts, Glu is more frequently found than most positively charged amino acids. Hence, overall, the negative charge of α -D-Neup5Ac has an effect on the kinds of amino acids with which interactions are formed, but this does not strongly affect all charged amino acids.

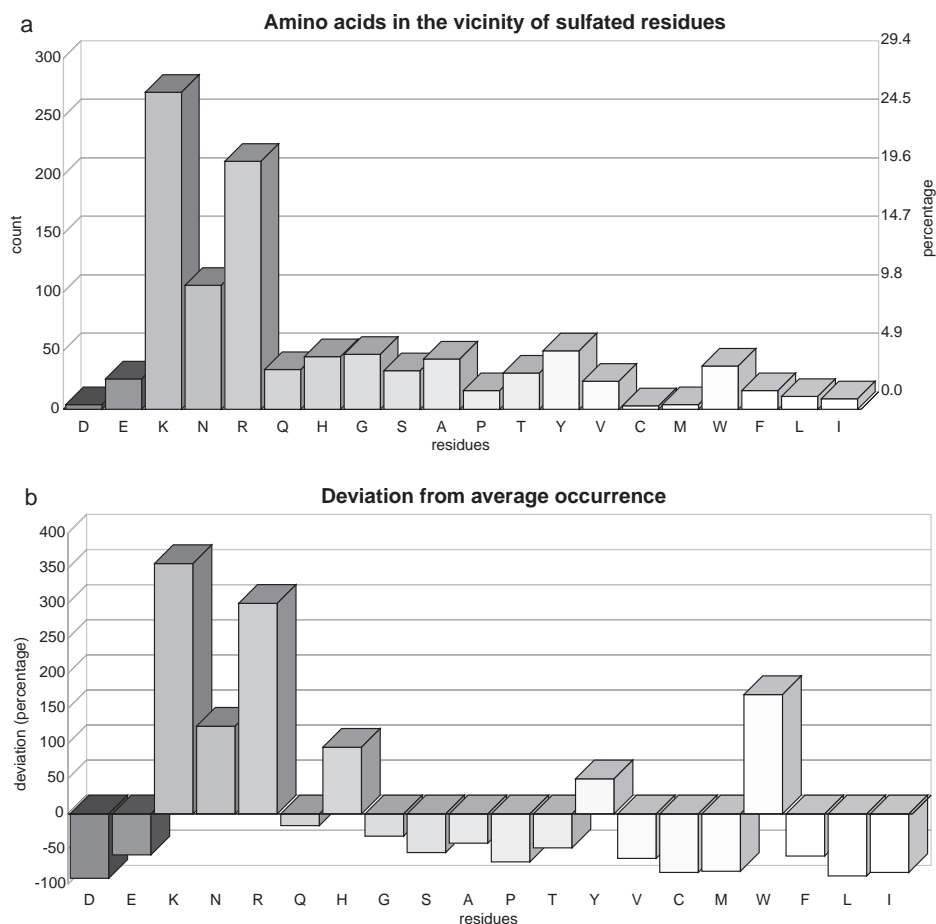


Figure 22.5 The negative charges of sulfate groups result in a clear dominance of the positively charged amino acids Lys (K) and Arg (R) in the vicinity of carbohydrate residues that feature these substituents, while the role of aromatic amino acids is less important than for other carbohydrate residues (see Figures 22.2–22.4).

The situation is different for sulfated carbohydrate residues, which bear multiple negative charges. Here, negatively charged amino acids are rarely present within the analyzed 4 Å radius, while the positively charged Lys and Arg and to a minor extent Asn and His are over-represented (Figure 22.5). Lys, which is under-represented around most other carbohydrate residues analyzed, is in fact the most frequent amino acid here. Among the non-polar residues, only Tyr and Trp are slightly over-represented. These findings indicate that the negative charges of the sulfate groups dominate the interactions of sulfated carbohydrates with proteins.

22.3.2 Interactions at the Atomic Level

If not only residues but also single atoms are examined, one can see that the interactions of polar residues with carbohydrates are primarily mediated by the polar atoms at the

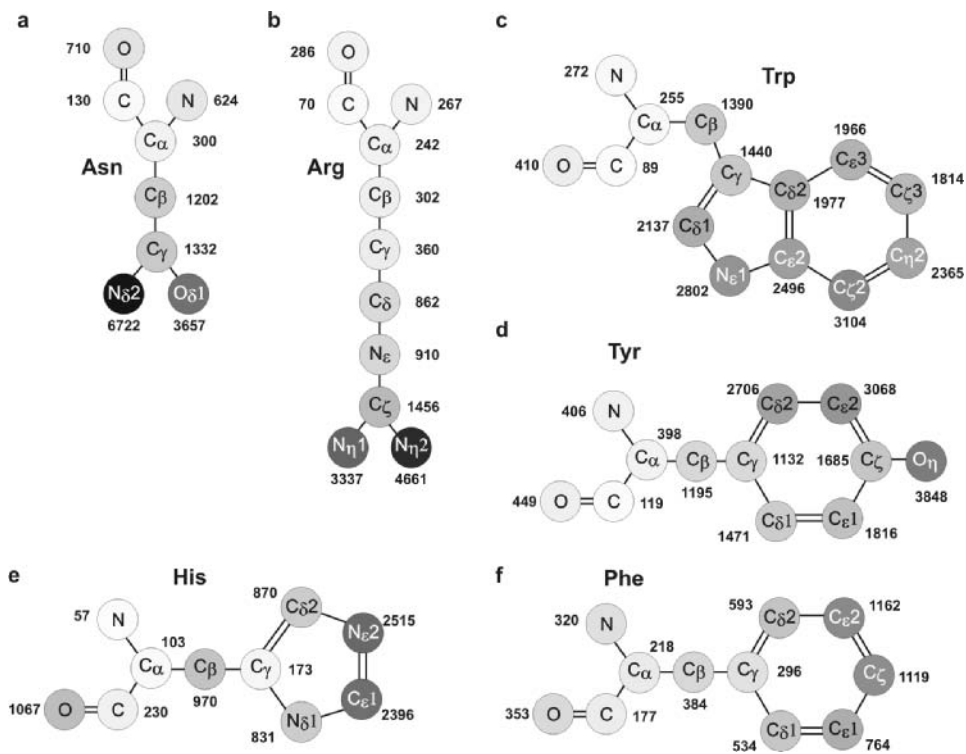


Figure 22.6 Interactions of single amino acid atoms with carbohydrate residues. The shades of gray indicate the ratio of interactions of each atom to the total number of amino acids of the respective type that participate in interactions. For polar amino acids there is a gradient in interaction counts from the polar atoms at the tips of the side-chains towards the protein backbone (a, b, e), whereas for aromatic amino acids the interactions are more evenly distributed on the aromatic rings (c, d, f). For Trp, no clear gradient can be observed (c). This indicates a different type of interactions between Trp and carbohydrate residues compared with the polar amino acids. Tyr contains a polar hydroxyl group, which is involved in more interactions than any other atom of this amino acid, but the difference from the non-polar atoms is not as distinct as in, for example, Asn or Arg (d).

tips of the amino acid side-chains, which mainly form hydrogen bonds to the hydroxyl groups of the carbohydrate residues. Interaction counts form a gradient from the tips of the side-chains to the backbone, with the fewest interactions found at the carbon atoms near or in the protein backbone (Figure 22.6). The polar backbone oxygen and nitrogen atoms show a slight increase in interactions compared with the surrounding carbons, but, since the backbone is the same for all amino acids, these interactions are non-specific.

The patterns observed with the polar residues are about the same for the different carbohydrate residues, no significant differences can be observed. Therefore, the only way for these amino acids to contribute to the specificity of carbohydrate binding proteins lies in their location on the protein surface. Depending on how the potential H-bond donors and acceptors are arranged within the carbohydrate recognition domain, different carbohydrates, which themselves provide different positions of the hydroxyl groups and thereby of the counterparts for H-bond formation, can be bound by the proteins.

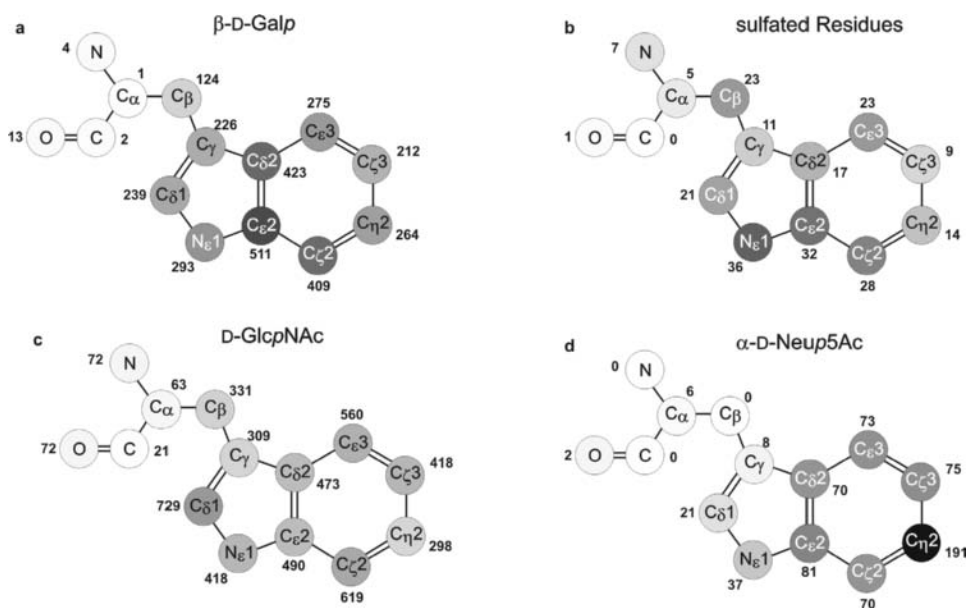


Figure 22.7 Interaction counts of Trp atoms for various carbohydrate residues. The variations in the distributions of count values result from different types of interactions that are preferably formed between Trp and the different carbohydrate residues.

Of the aromatic amino acids, Tyr can also form hydrogen bonds to carbohydrates via the hydroxyl group at the side-chain end. The Trp side-chain, however, contains no such atoms that could serve as strong hydrogen donors or acceptors. Nevertheless, Trp is strongly over-represented in the adjacent amino acids for all types of carbohydrates analyzed here. Consequently, this amino acid must interact with carbohydrate residues in a different way.

In contrast to the polar amino acids, the patterns of interactions with Trp atoms vary between different types of carbohydrate residues. Interactions with β -D-Galp are fairly evenly distributed over the indole ring of Trp, with a slight preference for the $C_{\delta 2}$ – $C_{\epsilon 3}$ – $C_{\zeta 2}$ axis (Figure 22.7a). This pattern mainly results from stacking ($\text{CH}-\pi$) interactions, where the pyranose ring of β -D-Galp and the indole ring of Trp are arranged parallel to each other [9, 19]. The non-polar hydrogen atoms attached to the C_3 – C_4 – C_5 – C_6 axis of β -D-Galp, which are part of the so-called “B-face” [20, 21], are located parallel to the axis of the indole ring mentioned above, which explains the increased number of interactions for these atoms. By this means, the Trp side-chain is not only involved in the binding of β -D-Galp residues but also determines the orientation of the carbohydrate residue. The importance of Trp for galactose binding can also be seen from the fact that the replacement of a His by Trp in the carbohydrate recognition domain of a rat serum mannose-binding protein results in an increased affinity for galactose [22]. For D-GlcpNAc, no significant preferences for any of the side-chain atoms can be observed (Figure 22.7c).

The majority of the interactions between Trp and α -D-Neup5Ac are observed at the $C_{\eta 2}$ atom of Trp (Figure 22.7d). The reason for this can be found by a visual inspection of the 3D structures of the carbohydrate–protein complexes, in which α -D-Neup5Ac interacts with Trp. In many cases, the α -D-Neup5Ac residue is linked to a β -D-Galp residue that

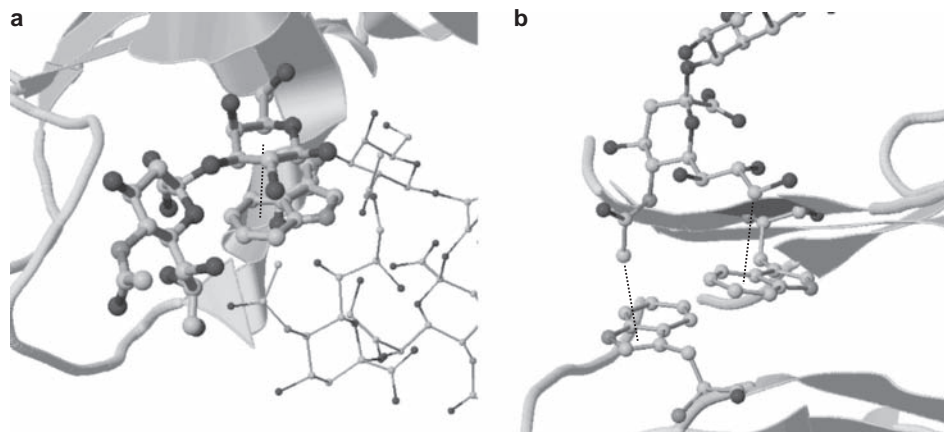


Figure 22.8 Many of the interactions between α -D-Neup5Ac found in the statistics result mainly from interactions between Trp and β -D-Galp residues, to which the α -D-Neup5Ac moieties are bonded (a). Direct interactions between Trp and α -D-Neup5Ac are mainly mediated by the *N*-acetyl group and the glycerol side-chain (b).

forms stacking interactions to the Trp side-chain as described above (Figure 22.8a). Direct interactions between Trp and α -D-Neup5Ac are mainly mediated by the *N*-acetyl group or the glycerol side-chain (Figure 22.8b). Only one complex is found in the PDB where the ring of an α -D-Neup5Ac residue is located parallel to a Trp indole ring. This can be explained by the fact that α -D-Neup5Ac does not have a non-polar B-Face comparable to that of β -D-Galp (Figure 22.9).

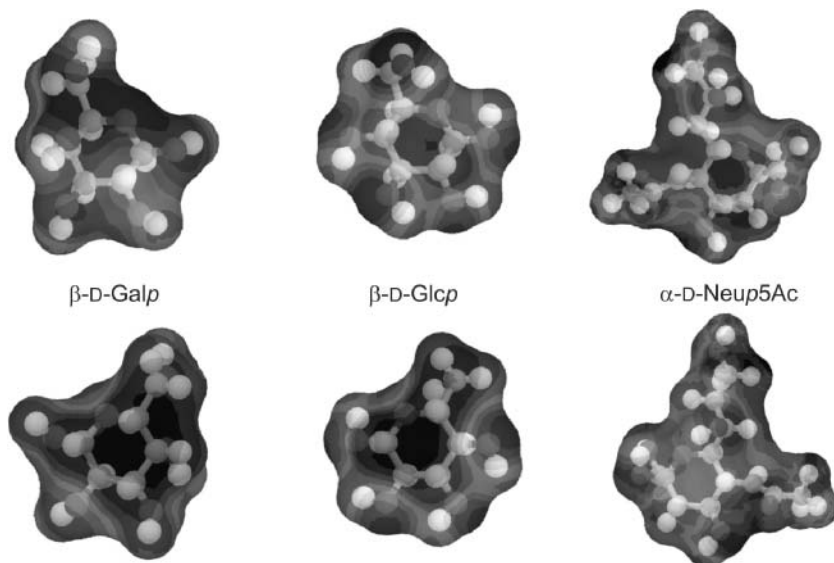


Figure 22.9 Electrostatic view of the surfaces of β -D-Galp, β -D-Glcp, and α -D-Neup5Ac. Top, “upper” ring side; bottom, “lower” ring side. A full-color version of this figure is included in the Plate section of this book.

The interactions between Trp and sulfated carbohydrate residues (Figure 22.7b) resemble those of Trp and β -D-Galp, with a small shift towards the $N_{\epsilon}1$ atom. At first sight, this is surprising as the total distributions of amino acids in the vicinity of these types of carbohydrate residues differ strongly from each other. However, the similarity of the interactions between Trp and these two classes of carbohydrates is not coincidental, as most sulfated residues which interact with Trp are sulfated galactoses. The sulfated carbohydrates also interact with Trp primarily via stacking interactions. The sulfate groups often are located at the side of the carbohydrate ring that is pointing away from the Trp side-chain and thus do not specifically interact with Trp. If the sulfate groups are directly interacting with the Trp side-chain, the $N_{\epsilon}1$ atom is often involved, which explains the increased number of interactions for this atom in comparison with β -D-Galp. The $N_{\epsilon}1$ atom can serve as an, albeit weak, hydrogen bond donor [23].

Slight differences in the stereochemistry of the ring carbon atoms and the corresponding differences in the distribution of electrostatic potentials on the surface of the monosaccharide unit can have a strong effect on the way the carbohydrate residue interacts with Trp. For example, β -D-Galp and β -D-Glcp only differ in the stereochemistry of the C4 atom. However, this means that within the C3–C4–C5–C6-axis all the hydroxyl groups point to one side of the β -D-Galp ring, which leads to a large non-polar region on the opposite ring side, whereas in the β -D-Glcp ring the hydroxyl group attached to the C4 atom is located within this region. Hence β -D-Glcp has not such a distinct non-polar B-face as β -D-Galp, but therefore the other side of the β -D-Glcp ring also exhibits a non-polar area on its surface (Figure 22.9).

The effect of this difference on the interactions with Trp is illustrated best by a superimposition of the carbohydrate rings of each type and the interacting Trp side-chain atoms. Whereas most of the Trp atoms that interact with β -D-Galp are found close to the hydrophobic area described above (the B-face or “lower” ring side), the Trp atoms that interact with β -D-Glcp are located on both sides of the carbohydrate ring, with a slight preference for the “upper” ring side (Figure 22.10). However, since most of the Trp side-chain atoms are found parallel to the carbohydrate ring and not accumulated around the hydroxyl groups, one can

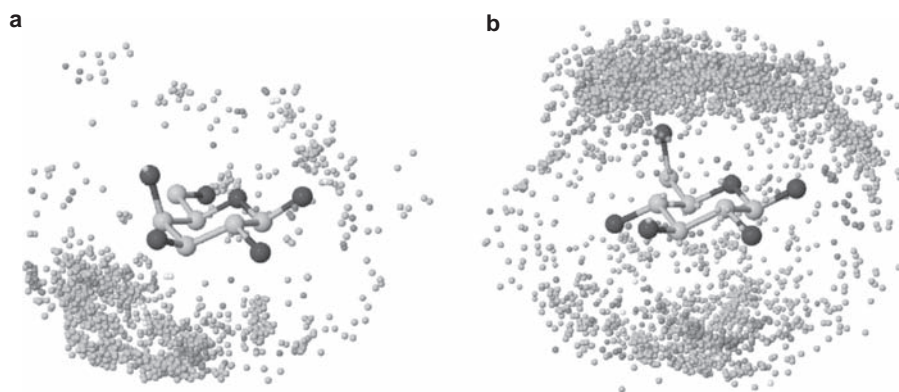


Figure 22.10 Superimposition of β -D-Galp (a) and β -D-Glcp (b) residues and the interacting Trp side-chain atoms. Whereas β -D-Galp mainly forms stacking interactions to Trp via the non-polar area at the lower side of the carbohydrate ring (a), both ring sides of β -D-Glcp interact with Trp. A slight preference for the upper side can be observed here (b).

conclude that Trp preferably forms stacking interactions to β -D-Glcp also. In contrast to β -D-Galp, where there is a strong preference for a certain orientation of the carbohydrate ring to the Trp side-chain, the interactions between Trp and β -D-Glcp are not position specific and therefore cannot be used to direct the carbohydrate on the protein surface.

22.4 Conclusion

Carbohydrate binding proteins are able to distinguish between different carbohydrate residues. The molecular basis for these distinctions can be revealed from statistical analyses of protein–carbohydrate complexes. With the exception of aromatic residues, polar amino acids are generally found more frequently than apolar amino acids in the vicinity of carbohydrates. The frequencies of the amino acids that are present in the neighborhood of non-covalently bound carbohydrates depend on the type of the carbohydrate residue. From the amino acid atoms and carbohydrate atoms that are frequently found in close proximity, conclusions can be drawn about the types of interactions that occur between the amino acids and the carbohydrate residues. Apart from the fact that charged residues attract residues bearing the opposite charge, interactions between polar amino acids and carbohydrates are not residue specific, so that these amino acids only provide specificity for certain carbohydrate residues by their relative positions within the carbohydrate binding domain of the protein. In contrast, aromatic amino acids, especially Trp, interact with carbohydrates in a residue-specific way. Therefore, these amino acids play a prominent role in the selective binding of carbohydrates.

References

1. Lis H, Sharon N: Lectins: carbohydrate-specific proteins that mediate cellular recognition. *Chem Rev* 1998, **98**: 637–674.
2. Helenius A, Aebi M: Intracellular functions of N-linked glycans. *Science* 2001, **291**: 2364–2369.
3. Dennis JW, Granovsky M, Warren CE: Protein glycosylation in development and disease. *Bioessays* 1999, **21**: 412–421.
4. Smith AE, Helenius A: How viruses enter animal cells. *Science* 2004, **304**: 237–242.
5. Dube DH, Bertozzi CR: Glycans in cancer and inflammation – potential for therapeutics and diagnostics. *Nat Rev Drug Discov* 2005, **4**: 477–488.
6. Loris R: Principles of structures of animal and plant lectins. *Biochim Biophys Acta* 2002, **1572**: 198–208.
7. Ünligil UM, Rini JM: Glycosyltransferase structure and mechanism. *Curr Opin Struct Biol* 2000, **10**: 510–517.
8. Elgavish S, Shaanan B: Lectin–carbohydrate interactions: different folds, common recognition principles. *Trends Biochem Sci* 1997, **22**: 462–467.
9. Sujatha MS, Balaji PV: Identification of common structural features of binding sites in galactose-specific proteins. *Proteins* 2004, **55**: 44–65.
10. Drickamer K: Making a fitting choice: common aspects of sugar-binding sites in plant and animal lectins. *Structure* 1997, **5**: 465–468.
11. Rini JM: Lectin structure. *Annu Rev Biophys Biomol Struct* 1995, **24**: 551–577.
12. Weis WI, Drickamer K: Structural basis of lectin–carbohydrate recognition. *Annu Rev Biochem* 1996, **65**: 441–473.

13. Flint J, Nurizzo D, Harding SE, *et al.*: Ligand-mediated dimerization of a carbohydrate-binding molecule reveals a novel mechanism for protein–carbohydrate recognition. *J Mol Biol* 2004, **337**: 417–426.
14. Kulkarni KA, Katiyar S, Surolia A, *et al.*: Structural basis for the carbohydrate-specificity of basic winged-bean lectin and its differential affinity for Gal and GalNAc. *Acta Crystallogr Sect D: Biol Crystallogr* 2006, **62**: 1319–1324.
15. Blanchard H, Yu X, Coulson BS, von Itzstein M: Insight into host cell carbohydrate-recognition by human and porcine rotavirus from crystal structures of the virion spike associated carbohydrate-binding domain (VP8*). *J Mol Biol.* 2007, **367**: 1215–1226.
16. Banerji S, Wright AJ, Noble M, *et al.*: Structures of the Cd44–hyaluronan complex provide insight into a fundamental carbohydrate–protein interaction. *Nat Struct Mol Biol* 2007, **14**: 234–239.
17. Berman HM, Westbrook J, Feng Z, *et al.*: The Protein Data Bank. *Nucleic Acids Res* 2000, **28**: 235–242.
18. Lehmann WD, Bohne A, von der Lieth CW: The information encrypted in accurate peptide masses – improved protein identification and assistance in glycopeptide identification and characterization. *J Mass Spectrom* 2000, **35**: 1335–1341.
19. Rao VS, Lam K, Qasba PK: Architecture of the sugar binding sites in carbohydrate binding proteins – a computer modeling study. *Int J Biol Macromol* 1998, **23**: 295–307.
20. Sundari CS, Balasubramanian D: Hydrophobic surfaces in saccharide chains. *Prog Biophys Mol Biol* 1997, **67**: 183–216.
21. Taylor ME, Drickamer K: *Introduction to Glycobiology*. New York: Oxford University Press; 2003.
22. Iobst S, Drickamer K: Binding of sugar ligands to Ca²⁺-dependent animal lectins. II. Generation of high-affinity galactose binding by site-directed mutagenesis. *J Biol Chem* 1994, **269**: 15512–15519.
23. Scheiner S, Kar T, Pattanayak J: Comparison of various types of hydrogen bonds involving aromatic amino acids. *J Am Chem Soc* 2002, **124**: 13257–13264.

Section 7: Appendices

Appendix 1: List of Available Websites

Name	Description	URL
Carbohydrate Web Portals		
CFG	Websites of the US Consortium for Functional Glycomics	www.functionalglycomics.org
EUROcarbDB	European Carbohydrate Database portal	www.eurocarbdb.org
GLYCOSCIENCES.de	Collection of databases and online tools for glycoscientists	www.glycosciences.de
KEGG Glycan	Glyco-related subpart of the KEGG portal	www.genome.jp/kegg/glycan/
RINGS	Resource for INformatics of Glycomes at Soka	rings.t.soka.ac.jp
Carbohydrate-specific Databases		
CCSDB/CarbBank	Complex Carbohydrate Structure Database	www.boc.chem.uu.nl/sugabase/carbbank.html
CFG Database	Database of the Consortium for Functional Glycomics	www.functionalglycomics.org/glycomics/common/jsp/firstpage.jsp
BCSDB	Bacterial Carbohydrate Structure Database	www.glyco.ac.ru/bcsdb/
GLYCOSCIENCES.de	Database of the GLYCOSCIENCES.de portal	www.glycosciences.de/sweetdb/
KEGG Glycan Database	Database of the KEGG Glycan portal	www.genome.jp/kegg/glycan/
GlycoconjugateDB:Structures	Carbohydrate 3D structures from the PDB	www.glycostructures.jp
DOUGAL	Glycoprotein structures database	www.cryst.bb.k.ac.uk/DOUGAL/
O-GlycBase	Database of O-glycosylation sites	www.cbs.dtu.dk/databases/OGLYCBASE/
ECCODAB	<i>E. coli</i> O-antigen database	www.casper.organ.su.se/ECCODAB/
Sugabase	Carbohydrate NMR database that combines CarbBank data with chemical shift values	www.boc.chem.uu.nl/sugabase/sugabase.html
GlycoMapsDB	Conformational maps of carbohydrates	www.glycosciences.de/modeling/glycomapsdb/
DisaccharideDB	Conformational maps of disaccharides	www.cermav.cnrs.fr/cgi-bin/di/di.cgi
CAZy	Carbohydrate Active enZymes database	www.cazy.org
KEGG Pathway	Pathways of carbohydrate metabolism	www.genome.jp/kegg/pathway.html#glycan
KEGG Orthology	KEGG Orthology (KO) groups for glycosyltransferases	www.genome.jp/kegg/glycan/GT.html
CFG GT Database	Glycosyltransferases database of the CFG	www.functionalglycomics.org/glycomics/molecule/jsp/glycoEnzyme/geMolecule.jsp
CFG GBP Database	Glycan Binding Proteins database of the CFG	www.functionalglycomics.org/glycomics/molecule/jsp/gbpMolecule-home.jsp
CFG Consortium data	Experimental data from the CFG; Glycan Profiling (MS), Mouse Phenotyping, Gene Microarray, and Glycan Array data	www.functionalglycomics.org/glycomics/publicdata/home.jsp
GlycopeptideDB	Database of carbohydrate recognition motifs, recognizing antibodies and glycoproteins/glycolipids carrying the motifs	www.glyco.is.ritsumei.ac.jp/epitope/

(Continues)

Name	Description	URL
GPI Anchor Biosynthesis Report	Database of enzymes for biosynthesis of GPI anchors	mendel.imp.ac.at/SEQUENCES/gpi-biosynthesis/
GlycoBase (Dublin)	Database of 2-aminobenzamide labeled glycans and exoglycosidase digestion pathways	glycobase.ucd.ie
GlycoBase (Lille)	Database of the glyco-biodiversity of different animal species	glycobase.univ-lille1.fr/base/
Elution Coordinate DB	Database of 2-aminopyridine labeled glycans	www.gak.co.jp/ECD/Hpg_eng/hpg_eng.htm
GlycomeDB	Meta-database to search various databases via one interface	www.glycome-db.org
MonoSaccharideDB	Reference database for monosaccharide notation	www.monosaccharidedb.org
SphingoMap	Sphingolipid synthesis pathways	www.sphingomap.org
GGDB	Human glyco-genes database	riodb.ibase.aist.go.jp/rcmg/ggdb/
LFDB	Lectin Frontier DataBase	riodb.ibase.aist.go.jp/rcmg/glycodb/LectinSearch
GMDB	GlycoMass DataBase, database of glycan mass spectra	riodb.ibase.aist.go.jp/rcmg/glycodb/Ms_ResultSearch
GPDB	GlycoProtDB, glycoprotein database	riodb.ibase.aist.go.jp/rcmg/glycodb/Glc_ResultSearch
SugarBindDB	Pathogen Sugar Binding Database	sugarbindb.mitre.org/
GlyAffinity	Lectin database	www.glycosciences.de/affinity/browse.action
SpeCarb	Database of Raman spectra of carbohydrates	www.models.kvl.dk/users/engelsen/specarb/specarb.html
Information on Carbohydrate-related Genes or Proteins in Other Databases		
UniProt	Protein information, successor of Swiss-Prot (manually annotated) and TrEMBL (automatically annotated)	www.uniprot.org
Brenda	Enzyme Information System	www.brenda-enzymes.info
PDB	Protein Data Bank, 3D structure database, contains some structures of glycoprotein or protein-carbohydrate complexes	www.pdb.org
EXPASY Enzyme	Enzyme nomenclature and cross-references	www.expasy.org/enzyme/
	2.4.-.-: Glycosyltransferases	
	3.2.-.-: Glycosylases	
	4.2.2.-: Lyases acting on polysaccharides	
	5.1.3.-: Epimerases acting on carbohydrates	
GenBank	Annotated collection of all publicly available DNA sequences	www.ncbi.nlm.nih.gov/Genbank/

(Continued)

Tools		
<i>Mass Spectrometry-related Sites</i>		
Glycofragment	Calculate and display the main fragments (B- and C-, Z- and Y-, A- and X-ions) of oligosaccharides that should occur in mass spectra	www.glycosciences.de/tools/GlycoFragments/
GlycoSearchMS	Search GLYCOSCIENCES.de database for structures matching a given set of mass peaks	www.glycosciences.de/sweetdb/start.php?action=form_ms_search
GlycoPeakfinder	Tool for fast annotation of glycan MS spectra, results can be used for advanced database search in GLYCOSCIENCES.de	www.glyco-peakfinder.org
GlycoWorkBench	Tool to assist the manual interpretation of mass spectra	www.eurocarbodb.org/applications/ms-tools/
GlycoMod	Prediction of possible oligosaccharide structures that occur on proteins from their experimentally determined masses	www.expasy.org/tools/glycomod/
GlycanMass	Calculation of the mass of an oligosaccharide structure from its composition	www.expasy.org/tools/glycomod/glycanmass.html
<i>NMR-related Sites</i>		
CASPER	Simulation of NMR spectra and carbohydrate sequence determination from NMR chemical shift values	www.casper.organ.su.se/casper/
GlyNest	Estimation of NMR chemical shifts	www.glycosciences.de/sweetdb/start.php?action=form_shift_estimation
ProSpectND	Integrated NMR data processing and inspection tool	www.eurocarbodb.org/applications/nmr-tools
CCPN	Website of the Collaborative Computing Project for NMR	www.ccpn.ac.uk
HPLC related sites		
AutoGU	Tool to assist the interpretation of HPLC data	glycibase.ucd.ie/cgi-bin/profile_upload.cgi
GALAXY	Visualization of HPLC 2D maps	www.glycoanalysis.info/ENG/index.html
<i>Sites Related to Glycosylation and Protein-Carbohydrate Interaction</i>		
GECS	KEGG Gene Expression to Chemical Structure, N-glycan prediction server	www.genome.jp/tools/gecs/
NetNGlyc	Prediction of N-glycosylation sites in protein sequences	www.cbs.dtu.dk/services/NetNGlyc
NetOGlyc	Prediction of O-glycosylation sites in protein sequences	www.cbs.dtu.dk/services/NetOGlyc
NetCGlyc	Prediction of C-mannosylation sites in protein sequences	www.cbs.dtu.dk/services/NetCGlyc
NetGlycate	Prediction of glycation of ε-amino groups of lysines in mammalian proteins	www.cbs.dtu.dk/services/NetGlycate
DictyOGlyc	Prediction of O-(α)-GlcNAc glycosylation sites	www.cbs.dtu.dk/services/DictyOGlyc

Name	Description	URL
YinOYang	Prediction of <i>O</i> -(B)-GlcNAc glycosylation and <i>Yin-Yang</i> sites	www.cbs.dtu.dk/services/YinOYang
GlySeq	Statistical analysis of amino acids in the neighbourhood of glycosylation sites in the PDB and Swiss-Prot	www.glycosciences.de/tools/glyseq/
GlyVicinity	Statistical analysis of protein-carbohydrate interactions data from the PDB	www.glycosciences.de/tools/glyvicinity/
<i>3D Structure-related Sites</i>		
Sweet-II	Generation of carbohydrate 3D structure models	www.glycosciences.de/modeling/sweet2/
Glycam Biomolecule Builder	Generation of carbohydrate 3D structure models	www.glycam.com/CCRC/biombuilder/biomb_index.jsp
pdb2linucs	Detection of carbohydrate moieties in PDB structures	www.glycosciences.de/tools/pdb2linucs/
pdb-care	Validation of carbohydrate 3D structures	www.glycosciences.de/tools/pdb-care/
CARP	Carbohydrate Ramachandran Plot	www.glycosciences.de/tools/carp/
GlyTorsion	Statistical analysis of torsion angles derived from the PDB	www.glycosciences.de/tools/glytorsion/
GlyProt	<i>In silico</i> glycosylation of proteins	www.glycosciences.de/modeling/glyprot/
Dynamic Molecules	Online molecular dynamics (MD) simulations	www.md-simulations.de/manager/
Glyco3D	Collection of carbohydrate 3D structural information	www.cermav.cnrs.fr/glyco3d/
<i>Notation-related Sites</i>		
IUPAC	Official IUPAC recommendations for carbohydrate notation	www.chem.qmul.ac.uk/iupac/2carb/
LINUCS	Linear notation for carbohydrate structures, conversion from IUPAC to LINUCS	www.glycosciences.de/tools/linucs/
GlycoCT	Connection-table based notation for carbohydrate structures	www.eurocarbdb.org/recommendations/encoding
LiGraph	Graphical representation of carbohydrate structures, conversion from LINUCS to IUPAC	www.glycosciences.de/tools/LiGraph/
KEGG CSM	Composite Structure Map, graphical query tool for KEGG databases	www.genome.jp/kegg-bin/draw_csm
Glycan Builder	Graphical input of carbohydrate structures	www.eurocarbdb.org/applications/structure-tools/
SuMo	Sugar Motif Search	www.glycosciences.de/tools/sumo/
Glyde-II	XML exchange format for carbohydrates	glycomics.ccr.cuga.edu/GLYDE-II/
KCF to LINUCS	Conversion from KEGG Chemical Format to LINUCS	rings.t.soka.ac.jp/kcf_to_linucs.html
LINUCS to KCF	Conversion from LINUCS to KEGG Chemical Format	rings.t.soka.ac.jp/linucs_to_kcf.html

Appendix 2: Glossary

Alditol: A *monosaccharide*, in which the carbonyl functional group is reduced to a hydroxyl group.

Artificial neural network (ANN): Mathematical or computational model that is used in decision-making processes, to model complex relationships between inputs and outputs or to find patterns in data. Most ANNs require a learning or training phase with known data before they can be used on new data (see Chapters 9 and 17).

Bacterial Carbohydrate Structure DataBase (BCSDB): A database aimed at the provision of structural, bibliographic and related information on bacterial carbohydrate structures.

CarbBank: Retrieval software to access the data contained in the Complex Carbohydrate Structure Database (CCSDB), which was developed and maintained by the Complex Carbohydrate Research Center of the University of Georgia (USA). It was the largest effort during the 1990s to collect glycan structures mainly through retrospective manual extraction from literature.

Carbohydrate binding modules (CBMs): These have no enzymatic activity *per se* but are known to potentiate the activity of many enzymes described above by targeting specific forms of the substrate. CBMs are most often associated with other carbohydrate-active enzyme catalytic modules in the same polypeptide and can target different substrate forms depending on different structural characteristics.

Carbohydrate esterases (CE): Carbohydrate-active enzymes that remove some ester-based modifications present in *poly-* and *oligosaccharides* and may facilitate the action of *glycoside hydrolases*.

CAZy (Carbohydrate-Active enZymes database): Provides an enzyme classification based on the conservation of sequence and structural features and covers different enzyme activity classes dealing with the formation and breakdown of glycosidic bonds and associated activities (see Chapter 5).

Chondroitin sulfate: A sulfated *glycosaminoglycan* (GAG) composed of a chain of alternating *N*-acetylgalactosamine and glucuronic acid: $[-3)\text{-D-GalpNAc-(}\beta\text{1-4)-D-GlcpA-(}\beta\text{1-)]}_n$.

Complex Carbohydrate Structure Database (CCSDB): See *CarbBank*.

Congenital disorders of glycosylation (CDG): Inherited diseases that are caused by malfunctions of enzymes involved in *glycosylation*, such as *glycosyltransferases*.

Consortium for Functional Glycomics (CFG): US research initiative to investigate the role of carbohydrate-protein interactions at the cell surface in cell–cell communication.

Data mining: The process of automatically searching large volumes of data for patterns.

Docking: A computer method to predict the binding mode of a small ligand in a protein binding site.

Dermatan sulfate: A sulfated *glycosaminoglycan* (GAG) composed of alternating *N*-acetylgalactosamine and iduronic acid: [–3)-D-GalpNAc-(β1–4)-L-IdopA-(β1–]_{*n*}. Some of the hexuronic acid units may be GlcPA.

Dynamic programming: A method used in computer science to solve problems by iteratively finding the optimal solution to sub-problems of the original problem such that intermediate solutions do not need to be recomputed. The optimal solution to the original problem is thus guaranteed to be found.

EUROCarbDB: European initiative to develop an open-access database for carbohydrate structures and analytical data (MS, HPLC and NMR).

Extracellular matrix: The extracellular part of tissues. It is composed of a mesh of fibrous proteins and *glycosaminoglycans*.

Fucose (Fuc): Trivial name for 6-deoxygalactose; predominantly present in the L-configuration.

Furanose: A *monosaccharide* having a five-membered ring (four carbon atoms and one oxygen atom).

GlycanBuilder: Glycan structure drawing tool developed by the *EUROCarbDB* project. GlycanBuilder is a Java application that can run locally or within a website in a platform-independent manner.

Glycan profiling: The analysis of the carbohydrate chains that are found in a specific organism and tissue. Comparison of different samples can be used, for example, to identify carbohydrate motifs that can serve as biomarkers for diseases.

Glyco-bioinformatics/Glycoinformatics: A branch of bioinformatics that deals with carbohydrates. It includes the development of carbohydrate-specific algorithms and also the application of classical bioinformatics tools such as sequence alignments to identify and classify genes which encode for carbohydrate-active enzymes. Because sugars are not encoded directly in genomes, the enzymes that assemble complex oligosaccharides, and the lectins that recognize them, provide essential links between genomes and *glycobiology*.

Glycobiology: The study of naturally occurring carbohydrates, their function, structure, biosynthesis, and interactions with proteins or with other (bio-)molecules.

Glycoconjugates: Generic term for carbohydrates that are linked to other components, such as proteins (forming *glycoproteins*) or lipids (forming *glycolipids* or *lipopolysaccharides*).

GlycoCT: Unique sequence format for carbohydrate structures.

Glycoforms: See *glycoproteins*.

Glycogen: A *polysaccharide* that is the principal storage form of glucose in animal organisms.

Glycogene: A gene that encodes for a protein, which is involved in sugar metabolism, such as *glycosyltransferases* or *glycoside hydrolases*. Not to be confused with *glycogen*.

Glycolipids: Lipids that are covalently linked to carbohydrates.

Glycome: The glycan complement of the cell or tissue as expressed by a genome in time and space. It includes all types of *glycoconjugates*: *glycoproteins*, *glycolipids*, *peptidoglycans*, *lipopolysaccharides*.

Glycomics: An integrated systems approach to study structure–function relationships of complex carbohydrates – the *glycome* – produced by an organism such as human or mouse. Rapid and sensitive high-throughput analytical methods employing *mass spectrometry* (MS) and *high-performance liquid chromatography* (HPLC) techniques are currently applied to provide information on the glycan repertoire of cells, tissues and organs. One of the aims of the emerging glycomics projects is to create a cell-by-cell catalog of *glycosyltransferase* expression and detected glycan structures.

Glycoproteins: Proteins that carry covalently linked carbohydrate chains. Glycoproteins generally exist as populations of glycosylated variants – called glycoforms – of a single polypeptide. Although the same glycosylation machinery is available to all proteins in a given cell, most glycoproteins emerge with characteristic glycosylation patterns and heterogeneous populations of glycans at each glycosylation site.

Glycoproteomics: The study of the *glycome* attached to proteins. Often the terms *glycomics* and *glycoproteomics* are used interchangeably. However, strictly glycoproteomics includes only the analysis of the glycan part attached to proteins, whereas the term *glycomics* includes all types of carbohydrates present in a cell including *glycolipids* and *glycosaminoglycans*.

Glycosaminoglycans (GAGs): Long, unbranched *polysaccharides* containing a repeating disaccharide unit. GAGs are highly negatively charged molecules, with an extended conformation that imparts high viscosity to the solution. The up to more than 100 individual *monosaccharide* units of a GAG chain can be sulfated in variable positions and quantities. GAGs are part of the *extracellular matrix* or are covalently attached to *proteoglycans*. GAGs of physiological significance are *hyaluronic acid*, *dermatan sulfate*, *chondroitin sulfate*, *heparin*, *heparan sulfate*, and *keratan sulfate*.

GLYCOSCIENCES.de portal: A collection of databases and tools related to glycobiology.

Glycosidases: See *glycoside hydrolases*.

Glycoside hydrolases (GH): Also called glycosidases. Enzymes that catalyze the hydrolysis of a glycosidic linkage, yielding two smaller carbohydrates.

Glycosphingolipid: A *glycolipid*, the lipid component of which is the amino alcohol sphingosine.

Glycosylation: Covalent linkage of carbohydrate chains to a protein. Classified into *N-glycosylation*, *O-glycosylation*, and *C-mannosylation* depending on the type of atom to which the carbohydrate is linked (see Section 8.1).

Glycosyltransferases (GT): Enzymes involved in glycan biosynthesis. They typically act by adding *monosaccharides* one at a time to specific positions on specific precursors. The structure of glycans is determined by the succession in which GT assemble monosaccharides into linear and branched sugar chains. Many, but not all, of these enzymes are found within the ER–Golgi pathway for export of newly synthesized glycoconjugates (see Chapter 5 and Section 8.1).

GlycoPeakFinder: Web-based software tool for semi-automatic composition analysis of carbohydrates from mass spectrometric data. This functionality is also available in GlycoWorkbench.

GlycoWorkbench: A standalone (Java) desktop application for the routine annotation and computer-assisted (manual) interpretation of mass spectra.

GLYDE-II: XML exchange format for carbohydrate data.

Golgi apparatus: An organelle of eukaryotic cells, which processes and packages macromolecules synthesized by the cell, mainly proteins and lipids. As part of the secretory pathway, it plays an important role in processing the carbohydrate moieties of *glycoproteins*.

Heparan sulfate (HS): A *glycosaminoglycan* (GAG) that is highly similar in structure to *heparin*. It consists of alternating *N*-acetylglucosamine and glucuronic acid residues: $[-4)\text{-D-GlcpNAc-(}\beta 1\text{-4)\text{-D-GlcpA-(}\beta 1\text{-)}_n$. Glucuronic acid can be epimerized to iduronic acid. Among other positions, the GlcpNAc units can be sulfated at position 2, i.e. the acetyl part of the *N*-acetyl group is replaced by a sulfate group, resulting in D-GlcpNSO_3 .

Heparin: A sulfated *glycosaminoglycan* (GAG) composed of alternating *N*-acetylglucosamine and iduronic acid: $[-4)\text{-D-GlcpNAc-(}\beta 1\text{-4)\text{-D-IdopA-(}\beta 1\text{-)}_n$. As in *heparan sulfate*, the GlcpNAc units can be sulfated at position 2 and the uronic acid component can be either iduronic acid or glucuronic acid, but in heparin iduronic acid is predominant. Heparin is used clinically as an anticoagulant. It has the highest density of negative charges of all known biological molecules.

Heptose: A *monosaccharide* with seven backbone carbon atoms.

Hexosamine: An amino-substituted *hexose*. In most cases, the amino group is situated at position 2 of the hexose.

Hexose: A *monosaccharide* with six backbone carbon atoms.

High-performance lipid chromatography (HPLC): Also called high-pressure lipid chromatography. Experimental technique used to quantify, separate, or identify compounds. In combination with *glycosidases* it can be used for sequencing carbohydrate chains (see Chapter 11).

Hyaluronic acid (hyaluronan): A *glycosaminoglycan* (GAG) composed of $[-3)\text{-D-GlcpNAc-(}\beta 1-4)\text{-D-GlcpA-(}\beta 1-)]_n$ repeating units. Hyaluronan is a main component of the *extracellular matrix* and the only GAG that is exclusively non-sulfated.

KegDraw: Glycan structure drawing tool developed by the KEGG project. KegDraw is a Java application, so it runs locally in a platform-independent manner.

KEGG Glycan: The carbohydrate related subpart of the Kyoto Encyclopedia of Genes and Genomes (KEGG) portal.

Keratan sulfate: A sulfated *glycosaminoglycan* (GAG) composed of alternating *N*-acetylgalactosamine and galactose: $[-4)\text{-D-GalpNAc-(}\beta 1-3)\text{-D-Galp-(}\beta 1-)]_n$. Keratan sulfate differs from other GAGs in that its repeating unit contains a *hexose* instead of a hexuronic acid in addition to the *hexosamine* residue.

Lectins: Carbohydrate-binding proteins that are neither enzymes nor antibodies. Most lectins specifically recognize a certain carbohydrate motif. In the microbial world lectins tend to be called by other names, such as hemagglutinins, adhesins, and toxins (see Chapter 21).

Lipopolysaccharide (LPS): A major component of the outer membrane of Gram-negative bacteria. LPS consists of a lipid and a polysaccharide moiety that are covalently linked. It protects the membrane from certain kinds of chemical attack and is an endotoxin that induces a strong immune response in most animals.

Mass spectrometry: Experimental technique to determine the composition of molecules such as carbohydrate chains (see Chapters 12 and 13).

Molecular dynamics (MD) simulation: Computer technique to simulate the motions of a molecular system at a given temperature over time (see Chapter 19). The result of an MD simulation is a trajectory, which represents a statistical conformational ensemble from which macroscopic molecular properties can be calculated. MD simulations are frequently performed in order to support the interpretation of NMR spectra (see Chapter 20).

Molecular modeling: Computer technique to generate three-dimensional (3D) models of molecules (see Chapter 19).

Microarray: Molecular biological technique, which allows thousands of single analytical experiments to be performed in parallel using small amounts of biological material. Depending on the molecules analyzed, they are classified as DNA-, peptide- or glycan-array.

Monosaccharide: The basic unit of carbohydrate chains, a polyhydroxy aldose or polyhydroxy ketose with at least three carbon atoms. Most naturally occurring monosaccharides have between five and nine backbone carbon atoms and can be modified by substituents.

Mucins: A highly *O*-glycosylated class of proteins. Mucins are a major component of mucosal surfaces such as mucous membranes, where the carbohydrate moieties are involved in binding of water.

Murein: See *peptidoglycan*

Neural network: See *artificial neural network*.

Neuraminic acid: Trivial name for 5-amino-3-deoxy-D-glycero-D-galactononulonic acid; its 5-*N*-acetylated form is the most common *sialic acid*.

***N*-Glycosylation:** Covalent linkage of a carbohydrate chain to a protein via the N δ 2 atom of an Asn side-chain (see Section 8.1).

Nonose: A *monosaccharide* with nine backbone carbon atoms.

Nuclear magnetic resonance (NMR): Experimental technique to determine molecular 3D structures (see Chapters 15–17 and 20).

Octose: A *monosaccharide* with eight backbone carbon atoms.

***O*-Glycosylation:** Covalent linkage of a carbohydrate chain to a protein via a free hydroxyl group of an amino acid side-chain (mainly Ser or Thr; see Chapter 8.1).

Oligosaccharide: A carbohydrate chain that consists of a number of *monosaccharide* units. This term is used for carbohydrates of a variable range of size, as the term “*polysaccharide*” is usually only used for carbohydrates that are composed of repeating units.

Oligosaccharyl transferase complex (OST): Complex of enzymes catalyzing the transfer of a dolichol linked *N*-glycan precursor to a protein as the initial step of *N*-glycosylation (see Section 8.1).

Pentose: A *monosaccharide* with five backbone carbon atoms.

Peptidoglycan/murein: A polymer that is found in the cell wall of *Eubacteria*. It consists of linear carbohydrate chains of alternating (1–4)-linked β -D-GlcpNAc and β -D-MurpNAc (muramic acid) residues. The muramic acid units of different chains are linked via oligopeptides, which gives the molecule a mesh-like structure.

Polysaccharide: A carbohydrate that is composed of repeating units, usually larger than 10 monosaccharides (see Chapter 14).

Polysaccharide lyases (PLs): Enzymes that cleave the glycosidic bonds of uronic acid-containing polysaccharides by a β -elimination mechanism. Many PLs have biotechnological and biomedical applications and are among the enzymes having most biochemically characterized examples present in CAZy (see Chapter 5).

Proteoglycans: A heavily glycosylated class of proteins. They consist of a core protein and a number of *O*-glycosidically bound *glycosaminoglycans* (GAGs). The carbohydrate mass can be significantly larger than the mass of the core protein. Proteoglycans are part of the *extracellular matrix* and a major component of cartilage.

Pyranose: A *monosaccharide* having a six-membered ring (five carbon atoms and one oxygen atom).

Rhamnose: Trivial name for 6-deoxymannose; predominantly present in the L-configuration.

Secretory pathway: Describes different methods that cells use to transport proteins to the outside, usually from the endoplasmic reticulum (ER) via the *Golgi apparatus* to vesicles that fuse with the cell membrane.

Sialic acids (Sia): Generic term for a family of acidic nonose monosaccharides. The most common sialic acid is *N*-acetylneuraminic acid (Neu5Ac, also sometimes called Neu5NAc, NeuAc, or NANA) (see Chapter 4).

Uronic acid: A *monosaccharide* with a carboxylate group at position 1.

Uronic acid: A *monosaccharide* with a carboxylate group at the terminal position (position 6 in the case of hexoses), such as glucuronic acid (GlcA) or iduronic acid (IdoA), usually negatively charged under physiological conditions.

Index

- α -D-GalNAc transferases 36–8, 45
- α -D-Manp 362, 366
- ab initio* methods 360, 364–5, 372
- adiabatic maps 343, 371–2
- advanced glycation end-products (AGEs) 183
- AFM *see* atomic force microscopy
- AGEs *see* advanced glycation end-products
- AMBER 344, 369, 375, 378, 381
- amino acids
 - glycosylation 153–7, 169–70
 - protein–carbohydrate interactions 435–44
- animal-specific lectins 417, 418–22
- anion exchange HPLC 211–12
- ANNs *see* artificial neural networks
- antigens
 - carbohydrate-active enzymes 101
 - carbohydrates 41
 - glycosylation 163
 - lectins 422
 - nuclear magnetic resonance 314–15
- API *see* Application Programming Interfaces
- Application Programming Interfaces (API) 127
- aromatic amino acids 153–4
- artificial neural networks (ANNs)
 - data processing 326–9
 - NeuroCarb 325–31
 - nuclear magnetic resonance 321–34
 - performance 331
 - prediction of glycosylation sites 164–5
 - supervised learning 322–5
 - test phase 329–30
 - training and validation phase 329
 - unsupervised learning 322–6
 - work phase 331
- atomic force microscopy (AFM) 346

- β -D-Galp 437–8, 443
- β -elimination of GAGs 272–3

- back-propagation algorithm 324–5
- bacterial
 - lectins 418
 - monosaccharides 30
 - oligosaccharides 76–7
 - polysaccharides 40–1, 296
- Bacterial Carbohydrate Structure Database (BCSDB) 50–1, 55–6, 66
- BCSDB *see* Bacterial Carbohydrate Structure Database
- big-PI 182
- BIOPSEL program 315
- BLAST-based analysis 95–6, 120, 126
- BLASTN searches 8
- BRENDA database 120, 122

- C-mannosylation 144, 171, 182–3
- C-type lectins 417, 419–20, 423, 425
- CabosML 50–1, 59–61
- calnexin (CNX) 146–8, 415, 419
- calreticulin (CRT) 146–7, 415
- Cambridge Structural Database (CSDB) 337, 390
- capillary electrophoresis (CE) 205–6, 210–11, 213, 281–3
- CarbBank 14–16, 50–1, 52–4, 299
- carbohydrate esterases (CEs) 93
- carbohydrate-active enzymes 91–118
 - abstract/symbolic representations 128
 - artificial neural networks 164–5
 - bioinformatics 94–6, 125–41
 - BLAST-based analysis 95–6, 120
 - BRENDA database 120, 122
 - C-mannosylation 171, 182–3
 - CAZy database 91–118, 120
 - classifications 92–4, 97–101
 - co-occurrence scores 135
 - Composite Maps 127–34, 138–40

- carbohydrate-active (*Contd.*)
- data collection 168–70
 - data mining 136–7
 - data-driven prediction methods 164–8
 - EC classification 93–4, 97
 - evaluation strategies 165–8
 - existing predictors 173–80, 182–3, 185
 - ExPASy ENZYME database 120, 122
 - family comparisons 101–14
 - FASTA-based analysis 120
 - genomics 137–8
 - glycation 170, 183–5
 - GlycoEnzyme 120–1
 - GlycoGene database 119–20
 - glycosylation 143–62, 163–92
 - glycosyltransferases 91–2, 97–115, 134–6, 139–40
 - GlySeq 151–9
 - GPI anchors 171, 180–2
 - KEGG database 119–20
 - KEGG GLYCAN 125–41
 - linkage-specific glycosylation 170–83
 - Markov models 136–7, 164
 - microarray analysis 136
 - Microarray Data 120–1
 - NCBI databases 120, 122
 - Protein Databank 120, 123
 - proteoglycans 150, 178–80
 - reaction libraries 134–5
 - sequence logos 169–70, 172–3, 175–7, 179–85
 - sequence motifs 172, 174–5, 177, 179–80, 182, 184
 - sequence/structure classification 99–101
 - stereochemical outcome 98–9, 102–14
 - structure prediction 134–6
 - Swiss-Prot/TrEMBL 120–1
- carbohydrate-binding molecules (CBMs) 93, 95, 101
- carbohydrates
- abstract/symbolic representations 49–50, 61–2
 - analytical complications 9–11
 - binding sites 422–3
 - bioinformatics 7–9, 195–201
 - biological roles 70–1, 72–3
 - biosynthetic pathways 5–7, 33, 76–8
 - composition 31–3
 - conformational analysis 337–57, 359–88, 389–412
 - databases 13–17, 33–4
 - descriptors and tools 61–4
 - digital representations 49–68, 80–2
 - disaccharide pairings 34
 - disease profile 13
 - drug and vaccine development 12–13
 - empirical structural classifications 34–44
 - evolution 78–9
 - experimental methods 195–201
 - extensible markup approaches 50–1, 58–61, 66
 - future developments 64–7, 80–2
 - glycolipids 38–9
 - glycomic mass spectrometry 257–68
 - high performance liquid chromatography 203–21
 - host defense mechanisms 74–6
 - host–pathogen interface 69–71, 74–8
 - lectins 415–31
 - life sciences research 3–4
 - linkage topologies 34
 - motifs 42–4
 - naturally occurring 28–30
 - nomenclature 24, 25–8
 - nuclear magnetic resonance 295–309, 311–12
 - pharmaceutical research 11–13
 - protein interactions 10
 - protein–carbohydrate interactions 415–31, 433–45
 - sequence formats 50–61
 - sialic acids 69–88
 - small molecule descriptors 62–4
 - species distribution 74
 - stereocodes 63–4
 - storage capabilities 50–1
 - structural heterogeneity 10
 - structure and diversity 23–47, 72–5
 - substituents 27–8
 - topological descriptors 51–2
 - see also* glycoproteins; monosaccharides; *N*-glycans; *O*-glycans; oligosaccharides; polysaccharides
- Cartoonist 245, 249
- CASPER program 301, 311–20, 344
- CAT *see* Conformational Analysis Tools
- Catalogue Library 246
- CAZy database 91–118, 120
- bioinformatics 94–6
 - BLAST-based analysis 95–6
 - classifications 92–4, 97–101
 - EC classification 93–4, 97
 - family comparisons 101–14

- glycosyltransferases 91–2, 97–115
 - sequence/structure classification 99–101
 - stereochemical outcome 98–9, 102–14
- CBMs *see* carbohydrate-binding molecules
- CCA *see* conformational clustering analysis
- CCSD *see* Complex Carbohydrate Structure Database
- CDGs *see* congenital disorders of glycosylation
- CE *see* capillary electrophoresis
- CEL-III 421
- cell–ECM interactions 269–70
- ceramide 38–9
- CEs *see* carbohydrate esterases
- CFG *see* Consortium for Functional Glycomics
- channels in conformational space analyzed by driver approach (CICADA) 345, 373
- CHARMM-type force fields 343–4, 375
- ChemDraw 126
- Chemical Function database 50–1, 57–9, 127
- chemically induced dynamic nuclear polarization (CIDNP) 346
- chondroitin polymerase 101
- chondroitin sulfate 42, 178, 271, 278–9
- CICADA 345, 373
- CIDNP *see* chemically induced dynamic nuclear polarization
- CNX *see* calnexin
- co-occurrence scores 135
- coarse-grain force fields 382
- Complex Carbohydrate Structure Database (CCSD) 14–16, 50–1, 52–4, 299
- Composite Maps 127–34, 138–40
- Composite Reaction Maps (CRM) 127–32, 138
- Composite Structure Maps (CSM) 129–34, 138–40
- compositional analysis tools 240–4
- conformational analysis
 - accessible conformational space 370–9
 - carbohydrates 337–57, 359–88, 389–412
 - comparisons with theoretical data 401–7
 - databases 390–402
 - distance mapping 390, 406
 - energy maps 340–2, 343, 364, 366–8, 370–3, 378, 402, 406
 - exo*-anomeric effects 342–3, 360, 363–5, 367
 - exocyclic groups 368–70, 393–5
 - glycoproteins 344–6, 379–82, 391
 - glycosidic linkages 339–40, 344, 363–8, 370, 378, 390, 395–400, 403
 - glycosylation 346, 379, 391–3
 - H–H distances 404–6
 - hydrogen bonding 368–70, 376–8
 - Karplus equations 390
 - Metropolis Monte Carlo simulations 372–4, 404
 - molecular dynamics 370, 372–3, 374–9, 404–5
 - monosaccharides 337, 339, 346, 361–3, 365, 395–7
 - oligosaccharides 337, 343–6, 360, 363, 368, 373, 377, 402–3
 - pioneering work 339–42
 - polysaccharides 337–8, 339–42, 358
 - predictive methods 359–88
 - torsion angles 403–4
- Conformational Analysis Tools (CAT) 372
- conformational clustering analysis (CCA) 373
- conformational ensembles 361
- congenital disorders of glycosylation (CDGs) 13, 140, 146–7
- Consortium for Functional Glycomics (CFG) 11, 61–2, 120–1
 - carbohydrate-active enzymes 132
 - glycomic mass spectrometry 244
 - glycosaminoglycans 283
- corporate identity 4
- CRM *see* Composite Reaction Maps
- cross-relaxation experiments 404–6
- cross-validation 165–6
- CRT *see* calreticulin
- CSEARCH 297
- CSM *see* Composite Structure Maps
- cyclic graphs 49
- cycloamyloses 337
- cyclodextrins 337
- cytoplasmic proteins 177–8

- data quality 200
- data-driven prediction methods 164–8
- deglycosylated glycopeptides 238–9
- 3-deoxy-D-*arabino*-heptulose-7-phosphate 79
- dermatan sulfate 42, 178, 271
- DictyOGlyc 177
- digraphs 49
- disaccharide pairings 34
- distance mapping 390, 406
- DKFZ *see* German Cancer Research Center
- DNA sequences 23–4

- EC *see* Enzyme Classification
- electron capture dissociation (ECD) 238

- electrospray ionization–mass spectrometry
 (ESI–MS) 228–9, 235–6, 250, 276–7
 endoglycosidases 81
 endoplasmic reticulum (ER) 5, 7
 carbohydrates 35, 37
 glycosylation 145–50, 177, 180
 lectins 416
 enzymatic digestion 199–200
 enzymatic sequencing 230, 234
 Enzyme Classification (EC) system 93–4, 97
 ENZYME database 120, 122
 Enzymes and Metabolic Pathways database 96
 EPO *see* erythropoietin
 ER *see* endoplasmic reticulum
 ERGIC-53 420–1, 424
 erythropoietin (EPO) 13
 ESI–MS *see* electrospray ionization–mass
 spectrometry
 eukaryotes 176–7, 415–16
 EUROCarbDB 15–17, 283
 see also Glyco-Peakfinder; GlycoWorkbench
exo-anomeric effects 342–3, 360, 363–5, 367
 exocyclic groups 368–70, 393–5
 exoglycosidases 207–10, 230, 234, 321
 ExPASy ENZYME database 120, 122
 extensible markup (XML) 50–1, 58–61, 66

 FAB *see* fast atom bombardment
 Factor H 77
 false negatives/positives 166–8
 fast atom bombardment (FAB) 224
 FASTA-based analysis 120
 FEP *see* free energy perturbation
 fibroblast growth factors (FGFs) 417, 421
 fluorescence HPLC
 experimental methods 198
 glycomic mass spectrometry 231–4
 glycoproteins 206–7, 209–10
 sialic acids 81
 Fourier transform (FT) mass spectrometry
 235, 238
 free energy perturbation (FEP) simulations 345
 FT *see* Fourier transform
 fungal lectins 416, 418, 425

 GAG–protein interactions 286–7
 GAGs *see* glycosaminoglycans
 galectins 416–17, 420, 425
 GalNAc
 carbohydrate-active enzymes 144, 149,
 151, 157
 glycomic mass spectrometry 224, 240
 glycosaminoglycans 279
 linkage-specific glycosylation 174–6
 GalNAc transferases 36–8, 45
 gas chromatography–mass spectrometry
 (GC–MS) 224
 GEGOP 345, 373
 gel electrophoresis 204–5, 229, 235, 237–9,
 274–6
 geometry of glycoproteins (GEGOP) 345, 373
 GEP *see* group epitope mapping
 German Cancer Research Center (DKFZ) 15, 54
 GHs *see* glycoside hydrolases
 GlcNAc
 conformational analysis 396–9, 402
 glycomic mass spectrometry 224, 240
 high performance liquid chromatography
 204, 208–9
 linkage-specific glycosylation 173,
 176–8, 185
 nuclear magnetic resonance 306
 protein–carbohydrate interactions 423, 426,
 437–8
 GlcNAc₂
 carbohydrate-active enzymes 144–50
 conformational analysis 362, 366, 378
 global minima 360
 glucans 40–1
 glucose units (gu) 215
 GlycamWeb 379
 Glycan Data Exchange *see* GLYDE II
 GLYCAN database
 carbohydrate-active enzymes 125–41
 conformational analysis 400
 glycomic mass spectrometry 241, 244
 glycosaminoglycans 283
 GlycanBuilder 262–3
 glycans *see* carbohydrates; glycosaminoglycans;
 N-glycans; *O*-glycans
 glycation 170, 183–5
 GLYCH 246
 Glyco-Peakfinder 257–62, 265–6
 GlycoComp 240
 GlycoCT 50–1, 65–6
 GlycoEnzyme 120–1
 glycoenzymes 4–9
 glycoforms 144, 196
 glycofragment mass fingerprinting (GMF)
 248–9
 GlycoGene database 8, 119–20
 glycogenes 4–9

- glycoglycerolipids 39
- glycoinformatics 24
- glycolipids
 - carbohydrate-active enzymes 138
 - conformational analysis 346
 - structure and diversity 38–9
- GlycomeDB 24, 30–1, 33–4
- glycomic mass spectrometry 223–56
 - compositional analysis 258–62, 265–6
 - compositional analysis tools 240–4
 - data handling/analysis 239–51
 - derivatization 229–34
 - downstream analysis of glycan release 227–34
 - fragment annotation 260–8
 - GlycanBuilder 262–3
 - Glyco-Peakfinder 257–62, 265–6
 - glycopeptides 227, 237–9
 - glycoproteomics 224, 234–5
 - GlycoWorkbench 257–8, 260–8
 - in silico* fragmentation 263–4
 - linkage analysis 249–51
 - monosaccharides 228, 240–1, 249
 - oligosaccharides 224, 226, 227–36, 240
 - parallel approach 239
 - peak annotation 265
 - sample preparation 224–7
 - semi-automatic interpretation 257–68
 - sequence analysis 244–9
 - validation 250–1
- GlycoMod 240, 245
- glycopeptides 227, 237–9
- glycoproteins
 - conformational analysis 344–6, 379–82, 391
 - intracellular trafficking 415–16
 - protein folding 145–8
 - protein sequences 23–4
 - protein–carbohydrate interactions 415–31
 - structure and diversity 35–8
- glycoproteomics 224, 234–5
- glycosaminoglycans (GAGs) 11–12
 - analytical methods 274–8
 - bioinformatics 278–86
 - biological function 269–70
 - carbohydrate-active enzymes 149–50
 - chemical modification 272–3
 - composition/mass profiles 285
 - databases 283–6
 - depolymerization/modification 272–4
 - enzymatic characterization 273–4, 275
 - future developments 286–7
 - linkage-specific glycosylation 178–9
 - property-encoded nomenclature 276, 278–83
 - sequence motifs 285–6
 - sequencing 278–86
 - structural characterization 270–2
 - structure 40, 42, 270
 - structure–function relationships 269–94
- GLYCOSCIENCES.de 15
 - conformational analysis 371, 379–80, 396
 - glycomic mass spectrometry 244–5, 260
 - glycosylation 151–9
 - nuclear magnetic resonance 301–6
 - sequence formats 54–5, 66
 - SWEET/SWEET-II 345, 406
- glycoside hydrolases (GHs) 92, 98–101, 113–14
- glycosphingolipids 38–9
- GlycoSuiteDB 15, 50–1, 56, 240–1, 244–5
- glycosylation 5–7
 - amino acids 153–7, 169–70
 - artificial neural networks 164–5
 - biological relevance 143–4
 - C*-mannosylation 144, 171, 182–3
 - carbohydrate-active enzymes 143–62, 163–92
 - conformational analysis 346, 379, 391–3
 - congenital disorders 140, 146–7
 - data collection 168–70
 - data-driven prediction methods 164–8
 - evaluation strategies 165–8
 - existing predictors 173–80, 182–3, 185
 - functions of *O*-glycans 150
 - glycation 170, 183–5
 - glycomic mass spectrometry 223–4, 227, 236, 239
- GlySeq 151–9
- GPI anchors 171, 180–2
- high performance liquid chromatography 203–4
- linkage-specific 170–83, 185
- Markov models 164
- nuclear magnetic resonance 312–14, 316–17, 321
- prediction methods 163–92
- protein folding 145–8
- proteoglycans 150, 178–80
- sequence dependence 151–9
- sequence logos 169–70, 172–3, 175–7, 179–85
- sequence motifs 172, 174–5, 177, 179–80, 182, 184
- types 144
- see also* *N*-glycosylation; *O*-glycosylation

- glycosylphosphatidylinositols (GPIs) 13, 39, 171, 180–2
- glycosyltransferases (GTs)
- biosynthetic pathways 5, 7–8
 - carbohydrate-active enzymes 134–6, 139–40, 145, 170–1, 177
 - carbohydrates 3
 - CAZy database 91–2, 97–115
 - EC classification 97–9
 - family comparisons 101–14
 - glycomic mass spectrometry 228
 - sequence/structure classification 99–101
 - stereochemical outcome 98–9, 102–14
- GlycoWorkbench 257–8, 260–8
- GLYDE-II 16–17, 50–1, 59–60, 66–7
- GlyProt 379–80
- GlySeq 151–9
- GlyVicinity 434–44
- GMF *see* glycofragment mass fingerprinting
- Golgi apparatus 5
- carbohydrates 35, 37
 - glycosylation 145–50
 - lectins 416
- GPI *see* glycosylphosphatidylinositols
- GPI-SOM 182
- GROMOS 344, 375
- group epitope mapping (GEP) 346
- GTs *see* glycosyltransferases
- gu *see* glucose units
- HA *see* human influenza virus
- hard-sphere *exo*-anomeric (HSEA) program 342–3, 345, 371
- Hassel–Ottar effect 340–1, 361
- HCA *see* hydrophobic cluster analysis
- HCGP-39 *see* human cartilage glycoprotein-39
- heparan sulfate
- glycosylation 149–50, 178
 - structure 42
 - structure–function relationships 270–1, 276, 278–82
- heparosan synthase 101
- hepatocyte growth factor/scatter factor (HGF/SF) 417, 420
- hidden Markov models (HMMs) 136–7, 164
- Hierarchical Organization of Spherical Environments (HOSE) 297–8
- high performance liquid chromatography (HPLC)
- capillary electrophoresis 205–6, 210–11, 213
 - carbohydrates 81, 203–21
 - data analysis 213–16
 - display options 215–16
 - experimental methods 196–7, 198–200, 204–10
 - glucose units 215
 - glycan labeling 205–6, 207
 - glycan profiling 206
 - glycan quantification 206–7
 - glycan release 204–5, 227–8, 231–4
 - glycan sequencing 207–10
 - glycosaminoglycans 277
 - glycosylation 203–4
 - reversed phase 213
 - separation techniques 211–13
- high-pH anion exchange with pulsed amperometric detection (HPAEC-PAD) 212–13
- high-temperature molecular dynamics (HTMD) 372
- HMMs *see* hidden Markov models
- HOSE *see* Hierarchical Organization of Spherical Environments
- host–pathogen interface 69–70
- HPAEC-PAD *see* high-pH anion exchange with pulsed amperometric detection
- HPLC *see* high performance liquid chromatography
- HSEA *see* hard-sphere *exo*-anomeric
- HTMD *see* high-temperature molecular dynamics
- human cartilage glycoprotein-39 (HCGP-39) 417, 419–20
- human influenza virus (HA) 10–11
- hyaluronan 42
- hyaluronan synthase 101
- hyaluronic acid 271
- hydrazinolysis 81
- hydrogen bonding 368–70, 376–8
- hydrophobic cluster analysis (HCA) 94
- I-type lectins 417, 422
- InChI 63
- independent linkage approximation 368, 370
- infrared multiphoton dissociation (IRMPD) 235, 238
- International Union of Biochemistry and Molecular Biology (IUBMB) 51–2, 122
- International Union of Pure and Applied Chemistry (IUPAC)
- carbohydrates 50, 51–3, 55–6, 59
 - conformational analysis 396

- high performance liquid chromatography 215–16
- nomenclature 24, 25–8
- nuclear magnetic resonance 299, 317–18
- invertebrate lectins 417
- IPSA *see* isolated spin-pair approximation
- IRMPD *see* infrared multiphoton dissociation
- isolated spin-pair approximation (IPSA) 405
- IUBMB *see* International Union of Biochemistry and Molecular Biology
- IUPAC *see* International Union of Pure and Applied Chemistry

- Karplus equations 390, 403–4
- KCaM 125–7, 129
- KCF database 50–1, 57–9, 127
- Kdn 79, 80
- Kdo 78–9
- KegDraw 126–7
- KEGG *see* Kyoto Encyclopedia of Genes and Genomes
- keratan sulfate 42, 271
- 2-keto-3-deoxy-D-manno-octulosonic acid 78–9
- KO database 120, 125
- Kohonen networks 325–6
- Kohonen self-organizing maps 164
- Kyoto Encyclopedia of Genes and Genomes (KEGG) 15–16
 - carbohydrate-active enzymes 119–20, 125–41
 - Chemical Function database 50–1, 57–9, 127
 - co-occurrence scores 135
 - Composite Maps 127–34, 138–40
 - data mining 136–7
 - GENES 134
 - genomics 137–8
 - GLYCAN database 125–41, 241, 244, 283, 400
 - glycosyltransferases 134–6
 - KCaM 125–7, 129
 - KegDraw 126–7
 - Markov models 136–7
 - microarray analysis 136
 - Orthology database 120, 125
 - PATHWAY database 5–7
 - reaction libraries 134–5
 - structure prediction 134–6

- laser-induced desorption (LID) 261
- LC-MS *see* liquid chromatography–mass spectrometry
- lectin binding assays 198

- lectins
 - animal-specific 417, 418–22
 - carbohydrate recognition 422–7
 - conformational analysis 363
 - divergent/covergent evolution 420–2
 - eukaryotes 415–16
 - fold types 418–22
 - glycoprotein targeting 415–16
 - glycosylation 145
 - plant-specific 416, 418–22, 424–6
 - prokaryotes 418
 - protein–carbohydrate interactions 415–31
 - quaternary structure and avidity 423–6
 - sialic acids 76–8
 - viruses 418
- legionaminic acid 79
- LID *see* laser-induced desorption
- linear interaction energy (LIE) 346
- Linear Notation for Unique Description of Carbohydrate Sequences (LINUCS) 50–1, 54, 395–6
- LinearCode 50–1, 56–8
- linkage analysis 249–51
- linkage descriptors 299
- linkage topologies 34
- linkage-specific glycosylation 170–83, 185
- LINUCS 50–1, 54, 395–6
- lipooligosaccharides (LOS) 76–7
- lipopolysaccharides (LPS)
 - host–pathogen interface 76, 79
 - nuclear magnetic resonance 314, 318
 - structure and diversity 41, 43
- liquid chromatography–mass spectrometry (LC-MS) 227, 235, 238–9, 248–9, 274–6
- liquid secondary ion–mass spectrometry (LSI-MS) 224
- LOS *see* lipooligosaccharides
- LPS *see* lipopolysaccharide
- LSI-MS *see* liquid secondary ion–mass spectrometry

- MALDI *see* matrix-assisted laser desorption/ionization
- mammalian glycosylation 174–6
- mammalian monosaccharides 29–30
- mass spectrometry (MS)
 - carbohydrates 3, 204, 207, 210
 - compositional analysis 240–4, 258–62, 265–6
 - data handling/analysis 239–51
 - derivatization 229–34

- mass spectrometry (*Contd.*)
 downstream analysis of glycan release
 227–34
 experimental methods 196, 198–200, 204,
 207, 210
 fragment annotation 260–8
 GlycanBuilder 262–3
 Glyco-Peakfinder 257–62, 265–6
 glycomics 223–56, 257–68
 glycopeptides 227, 237–9
 glycoproteomics 224, 234–5
 glycosaminoglycans 276–7
 GlycoWorkbench 257–8, 260–8
in silico fragmentation 263–4
 linkage analysis 249–51
 monosaccharides 228, 240–1, 249
 oligosaccharides 224, 226, 227–36, 240
 parallel approach 239
 peak annotation 265
 sample preparation 224–7
 semi-automatic interpretation 257–68
 sequence analysis 244–9
 validation 250–1
- matrix-assisted laser desorption/ionization
 (MALDI) 229, 231, 234–9, 250, 259,
 276–7, 280–1
- Matthews correlation coefficient
 166, 173, 185
- MC *see* Monte Carlo
- MCMM *see* Monte Carlo Multi Minimum
- MD *see* molecular dynamics
- Medline 168
- metabolic oligosaccharide engineering 13
- Metropolis Monte Carlo (MMC) simulations
 372–4, 404
- microarray analysis 11, 136
- Microarray Data 120–1
- microbial oligosaccharides 77
- milk oligosaccharides 75–6
- MM2CARB method 343–4
- MMC *see* Metropolis Monte Carlo
- molecular dynamics (MD) 342, 344–7, 370,
 372–9, 404–5
- monoclonal antibodies 321
- MonosaccharideDB 64–5
- monosaccharides
 carbohydrate-active enzymes 132, 137, 149
 conformational analysis 337, 339, 346,
 361–3, 365, 395–7
 digital representations 56–7, 60, 62–6
 glycomic mass spectrometry 228, 240–1, 249
 high performance liquid chromatography
 207–8, 215
 naturally occurring 28–30
 nomenclature 25–8
 nuclear magnetic resonance 295–6, 306,
 311–13, 321, 326, 328–9, 332
 protein–carbohydrate interactions 422–3
 small molecule descriptors 62–4
 structure and diversity 23–4, 25–30
- Monte Carlo (MC) simulations 345, 372–4, 404
- Monte Carlo Multi Minimum (MCMM) 373
- MS *see* mass spectrometry
- MS/MS *see* tandem mass spectrometry
- mucins 75–6, 174–6
- N*-acetylneuraminic acid (Neu5Ac) 72–3,
 76–9, 80
- N*-glycans 4
 conformational analysis 359, 362, 368, 395,
 397–9
 glycomic mass spectrometry 227–9, 240,
 244, 249
 nuclear magnetic resonance 302–8
 protein–carbohydrate interactions 415–16
 structure and diversity 35–6, 45
- N*-glycosylation 5–7
 carbohydrate-active enzymes 128, 130–1,
 138, 143–4, 145–8, 153–9
 conformational analysis 391–2
 linkage-specific 171, 172–4
- National Center for Biotechnology Information
 (NCBI) 120, 122
- negative hyperconjugation 365
- NetCGlyc 183
- NetGlycate 185
- NetNGlyc 173–4
- NetOGlyc 175
- NetPGlyc 179–80
- Neu5Ac *see* *N*-acetylneuraminic acid
- neural networks *see* artificial neural networks
- NeuroCarb 325–32
 data preprocessing 326–9
 performance 331
 test phase 329–31
 training and validation 329
 work phase 330–1
 workflows 325, 327
- nitrous acid cleavage 272–3
- NMR *see* nuclear magnetic resonance
- NOEs *see* nuclear Overhauser effects
- nuclear magnetic resonance (NMR)

- advantages/disadvantages 295–6
- artificial neural networks 321–34
- automatic procedures 308, 311–20
- bioinformatics 296–9
- carbohydrates 295–309, 311–12
- CASPER program 301, 311–20, 344
- chemical shift estimation 301–8, 312–14, 328–9
- comparisons with theoretical data 402–7
- conformational analysis 338, 343–7, 365, 379, 389–90, 402–7
- coupling constants 403–4
- cross-relaxation experiments 404–6
- data handling/analysis 315–16
- databases 296, 298–9, 316–17
- experimental methods 197–200
- glycomic mass spectrometry 249
- glycosaminoglycans 274, 277, 281–3
- glycosylation 312–14, 316–17, 321
- H–H distances 404–6
- monosaccharides 295–6, 306, 311–13, 321, 326, 328–9, 332
- NeuroCarb 325–32
- oligosaccharides 306, 312–14, 316, 318–19, 321, 330–2
- performance measures 318
- polysaccharides 296, 306, 312–14, 319
- protein–carbohydrate interactions 433
- spectral matching 308
- spectral searches 301
- stereochemistry 297–8
- structure generation 314–15, 319
- SugaBase 299–301
- torsion angles 403–4
- nuclear Overhauser effects (NOEs) 277, 338, 345–6, 374, 389–90, 403, 405–6
- nuclear proteins 177–8
- NXS sequons 157, 169, 172–3
- NXT sequons 157, 169, 172–3
- O*-glycans
 - conformational analysis 359, 395, 397–9
 - glycomic mass spectrometry 227–9, 239–40, 244
 - nuclear magnetic resonance 308
 - structure and diversity 36–7
- O-GLYCOBASE 152–3, 168
- O*-glycosylation 5–7, 36–8
 - carbohydrate-active enzymes 138, 144, 148–50, 153–9
 - data collection 168
 - linkage-specific 171–2, 174–8
- O*-mannosylglycans 149
- Oligosaccharide Subtree Constraint Algorithm (OSCAR) 247, 250
- oligosaccharides
 - assembly level 30–4
 - composition 31–3
 - conformational analysis 337, 343–6, 360, 363, 368, 373, 377, 402–3
 - database analysis 33–4
 - digital representations 52–6, 57–61
 - empirical structural classifications 34–44
 - experimental methods 195–6
 - glycomic mass spectrometry 224, 226, 227–36, 240
 - glycosylation 145, 148–9, 151
 - host–pathogen interface 69, 75–8, 81
 - nuclear magnetic resonance 306, 312–14, 316, 318–19, 321, 330–2
 - protein–carbohydrate interactions 422–4
 - structure and diversity 30–44
 - see also* glycosaminoglycans
- oligosaccharyl transferase (OST) 35–6, 144, 146
- ordered tree Markov model (OTMM) 137
- Orthology database 120, 125
- OSCAR *see* Oligosaccharide Subtree Constraint Algorithm
- OST *see* oligosaccharyl transferase
- OTMM *see* ordered tree Markov model
- Oxford GlycoSciences 249
- PAD *see* pulsed amperometric detection
- particle-mesh Ewald (PME) 381
- PATHWAY database 5–7
- PBC *see* periodic boundary conditions
- PCR *see* polymerase chain reaction
- PDB *see* Protein Data Bank
- PDI *see* protein disulfide isomerase
- PEN *see* property-encoded nomenclature
- pentraxins 420, 425
- peptidoglycans 40–1
- peracetylation 229–31
- periodate oxidation 230–1, 272–3
- periodic boundary conditions (PBC) 381
- permethylation 229–31
- PES *see* potential energy surfaces
- PFOS *see* potential function for oligosaccharides
- PGs *see* proteoglycans
- plant polysaccharides 41–2
- plant-specific lectins 416, 418–22, 424–6

- PLs *see* polysaccharide lyases
- PME *see* particle-mesh Ewald
- PNGase F/A 81, 227–8, 239
- polymerase chain reaction (PCR) 7, 9, 296
- polypeptides 145
- polysaccharide lyases (PLs) 93
- polysaccharides
- conformational analysis 337–8, 339–42, 358
 - digital representations 52–6, 57–61
 - host–pathogen interface 76
 - nuclear magnetic resonance 296, 306, 312–14, 319
 - structure and diversity 24, 33, 39–42
 - see also* glycosaminoglycans
- positive predictive value (PPV) 166
- post-source decay (PSD) 231, 235
- post-translational modification (PTM) 23
- potential energy surfaces (PES) 373
- potential function for oligosaccharides (PFOS) 371
- PPV *see* positive predictive value
- probabilistic sibling-dependent tree Markov model (PSTMM) 137
- prokaryotic lectins 418
- property-encoded nomenclature (PEN) 276, 278–83, 285
- Protein Data Bank (PDB)
- algorithm 396–7
 - amino acids 435–44
 - atomic level interactions 439–44
 - carbohydrate-active enzymes 120, 123, 152–3
 - carbohydrates 4, 63
 - comparisons with theoretical data 401–2
 - conformational analysis 338, 347, 379, 390–402
 - data handling/analysis 434–44
 - dataset generation 434
 - erroneous entries 399–400
 - glycosylation 152–3
 - GlyVicinity 434–44
 - informational resources 390–3
 - overview of detected structures 397–400
 - protein–carbohydrate interactions 433–45
 - searching 3D structures 393–7
- protein disulfide isomerase (PDI) 147
- Protein Mutant database 96
- protein–carbohydrate interactions
- amino acids 435–44
 - atomic level interactions 439–44
 - data handling/analysis 434–44
 - dataset generation 434
- GlyVicinity 434–44
- lectins 415–31
- Protein Data Bank 433–45
- statistical analysis 433–45
 - see also* glycoproteins
- proteoglycans (PGs) 40, 42, 150, 178–80
- proteomic proteolytic digest mass spectrometry 228
- PSD *see* post-source decay
- pseudaminic acid 79
- PSTMM *see* probabilistic sibling-dependent tree Markov model
- PTM *see* post-translational modification
- public WEB-Medline (PubMed) 14
- pulsed amperometric detection (PAD) 205
- quantum mechanics (QM) 360, 364–5, 372, 404
- random array analysis method (RAAM) 210
- Random Molecular Mechanics (RAMM) 344, 372–3
- reaction libraries 134–5
- receiver operating characteristic (ROC) curves 167, 173, 175
- Red Queen Effect 78
- reducing end derivatization 230–4
- relaxed maps 342, 371
- reproductive and respiratory syndrome virus (RRSR) 78
- reversed phase HPLC 213
- ricin 416, 419, 421
- rigid energy maps 342
- rigid tissue approximation 340
- rigid-residue approach 371
- ring puckering 361
- RNA sequences 23–4
- ROC *see* receiver operating characteristic
- rotational nuclear Overhauser effect spectroscopy (ROESY) 277
- RRSR *see* reproductive and respiratory syndrome virus
- Saccharide Topology Analysis Tool (STAT) 246, 250
- saccharides *see* monosaccharides; oligosaccharides; polysaccharides
- SAX *see* strong anion exchange
- SD *see* stochastic dynamics
- SDS-PAGE *see* sodium dodecyl sulfate polyacrylamide gel electrophoresis
- Secondary Red Queen Effect 78

- self-organizing maps (SOMs) 325
- sensitivity testing 165–7
- sequence analysis
 - CAZy database 99–101
 - glycomic mass spectrometry 244–9
 - LINUCS 50–1, 54, 395–6
- sequence formats 50–61
- sequence logos 169–70, 172–3, 175–7, 179–85
- sequence motifs 172, 174–5, 177, 179–80, 182, 184, 285–6
- Sequence Retrieval System (SRS) 168
- sialic acids 69–88
 - biological roles 70–1, 72–3
 - biosynthetic pathways 76–8
 - digital representations 80–2
 - esterification 230–1, 234
 - evolution 78–9
 - future directions 80–2
 - host defense mechanisms 74–6
 - pathogens 70–1
 - species distribution 74
 - structure and diversity 72–5
- siglecs 77–8, 417, 422
- SignalP 182
- simulated annealing 345
- size exclusion chromatography 209
- small molecule descriptors 62–4
- SMILES 62–3
- snowdrop lectin 419, 426–7
- sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) 204–5, 229, 237–8, 276
- SOMs *see* self-organizing maps
- Spanish flu 10–11
- specificity testing 165–7
- spermadhesins 417–19, 426
- sphingosine 39
- SPR *see* surface plasmon resonance
- SRS *see* Sequence Retrieval System
- STAT *see* Saccharide Topology Analysis Tool
- stem names 26–7
- stereocodes 63–4
- stochastic dynamics (SD) 345, 372
- StrOligo 247
- strong anion exchange (SAX) 211
- SugaBase 299–301
- supervised learning 322–5
- Support Vector machines 164
- surface plasmon resonance (SPR) 346
- SWEET/SWEET-II 345, 406
- Swiss-Prot 4
 - carbohydrate-active enzymes 152–3
 - glycosylation 152–3, 168, 173, 175–6
 - TrEMBL 120–1
- tandem mass spectrometry (MS/MS)
 - 226, 229, 234–5, 238, 244–50, 260, 263
- TBA *see* thiobarbituric acid
- teichoic acids 41
- terminal sialic acid esterification 230–1, 234
- thiobarbituric acid (TBA) assays 81
- time-of-flight (TOF) analysis 231, 235
- TINKER/MM3 364, 366, 372, 375–6, 406
- TOF *see* time-of-flight
- transferred nuclear Overhauser enhancement (TR-NOE) 346
- TrEMBL 120–1
- trisaccharides 337
- trypsin 237
- UDP-Glc:glycoprotein glucosyltransferase (UGGT) 147–8, 174
- unsupervised learning 322–6
- uridine triphosphate (UTP) 5
- vaccines 12–13
- vertebrates 76–7, 417
- viruses 418
- weak anion exchange (WAX) 211
- wheat germ agglutinin (WGA) 425, 427
- X-ray crystallography 337, 389, 407, 433
- X-ray diffraction 337
- XML *see* extensible markup
- YinOYang 177–8
- zanamivir 345

This index was prepared by Neil Manley

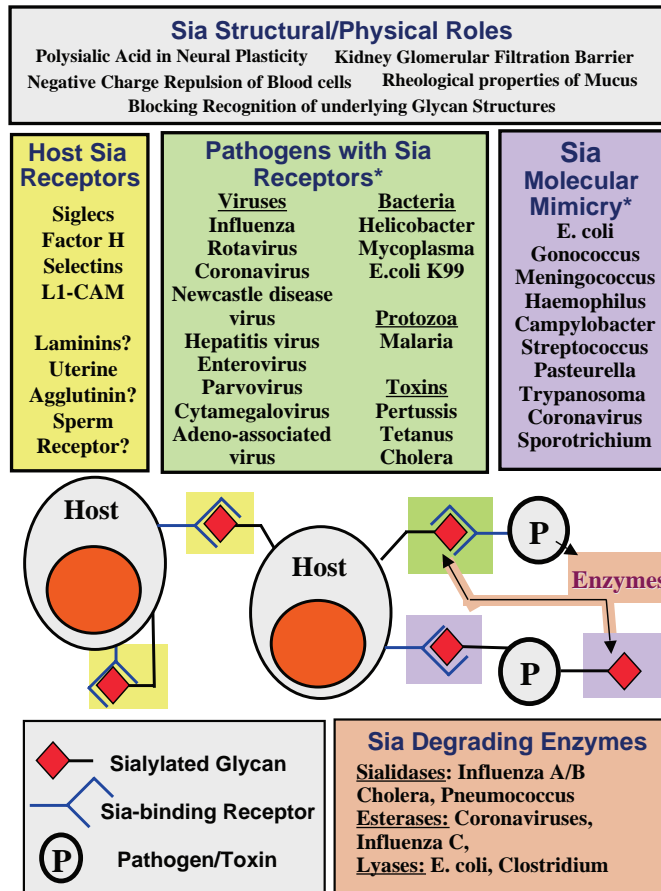


Plate 4.1 Biological roles of sialic acids. The diverse biological roles of Sias include structural/physical functions (shown in gray), or functions that require Sia-recognizing proteins. In the latter category, intrinsic (host) Sia-recognizing proteins are typically involved in endogenous functions (yellow), while extrinsic (pathogen) Sia-recognizing proteins mediate mechanisms of host cell adherence or entry (green). Microbial pathogens can also decorate themselves with Sia to hijack host processes (blue). In addition, some microbes express enzymes that degrade Sias (pink). A schematic representation of these relationships is color-coded as above. See the text and the cited literature for details about the biological functions of Sias depicted in this figure. Note: most pathogens express Sias or Sia-binding proteins in a strain-specific manner; hence these properties do not apply to all members of the indicated pathogen classes. Also, this figure represents an incomplete listing of most of the categories above. See [17] for details.

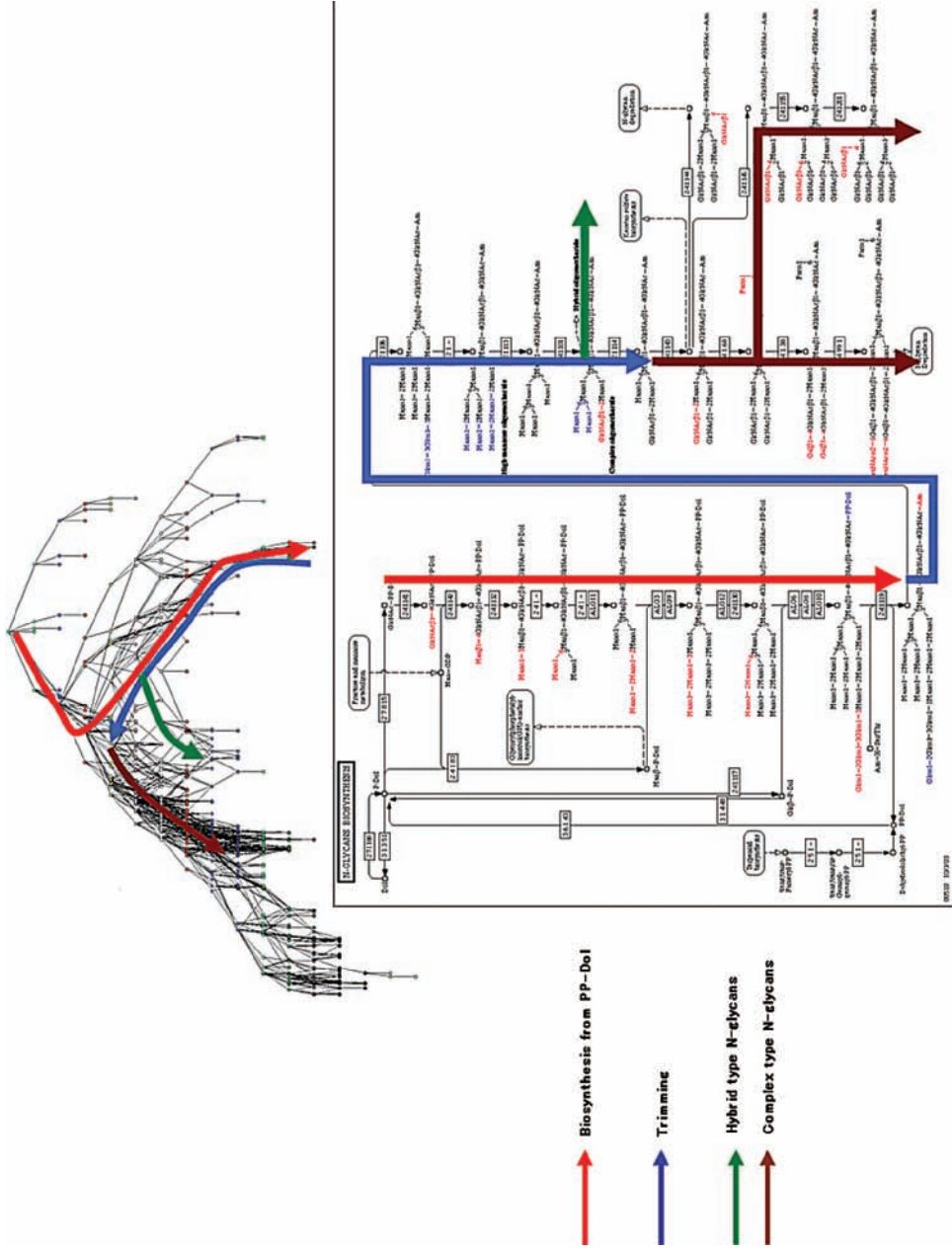


Plate 7.5 Correspondence of *N*-glycan biosynthesis pathway with the CRM.

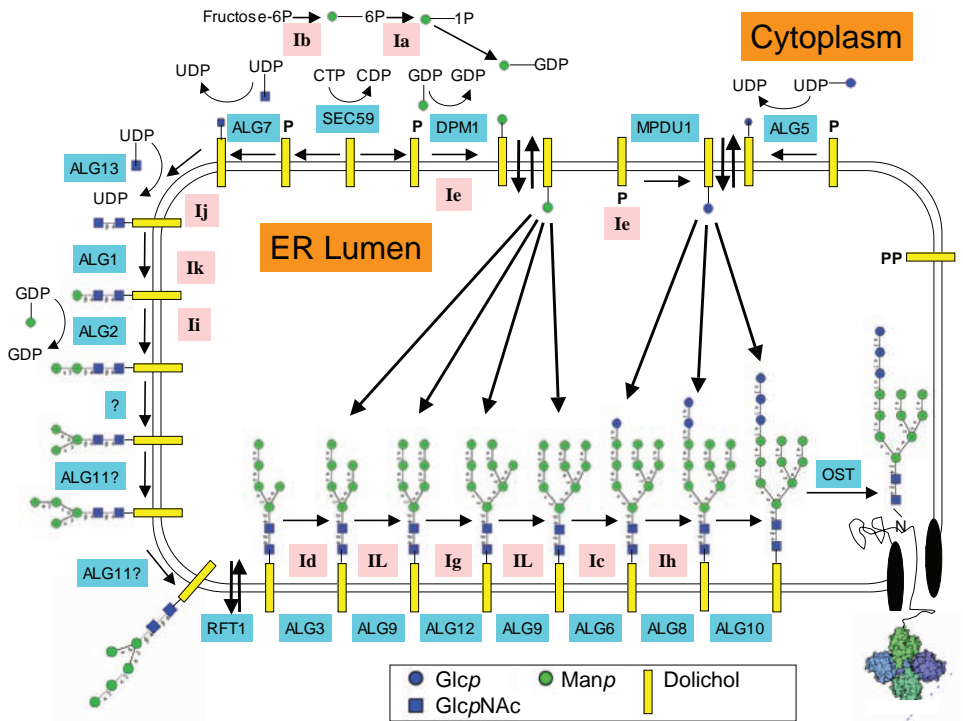


Plate 8.1 *N*-Glycan biosynthetic pathway – biosynthesis of lipid-bound oligosaccharide and transfer to a nascent polypeptide in the ER. The evolutionarily conserved *ALG* (asparagine-linked glycosylation) enzymes and other yeast loci involved in this pathway are displayed on a cyan background. The congenital disorders of glycosylation (CDG) disease classification (Ia–IL, see [38]) assigned to an enzyme’s malfunction is depicted on a pink background. Synthesis starts at the cytoplasmic face of the ER with UDP-GlcNAc and GDP-Man as glycosyl donors, transferring sugar residues onto dolichol (Dol). The $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ is then transferred to the luminal side with the help of *Rft1*, and elongated to the full-length lipid-linked oligosaccharide $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2\text{-PP-Dol}$ using Dol-P-Man and Dol-P-Glc. The oligosaccharide is subsequently linked by the oligosaccharyl transferase (OST) to the side-chain amido group of an asparagine residue within the consensus sequence Asn–Xaa–Ser/Thr of nascent secretory proteins. The glycosylation occurs while the polypeptide is still unfolded. Hence it can be classified as a co-translational protein modification. Adapted from [23] with slight modification.

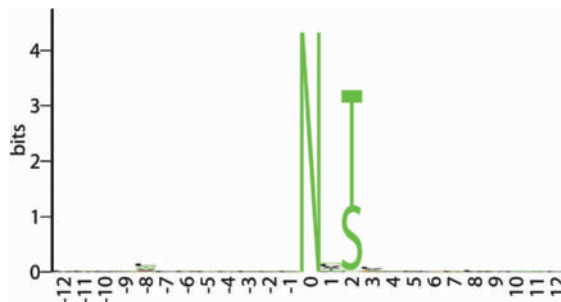


Plate 9.3 Shannon sequence logo for *N*-glycosylation sites. Please see page 169 for full caption and information.

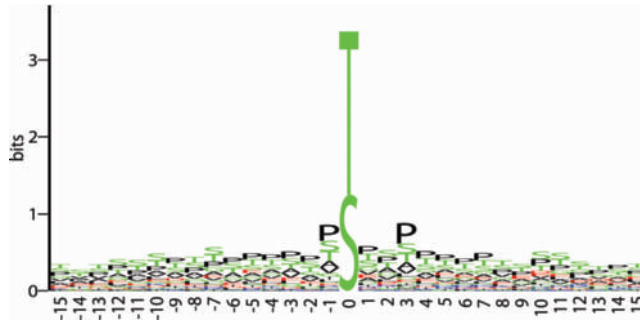


Plate 9.5 Shannon sequence logo for mammalian mucin-type *O*-glycosylation sites. Please see page 175 for full caption and information.

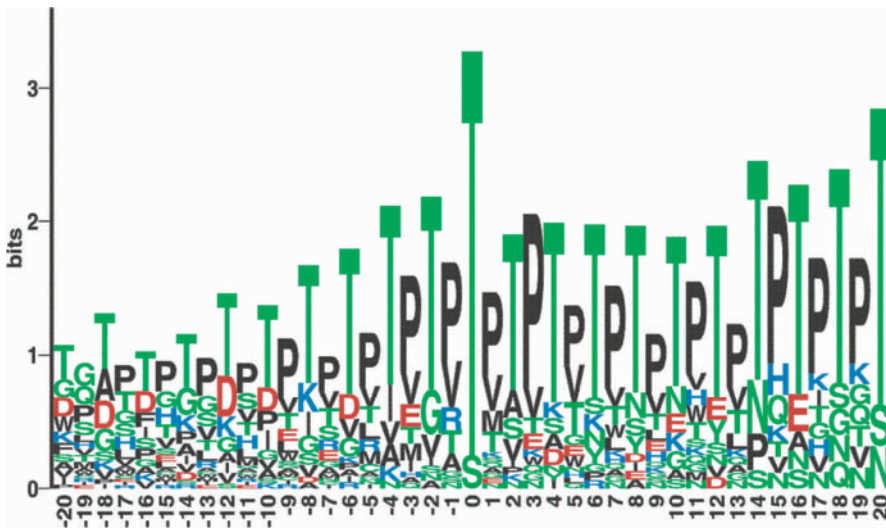


Plate 9.6 Shannon sequence logo for *O*- α -GlcNAc glycosylation sites in simple eukaryotes. Please see page 176 for full caption and information.

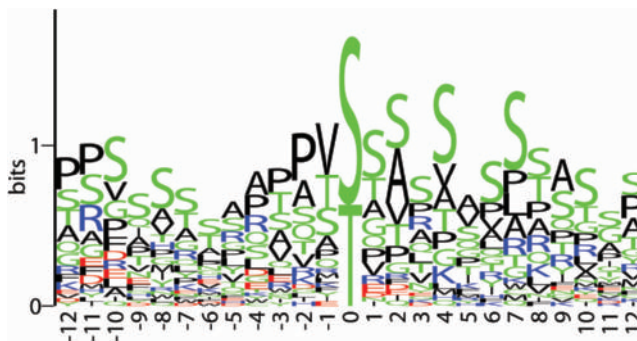


Plate 9.7 Shannon sequence logo for *O*- β -GlcNAc glycosylation sites in cytoplasmic/nuclear proteins. No clear consensus emerges around the acceptor serine/threonine except the usual high occurrence of proline, valine, and other serine/threonine residues.

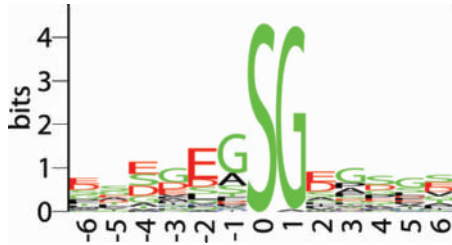


Plate 9.8 Shannon sequence logo for 95 proteoglycan sites in proteins. Please see page 179 for full caption and information.

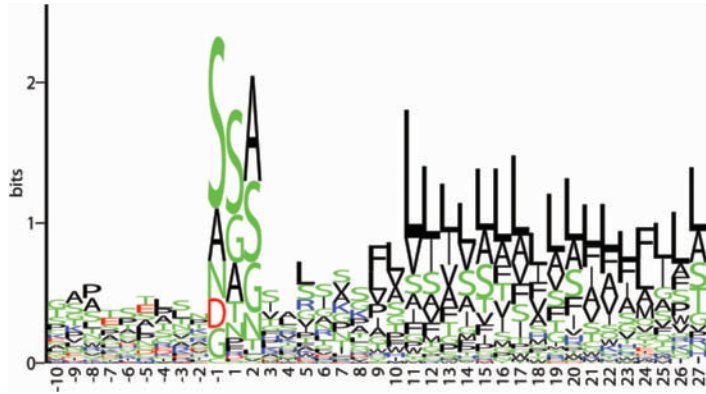


Plate 9.10 Shannon sequence logo for GPI-anchor sites. Please see page 181 for full caption and information.

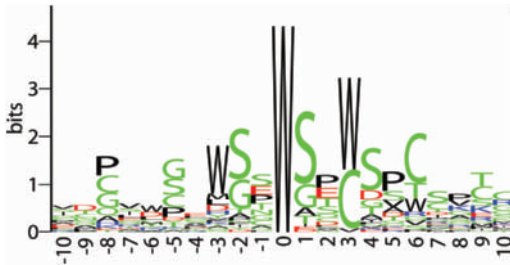


Plate 9.11 Shannon sequence logo for C-mannosylation sites. Please see page 183 for full caption and information.

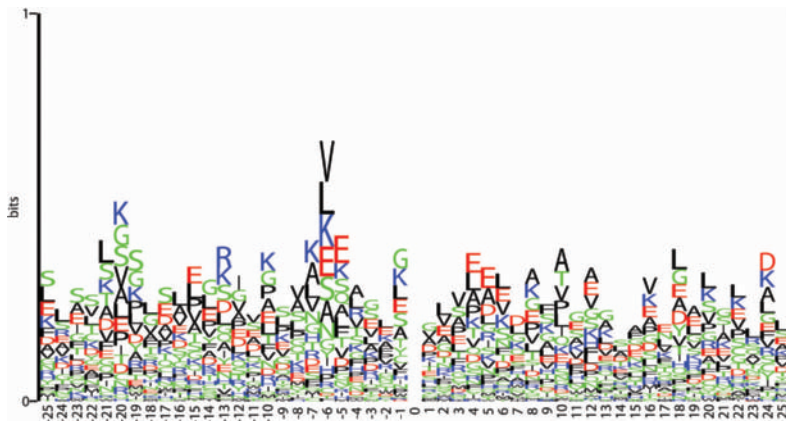


Plate 9.12 Shannon sequence logo for lysine glycation sites. Please see page 184 for full caption and information.

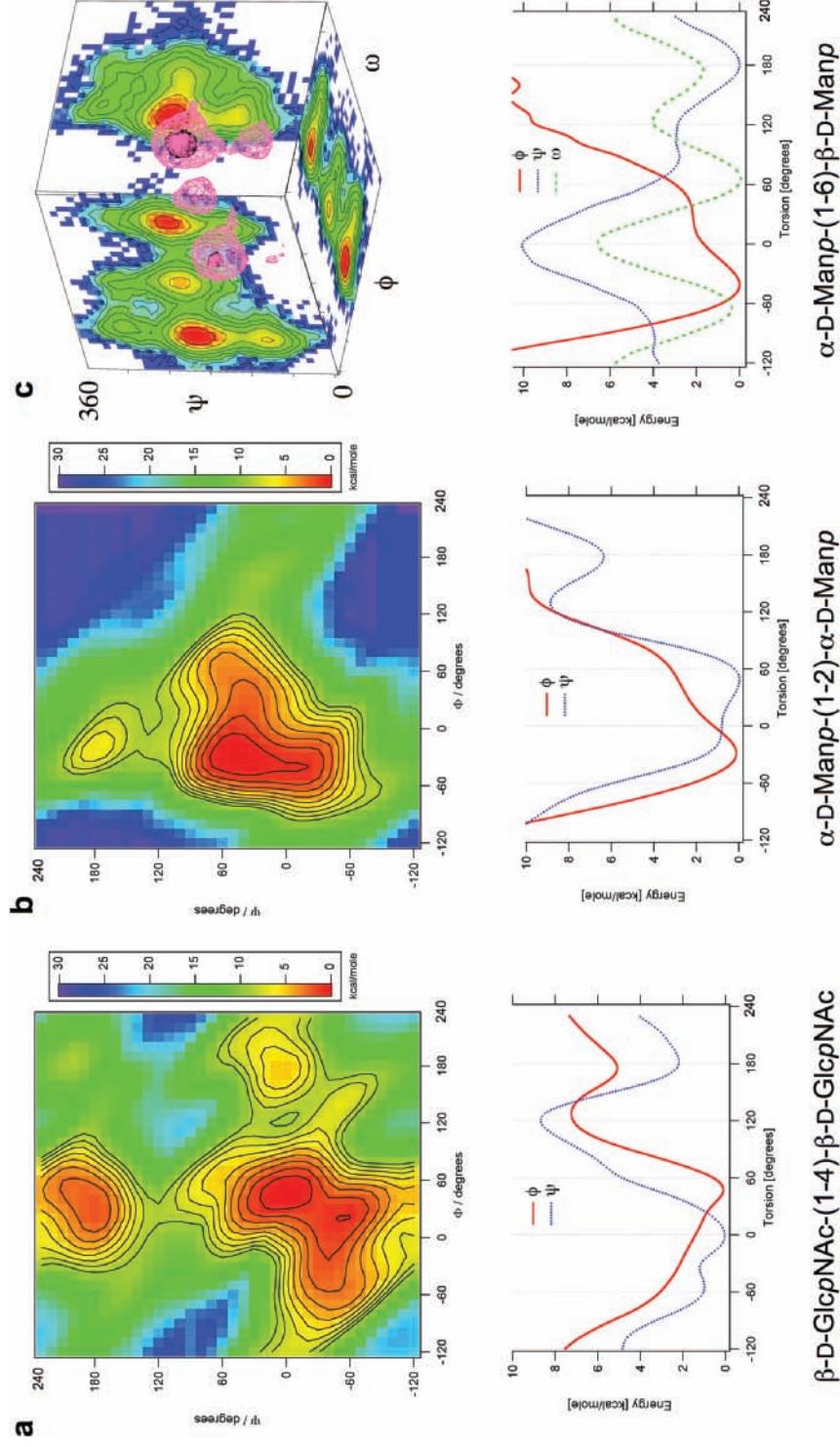


Plate 19.4 Conformational energy maps (top) and rotation profiles (bottom) of the glycosidic torsions of disaccharide fragments that are typically found in *N*-glycans. (a) Adiabatic map of β -D-GlcpNAc-(1-4)- β -D-GlcpNAc (*N,N'*-diacetylchitobiose) with a global minimum at $\phi_H/\psi_H \approx 45^\circ/0^\circ$ and secondary minima at $\phi_H/\psi_H \approx 20^\circ/-50^\circ$; $30^\circ/185^\circ$; $180^\circ/5^\circ$. (b) Adiabatic map of α -D-Manp-(1-2)- α -D-Manp with a global minimum at $\psi_H/\psi_H \approx -30^\circ/40^\circ$ and secondary minima at $\phi_H/\psi_H \approx -40^\circ/0^\circ$; $-20^\circ/170^\circ$. (c) Free energy map of α -D-Manp-(1-6)- β -D-Manp derived from an MD simulation with a global minimum at $\psi_H/\psi/\omega \approx -40^\circ/180^\circ/60^\circ$ break ($\psi = C1-O6-C6-C5$). Please see page 366 for full caption and further information.

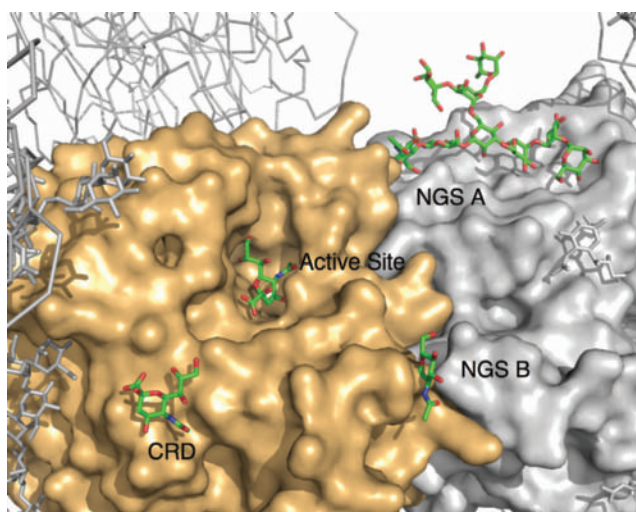


Plate 20.1 X-ray structure of avian influenza virus neuraminidase (PDB code 1MWE) [90]. Neuraminidase is an enzyme that cleaves terminal sialic acids (Neu5Ac) from glycoconjugates found on cell surfaces. Please see page 391 for further information and full caption.

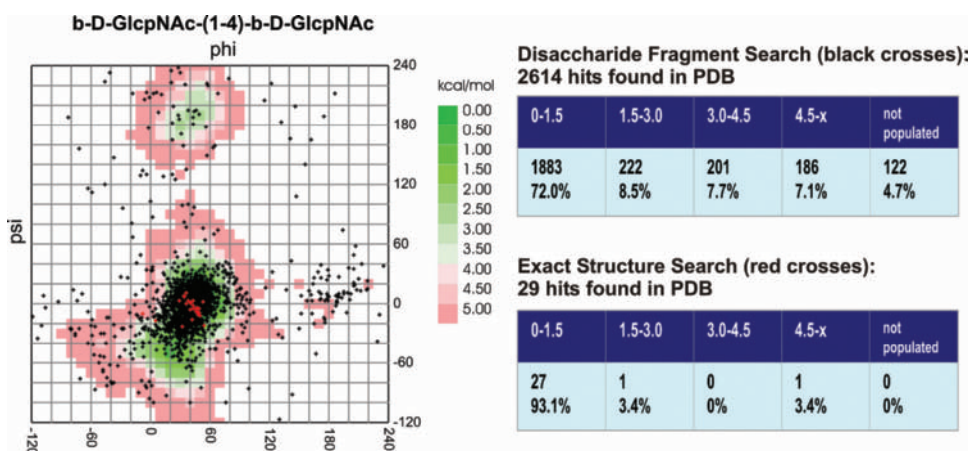


Plate 20.8 Comparison of experimental and theoretical torsion angle data for the β -D-GlcpNAc-(1-4)- β -D-GlcpNAc linkage. Comparison of experimental data with calculated conformational energy maps can be done by plotting the experimentally determined torsions on to the energy map. In those cases where the experimental torsions were obtained from carbohydrate chains that exactly match the one that was used to calculate the energy map (red crosses), more than 90% of all torsions are located in low-energy areas of 0–1.5 kcal mol⁻¹. Of those torsions that were derived from disaccharide fragments that are identical with the one from which the map was computed, but are part of a different carbohydrate chain (black crosses), still more than 70% are found in low-energy areas of 0–1.5 kcal mol⁻¹, and energy areas of 0–3 kcal mol⁻¹ cover more than 80% of the β -D-GlcpNAc-(1-4)- β -D-GlcpNAc linkage torsions that are present in the PDB.

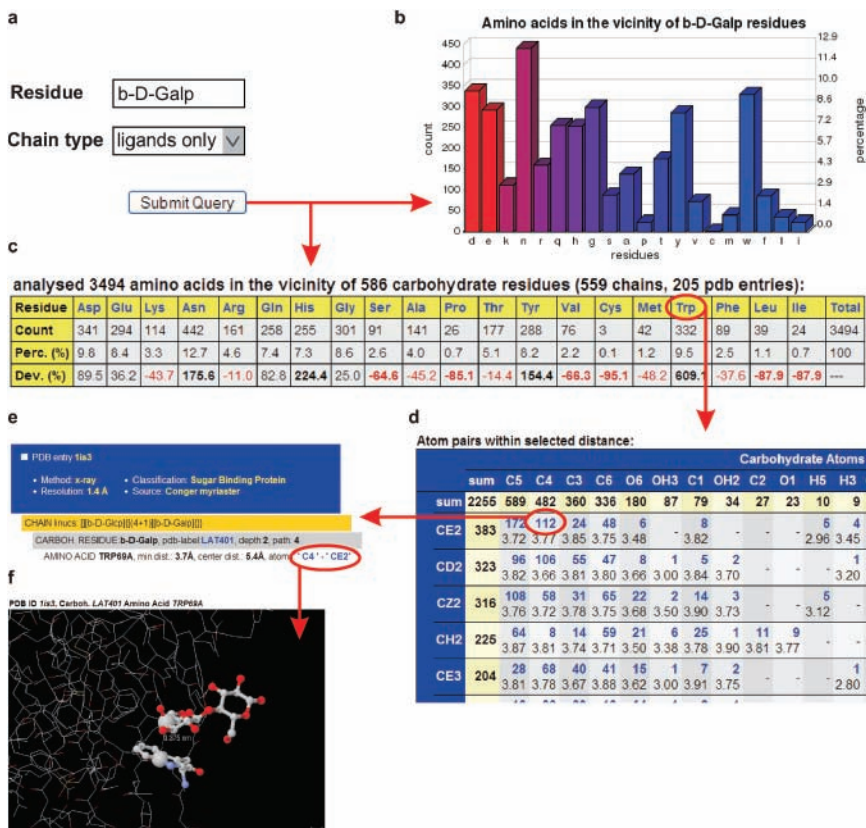


Plate 22.1 Workflow in the GlyVicinity web interface. After selecting a carbohydrate residue and the type of chains (glycan, ligand, or all) to be analyzed (a), data about the amino acids within a 4 Å radius around the selected carbohydrate residues are displayed graphically as diagrams (b) and numerically as tables (c). Statistics about the atoms involved in interactions can be displayed separately for each amino acid (d). Detailed data on the PDB entries, the carbohydrate chains and the residues from which the data were obtained (e) can be accessed in addition to the 3D structures with the selected atoms and residues highlighted (f).

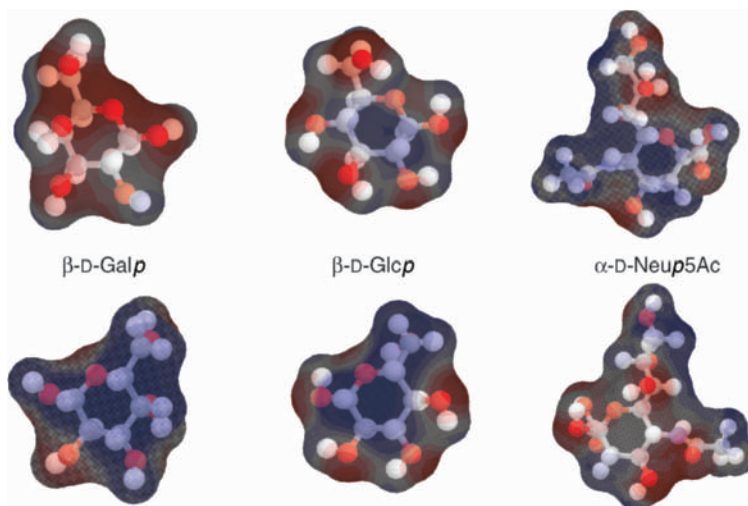












Plate 22.9 Electrostatic view of the surfaces of β -D-Galp, β -D-Glcp, and α -D-Neup5Ac. Top, "upper" ring side; bottom, "lower" ring side.













Legend to the Residue Symbols as used in this Book

Residue **Color Symbol** **Grayscale Symbol**

Hex (circles) / HexNAc (squares) / HexN (squares divided vertically)

Glc / GlcNAc / GlcN		
Gal / GalNAc / GalN		
Man / ManNAc / ManN		
Fuc		
Xyl		

Acidic monosaccharides (Diamonds):

GlcA		
GalA		
ManA		
IdoA		
NeuAc		
NeuGc		
KDN	