Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences

Edited by

Navneet Sharma

Himanshu Ojha

Pawan Kumar Raghav

Ramesh K. Goyal

ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-821748-1

Publisher: Andre Wolff
Acquisitions Editor: Erin Hill-Parks
Editorial Project Manager: Billie Jean Fernandez
Production Project Manager: Maria Bernadette Vidhya
Cover Designer: Mark Rogers

Typeset by TNQ Technologies

Contributors

Tanmay Arora
School of Chemical and Life Sciences (SCLS), Jamia Hamdard, New Delhi, Delhi, India; Division of CBRN Defence, Institute of Nuclear Medicine & Allied Sciences, DRDO, New Delhi, Delhi, India

Shereen Bajaj
Division of CBRN Defence, Institute of Nuclear Medicine & Allied Sciences, DRDO, New Delhi, Delhi, India

Prerna Bansal
Department of Chemistry, Rajdhani College, University of Delhi, New Delhi, Delhi, India

Aman Chandra Kaushik
Wuxi School of Medicine, Jiangnan University, Wuxi, Jiangsu, China

Raman Chawla
Division of CBRN Defence, Institute of Nuclear Medicine & Allied Sciences, DRDO, New Delhi, Delhi, India

Gurudutta Gangenahalli
Stem Cell and Gene Therapy Research Group, Institute of Nuclear Medicine & Allied Sciences (INMAS), Defence Research and Development Organisation (DRDO), New Delhi, Delhi, India

Srishty Gulati
Nucleic Acid Research Lab, Department of Chemistry, University of Delhi, North Campus, New Delhi, Delhi, India

Monika Gulia
School of Medical and Allied Sciences, GD Goenka University, Gurugram, Haryana, India

Vikas Jhawat
School of Medical and Allied Sciences, GD Goenka University, Gurugram, Haryana, India

Divya Jhinjharia
School of Biotechnology, Gautam Buddha University, Greater Noida, India

Jayadev Joshi
Genomic Medicine, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH, United States

Rita Kakkar
Computational Chemistry Laboratory, Department of Chemistry, University of Delhi, New Delhi, Delhi, India

Aman Chandra Kaushik
Wuxi School of Medicine, Jiangnan University, Wuxi, Jiangsu, China

Shrikant Kukreti
Nucleic Acid Research Lab, Department of Chemistry, University of Delhi, North Campus, New Delhi, Delhi, India

Shweta Kulshrestha
Division of CBRN Defence, Institute of Nuclear Medicine & Allied Sciences, DRDO, New Delhi, Delhi, India

Rajesh Kumar
Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, Delhi, India; Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India

Subodh Kumar
Stem Cell and Gene Therapy Research Group, Institute of Nuclear Medicine & Allied Sciences (INMAS), Defence Research and Development Organisation (DRDO), New Delhi, Delhi, India

Hirdesh Kumar
Laboratory of Malaria Immunology and Vaccinology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, United States

Vinod Kumar
Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, Delhi, India; Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India

Anjali Lathwal
Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, Delhi, India

Asrar A. Malik
School of Chemical and Life Sciences (SCLS), Jamia Hamdard, New Delhi, Delhi, India

Gandharva Nagpal
Department of BioTechnology, Government of India, New Delhi, Delhi, India

Himanshu Ojha
CBRN Protection and Decontamination Research Group, Division of CBRN Defence, Institute of Nuclear Medicine and Allied Sciences, New Delhi, Delhi, India

Mallika Pathak
Department of Chemistry, Miranda House, University of Delhi, New Delhi, Delhi, India

Pawan Kumar Raghav
Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, Delhi, India

Shakti Sahi
School of Biotechnology, Gautam Buddha University, Greater Noida, Uttar Pradesh, India

Manisha Saini
CBRN Protection and Decontamination Research Group, Division of CBRN Defence, Institute of Nuclear Medicine and Allied Sciences, New Delhi, Delhi, India

Manisha Sengar
Department of Zoology, Deshbandhu College, University of Delhi, New Delhi, Delhi, India

Mamta Sethi
Department of Chemistry, Miranda House, University of Delhi, New Delhi, Delhi, India

V.G. Shanmuga Priya
Department of Biotechnology, KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belagavi, Karnataka, India

Vidushi Sharma
Delhi Institute of Pharmaceutical Education and Research, New Delhi, Delhi, India

Anil Kumar Sharma
School of Medical and Allied Sciences, GD Goenka University, Gurugram, Haryana, India

Malti Sharma
Department of Chemistry, Miranda House, University of Delhi, New Delhi, Delhi, India

Navneet Sharma
Department of Textile and Fiber Engineering, Indian Institute of Technology, New Delhi, Delhi, India

Md Shoaib
Nucleic Acid Research Lab, Department of Chemistry, University of Delhi, North Campus, New Delhi, Delhi, India

Anju Singh
Nucleic Acid Research Lab, Department of Chemistry, University of Delhi, North Campus, New Delhi, Delhi, India; Department of Chemistry, Ramjas College, University of Delhi, New Delhi, Delhi, India

Jyoti Singh
Department of Chemistry, Hansraj College, University of Delhi, New Delhi, Delhi, India

Kailas D. Sonawane
Structural Bioinformatics Unit, Department of Biochemistry, Shivaji University, Kolhapur, Maharashtra, India; Department of Microbiology, Shivaji University, Kolhapur, Maharashtra, India

Rakhi Thareja
Department of Chemistry, St. Stephen's College, University of Delhi, New Delhi, Delhi, India

Nishant Tyagi
Stem Cell and Gene Therapy Research Group, Institute of Nuclear Medicine & Allied Sciences (INMAS), Defence Research and Development Organisation (DRDO), New Delhi, Delhi, India

Yogesh Kumar Verma
Stem Cell and Gene Therapy Research Group, Institute of Nuclear Medicine & Allied Sciences (INMAS), Defence Research and Development Organisation (DRDO), New Delhi, Delhi, India

Sharad Wakode
Delhi Institute of Pharmaceutical Education and Research, New Delhi, Delhi, India

CHAPTER

1 Impact of chemoinformatics approaches and tools on current chemical research

Rajesh Kumar[1,3,a], Anjali Lathwal[1,a], Gandharva Nagpal[2], Vinod Kumar[1,3], Pawan Kumar Raghav[1,a]

[1]*Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, Delhi, India;* [2]*Department of BioTechnology, Government of India, New Delhi, Delhi, India;* [3]*Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India*

1.1 Background

Biological research remains at the core of fundamental analysis in the quest to understand the molecular mechanism of living things. Biological researchers produce enormous amounts of data that critically need to be analyzed. Bioinformatics is an integrative science that arises from mathematics, chemistry, physics, statistics, and informatics, which provides a computational means to explore a massive amount of biological data. Also, bioinformatics is a multidisciplinary science that includes tools and software to analyze biological data such as genes, proteins, molecular modeling of biological systems, molecular modeling, etc. It was Pauline Hogeweg, a Dutch system biologist, who coined the term bioinformatics. After the advent of user-friendly Swiss port models, the use of bioinformatics in biological research has gained momentum at unparalleled speed. Currently, bioinformatics has become an integral part of all life science research that assists clinical scientists and researchers in identifying and prioritizing candidates for targeted therapies based on peptides, chemical molecules, etc.

Chemoinformatics is a specialized branch of bioinformatics that deals with the application of developed computational tools for easy data retrieval related to chemical compounds, identification of potential drug targets, and performance of simulation studies. These approaches are used to understand the physical, chemical, and biological properties of chemical compounds and their interactions with the biological system that can have the potential to serve as a lead molecule for targeted therapies. Although the sensitivity of the computational methods is not as reliable as experimental studies, these tools provide an alternative means in the discovery process because experimental techniques are time consuming and expensive. The primary

[a] Equal contribution

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00001-4

application of advanced chemoinformatics methods and tools is that they can assist biological researchers to arrive at informed decisions within a shorter timeframe. A molecule with drug-likeness properties has to pass physicochemical properties such as the Lipinski rule of five and absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties before submitting it for clinical trials. If any compound fails to possess reliable ADMET properties, it is likely to be rejected. So, in the process of accelerating the drug discovery process, researchers can use different in silico chemoinformatics computational methods for screening a large number of compounds from chemical libraries to identify the most druggable molecule before launching into clinical trials. A similar approach can be employed for designing subunit vaccine candidates from a large number of protein sequences of pathogenic bacteria.

In the literature, several other review articles focus on specialized parts of bioinformatics, but there is no such article describing the use of bioinformatics tools for nonspecialist readers. This chapter describes the use of different biological chemoinformatics tools and databases that could be used for identifying and prioritizing drug molecules. The key areas included in this chapter are small molecule databases, protein and ligand databases, pharmacophore modeling techniques, and quantitative structure−activity relationship (QSAR) studies. Organization of the text in each section starts from a simplistic overview followed by critical reports from the literature and a tabulated summary of related tools.

1.2 Ligand and target resources in chemoinformatics

Currently, there has been an enormous increase in data related to chemicals and medicinal drugs. The available experimentally validated data can be utilized in computer-aided drug design and discovery of some novel compounds. However, most of the resources having such data belongs to private domains and large pharmaceutical industries. These resources mainly house data in form of chemical descriptors that may be used to build different predictive models. A complete overview of the chemical descriptors/features and databases can be found in Tables 1.1 and 1.2. A brief description of each type of database can be found in the subsequent subsections of this chapter.

1.2.1 Small molecule compound databases

Small molecule compound databases hold information on active organic and inorganic substances, which can show some biological effect. The largest repository of active small molecule compounds is the Available Chemical Directory (ACD), which stores almost 300,000 active substances. The ACD/Labs database provides information on the physicochemical properties such as logP, logS, and pKa values of active compounds. Another such database is the SPRESIweb database containing more than 4.5 million compounds and 3.5 million reactions. Another database, CrossFire Beilstein, has more than 8 million organic compounds and 9 million active biochemical reactions along with a variety of properties, including various physical properties, pharmacodynamics, and environmental toxicity.

Table 1.1 Table representing standard features and their type utilized in quantitative structure–activity relationship.

Descriptor type	Basis	Example
Theoretical descriptors		
0D	Structural count	Molecular weight, number of bonds, number of hydrogen bonds, aromatic and aliphatic bonds
1D	Chemical graph theory	Numbers of functional groups, fragment counts, disulfide bonds, ammonium bond
2D	Topological properties	Randic index, wiener index, molecular walk count, kappa shape index
3D	Geometrical structural properties	Autocorrelation, 3D-Morse, fingerprints
4D	Conformational	GRID, raptor, sample conformation
Experimental descriptors		
Electronic	Electrostatic properties	Dissociation constant, hammett constant
Steric	Steric properties	Charton constant
Hydrophobic	Hydrophobic properties	logP, hydrophobic constant

Table 1.2 Commonly used tools and software categorized on algorithms/scoring functions/description availability with uniform resource locator (URL) and supported platforms.

Databases/ tools	Algorithm/scoring functions/description	Website URL	PMIDs
Commonly used databases in chemoinformatics			
Available Chemical Directory (ACD)	Access to meticulously examined experimental Nuclear Magnetic Resonance (NMR) data, complete with assigned structures and references of millions of chemical compounds	https://www.acdlabs.com/products/dbs/nmr_db/index.php	32681440
CrossFire Beilstein	Data on more than 320 million scientifically measured properties of chemical compounds. The largest database in organic chemistry.	www.crossfirebeilstein.com	11604014
SpresiWeb	Data regarding millions of chemical molecules and reactions extracted from research articles	https://www.spresi.com/indexunten.htm	24160861

Continued

Table 1.2 Commonly used tools and software categorized on algorithms/scoring functions/description availability with uniform resource locator (URL) and supported platforms.—*cont'd*

Databases/ tools	Algorithm/scoring functions/description	Website URL	PMIDs
ChEMBL	Approximately 2.1 million chemical compounds from nearly 1.4 million assays	https://www.ebi.ac.uk/chembl/	21948594
PubChem	Contains 9.2 million compounds with activity information	https://pubchem.ncbi.nlm.nih.gov	26400175
CARLSBAD	Contains activity information of 0.43 million active compounds	http://carlsbad.health.unm.edu/carlsbad/	23794735
Drugcentral	Provides information on 4,444 pharmaceutical ingredients with 1,605 human protein targets	http://drugcentral.org	27789690
repoDB	A standard database for drug repurposing	http://apps.chiragjpgroup.org/repoDB/	28291243
PharmGKB	A database for exploring the effect of genetic variation on drug targets	https://www.pharmgkb.org	23824865
ZINC	A commercially available database for virtual screening	http://zinc.docking.org	15667143
Databases for exploring protein–ligand interaction			
Protein Data Bank (PDB)	Provides information on 166,301 crystallographic identified structures of macromolecules	https://www.rcsb.org/	10592235
Cambridge Structural Database (CSD)	Provides information on nearly 0.8 million compounds	https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/	27048719
Protein Ligand Interaction Database (PLID)	A resource for exploring the protein–ligand interaction from PDB	http://203.199.182.73/gnsmmg/databases/plid/	18514578
Protein Ligand Interaction Clusters (PLIC)	A repository for exploring nearly 84,846 protein–ligand interactions derived from PDB	http://proline.biochem.iisc.ernet.in/PLIC	24763918
CREDO	A resource providing protein–ligand interaction information for drug discovery	http://www-cryst.bioc.cam.ac.uk/credo	19207418

Table 1.2 Commonly used tools and software categorized on algorithms/scoring functions/description availability with uniform resource locator (URL) and supported platforms.—*cont'd*

Databases/ tools	Algorithm/scoring functions/description	Website URL	PMIDs
PDBbind	A resource for the binding affinity of nearly 5,897 protein–ligand complexes	http://www.pdbbind.org/	15943484
Database for exploring macromolecular interactions			
DOMININO	A database for exploring the interaction between protein domains and interdomains	http://dommino.org	22135305
PIMAdb	A resource for exploring interchain interaction among protein assemblies	http://caps.ncbs.res.in/pimadb	27478368
PDB-eKB	A community-driven knowledgebase for functional annotation and prediction of PDB data	https://www.ebi.ac.uk/pdbe/pdbe-kb	31584092
CATH	A database for classification of protein domains	http://www.cathdb.info	20368142
LIGAND	A composite database of chemical compounds, reactions, and enzymatic information	http://www.genome.ad.jp/ligand/	11752349
The Molecular Interaction Database (MINT)	Provides information on experimentally verified protein–protein interactions	https://mint.bio.uniroma2.it/mint/	17135203
Database of Interacting Proteins (DIP)	A database for exploring, prediction, and evolution of protein–protein interaction, and identification of a network of interactions	http://dip.doe-mbi.ucla.edu	10592249
The Biomolecular Interaction Network Database (BIND)	A database for exploring biomolecular interactions	www.bind.ca	2519993
Software used for pharmacophore modeling			
Pharmer	A computational tool for pharmacophore searching using bloom fingerprint	smoothdock.ccbb.pitt.edu/pharmer/	21604800
PharmaGist	A server for ligand-based pharmacophore searching by utilizing the Mtree algorithm	http://bioinfo3d.cs.tau.ac.il/PharmaGist/	18424800

Continued

Table 1.2 Commonly used tools and software categorized on algorithms/scoring functions/description availability with uniform resource locator (URL) and supported platforms.—*cont'd*

Databases/ tools	Algorithm/scoring functions/description	Website URL	PMIDs
LiSiCA	A software for ligand-based virtual screening	http://insilab.org/lisica/	26158767
ZINCPharmer	A tool for pharmacophore searching from the ZINC database	http://zincpharmer.csb.pitt.edu/	22553363
LigandScout	A tool for generating a 3D pharmacophore model using six types of chemical features	http://www.inteligand.com/download/InteLigand_LigandScout_4.3_Update.pdf	15667141
Schrodinger	The phase function of schrodinger can be utilized for ligand and structure-based pharmacophore modeling	https://www.schrodinger.com/phase	32860362
VirtualToxLab	Allows rationalizing prediction at the molecular level by analyzing the binding mode of the tested compound for target proteins in real-time 3D/4D	http://www.biograf.ch/index.php?id=projects&subid=virtualtoxlab	32244747
Tools used in QSAR model development			
DPubChem	Software for automated generation of a QSAR model	www.cbrc.kaust.edu.sa/dpubchem	29904147
QSAR-Co	Open-source software for QSAR-based classification model development	https://sites.google.com/view/qsar-co	31083984
DTClab	A suite of software for curating and generating a QSAR model for virtual screening	https://dtclab.webs.com/software-tools	31525295
Ezqsar	A standalone program suite for QSAR model development	https://github.com/shamsaraj/ezqsar	29387275
DataWarrior	An integrated computer tool for generation and virtual screening of a QSAR model	http://www.openmolecules.org/datawarrior/	30806519

Table 1.2 Commonly used tools and software categorized on algorithms/scoring functions/description availability with uniform resource locator (URL) and supported platforms.—*cont'd*

Databases/ tools	Algorithm/scoring functions/description	Website URL	PMIDs
Feature selection algorithm used in building a QSAR model			
Waikato Environment for Knowledge Analysis (WEKA)	A general-purpose environment for automatic classification, regression, clustering, and feature selection of common data mining problems in bioinformatics research	http://www.cs.waikato.ac.nz/ml/weka	15073010
DWFS	A web-based tool for feature selection	https://www.cbrc.kaust.edu.sa/dwfs/	25719748
SciKit	A Python-based framework for feature selection and model optimization	https://scikit-learn.org/stable/modules/feature_selection.html	32834983
Docking software commonly used in chemoinformatics			
Autodock4	GA; LGA; SA/empirical free energy forcefield	http://autodock.scripps.edu	19399780
Autodock Vina	GA, PSO, SA, Q-NM/X-Score	http://vina.scripps.edu	19499576
BDT	AutoGrid and AutoDock	http://www.quimica.urv.cat/~pujadas/BDT/index.html	16720587
BetaDock	GA	http://voronoi.hanyang.ac.kr/software.htm	21696235
CDocker	SA	http://accelrys.com/services/training/life-science/StructureBasedDesignDescription.html	11922947
DARWIN	GA	http://darwin.cirad.fr/product.php	10966571
DOCK	IC/Chem Score, SA solvation scoring, DockScore	http://dock.compbio.ucsf.edu	19369428
DockoMatic	AutoDock	https://sourceforge.net/projects/dockomatic/	21059259

Continued

Table 1.2 Commonly used tools and software categorized on algorithms/scoring functions/description availability with uniform resource locator (URL) and supported platforms.—*cont'd*

Databases/ tools	Algorithm/scoring functions/description	Website URL	PMIDs
DockVision	MC, GA	http://dockvision.sness.net/overview/overview.html	1603810
eHiTS	RBD of fragments followed by reconstruction/eHiTS	www.simbiosys.cs/ehits/index.html	16860582
FINDSITE-COMB	SP-score	http://cssb.biology.gatech.edu/findsitelhm	19503616
FITTED	GA/RankScore	http://www.fitted.ca	17305329
Fleksy	Flexible approach to IFD	http://www.cmbi.ru.nl/software/fleksy/	18031000
FlexX	IC/FlexXScore, PLP, Screen Score, Drug Score	https://www.biosolveit.de	10584068
FlipDock	GA	http://flipdock.scripps.edu	17523154
FRED	RBD/Screen Score, PLP, Gaussian shape score, ChemScore, ScreenScore, Chemgauss4 scoring function	https://docs.eyesopen.com/oedocking/fred.html	21323318
GalaxyDock	GalaxyDock BP2 Score	http://galaxy.seoklab.org/softwares/galaxydock.html	23198780, 24108416
GEMDOCK	EA/empirical scoring function	http://gemdock.life.nctu.edu.tw/dock/	15048822
GlamDock	MC/SA	http://www.chil2.de/Glamdock.html	17585857
Glide	Hierarchical filters and MC/ Glide Score, glide comp	https://www.schrodinger.com/glide/	15027865
GOLD	GA/Gold Score, chem Score	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/	12910460
GriDock	AutoDock4.0	http://159.149.85.2/cms/index.php?Software_projects:GriDock	20623318

Table 1.2 Commonly used tools and software categorized on algorithms/scoring functions/description availability with uniform resource locator (URL) and supported platforms.—*cont'd*

Databases/ tools	Algorithm/scoring functions/description	Website URL	PMIDs
HADDOCK	SA/HADDOCK Score	http://haddock.science.uu.nl/services/HADDOCK2.2/	12580598
HYBRID	CGO/Ligand-based scoring function	https://docs.eyesopen.com/oedocking/hybrid.html	17591764
iGEMDOCK	GA/Simple empirical scoring function and a pharmacophore-based scoring function	http://gemdock.life.nctu.edu.tw/dock/igemdock.php	15048822
Lead Finder	GA	http://moltech.ru	19007114
LigandFit	Monte Carlo sampling/Lig Score, PLP, PMF, hammerhead	https://www.phenix-online.org/documentation/reference/ligandfit.html	12479928
Mconf-DOCK	DOCK5	http://www.mti.univ-paris-diderot.fr/recherche/plateformes/logiciels	18402678
MOE	Gaussian function	http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm	19075767
Molegro Virtual Docker	Evolutionary algorithm	http://www.scientificsoftware-solutions.com/product.php?productid=17625	16722650
POSIT	SHAPEFIT	https://docs.eyesopen.com/oedocking/posit_usage.html	21323318
Rosetta Ligand	Rosetta script	https://www.rosettacommons.org/software	22183535

Continued

Table 1.2 Commonly used tools and software categorized on algorithms/scoring functions/description availability with uniform resource locator (URL) and supported platforms.—*cont'd*

Databases/ tools	Algorithm/scoring functions/description	Website URL	PMIDs
Surflex-Dock	IC/Hammerhead	https://omictools.com/surflex-dock-tool	22569590
VLifeDock	GA/PLP score, XCscore, and Steric + Electrostatic score	http://www.vlifesciences.com/products/VLifeMDS/VLifeDock.php	30124114
Commonly used MD simulations tools			
Abalone	Suitable for long simulations	http://www.biomolecular-modeling.com/Abalone/index.html	26751047
ACEMD	Fastest MD engine	https://www.acellera.com/products/molecular-dynamics-software-GPU-acemd/	26616618
AMBER	Used for simulations	http://ambermd.org/	16200636
CHARMM	Allows macromolecular simulations	http://yuri.harvard.edu/	31329318
DESMOND	Performs high-performance MD simulations	https://www.deshawresearch.com/resources_desmond.html	16222654
GROMACS	Widely used with excellent performance	http://www.gromacs.org/	21866316
LAMMPS	A coarse-grain tool, specifically designed for material MD simulations	http://lammps.sandia.gov/	31749360
MOIL	A complete suite for MD simulations and modeling	http://clsbweb.oden.utexas.edu/moil.html	32375019
NAMD	Provides a user-friendly interface and plugins to perform large simulations	http://www.ks.uiuc.edu/Research/namd/	29482074
TINKER	Performs biomolecule and biopolymer MD simulations	http://dasher.wustl.edu/tinker/	30176213

1.2.2 Protein and ligand information databases

3D information of a ligand and its binding residues within the pocket of its target protein is an essential requirement while developing 3D-QSAR-based models. Thus, the databases holding information about macromolecule structures are of great importance for pharmaceutical industries and researchers. The Protein Data Bank (PDB) (Rose et al., 2017) is one such open-source large repository containing structural information identified via crystallographic and Nuclear Magnetic Resonance (NMR) experimental techniques. The current version of PDB holds structural information on 166,301 abundant macromolecular compounds. The PDB is updated weekly with a rate of almost 100 structures. Another such extensive database is the Cambridge Structural Database (Groom et al., 2016), which provides structural information on large macromolecules such as proteins.

1.2.3 Databases related to macromolecular interactions

Often the biological activity of a protein can be modulated by binding a ligand molecule within its active site. Thus, identification of molecular interactions among ligand−protein and protein−protein is of utmost importance. Moreover, the biological pathways and chemical reactions occurring at the protein−ligand interface are also essential in understanding disease pathology. LIGAND is a database that provides information on enzymatic reactions occurring at the macromolecular level (Goto et al., 2000). Several other databases, such as the Database of Interacting Proteins, Biomolecular Interaction Network Database, and Molecular Interaction Network, are also present in the literature, which includes information on protein−protein interactions.

1.3 Pharmacophore modeling

The process of drug designing dates back to 1950 (Newman and Cragg, 2007). Historically, the process of drug designing follows a hit-and-miss approach. It has been observed that only one or two tested compounds out of 40,000 reach clinical settings, suggesting a low success rate. Often the developed lead molecule lacks potency and specificity. The traditional drug design process may take up to 7−12 years, and approximately $1−2 billion in launching a suitable drug into the market. All this suggests that finding a drug molecule is time consuming, expensive, and needs to be optimized in a different way to identify the correct lead molecule. These limitations also signify that there should be some novel alternative ways to identify hits that may lead to drug molecules. Soon after discovering computational methods to design and screen large chemical databases, the process of drug discovery has primarily shifted from natural to synthetic (Lourenco et al., 2012). The rational strategies for creating active pharmaceutical compounds have become an exciting area of research. Industries and research institutions are continuously developing new

tools that can accelerate and speed up the drug discovery process. The methodology involves identifying active molecules via ligand optimization known as pharmacophore modeling or the structure–activity relationship approach. This section of the chapter describes ligand-based pharmacophore modeling in detail to find the active compound with desired biological effects.

A pharmacophore is simply a representation of the ligand molecules' structural and chemical features that are necessary for its biological activity. According to the International Union of Pure and Applied Chemistry, a pharmacophore is an ensemble of steric and electrostatic features required to ensure optimal interactions with specific biological targets to block its response. The pharmacophore is not a real lead molecule, but an ensemble of common molecular descriptors shared by active ligands of diverse origins. This way, pharmacophore modeling can help identify the active functional groups within ligand binding sites of target proteins and provide clues on noncovalent interactions. The active pharmacophore feature includes hydrogen bond donor, acceptor, cationic, aromatic, and hydrophobic components of a ligand molecule, etc. The characteristic features of active ligands are often described in 3D space by torsional angle, location distance, and other features. Several software tools are available to design the pharmacophore model, such as the catalyst, MOE, LigandScout, Phases, etc.

1.3.1 Types of pharmacophore modeling

Pharmacophore modeling is broadly classified into two categories: ligand-based and structure-based pharmacophore modeling. A brief about the methodology adopted by each type of modeling is shown in Fig. 1.1. However, structure-based pharmacophore modeling exclusively depends on the generation of pharmacophore models based on the receptor-binding site. Still, for ligand-based pharmacophore modeling, the bioactive conformation of the ligand is used to derive the pharmacophore model. The best approach is to consider the receptor–ligand complex and generate the pharmacophore models from there. This provides exclusion volumes that restrict the ligand during virtual screening to the target site and thus is quite successful in virtual screening of large chemical database libraries.

1.3.2 Scoring scheme and statistical approaches used in pharmacophore modeling

Several parameters assess the quality of developed pharmacophore models, such as predictive power, identifying novel compounds, cost function, test set prediction, receiver operating characteristic (ROC) analysis, and goodness of fit score. Generally, a test set approach is used to estimate the predictive power of a developed pharmacophore model. A test set is a group of the external dataset of structurally diverse compounds. It checks whether the developed model can predict the unknown instance. A general observation is that if a developed model shows a correlation coefficient greater than 0.70 on both training and test set, it is of good quality.

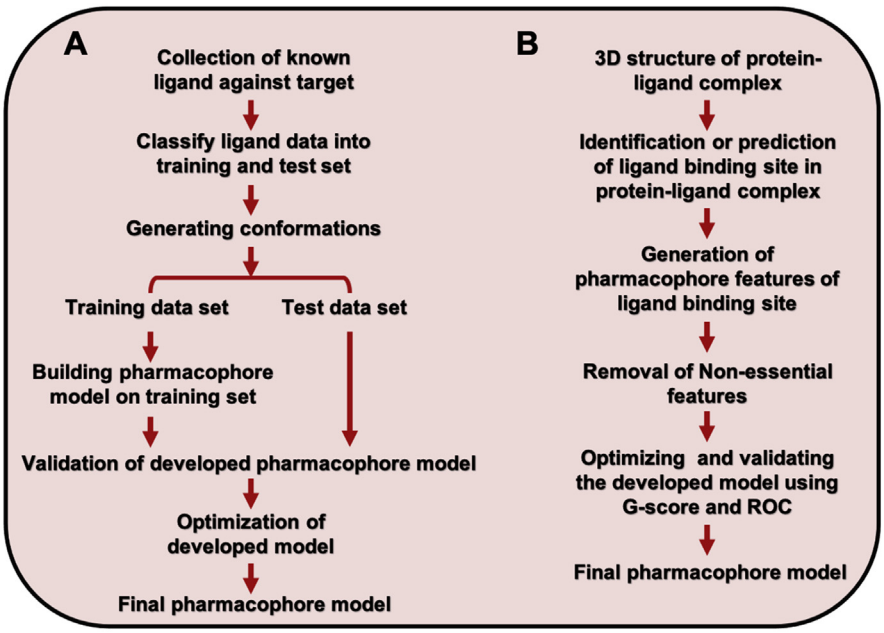


FIGURE 1.1

Overall workflow of the methodology used in developing the pharmacophore model. (A) Ligand-based pharmacophore model. (B) Structure-based pharmacophore model. *ROC*, Receiver operating characteristic.

The commonly used statistical parameter, cost−function analysis, is integrated into the HypoGen program to validate the predictive power of the developed model. The optimal quality pharmacophore model generally has a cost difference between 40 and 60 bits. The cost value signifies the percentage of probability of correlating the data points. The value between 40 and 60 bits means that the developed pharmacophore model shows a 75%−90% probability of correlating the data points. The ROC plot gives visual as well as numerical representation of the developed pharmacophore model. It is a quantitative measure to assess the predictive power of a developed pharmacophore model. The ROC curve depends on the true positive, true negative, false positive, and false negative predicted by the developed model. The ROC plot can be plotted using 1-specificity (false positive rate) on the X-axis and sensitivity (true positive rate) on the Y-axis of the curve.

The developed pharmacophore model has huge therapeutic advantages in the screening of large chemical databases. The identified pharmacophore utilized by the methodology just mentioned and statistical approaches may serve the basis of designing active compounds against several disorders. Successful examples include novel CXCR2 agonists against cancer (Che et al., 2018), a cortisol synthesis inhibitor designed against Cushing syndrome (Akram et al., 2017), designing of ACE2

inhibitors (Rella et al., 2006), and chymase inhibitors (Arooj et al., 2013). Various software tools that are available for designing the correct pharmacophore are shown in Table 1.2. Overall, we can say that medicinal chemists and researchers can use pharmacophore approaches as complementary tools for the identification and optimization of lead molecules for accelerating the drug designing process.

A QSAR model can be developed using essential statistics such as regression coefficients of QSAR models with significance at the 95% confidence level, the squared correlation coefficient (r^2), the cross-validated squared correlation coefficient (Q^2), the standard deviation (SD), the Fisher's F-value (F), and the root mean squared error. These parameters suggest better robustness of the predicted QSAR model based on different algorithms like simulated annealing and artificial neural network (ANN). The algorithm-based acceptable QSAR model is required to have statistical parameters of higher value for the square of correlation coefficient (r^2 near to 1), and Fisher's F-value (F = max), while the value is lower for standard deviation (SD = low). The intercorrelation of these independent parameters generated for descriptors is required to develop the QSAR model.

1.4 QSAR models

It is of utmost importance to identify the drug-likeness of the compounds obtained after pharmacophore modeling and virtual screening of the chemical compound databases. QSAR-based machine learning models are continuously being used by the pharmaceutical industries to understand the structural features of a chemical that can influence biological activity (Kausar and Falcao, 2018). The QSAR-based model solely depends on the descriptors of the chemical compound. Descriptors are the numerical features extracted from the structure of a compound. The QSAR model attempts to correlate between the descriptors of the compounds with its biological activity. A brief overview of the QSAR methodology used in pharmaceutical industries and research laboratories follows.

1.4.1 Methodologies used to build QSAR models

The primary goal of all QSAR models is to analyze and detect the molecular descriptors that best describe the biological activity. The descriptors of chemical compounds are mainly classified into two categories: theoretical descriptors and experimental descriptors (Lo et al., 2018).

The theoretical descriptors are classified into 0D, 1D, 2D, 3D, and 4D types, whereas the experimental descriptors are of the hydrophobic, electronic, and steric parameter types. A brief description of descriptor types is shown in Table 1.1.

The descriptors used as input for the development of machine learning-based models predict the property of the chemical compound. QSAR methods are named after the type of descriptors used as input, such as 2D-QSAR, 3D-QSAR, and 4D-QSAR methods. A brief description of each QSAR method follows.

1.4.2 Fragment-based 2D-QSAR

In recent years, the use of 2D-QSAR models to screen and predict bioactive molecules from large databases has gained momentum in pharmaceutical industries due to their simple, easy-to-use, and robust nature. It allows the building of QSAR models even when the 3D structure of the target is mainly unknown. A hologram-based QSAR model was the first 2D-QSAR method developed by researchers that did not depend on the alignment between the calculated descriptors of a compound. First, the input compound is split into all possible fragments fed to the CRC algorithm, which then hashes the fragments into bins. The second step involves the correlation analysis of generated fragment bins with the biological activity. The basis of the final model is partial least regression that identifies the correlation of fragment bins with biological activity (IC_{50}, V_{max}).

1.4.3 3D-QSAR model

3D-QSAR models are computationally intensive, bulky, and implement complex algorithms. They are of two types: alignment dependent and alignment independent, and both types require 3D conformation of the ligand to build the final model. Comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) are the popularly used 3D-QSAR methods utilized by pharmaceutical industries for model building. The CoMFA method considers the electrostatic and steric fields in the generation and validation of a 3D model, while the CoMSIA utilizes hydrogen bond donor−acceptor interactions. Then, steric and electrostatic interactions are measured at each grid point. Subsequently, partial least squares regression analysis correlates the molecular descriptors of the ligand with the biological activities to make a final QSAR model.

1.4.4 Multidimensional or 4D-QSAR models

To tackle the limitations of 3D-QSAR methods, multidimensional QSAR models are heavily used in the pharmaceutical industries. The essential requirement for the development of 4D-QSAR methods is the 3D geometry of the receptors and ligand. One such 4D-QSAR method is Hopfinger's, which is dependent on the XMAP algorithm. The commonly used software tools for developing multidimensional QSAR models are Quasar and VirtualToxLab software.

Before applying machine learning-based QSAR modeling, the feature selection process for dimensionality reduction must ensure that only relevant and best features should be used as input in the machine learning process. Otherwise, the developed QSAR model on all relevant and irrelevant features will decrease the model's performance. The most widely used open-source feature selection tools are WEKA, scikit in Python, DWS, FEAST in Matlab, etc. A complete list of feature extraction algorithms commonly used in pharmaceutical industries is shown in Table 1.2. The selected features of the active and inactive compounds were used as input features for developing the QSAR-based machine learning model.

Machine learning-based strategies try to learn from the input structural features and predict the compounds' biological properties. The final developed QSAR model can be applied to the large chemical compound libraries to screen the compounds and predict their biological properties. All the feature selection programs utilize one or other algorithms, namely stepwise regression, simulated annealing, genetic algorithm, neural network pruning, etc.

1.4.5 Statistical methods for generation of QSAR models

The machine learning-based QSAR modeling approach has two subcategories. The first one includes regression-based model development, and the second one provides classification techniques based on the properties of the data. The regression-based statistical methods implement algorithms, such as multivariate linear regression (MLR), principal component analysis, partial least square, etc. At the same time, classification techniques include linear discriminant analysis, *k*-nearest neighbor algorithm, ANN, and cluster analysis that link qualitative information to arrive at property–structure relationships for biological activity. Each algorithm has its unique function and scoring scheme for building the predictive QSAR model (Hao et al., 2010). The general workflow and statistical details of MLR are shown in Fig. 1.2.

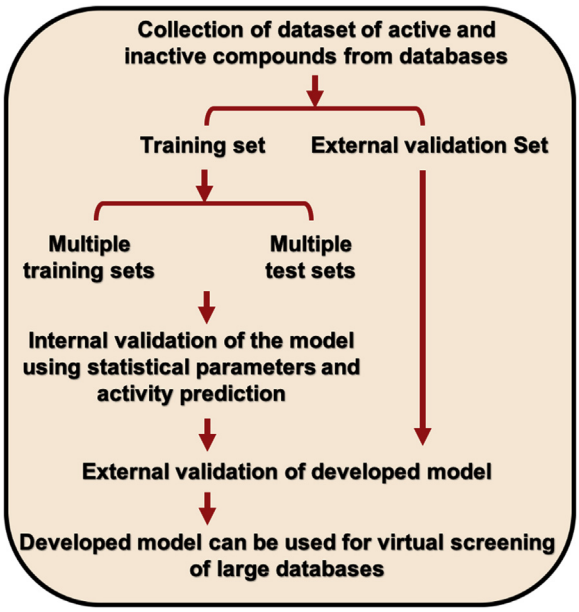


FIGURE 1.2

Overall workflow of the predictive quantitative structure–activity relationship model development.

1.4.6 Multivariate linear regression analysis

The regression analysis module of the MLR algorithm estimates the correlation between the biological activities of ligands/compounds with their molecular chemical descriptors. The essential and first step includes the finding of data points from descriptors that best suit the performance of the QSAR model. Next, a series of stepwise filters is applied, which reduces the dimensionality of descriptors to arrive at minimum descriptors that best fit the model. This will increase the predictive power of the algorithm as well as make it less computationally exhaustive. Cross-validation estimates the predictive power of the developed model. The mathematical details of the procedure, as already mentioned, are described as follows. Let X be the data matrix of descriptors (independent variable), and Y be the data vectors of biological activity (dependent variable). Then, regression coefficient b can be calculated as:

$$b = (\mathrm{X}'\mathrm{X})-1\mathrm{X}'\mathrm{Y}$$

The statistical parameter total sum of squares is a way of representing the result obtained from MLR analysis. An example set here shows all the mathematical equations. For example, the development of a QSAR model for predicting the antiinflammatory effects of the COX2 compound is done with the help of the Scigress Explore method. The correlation between the actual inhibitory value ($r^2 = 0.857$) and predicted inhibitory values ($r^2CV = 0.767$) is good enough, proving that the predicted model is of good quality. The features used in developing the predictive models are as explained in the following equation:

Predicted antiinflammatory activity $\log(\mathrm{LD}_{50}) = +0.167357 \times$ Dipole vector $\times$ (Debye) $+ 0.00695659 \times$ Steric energy (kcal/mol) $- 0.00249368 \times$ Heat of formation (kcal/mol) $+ 0.852125 \times$ Size of smallest ring $- 1.1211 \times$ Group count (carboxyl) $- 1.24227$

Here, r^2 defines the regression coefficient. For better QSAR model development, the mean difference between actual and predicted values should be minimum. If the value of r^2 varies a lot, then the model is overfitted. A brief of the general methodology used in building the QSAR model is illustrated in Fig. 1.2.

Traditional QSAR-based modeling only predicts the biological nature of the compound and is capable of screening the new molecule based on the learning. However, this approach has several limitations; all the predicted compounds do not fit into the criteria of the Lipinski rule of five and thus may have cytotoxic properties, etc. Modern QSAR-based strategies should employ various other filtration processes such as the incorporation of empirical rules, pharmacokinetic and pharmatoxicological profiles, and chemical similarity cutoff criteria to handle the aforementioned issues (Cherkasov et al., 2014). This way, a ligand with potential druggability and ADMET properties can be made in a time-efficient manner. Several software tools like click2drug, SWISS-ADME, and ADMET-SAR can solve the user's problems in predicting the desired ADMET properties of a compound.

1.5 Docking methods

Docking is an essential tool in drug discovery that predicts receptor–ligand interactions by estimating its binding affinity (Meng et al., 2012), due to its low cost and time saving that works well on a personal computer compared to experimental assays. The significant challenges in docking are a representation of receptor, ligand, structural waters, side-chain protonation, flexibility (from side-chain rotations to domain movement), stereoisomerism, input conformation, solvation, and entropy of binding (Torres et al., 2019). However, recent advances in the field of drug designing have been reported after the advent of docking and virtual screening (Lounnas et al., 2013). Receptor–ligand complex structure generation using in silico docking approaches involves two main components: posing and scoring. Docking is achieved through ligand orientational and conformational sampling in the receptor-active site, wherein scoring predicts the best native pose among the rank ligands (Chaput and Mouawad, 2017). Docking involves the structure of ligands for pose identification and ligand binding tendency to predict affinity (Clark et al., 2016). This implies that search methods of ligand flexibility are categorized into systematic strategies based on incremental construction (Rarey et al., 1996), conformational search, and databases (DOCK and FlexX). The stochastic or random approaches use genetic, Monte Carlo, and tabu search algorithms implemented in GOLD, AutoDock, and PRO_LEADS, respectively. At the same time, simulation methods are associated with molecular dynamics (MD) simulations and global energy minimization (DOCK) (Yuriev et al., 2011).

The receptor is represented as a 3D structure in docking obtained from NMR, X-ray crystallography, threading, homology modeling, and de novo methods. Nevertheless, ligand binding is a dynamic event instead of a static process, wherein both ligand and protein exhibit conformational changes.

Several docking software and virtual screening tools (Table 1.2) are available and widely used. Nonetheless, one such software that explicitly addresses receptor flexibility is RosettaLigand, which uses the stochastic Monte Carlo approach, wherein a simulated annealing procedure optimizes the binding site side-chain rotamers (Davis et al., 2009). Another software, Autodock4, completely models the flexibility of the selected protein portion in which selected side chains of the protein can be separated and explicitly treated during simulations that enable rotation throughout the torsional degree of freedom (Bianco et al., 2016). Alternatively, the protein can be made flexible by the Insight II side-chain rotamer libraries (Wang et al., 2005). Besides, the Induced Fit Docking (IFD) workflow of Schrodinger software relies on rigid docking using the Glide module combined with the minimization of complexes and homology modeling. IFD has been used for kinases (Zhong et al., 2009), HIV-1 integrase (Barreca et al., 2009), heat shock protein 90 (Lauria et al., 2009), and monoacylglycerol lipase (King et al., 2009) studies. Furthermore, atom receptor flexibility into docking was introduced using MD simulations, which measured its effect on the accuracy of this tool by cross-docking (Armen et al., 2009). The best complex models are obtained based on flexible side chains and multiple flexible backbone segments.

In contrast, the binding of docked complexes containing flexible loops and entirely flexible targets was found less accurate because of increased noise that affects its scoring function. Internal Coordinate Mechanics (ICM), a 4D-docking protocol, was reported where the fourth dimension represents receptor conformation (Abagyan and Totrov, 1994). ICM accuracy was found to be increased using multiple grids that described multiple receptor conformations compared to single grid methods. A gradient-based optimization algorithm was implemented in a local minimization tool used to calculate the orientational gradient by adjusting parameters without altering molecular orientation (Fuhrmann et al., 2009). The docking approaches are computationally costly for creating docker ligand libraries, receptor ensembles, and developing individual ligands against larger ensembles (Huang and Zou, 2006). Normal mode analysis used to generate receptor ensembles is one of the best alternatives to MD simulations (Moroy et al., 2015). The elastic network model (ENM) method induces local conformational changes in the side chains and protein backbone, which signifies its importance more efficiently than MD simulations.

A small change in the ligand conformation causes significant variations in the scores of docked poses and geometries. This suggests that no method or ligand geometry produces the most precise docking pose (Meng et al., 2012). Ligand conformational treatment has been precomputed through several available methods like the generation of ligand conformations (TrixX Conformer Generator) (Griewel et al., 2009), systematic sampling (MOLSDOCK and AutoDock 4) (Viji et al., 2012), incremental construction (DOCK 6), genetic algorithms (Jones et al., 1997), Lamarckian genetic algorithm (FITTED and AutoDock), and Monte Carlo (RosettaLigand and AutoDock-Vina).

1.5.1 Scoring functions

Docking software and webservers are validated by producing "correct" binding modes based on the ranking, which identifies active and inactive compounds still under study. Thus, several attempts have been made to improve scoring functions like entropy (Li et al., 2010), desolvation effects (Fong et al., 2009), and target specificity. Mainly, four types of scoring functions have been categorized and implemented in forcefields: classical (D-Score, G-Score, GOLD, AutoDock, and DOCK) (Hevener et al., 2009); empirical (PLANTSCHEMPLP, PLANTSPLP) (Korb et al., 2009), RankScore 2.0, 3.0, and 4.0 (Englebienne and Moitessier, 2009), Nscore (Tarasov and Tovbin, 2009), LUDI, F-Score, ChemScore, and X-SCORE (Cheng et al., 2009); knowledge (ITScore/SE) (Huang and Zou, 2010), PoseScore, DrugScore (Li et al., 2010), and MotifScore based; and machine learning (RF-Score, NNScore) (Durrant and McCammon, 2010).

Docking calculations of entropies are included within the Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA), wherein it is a modified form of framework, and the entropy loss is calculated. This is correspondingly assessed after ligand−receptor binding based on the loss of rotational, torsional, translational, vibrational, and free energies. The modification includes the free energy change

of ligand in free or bound states. In contrast, the reorganization energy of ligand requires the prediction of native binding affinities included in the new scoring function.

In terms of specificity and binding affinity, water molecules play an essential role in receptor–ligand complexes. Thus it is necessary to consider specific water molecules to predict the effect of solvation in docking. An empirical solvent-accessible surface area energy function gave an improved success rate in pose prediction compared to native experimental binding scores. Still, it failed for receptors where electrostatic interactions are considered. On the contrary, in silico, MM/PBSA, and Molecular Mechanics/Generalized Born and Surface Area (MM/GBSA) calculations are performed for an ensemble if a receptor–ligand complex correlates with experimentally measured binding free energies (Hou et al., 2011).

Besides, the scoring functions based on Molecular Mechanical/Quantum Mechanical (MM/QM) have been considered for the treatment of ligand in combination with GoldScore, ChemScore, and AMBER to predict the right poses based on three essential functions: AM1d, HF/6-31G, and PM3 (Fong et al., 2009). Furthermore, cross-docking was also performed using a combination of Universal Force Field and B3LYP/6-31G. Similarly, an MM/QM-based docking program, QM-Polarized Ligand Docking with SiteMap, was developed to identify binding sites that predict improved scoring compared to Glide in terms of hydrophilic, hydrophobic, and metalloprotein binding sites (Chung et al., 2009). The statistical parameters of receptor–ligand complex structures are summarized in knowledge-based scoring functions that can handle two crucial tasks: pose prediction and ligand ranking (Charifson et al., 1999). Consensus scoring predicts the binding affinities and evaluates multiple-docked pose rescoring combined with specific scoring functions. The four universal forcefield energy functions have been applied in consensus scoring of fragment-based virtual screening to estimate binding free energy: CHARMm electrostatic interaction energy, Van der Waals efficiency, TAFF interaction energy, and linear interaction energy with continuum electrostatics (Friedman and Caflisch, 2009). However, a combination of ASP, ChemScorePLP, LigScore, GlideScore, and DrugScorein scoring function was considered (Li et al., 2014). Virtual screening against kinases (Brooijmans and Humblet, 2010) was successfully applied using consensus scoring such as VoteDock development (Plewczynski et al., 2011), a knowledge-based approach combining the quantitative structure and binding affinity relationship, and MedusaScore, a forcefield-based method is a combination of GOLD and MCSS docking with fragments rescoring using MM/GBSA, HarmonyDOCK (Plewczynski et al., 2014), a combination of AutoDock4 and Vina, PMF (Okamoto et al., 2010), DOCK4 (Ewing et al., 2001), and FlexX. Not all functions of scoring are accurate to identify correct binding affinity. Consequently, machine learning is currently considered essential to develop a new neural network-based scoring function, NNScore, which is found to be very fast and accurate (Durrant and McCammon, 2010). NNScore distinguished precisely between active and decoy

ligands using pKd values. Similarly, RF-Score based on interacting atom pair counts of ligand and receptor using the machine learning-based approaches suggested a new scoring function with correct binding affinity prediction. Similarly, a support vector machine-based model demonstrated improved affinity prediction using docking energy and native binding affinities (Kinnings et al., 2011). Subsequently, a regression model and a classification model, trained on IC_{50} values from BindingDB and active compounds, respectively, and decoys from the DUD database were used. Afterward, scoring prediction was improved with interaction fingerprints and profile-based methods, wherein Glide XP, a new precision scoring function descriptor, was used to identify standard pharmacophoric features of the docked fragments.

1.5.2 Pose prediction

Docking methods rank the predicted binding affinities and poses based on their scoring functions. However, docking-based prediction of the binding mode is not always reliable, and indicates that there is no universal docking method. Since the docking technique works best in small ligands and controls binding sites (Kolb and Irwin, 2009), it was used in combination with pharmacophore modeling to predict the correct pose. Also, the associated locations of pairs of interacting atoms were taken into account as a new atom pair IF-based method that demonstrated the improved pose prediction (Perez-Nueno et al., 2009). The entropic term ($-T\Delta S$) was used in the analysis of MM/PBSA to identify the highly stable docking pose (Yasuo et al., 2009). An in silico fragment-based approach was developed through searching local similarity of a protein. A database of MED portions containing experimental protein−ligand structures was combined with MED-SuMo, a superimposition tool, and MED-Hybridize, a tool for linking chemical moieties to known ligands, which retrieved similar matching portions of ligands for a query. Likewise, the fragment mapping approach (FTMap) successfully identified protein hotspots suitable for drug targeting (Landon et al., 2009).

In contrast, machine/deep learning techniques were found to be better at predicting receptor−ligand binding poses. This represented the convolutional neural network (CNN)-based scoring functions, which utilized 3D receptor−ligand complex structure as input. The scoring function of CNN learns the characteristics of protein−ligand binding automatically. The trained CNN scoring functions separate the correct binding poses from incorrect and known binders from nonbinders with better accuracy as compared to AutoDock Vina. The native ligand pose prediction of docked and experimental binding modes is validated by measured root mean square deviation within a range of 2 Å, thus gaining useful information and a potential pose. The best scoring function always obtained the correct binding pose by considering lower $\Delta\Delta G$ that demonstrated the most stable protein−ligand complex (Ferrara et al., 2004).

1.5.3 MD simulations

The flexibility of receptor atoms into docking was introduced using MD simulations, which measured docking accuracy through cross-docking (Armen et al., 2009). The obtained ligand−receptor complex is considered the best model, which includes flexible side chains and multiple flexible backbones. In contrast, accuracy was found reduced in complexes, which provided flexible loops and entirely flexible targets because of the increased noise that affects scoring function. This noise can be overcome using developed ICM, a 4D-docking protocol, wherein the receptor conformation was considered the fourth dimension (Abagyan and Totrov, 1994). ICM accuracy was increased using multiple grids that represent multiple receptor conformations. Besides, a gradient-based optimization algorithm helped to calculate the orientational gradient. This calculation was achieved by adopting a local minimization algorithm to modify the orientational parameters but maintaining its molecular orientation (Fuhrmann et al., 2009). The computational approach is utilized in developing receptor ensembles and docked ligand databases, while normal mode analysis is one of the best alternatives to MD for generating receptor ensembles (Moroy et al., 2015). Additionally, an ENM method was found to be more efficient than MD simulations in identifying local side-chain conformational changes and protein backbone movement. Several MD simulation tools have been widely used (Table 1.2), in which Gromacs and NAMD are two useful tools commonly used to predict protein structure and docked complex stability near-native states (Raghav et al., 2012a,b; Raghav et al., 2018).

1.6 Conclusion

Modern drugs to combat diseases are based on chemical compounds. The identification of a target drug for a disease is still a challenging task for medical researchers due to complex behaviors and patterns of interaction of the cellular entities inside the body. In this regard, designing and developing new effective molecules with few or no side effects is becoming mandatory for the advancement of the human lifestyle. Chemoinformatics is a branch of bioinformatics that deals with the design, analysis, management, and visualization of small molecules data for the drug discovery process. It involves the use of tools and databases for the retrieval of information from chemical compounds with the intended aim of making better decisions in areas of drug discovery and lead identification. Several tools and databases are available in the literature to represent the chemical structure, perform the QSAR study, and predict the chemical, physical, and biological properties of chemical compounds. Chemoinformatics is an advanced field in the modern drug discovery process, which is used for the understanding of complex patterns of chemicals and biomolecules. In this regard, the range of applications of chemoinformatics is rich; indeed, any field of chemistry can profit from its methods. Therefore, this chapter focused on the collection and compilation of essential chemoinformatics tools and databases, which

are commonly used in studies and industries for the advancement of drug discovery. Also, the chapter would likely help chemists and researchers in understanding the flow of modern drug discovery processes.

Acknowledgments

PKR would like to acknowledge the Department of Science and Technology-Science and Engineering Research Boards (DST-SERB), India, for providing a financial grant (PDF/2016/003387) for this work.

References

Abagyan, R., Totrov, M., 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235 (3), 983−1002. https://doi.org/10.1006/jmbi.1994.1052.

Akram, M., et al., 2017. Pharmacophore modeling and in Silico/in Vitro screening for human cytochrome P450 11B1 and cytochrome P450 11B2 inhibitors. Front. Chem. Front. Media S. A 5 (DEC). https://doi.org/10.3389/fchem.2017.00104.

Armen, R.S., Chen, J., Brooks, C.L., 2009. An evaluation of explicit receptor flexibility in molecular docking using molecular dynamics and torsion angle molecular dynamics. J. Chem. Theory Comput. 5 (10), 2909−2923. https://doi.org/10.1021/ct900262t.

Arooj, M., et al., 2013. A combination of receptor-based pharmacophore modeling & QM techniques for identification of human chymase inhibitors. PLoS One 8 (4). https://doi.org/10.1371/journal.pone.0063030.

Barreca, M.L., et al., 2009. Induced-fit docking approach provides insight into the binding mode and mechanism of action of HIV-1 integrase inhibitors. ChemMedChem 4 (9), 1446−1456. https://doi.org/10.1002/cmdc.200900166.

Bianco, G., et al., 2016. Covalent docking using autodock: two-point attractor and flexible side chain methods. Protein Sci. 25 (1), 295−301. https://doi.org/10.1002/pro.2733. Blackwell Publishing Ltd.

Brooijmans, N., Humblet, C., 2010. Chemical space sampling in virtual screening by different crystal structures. Chem. Biol. Drug Des. 76 (6), 472−479. https://doi.org/10.1111/j.1747-0285.2010.01041.x.

Chaput, L., Mouawad, L., 2017. Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. J. Cheminf. 9 (1) https://doi.org/10.1186/s13321-017-0227-x. BioMed Central Ltd.

Charifson, P.S., et al., 1999. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J. Med. Chem. 42 (25), 5100−5109. https://doi.org/10.1021/jm990352k.

Che, J., et al., 2018. Ligand-based pharmacophore model for the discovery of novel CXCR2 antagonists as anti-cancer metastatic agents. Royal Soci. Open Sci. 5 (7) https://doi.org/10.1098/rsos.180176. Royal Society Publishing.

Cheng, T., et al., 2009. Comparative assessment of scoring functions on a diverse test set. J. Chem. Inf. Model. 49 (4), 1079−1093. https://doi.org/10.1021/ci9000053.

Cherkasov, A., et al., 2014. QSAR modeling: where have you been? Where are you going to? J. Med. Chem. Am. Chem. Soc. 4977−5010. https://doi.org/10.1021/jm4004285.

Chung, J.Y., Hah, J.M., Cho, A.E., 2009. Correlation between performance of QM/MM docking and simple classification of binding sites. J. Chem. Inform. Model. 49 (10), 2382−2387. https://doi.org/10.1021/ci900231p.

Clark, A.J., et al., 2016. Prediction of protein-ligand binding poses via a combination of induced fit docking and metadynamics simulations. J. Chem. Theory Comput. Am. Chem. Soc. 12 (6), 2990−2998. https://doi.org/10.1021/acs.jctc.6b00201.

Davis, I.W., et al., 2009. Blind docking of pharmaceutically relevant compounds using RosettaLigand. Protein Sci. Protein Sci. 18 (9), 1998−2002. https://doi.org/10.1002/pro.192.

Durrant, J.D., McCammon, J.A., 2010. NNScore: a neural-network-based scoring function for the characterization of protein-ligand complexes. J. Chem. Inform. Model. 50 (10), 1865−1871. https://doi.org/10.1021/ci100244v.

Englebienne, P., Moitessier, N., 2009. Docking ligands into flexible and solvated macromolecules. 5. Force-field-based prediction of binding affinities of ligands to proteins. J. Chem. Inf. Model. 49 (11), 2564−2571. https://doi.org/10.1021/ci900251k.

Ewing, T.J.A., et al., 2001. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput. Aid. Mol. Des. 15 (5), 411−428. https://doi.org/10.1023/A:1011115820450.

Ferrara, P., et al., 2004. Assessing scoring functions for protein-ligand interactions. J.Med. Chem. 47 (12), 3032−3047. https://doi.org/10.1021/jm030489h.

Fong, P., et al., 2009. Assessment of QM/MM scoring functions for molecular docking to HIV-1 protease. J. Chem. Inf. Model. 49 (4), 913−924. https://doi.org/10.1021/ci800432s.

Friedman, R., Caflisch, A., 2009. Discovery of plasmepsin inhibitors by fragment-based docking and consensus scoring. ChemMedChem 4 (8), 1317−1326. https://doi.org/10.1002/cmdc.200900078.

Fuhrmann, J., et al., 2009. A new method for the gradient-based optimization of molecular complexes. J. Comput. Chem. 30 (9), 1371−1378. https://doi.org/10.1002/jcc.21159.

Goto, S., Nishioka, T., Kanehisa, M., 2000. LIGAND: chemical database of enzyme reactions. Nucleic Acid. Res. 28 (1), 380−382. https://doi.org/10.1093/nar/28.1.380.

Griewel, A., et al., 2009. Conformational sampling for large-scale virtual screening: accuracy versus ensemble size. J. Chem. Inf. Model. 49 (10), 2303−2311. https://doi.org/10.1021/ci9002415.

Groom, C.R., et al., 2016. The Cambridge structural database', *acta crystallographica section B: structural science, crystal Engineering and materials*. Int. Union Crystal. 72 (2), 171−179. https://doi.org/10.1107/S2052520616003954.

Hao, M., et al., 2010. Prediction of PKCθ inhibitory activity using the random forest algorithm. Int. J. Mol. Sci. 11 (9), 3413−3433. https://doi.org/10.3390/ijms11093413.

Hevener, K.E., et al., 2009. Validation of molecular docking programs for virtual screening against dihydropteroate synthase. J. Chem. Inf. Model. 49 (2), 444−460. https://doi.org/10.1021/ci800293n.

Hou, T., et al., 2011. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J. Chem. Inf. Model. 51 (1), 69−82. https://doi.org/10.1021/ci100275a.

Huang, S.-Y., Zou, X., 2006. Efficient molecular docking of NMR structures: application to HIV-1 protease. Protein Sci. 16 (1), 43−51. https://doi.org/10.1110/ps.062501507. Wiley.

Huang, S.Y., Zou, X., 2010. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. J. Chem. Inf. Model. 50 (2), 262−273. https://doi.org/10.1021/ci9002987.

Jones, G., et al., 1997. Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. 267 (3), 727−748. https://doi.org/10.1006/jmbi.1996.0897. Academic Press.

Kausar, S., Falcao, A.O., 2018. An automated framework for QSAR model building. J. Cheminf. 10 (1) https://doi.org/10.1186/s13321-017-0256-5. BioMed Central Ltd.

King, A.R., et al., 2009. Discovery of potent and reversible monoacylglycerol lipase inhibitors. Chem. Biol. 16 (10), 1045−1052. https://doi.org/10.1016/j.chembiol.2009.09.012.

Kinnings, S.L., et al., 2011. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. J. Chem. Inf. Model. 51 (2), 408−419. https://doi.org/10.1021/ci100369f.

Kolb, P., Irwin, J., 2009. Docking screens: right for the right reasons? Curr. Top. Med. Chem. 9 (9), 755−770. https://doi.org/10.2174/156802609789207091. Bentham Science Publishers Ltd.

Korb, O., Stützle, T., Exner, T.E., 2009. Empirical scoring functions for advanced Protein-Ligand docking with PLANTS. J. Chem. Inf. Model. 49 (1), 84−96. https://doi.org/10.1021/ci800298z.

Landon, M.R., et al., 2009. Detection of ligand binding hot spots on protein surfaces via fragment-based methods: application to DJ-1 and glucocerebrosidase. J. Comput. Aid. Mol. Des. 23 (8), 491−500. https://doi.org/10.1007/s10822-009-9283-2.

Lauria, A., Ippolito, M., Almerico, A.M., 2009. Inside the Hsp90 inhibitors binding mode through induced fit docking. J. Mol. Graph. Model. 27 (6), 712−722. https://doi.org/10.1016/j.jmgm.2008.11.004.

Li, Y., et al., 2014. Comparative assessment of scoring functions on an updated benchmark: 1. compilation of the test set. J. Chem. Inform. Model. Am. Chem. Soci. 54 (6), 1700−1716. https://doi.org/10.1021/ci500080q.

Li, Y., Liu, Z., Wang, R., 2010. Test MM-PB/SA on true conformational ensembles of protein-ligand complexes. J. Chem. Inf. Model. 50 (9), 1682−1692. https://doi.org/10.1021/ci100036a.

Lo, Y.C., et al., 2018. Machine learning in chemoinformatics and drug discovery. Drug Discov. Today 1538−1546. https://doi.org/10.1016/j.drudis.2018.05.010. Elsevier Ltd.

Lounnas, V., et al., 2013. Current progress in structure-based rational drug design marks a new mindset in drug discovery. Comput. Struct. Biotechnol. J. e201302011. https://doi.org/10.5936/csbj.201302011. Research Network of Computational and Structural Biotechnology.

Lourenco, A.M., Ferreira, L.M., Branco, P.S., 2012. Molecules of natural origin, semi-synthesis and synthesis with anti-inflammatory and anticancer utilities. Curr. Pharmaceut. Des. 18 (26), 3979−4046. https://doi.org/10.2174/138161212802083644. Bentham Science Publishers Ltd.

Meng, X.-Y., et al., 2012. Molecular docking: a powerful approach for structure-based drug discovery. Curr. Comput. Aided Drug Des. 7 (2), 146−157. https://doi.org/10.2174/157340911795677602. Bentham Science Publishers Ltd.

Moroy, G., et al., 2015. Sampling of conformational ensemble for virtual screening using molecular dynamics simulations and normal mode analysis. Fut. Med. Chem. Fut. Sci. 7 (17), 2317−2331. https://doi.org/10.4155/fmc.15.150.

Newman, D.J., Cragg, G.M., 2007. Natural products as sources of new drugs over the last 25 years. J. Nat. Prod. 461−477. https://doi.org/10.1021/np068054v.

Okamoto, M., et al., 2010. Evaluation of docking calculations on X-ray structures using CONSENSUS-DOCK', *Chemical and Pharmaceutical Bulletin*. Chem. Pharm. Bull. 58 (12), 1655−1657. https://doi.org/10.1248/cpb.58.1655.

Perez-Nueno, V.I., et al., 2009. APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening. J. Chem. Inform. Model. 49 (5), 1245−1260. https://doi.org/10.1021/ci900043r.

Plewczynski, D., et al., 2011. VoteDock: consensus docking method for prediction of protein-ligand interactions. J. Comput. Chem. 32 (4), 568−581. https://doi.org/10.1002/jcc.21642.

Plewczynski, D., et al., 2014. HarmonyDOCK: the structural analysis of poses in protein-ligand docking. J. Comput. Biol. 21 (3), 247−256. https://doi.org/10.1089/cmb.2009.0111.

Raghav, P.K., Singh, A.K., Gangenahalli, G., 2018. A change in structural integrity of c-Kit mutant D816V causes constitutive signaling. Mutat. Res. Fund Mol. Mech. Mutagen 808, 28−38. https://doi.org/10.1016/j.mrfmmm.2018.02.001. Elsevier B.V.

Raghav, P.K., Verma, Y.K., Gangenahalli, G.U., 2012. Molecular dynamics simulations of the Bcl-2 protein to predict the structure of its unordered flexible loop domain. J. Mol. Model. 18 (5), 1885−1906. https://doi.org/10.1007/s00894-011-1201-6.

Raghav, P.K., Verma, Y.K., Gangenahalli, G.U., 2012. Peptide screening to knockdown Bcl-2's anti-apoptotic activity: implications in cancer treatment. Int. J. Biol. Macromol. 50 (3), 796−814. https://doi.org/10.1016/j.ijbiomac.2011.11.021.

Rarey, M., et al., 1996. A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. 261 (3), 470−489. https://doi.org/10.1006/jmbi.1996.0477. Academic Press.

Rella, M., et al., 2006. Structure-based pharmacophore design and virtual screening for novel Angiotensin Converting Enzyme 2 inhibitors. J. Chem. Inform. Model. 708−716. https://doi.org/10.1021/ci0503614.

Rose, W., et al., 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. 45 (D1), D271−D281. https://doi.org/10.1093/nar/gkw1000, 2017.

Tarasov, D., Tovbin, D., 2009. How sophisticated should a scoring function be to ensure successful docking, scoring and virtual screening? J. Mol. Model. 15 (3), 329−341. https://doi.org/10.1007/s00894-008-0390-0.

Torres, P.H.M., et al., 2019. Key topics in molecular docking for drug design. Int. J. Mole. Sci. https://doi.org/10.3390/ijms20184574. MDPI AG.

Viji, S.N., Balaji, N., Gautham, N., 2012. Molecular docking studies of protein-nucleotide complexes using MOLSDOCK (mutually orthogonal Latin squares DOCK). J. Mol. Model. 18 (8), 3705−3722. https://doi.org/10.1007/s00894-012-1369-4.

Wang, C., Schueler-Furman, O., Baker, D., 2005. Improved side-chain modeling for protein-protein docking. Protein Sci. 14 (5), 1328−1339. https://doi.org/10.1110/ps.041222905. Wiley.

Yasuo, K., et al., 2009. Structure-based CoMFA as a predictive model - CYP2C9 inhibitors as a test case. J. Chem. Inf. Model. 49 (4), 853−864. https://doi.org/10.1021/ci800313h.

Yuriev, E., Agostino, M., Ramsland, P.A., 2011. Challenges and advances in computational docking: 2009 in review. J. Mol. Recognit. 149−164. https://doi.org/10.1002/jmr.1077.

Zhong, H., Tran, L.M., Stang, J.L., 2009. Induced-fit docking studies of the active and inactive states of protein tyrosine kinases. J. Mole. Graph. Model. 28 (4), 336−346. https://doi.org/10.1016/j.jmgm.2009.08.012.

CHAPTER

2 Structure- and ligand-based drug design: concepts, approaches, and challenges

Vidushi Sharma[1], **Sharad Wakode**[1], **Hirdesh Kumar**[2]

[1]*Delhi Institute of Pharmaceutical Education and Research, New Delhi, Delhi, India;* [2]*Laboratory of Malaria Immunology and Vaccinology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, United States*

2.1 Introduction

The bench-to-bedside journey of a drug costs more than a billion US dollars and takes more than 10 years (Steven et al., 2010; Bernard, 2009). The classical approach to identify hit molecules against a disease is high-throughput screening (HTS). HTS involves the screening of hundreds of thousands of small molecules to identify active molecules that can elicit desired biological response. This brute-force approach often fails due to instability, toxicity, or other poor pharmacokinetic/pharmacodynamic properties of hit molecules. Among other approaches, computer-aided drug design (CADD) is effective in reducing the cost, duration, and attrition rate of the drug discovery process. CADD involves predictive algorithms, computing resources, and 3D visualization tools to design, optimize, and develop small molecule therapeutics against diseases. The early detection of undesired molecules reduces the cost and workload of HTS without compromising the success rate. For example, in a comparative case study, a group performed virtual screening (VS) of small molecules against tyrosine phosphatase-1B, a therapeutic target in diabetes mellitus. They reported 365 compounds, among which 127 (35%) showed an effective inhibition. In parallel, the group performed traditional HTS and among 400,000 compounds tested, only 81 (0.021%) showed effective inhibition (Thompson et al., 2002). This study showed the power of CADD, which has become an integral part of the drug discovery process. In contrast to HTS or combinatorial chemistry, CADD involves much more targeted searching and therefore results in higher success rates. Table 2.1 summarizes the success case studies of drugs that have been identified by CADD.

There are three major roles of CADD in pharmaceutical industries: (1) the screening of large libraries of molecules to predict minimal best small molecules to further test in actual experiments; (2) lead identification by designing novel

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00004-X

Table 2.1 Success case studies of drug discovery by computer-aided drug design.

Drug	Target	Diseases	References
Amprenavir	Human immunodeficiency virus (HIV) protease	HIV	Wlodawer and Vondrasek (1998)
Captopril	Angiotensin-converting enzyme	Blood pressure	Redshaw (1993)
Dorzolamide	Carbonic anhydrase	Glaucoma	Baldwin et al. (1989)
Raltitrexed	Thymidylate synthase	HIV	
Isoniazid	InhA	Tuberculosis	Hedia et al. (2000)
Inhibitors	Pim-1 kinase	Cancer	Ji-Xia et al. (2011)
Epalrestat 2	Aldose reductase	Diabetic neuropathy	Ling et al. (2013)
Flurbiprofen	Cyclooxygenase-2	Rheumatoid arthritis	Zachary et al. (2015)
STX-0119	STAT3	Lymphoma	Kenji et al. (2010)
Norfloxacin	Topoisomerase II, IV	Urinary infection	
Dorzolamide	Carbonic anhydrase	Glaucoma	

small molecules; and (3) lead optimization for affinity or pharmacokinetic/pharmacodynamic (PK/PD) properties. In this structural genomic era, the traditional usage of CADD has been extended upstream to target identification/validation and downstream for absorption, distribution, metabolism, excretion, and toxicity (ADME/T) predictions in preclinical studies. This chapter excludes bioinformatics techniques, which are crucial in identification of a drug or vaccine target (Kumar et al., 2015, 2019). Fig. 2.1 illustrates the role of CADD in the drug discovery pipeline.

2.1.1 Advantages of CADD

The main advantages of CADD are:

- Screening of millions of small molecules to preselect a handful of potential candidates for further evaluation in experimental testing.
- Design of novel, patentable compounds through de novo, fragment-based, or scaffold-hopping approaches.
- Minimization of experiments in animals/humans.
- Prediction of the PK/PD parameters of lead molecules.
- Construction and usage of high-quality datasets and libraries to optimize lead molecules for diversity or similarity.
- Reduction and overcoming of drug resistance.

Although a CADD scientist can aid in the discovery of novel drug candidates by making the process quick and efficient, there are several hurdles and subtleties in the

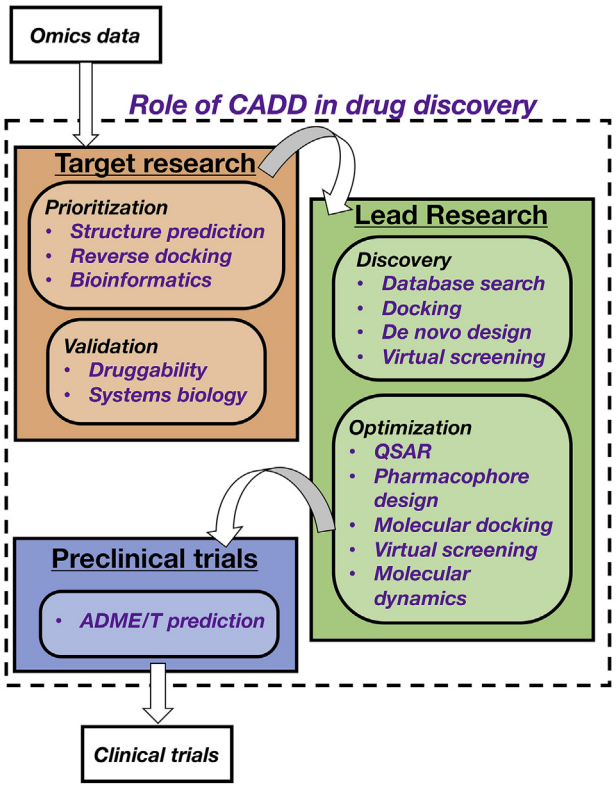


FIGURE 2.1

Role of computer-aided drug design (CADD) in drug discovery.
ADME/T, Absorption, distribution, metabolism, excretion, and toxicity; *QSAR*, quantitative structure—activity relationship.

drug discovery process. Therefore a multidisciplinary team is always needed to identify/validate a therapeutic target/drug candidate in preclinical and clinical studies. CADD can be broadly classified in two categories on the basis of any structural information available for therapeutic targets: ligand-based drug design (LBDD) and structure-based drug design (SBDD). LBDD exploits experimental knowledge of active/inactive molecules, whereas SBDD demands the structural knowledge of the therapeutic target.

2.2 Ligand-based drug design

In the absence of any structure information available for the therapeutic target, the alternative approach is LBDD. Unlike SBDD, LBDD does not require a priori knowledge of mechanisms of action and only needs structural information and

bioactivity data for small molecules. The principle of LBDD is that structurally similar molecules are likely to have similar properties (Hendrickson, 1991). An imperative step in LBDD is to retrieve and prepare small molecule libraries. Chemical structures are usually created, processed, and utilized as *molecular graphs*. A molecular graph is a combination of *nodes* and *edges* in which atoms and bonds are represented as nodes and edges, respectively. *Connection tables* and *linear notations* are two common ways to communicate with molecular graphs. A connection table contains sections with information about atom types, connection types, and coordinates. Examples of connection tables include: mol2, sdf, pdb, etc. file formats. A linear notation is a combination of alphanumeric characters. Examples of linear notations include simplified molecular input line entry specification (SMILE) and Wiswesser line notation. Linear notations are more compact than connection tables and therefore are preferred while storing or transferring millions of small molecules. A list of different small molecular databases is given in Table 2.2.

The most common LBDD techniques include molecular similarity-based search, quantitative structure−activity relationship (QSAR), and pharmacophore modeling. These techniques are discussed in the next sections.

Table 2.2 List of small molecule resources.

Name	Weblink
AffinDB	http://pc1664.pharmazie.uni-marburg.de/affinity/
Aureus	http://www.aureuspharma.com/Pages/Products/Aurscope.php
BindingDB	http://www.bindingdb.org
BindingMOAD	http://www.bindingmoad.org
BioPrint	http://www.eidogen-sertanty.com/products_kinasekb.html
ChEMBLdb	http://www.ebi.ac.uk/chembldb/
CTD	http://ctd.mdibl.org
DrugBank	http://www.drugbank.ca
Eidogen-Sertanty	http://www.eidogen-sertanty.com/products_kinasekb.html
GLIDA	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/
GVKBio	http://www.gvkbio.com/informatics.html
IUPHARdb	http://www.iuphar-db.org
KEGG	http://www.genome.jp/kegg/
NikkajiWeb	http://nikkajiweb.jst.go.jp
PDSP	http://pdsp.med.unc.edu
PharmGKB	http://www.pharmgkb.org
PubChem	http://pubchem.ncbi.nlm.nih.gov
STITCH	http://stitch.embl.de
Symyx	http://www.symyx.com/products/databases/bioactivity/
WOMBAT	http://www.sunsetmolecular.com

2.2.1 Molecular similarity-based search

2.2.1.1 Concept

Molecular similarity-based search is the simplest LBDD technique to identity desired small molecules. Molecular similarity-based search is an independent as well as integral part of other LBDD and SBDD techniques in which small molecule libraries are searched using molecular descriptors. Molecular descriptors are characteristic numerical values that represent small molecules and range from simple physicochemical properties to complex structural properties. Examples of molecular descriptors include molecular weight, atom types, bond distances, surface area, electronegativities, atom distributions, aromaticity indices, solvent properties, and many others (Leach and Valerie, 2007). Molecular descriptors are derived through experiments, quantum-mechanical tools, or previous knowledge. Depending on the "dimensionality," molecular descriptors can be a 1D, 2D, or 3D descriptor. 1D descriptors are scalar physicochemical properties of a molecule such as molecular weight, logP values, and molar refractivity. 2D descriptors are derived from molecular constitution or configuration and include topological indices and 2D fingerprints. 3D descriptors are derived from the conformation of molecules. 3D descriptors include 3D fingerprints, dipole moments, highest occupied molecular orbital/lowest unoccupied molecular orbital energies, electrostatic potentials, etc. A list of software that predicts molecular descriptors of small molecules is given in Table 2.3.

Table 2.3 Common software to predict molecular descriptors.

Software	Total numbers (types of predicted descriptors)
ADAPT	>260 (topological, geometrical, electronic, physicochemical)
ADMET Predictor	>290 (constitutional, functional group counts, topological, E-state, moriguchi descriptors, meylan flags, molecular patterns, electronic properties, 3D descriptors, hydrogen bonding, acid–base ionization, empirical estimates of quantum descriptors)
CODESSA	>1500 (constitutional, topological, geometrical, charge related, semiempirical, thermodynamical)
DRAGON	>5200 (constitutional, topological, 2D autocorrelations, geometrical, WHIM, GETAWAY, RDF, functional groups, properties, 2D binary and 2D frequency fingerprints, etc.)
MARVIN Beans	>500 (physicochemical, topological, geometrical, fingerprints, etc.)
MOE	>300 (topological, physical properties, structural keys, etc.)
MOLGEN-QSPR	>700 (constitutional, topological, geometrical, etc.)
PreADMET	>955 (constitutional, topological, geometrical, physicochemical, etc.)

Molecular descriptors allow a rapid comparison of structural and/or physicochemical features of small molecules. Tanimoto coefficient, T, is the most popular tool for measuring the similarity between the two molecules. Although $T > 0.85$ suggests a good fit, it does not reflect biosimilarity between the two molecules.

2.2.1.2 Workflow

The different steps involved in a molecular similarity-based search are:

- *Standard formatting:* At first, molecules are read and converted to the standard formats for further steps. The standard formats are mol, mol2, SDF, pdb, etc. The step is critical to reject ambiguous structures such as free radicles, wrong valences, polymers, etc.
- *Filtering:* In addition to desired small molecules, small molecule databases cover a large number of problematic molecules containing isotype atoms, inorganic atoms, or charged carbon atoms. It is imperative to reject such undesired small molecules in a similarity-based search using molecular filters. One of the most popular criteria to filter small molecule libraries is drug-likeliness. Drug-likeliness is evaluated by applying the Lipinski et al. (2001) rule of five, which states that a drug-like candidate should have (1) <5 hydrogen bond donor atoms, (2) <10 oxygen or nitrogen atoms, (3) a molecular mass of <500 Da, and (4) an octanol−water partition coefficient of <5. Violation of two or more of these rules leads to poor absorption. Other commonly used screening filters include extended drug-like filters (Daniel et al., 2002), fragment-like filter (Simon et al., 1999), Egan filter (Egan et al., 2000), Veber filter (Daniel et al., 2002), etc.
- *Remove duplicates:* A molecule can have different protonation states and thus different tautomers. It is the user's responsibility to decide if such duplicates are important for study or not.

2.2.1.3 Applications

- *Identification of novel targets based on chemical similarities of small molecules:* Keiser et al. correlated different receptors on the basis of ligand similarities and annotated 65,000 ligands into sets of hundreds of drug targets. The authors built minimal spanning trees solely based on ligand similarity, and predicted and validated novel biological targets for ligands (Keiser et al., 2007).
- *Off-target prediction:* As previously stated, molecular similarity measures such as Tanimoto coefficient are efficient tools to cluster and build networks of similar small molecules. Recently, chemical similarity measures such as Tanimoto coefficients are used to predict binding to multiple therapeutic targets and thus to predict off-targets and adverse drug reactions.

2.2.1.4 Challenges

There is a trade-off between 3D descriptors and speed, and it is the user's decision to include 3D descriptors in a study or not.

2.2.2 Quantitative structure–activity relationship

2.2.2.1 Concept

QSAR, as the name suggests, is the computational technique to establish the correlation between chemical structures and biological activity. In general, the QSAR technique is implemented in rational drug design; however, the technique is widely accepted to predict other physicochemical properties and therefore is also termed quantitative structure property relationship. QSAR is based on the hypothesis that similar structural compounds may possess similar biological activities (Miki, 2002). Chemical features of molecules, also known as molecular descriptors, are correlated with the observed activity by the mean of statistical analysis. Data type decides the statistical approach to implement the building of the model. For example, quantitative data are processed by regression-based methods, while graded response data are processed by means of classification-based methods. Thus built models need to be validated before being used further for drug designing.

Among regression-based methods, multiple linear regression (MLR) is the most commonly used method to build a regression-based QSAR model. MLR is a simple regression method that assumes a linear relation between multiple independent variables (descriptors) and a dependent variable (biological activity). MLR involves stepwise regression to find the best fit model and therefore it can be time consuming for a large number of descriptors. Principal component analysis (PCA) covers such a drawback and can reduce information from a large number of variables into a smaller subset of unique variables. The major drawback of the PCA method is the difficulty in extracting details of molecular descriptors that contribute to the biological activity (Wold et al., 1987). A solution to problems associated with MLR and PCA is partial least square (PLS) analysis. In PLS, the dependent variable, i.e., biological activity values, are also extracted into new variables to improve the correlations (Geladi and Kowalski, 1986). MLR, PCA, and PLS are three commonly used methods to build linear QSAR models. Nevertheless, biological systems often show a nonlinear regression relationship between molecular descriptors and biological activities. A neural network is the most widely used approach to deal with nonlinear regression.

An imperative aspect of QSAR is to validate the newly built model. In a QSAR study, the group of molecules used to build a model is known as the "training set," while the group of molecules used to predict the model is known as the "test set." There are two types of validation methods available to validate the QSAR model: (1) internal validation and (2) external validation. Leave-one-out is the most common type of internal validation in which one of the molecules is kept in the test set while rest of the molecules, i.e., the training set, are used to estimate the coefficients of different descriptors in a QSAR model. Next, the test set molecule is used to predict its activity using the model built on the training set. The process is repeated multiple times until all the molecules of the training set have served as the test set molecule. In contrast to internal validation, external validation involves

Table 2.4 Commonly used quantitative structure–activity relationship (QSAR) methods and their descriptions.

Method	Descriptor type	Description
HQSAR	2D	Hologram QSAR is the technique in which molecular substructures are represented as binary patterns and fingerprints that are combined to generate molecular holograms and are correlated with biological activities
CoMFA	3D	Comparative molecular field analysis relates steric and electrostatic properties of molecules to their biological activities
CoMSIA	3D	In addition to steric and electrostatic contribution, comparative molecular similarity indices include steric, H-bond donor/acceptor, and hydrophobic terms
COMBINE	3D	Comparative binding energy analysis estimates the binding affinities of ligands

prediction of a QSAR model using a test set that was never used to build the model (Gramatica, 2007). The most widely used QSAR methods are summarized in Table 2.4.

2.2.2.2 Workflow

- Retrieve a congeneric series of ligands that have been evaluated in similar biological assay and showed diverse variation in activity.
- Identify and determine the molecular descriptors associated with physiochemical properties of the molecules.
- Randomly divide the molecules into training set and test set.
- Use the training set to identify and calculate the correlation coefficient that can explain the relationship between the descriptor values and the biological activities.
- Evaluate the stability of the statistical equation using the test set molecules.
- Use the statistical model to predict the biological activity of new molecules.

2.2.2.3 Tools

Table 2.5 contains a list of the most commonly used QSAR software.

2.2.2.4 Applications

- *Combining 2D and 3D descriptors:* Kumar et al. (2011) used a combined holograph QSAR and comparative molecular similarity indices analysis to develop a robust QSAR model to predict topological features for protein kinase C βII inhibition.

Table 2.5 List of quantitative structure–activity relationship packages and their sources.

Package	Source
HQSAR, CoMFA, CoMSIA, Volsurf	Tripos, SYBYL
Molecular field-based 3D QSAR	PHASE, Schrodinger
HipHop, HypoGen	CATALYST, Serius
QSAR Toolbox	Oasis
TOPKAT	Accelrys
Derek	Lhasa

- *Integrating QSAR model with VS:* Sharma et al. (2016) applied atom-based 3D-QSAR modeling with ligand-based pharmacophore mapping to identify novel, selective phosphodiesterase 4B (PDE4B) inhibitors through VS and molecular docking. The authors confirmed the stability of new molecules though molecular dynamics (MD) simulations and evaluated their predictions through in vitro enzymatic assay.

2.2.2.5 Challenges

- *Replication of molecules in training and test sets:* It is a common mistake to have duplicated molecules in the training and test sets. Such duplications falsely improve the prediction power of QSAR models.
- *Experimental and descriptor-associated errors:* It is imperative to include the standard errors details associated with biological activities and molecular descriptors to minimize model-associated errors.
- *Poor transferability:* The developed QSAR model from a research group is rarely (efficiently) used by other research groups for predictive purposes.
- Simple linear regression or MLR methods are easy to calculate but are inefficient if the number of independent variables (descriptors) is comparable or higher than the number of total molecules.
- PLS can handle "*n*" number of independent variables but builds only linear relationships.

2.2.3 Ligand-based pharmacophore

2.2.3.1 Concept

The International Union of Pure and Applied Chemistry defines a pharmacophore as "The ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" (Wermuth et al., 1998). Ligand-based pharmacophore (LBP) is the pharmacophore-based method of choice in the absence of any structural information available for the therapeutic target. The aim

of LBP is to identify the largest 3D pattern of features, which is imperative for most of the input ligands to bind to the receptor. The task becomes more complicated as the number of input ligands and their flexibilities increases. Therefore conformational search is an important and costly step in LBP. Programs like RAPID, MPHIL, Phase, DISCO, HipHop, and HypoGen calculate all possible conformations of input molecules, while programs like GALAHAD, GASP screen ligand conformations with respect to pattern constraint. A critical step in LBP is to identify a "bioactive" conformation of an active molecule to align the rest of the molecules. In the absence of 3D bioactive conformation, databases are searched to find a conformation of a molecule that is similar to the input ligands, otherwise, the most active molecule is geometrically optimized and thus the obtained minimal energy conformation is considered as the bioactive conformation to align the rest of the molecules. It is imperative to validate the developed LBP before using it further, and similar to QSAR, a separate test set of molecules is created to validate the prediction power of the LBP.

2.2.3.2 Workflow

- Selection of active and inactive molecules in the training sets.
- Structural optimization of all molecules using a suitable forcefield.
- Superposition of all molecules to the bioactive conformation/or minimized conformation of most active molecules.
- Model validation.
- Database screening using developed LBP.

2.2.4 Tools

Table 2.6 lists the software used for pharmacophore modeling.

2.2.4.1 Applications

- *3D-LBP and VS:* Researchers have used a cocrystallized ligand as a bioactive molecule to develop an LBP model. Using validated LBP, the authors screened a large dataset to identify novel, selective, and submicromolar-range active inhibitors (Sharma et al., 2016; Al-Sha'er and Taha, 2010).

Table 2.6 List of software used for ligand-based pharmacophore modeling.

Name	Source
LigandScout	Inteligand
MOE	Chemical Computing Group
Phase	Schrodinger
Unity	Certara
Quasi	Denovopharma

2.2.5 Challenges

- *Binding characteristics:* In an ideal LBP study, the cocrystallized ligand is used as a bioactive conformation to align the rest of the molecules. LBP is based on the assumption that all molecules bind receptor at a single site and have similar binding characteristics. However, even similar molecules may bind in different ways.

2.3 Structure-based drug design

2.3.1 Homology modeling

2.3.1.1 Concept

A preliminary requirement for any SBDD technique is the 3D structure of the target. Various integrative structure biology techniques are available to determine the 3D structure of molecules: X-ray crystallography, nuclear magnetic resonance spectroscopy, or single-particle cryoelectron microscopy. However, the structure of a therapeutic protein is difficult to solve due to technical difficulties in expressing, purifying, or characterizing proteins. In the absence of any experimental structures, in silico methods are used to predict the 3D structure of a target: homology modeling, threading, or ab initio modeling. Among the three computational methods, homology modeling is the most reliable method to predict the 3D structure of a target because it uses the structural information of a similar protein with $>40\%$ identity (known as template structure). Homology modeling is based on the hypothesis that the two highly similar sequences have similar structures. Threading or fold recognition is the method-of-choice if the target sequences have same protein fold as that of known structures but there is no template ($>40\%$ sequence identity) structure available in the protein structure database. Ab initio is the method-of-choice to predict a protein structure when the target sequence lacks any similar known structure or similar fold in the structure database. Ab initio modeling considers the physicochemical properties of amino acids to predict the least energy and stable conformation, and is currently limited to small proteins (<120 amino acids).

2.3.1.2 Workflow

- *Template search:* Using a target sequence-of-interest, a protein structure database is screened using protein BLAST to identify template sequences that are highly similar to the target sequence.
- *Global sequence alignment:* The best aligned template sequence is chosen to align against the desired target sequence. The resulting global sequence alignment is evaluated and corrected to confirm the conservation of the functional domain.
- *Model building:* The first step is to generate the backbone of the target using structural information from the template structure. During backbone modeling, the target residues that correspond to inserts and gaps in the multiple sequence alignment are deleted. Such deleted residues are modeled in the next step, loop

modeling. If a similar loop region is available in the structure database (like Research Collaboratory for Structural Bioinformatics (RCSB), the loop is modeled using the knowledge-based method. Otherwise, the energy-based method is used to minimize the energy of the loop and to model missing residues. Next, the missing side chains are modeled by finding hit rotamers from the rotamer library. Finally, the modeled structure is optimized to remove any steric clashes or other structural issues. MD simulations or Monte Carlo (or a combination of both methods) are used to optimize the homology model and to predict the low-energy, native-like conformation.

- *Model validation:* Newly modeled structures may contain errors for the following reasons: (1) experimental errors in the template structure, and (2) percentage identity value of $<100\%$. Therefore it is imperative to evaluate the newly designed model for different errors before proceeding to any actual CADD. Ramachandran plot and favorable energies are evaluated to validate and optimize the model to generate the final, stable homology structure of the desired target sequence (Ramachandran and Sasisekharan, 1968). The model structure must have minimal residues (or no residue at all) in the outlier region of the Ramachandran plot and must have minimal energy among all possible conformations.

2.3.1.3 Tools

Table 2.7 enlists databases and web-servers to predict protein structures.

Table 2.7 List of databases and webservers to predict protein structures.

Software/webserver	URL
Homology modeling	
MODELLER	https://salilab.org/modeller/
SWISS-MODEL	https://swissmodel.expasy.org/
PRIMO	https://primo.rubi.ru.ac.za/
PyMod	http://schubert.bio.uniroma1.it/pymod/index.html
MaxMod	http://www.immt.res.in/maxmod/
Fold recognition	
GenTHREADER	http://bioinf.cs.ucl.ac.uk/psipred/
Phyre2	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
MUSTER	http://zhanglab.ccmb.med.umich.edu/MUSTER/
ORION	http://www.dsimb.inserm.fr/orion/
DN-Fold	http://iris.rnet.missouri.edu/dnfold/
Ab initio structure prediction	
Robetta	http://www.robetta.org/submit.jsp
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
QUARK	http://zhanglab.ccmb.med.umich.edu/QUARK/
BHAGEERATH	http://www.scfbio-iitd.res.in/bhageerath/index.jsp
EVfold	http://evfold.org/evfold-web/evfold.do
CABS-fold	http://biocomp.chem.uw.edu.pl/CABSfold/

2.3.2 Applications

- *Rational drug design:* The modeled structure can be used to perform CADD to identify potent small molecule inhibitors. For example, Liu et al. (2005) modeled flexible severe acute respiratory syndrome 3C-like proteinase A and used the homology model to screen 630,000 small molecules from different databases. From the study, the authors reported 40 bioactive inhibitors, including calmidazolium that inhibited the enzyme in micromolar concentration.
- *Chimeric models:* Multiple templates are used to build the structure of a protein containing disordered or flexible regions. Chimeric modeling is imperative when the missing region plays a crucial functional role in a protein. For example, B-cell lymphoma-2 (Bcl-2) protein is an antiapoptotic member of the Bcl-2 family. Bcl-2 lacks structural details of a functionally important flexible loop domain (FLD) region. Raghav et al. (2012a) built the high-quality structure of an FLD domain using loop modeling and further refined the structure using MD simulations.

2.3.3 Challenges

- *Resolution of template structure:* The quality of a model is directly related to the quality of the input template structure. Therefore a high-resolution structure should be used whenever possible.
- *Low sequence similarity:* The model is not reliable if the sequence identity between the target and template is $<30\%$.

2.3.4 Molecular docking

2.3.4.1 Concept

Molecular docking is a powerful tool to predict favorable, low-energy, binding modes of a ligand in the active site of a receptor. In rational drug design, the ligand corresponds to a small molecule and the receptor corresponds to a protein. The most suitable interactions in rational drug design are noncovalent interactions and include hydrogen bonds, van der Waals bonds, or any possible electrostatic attractions. The concept of molecular docking is based on the hypothesis that the enhanced affinity between protein−ligand interactions is directly correlated with inhibition of therapeutic enzymes. Molecular docking is the most suitable structure-based drug design technique when a high-resolution structure of a receptor is available. Molecular docking is broadly divided into three categories depending on the flexibility of the components: (1) rigid docking, in which both the ligand and the receptor remain fixed; (2) flexible docking, in which the ligand is flexible but the receptor is fixed; and (3) full flexible docking, in which both the receptor and the ligand are flexible. Among the three docking types, flexible docking is the most widely used docking technique in academia and industry. Docking algorithms are search parameters that are implemented to predict binding modes of ligands in receptors. A combination of one or more algorithms is often used to accurately predict the binding mode of a ligand.

The complexity and effectiveness of a docking algorithm depend on the degrees of freedom that it can handle. The translation and rotation motion of a molecule results in six degrees of freedom. An additional degree of freedom is the rotational motion of the ligand. The simplest docking algorithm consists of only rotational and translational degrees of freedom, and was the basis of DOCK software. Other software can handle rotational and translational degrees of freedom in parallel and are briefly divided into three groups: Monte Carlo, genetic algorithm, and incremental construction (Leach and Valerie, 2007). Scoring functions complement docking algorithms to predict ligand–receptor complexes. Scoring functions are mathematical functions that are used to predict the binding affinity between the ligand and the receptor in a post-docking conformation (Huang et al., 2010). Scoring functions are of four types: (1) forcefield, (2) empirical, (3) knowledge based, and (4) machine learning (Leach and Valerie, 2007). Forcefield-based scoring functions calculate binding affinities on the basis of the strength of intermolecular noncovalent interactions such as van der Waals and electrostatic functions. Empirical scoring functions are regression-based scoring functions, which are based on the correlation of nonrelated variables and are used to mimic experimental binding affinities. Knowledge-based scoring functions are derived from the statistical analysis of experimental 3D structures of biomolecules. Lastly, machine learning scoring functions are built by training on the dataset and thus, unlike the rest of the three scoring functions, can predict the binding interactions implicitly.

2.3.4.2 Workflow

- *Ligand preparation:* The ligand molecule/s are sketched or retrieved from the database and converted into 3D structures. The structure is processed to generate different tautomers and stereoisomers (if any). A suitable charge type is added to the ligand molecule. Commonly used changes for small molecules include Gasteiger, AM1-BCC, Mulliken, GAFF, OPLS, etc.
- *Receptor preparation:* The 3D structure of the receptor is processed to add missing hydrogens, add missing side chains, correct tautomeric states of ionizable residues, and briefly minimize to remove any steric clashes. A suitable charge type is added to protein residues. Commonly used charges for receptor include Amber, CHARMM, or OPLS/AA.
- *Grid generation:* The probable ligand-binding site is defined either by selecting cocrystallized ligand or by active site residues or a combination of both. In a blind docking, the entire receptor is kept under the grid such that the program can perform a time-consuming but detailed search of a ligand in the receptor.
- *Docking:* Next, the actual docking begins. During docking, the program tries to find an optimal binding pose of a ligand in the receptor on the basis of docking using algorithms and scoring functions. First and foremost, it is imperative to optimize docking parameters by redocking the extracted cocrystallized ligand into the protein structure. The minimal root mean square deviation (RMSD) value between the docked and cocrystallized pose suggests proper docking

parameters. Next, actual docking is performed using this established list of docking parameters. In the absence of any cocrystallized ligand, literature-based information is used to consider the favorable binding interactions with the active site residues of interest.

- *Posttrajectory analysis*: Docked ligands and the grid containing receptors are visualized in graphical user display software to study interactions and binding poses of the docked ligands in the active site of the receptor. In particular, minimal binding energy, molecular mechanics/Poisson–Boltzmann (generalized born) surface area (MM/PB(GB)SA) energies, and interactions in the desired cavity regions are considered to shortlist desired poses of molecules. Virtual ligands having a similar scaffold to that of a previously cocrystallized ligand may bind in a similar manner.

2.3.4.3 Tools and software

Table 2.8 shows is a list of widely used docking software for protein–ligand docking.

In addition to predicting the binding characteristics between protein and ligands, molecular docking can also be performed between two macromolecules. The most widely used protein–protein docking tools include HDOCK, ZDOCK, CluPro, PatchDock, FireDock, InterEvDock2, SOAP PP, and FRODOCK2. However, these tools are beyond the scope of this chapter and are described elsewhere (Raghav et al., 2019).

2.3.4.4 Applications

- *Integration with other LBDD or SBDD techniques:* Sharma and Wakode (2017) designed a QSAR-based pharmacophore to screen small molecule databases against PDE4B, a therapeutic target in inflammatory diseases. Thus screened molecules were docked in a PDE4B crystal structure to identify novel, potent PDE4B inhibitors. The authors confirmed their predictions by in vitro enzymatic assays. Kumar et al. (2012) combined molecular docking with VS to identify novel, selective, and specific aldose reductase inhibitors.

Table 2.8 Widely used docking software.

Software	Algorithm	Scoring function	Webpage
AutoDock	GA	Forcefield + empirical	http://autodock.scripps.edu
DOCK	IC	Forcefield	http://dock.compbio.ucsf.edu
FlexX	IC	Empirical	https://www.biosolveit.de/FlexX/
GOLD	GA	Empirical + knowledge based	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/
Glide	SA/IC	Empirical + knowledge based	https://www.schrodinger.com/glide

GA, *Genetic algorithm;* IC, *incremental construction;* SA, *simulated annealing.*

- *Target fishing:* In an approach called "reverse docking (RD)," a biological target is predicted using a ligand of interest. Park and Cho (2017) screened 26 ginsenosides against 1078 targets and identified hit targets on the basis of docking score.
- *Polypharmacology:* A high rate of failure in clinical trials has moved modern drug discovery to design and develop small molecule inhibitors that can bind to multiple targets. For example, Anighoro et al. (2017) designed first-in-class dual inhibitors that can inhibit heat shock protein 90 and serine/threonine kinase B-Raf.
- *Improving PK/PD properties of therapeutic biomolecules:* Bcl-2 is therapeutic target in several types of cancer. Raghav et al. (2012b) modified immunoglobulin D (IGD), a poorly binding peptide to Bcl-2, using deprotonation, amidation, acetylation, benzoylation, benzylation, and addition of phenyl, deoxyglucose, and glucose fragments. Such modification not only improved the binding affinity with the Bcl-2 but also improved the PK/PD profile of IGD peptide.
- *Adverse drug reactions:* An early detection of possible toxicity/side effects associated with a small molecule can save time and money in the drug discovery process. For example, Ji et al. (2006) performed reverse docking of 11 marketed antihuman immunodeficiency virus (HIV) drugs and identified drug adverse effects that were previously reported in the literature.

2.3.4.5 Challenges

- *Target structure:* Molecular docking must have a 3D structure of target. Molecular docking cannot be applied if any experimental or modeled structure of the therapeutic target is not known.
- *Target flexibility:* Biomolecules are always in dynamic motion inside the cell and may occupy multiple conformations that are otherwise difficult to detect using experimental methods. Most of the docking programs ignore target flexibility.
- *Solvent molecules:* Solvent may stabilize ligand−receptor interactions, which are often ignored while preparing the receptor for docking.

2.3.5 Virtual screening

2.3.5.1 Concept

VS is a robust technique to identify lead molecules. As the name suggests, the approach involves *virtual* selection of designed small molecules from large databases through computational tools without physically testing them in the laboratory.

2.3.5.2 General workflow

Structure-based VS encompasses the preparation of target and small molecule database, docking, and postdocking analysis. The steps are:

- *Target preparation:* The foremost step is to retrieve a pdb file from the structure database (such as RCSB) or to model the structure of the target protein. The retrieved protein structure is protonated and relaxed to avoid any steric clashes.

The correct ionization states are determined for tautomeric residues such as His or Asp. The user decides to keep or delete the cocrystallized water molecules depending on the role of water molecules in target inhibition.

- *Database preparation:* The initial dataset is converted to standard file format and reduced in size by applying several molecular filters to preselect the drug-like molecules. A special care is given to decide tautomeric state or a stereoisomer during conversion of 2D representations (such as SMILE) to 3D structures. Enantiomers are generated and stored as separate molecules. It is imperative to assign partial atomic charge before proceeding to molecular docking. The common charges include Gasteiger and Marsili (1980) or MMFF94 (Halgren, 1996).
- *Molecular docking:* Next, the prepared small molecule databases are docked in the prepared target. Details of molecular docking are described in the previous section.
- *Postdocking analysis:* It is important that the user understands the difference between docking and scoring. While docking predicts the binding pose of a small molecule in the receptor, the scoring function is related to the free energy of association between the ligand and the receptor. Ideally, a good correlation is desired between the docking and the scoring function to predict the best binding pose with the best score. But the correlation is not always straightforward. Therefore it is recommended that the user couple the scoring results with the 3D interactive visualization to finalize the list of active molecules.

2.3.5.3 Tools

Table 2.9 is a list of commonly used small molecule databases that are widely used for VS.

Table 2.10 is a list of software that is widely used for VS.

Table 2.9 A list of small molecule databases.

Database	Molecules (#)	Website
ZINC	230 million	https://zinc.docking.org
eMolecules	7.3 million	https://www.emolecules.com
Enamine	2.7 million	https://enamine.net
ChEMBL	1.9 million	https://www.ebi.ac.uk/chembl/
ChemBridge	1.3 million	https://www.chembridge.com/screening_libraries/
SPECS	350,000+	https://www.specs.net
NCI	250,000+	https://cactus.nci.nih.gov/index.html
Maybridge	500,000	https://www.fishersci.com/us/en/brands/I9C8LZ4U/maybridge.html
DrugBank	13,500	https://www.drugbank.ca

Table 2.10 Common docking software used for structure-based virtual screening.

Software	Ligand sampling	References
Dock	Incremental build	Ewing et al. (2001)
FlexX	Incremental build	Rarey et al. (1996)
Gold	Genetic algorithm	Jones et al. (1995)
Glide	Exhaustive search	Alogheli et al. (2017)
AutoDock	Genetic algorithm	Morris (2020)

2.3.5.4 Applications

- VS is significantly faster and requires fewer resources than traditional high-throughput assays.
- A user can prioritize and rationalize molecules that he/she wants to buy by evaluating millions of molecules through VS.
- The accuracy of structure-based VS can be further enhanced by complementing with ligand-based activity data.

2.3.5.5 Challenges

- In a VS workflow, the desired molecules are rapidly selected due to VS-associated coarse filters. Such coarse filters are less reliable to predict accurate binding energies and thus result in rejection of many good molecules.
- Receptor flexibility is poorly handled in VS.
- The success of VS depends on the accuracy of the input structural model. Therefore a poorly predicted homology model structure may result in false positive/negative hits.
- There is a trade-off between accuracy and speed in VS. An algorithm that is meant to rapidly screen millions/billions of small molecules may not successfully handle bound metals or crystallized water molecules.

2.3.6 Receptor-based pharmacophore modeling

2.3.6.1 Concept

As the name suggests, receptor-based pharmacophore (RBP) modeling utilizes the 3D structure of the receptor and depending on the availability of ligand structure, RBP is divided into two categories: receptor—ligand complex-based pharmacophore, and RBP. In both cases, complementary chemical features of the active site and their spatial arrangements are defined. Pharmacophore features are defined by either of three methods: (1) a molecular probe-based characterization of potential interaction energy (*molecular field*); or a search for (2) *substructure patten* or (3) chemical *features* that can fulfill the interaction requirements. Receptor—ligand-based pharmacophore modeling has an additional advantage of exclusion-volume constraint.

Exclusion-volume is a set of spheres that represents receptor residues and thus imposes a constraint for ligand binding. The constraint is helpful to filter false positive ligand hits that may otherwise pass through a ligand-only pharmacophore model. To summarize, a receptor (or receptor−ligand)-based pharmacophore creates a virtual 3D mold on the basis of spatial arrangement of nonbonded interactions in the active site. This virtual 3D mold is computationally less complex compared to the 3D all-atom model of the receptor−ligand complex and thus is efficient in performing a rapid search against a large dataset of small molecules.

2.3.6.2 Workflow

- *Prepare the input files:* Retrieve the 3D structure of the receptor (*preferably*) complexed with the ligand and correct the structure for bond orders, H-bonds, steric clashes, ionization states, etc.
- *Define the binding site:* Utilize the bound ligand (if available) to define the active site region. In case of apo-structure, define the active site manually.
- *Generate the pharmacophore map:* Pharmacophore features are derived from the bound ligand or from the active site cavity of the apo-enzyme.
- *Select the optimal pharmacophore features:* Among all possible features, the most optimal features are shortlisted on the basis of interaction energies or using receptor−ligand interaction information, or by training the pharmacophore with the active/inactive ligands. It is a bonus to include the volume restraint to define a receptor-based pharmacophore model.
- *Validation:* Before the pharmacophore is used for drug designing, it must be validated by screening and scoring the active/inactive ligands.

2.3.6.3 Tools

Table 2.11.enlists different software that are commonly used to build 3D pharmacophore models.

2.3.6.4 Applications

- *Pharmacophore-based VS:* Pirard et al. generated a homology model of voltage-dependent potassium channel Kv1.5 and utilized this model to develop a

Table 2.11 3D pharmacophore model software and their details.

Software	Input	Method of identification
FLAP	Ligand, complex, apo	Molecular field
Pharmer	Ligand, complex	Substructure pattern, feature
LigandScout	Ligand, complex, apo	Substructure pattern, feature, molecular field
Catalyst	Ligand, complex, apo	Substructure pattern, feature, molecular field
MOE	Ligand, complex, apo	Substructure pattern, feature, molecular field
PHASE	Ligand, complex, apo	Substructure pattern, feature, molecular field
UNITY	Ligand, complex	Substructure pattern, feature

receptor-based pharmacophore model. Next, the authors used the pharmacophore model to perform VS against their in-house small molecule database and identified 19 inhibitors. Among these 19 inhibitors, five molecules had $IC_{50} < 10$ μM (Pirard et al., 2005).

- *De novo drug design:* Ajay et al. used the apo-structure of a protein-tyrosine phosphatase leukocyte antigen-related receptor and designed novel inhibitors using LUDI (Ajay and Sobhia, 2011).
- *Lead optimization:* Boehm et al. utilized the available 3D structure of DNA gyrase and synthetic-aperture radar data to optimize hit molecules and obtained a 3,4-disubstituted indazole molecule that was 10 times more potent than novobiocin, the standard DNA gyrase inhibitor (Boehm et al., 2000).
- *Polypharmacology:* Wei et al. combined an RBP model with molecular docking to identify dual target inhibitors against human leukotriene A4 hydrolase and the human nonpancreatic secretory phospholipase A2, two therapeutic enzymes of the arachidonic acid metabolism pathway. In brief, the authors screened small molecules against a common pharmacophore that represented both the therapeutic targets and thus identified dual-target inhibitors (Wei et al., 2008).

2.3.6.5 Challenges

- *Minimal pharmacophore features:* Reduction of large numbers of features to identify the least number of pharmacophoric features is a challenge. It may cause rejection of true positive features.
- *Receptor flexibility:* An active site of a receptor may bind differently to a diverse set of ligands and therefore a single pharmacophore is not enough to perform drug design against flexible targets.

2.3.7 Molecular dynamics simulations

2.3.7.1 Concept

Experimentally solved 3D structures are just snapshots of highly mobile biomolecules. MD simulation calculates the time-dependent motion of a biomolecule and thus provides detailed insights into the flexibility or conformational rearrangements of the system. MD simulation is based on Newton's second law of motion, $F=ma$. To begin with, the force on the starting static structure is calculated using the coordinate and potential energies at time t_0. The equation of motion is deterministic, i.e., the user can determine the velocities and coordinates at a time t_1, considering that the values are known at time t_0. Numerous numerical algorithms are known that can be used to integrate the equations of motions, such as Verlet (1967), Leap-frog, Velocity, and Beeman's algorithms (Mcquarrie, 1976). Based on the calculated force at time t_0, the coordinates of all atoms of a molecule are calculated at time t_1. The process is repeated millions of times to generate the trajectory from the MD simulation of a biomolecule. The usual time step in classical MD simulation is 1−2 fs and is enough to measure the dynamics at the atomic level.

2.3.7.2 Workflow

- *System preparation:* A fundamental prerequisite for an MD simulation is the 3D structure of a biomolecule. The 3D structure is prepared as previously described in the section on molecular docking. In addition, any further changes are incorporated as per the experiment such as point mutations, truncations, etc. Next, the software compatible forcefields are applied for the biomolecule and other nonstandard molecules. Nonstandard molecules include bound inhibitor, cofactor, glycosylation, nucleic acids, other posttranslation modifications, etc. Next, the unit cell is solvated in the periodic boundary conditions, and salt atoms are added to the desired concentration.
- *Minimization:* The prepared system may contain steric clashes or unusual geometry that may artificially raise the energy of the system. Therefore the system is first relaxed by performing a small minimization. A harmonic restraint is applied for nonwater atoms, which is gradually removed during the next steps of heating and equilibration.
- *Heating and equilibration:* The minimized system gradually heads from 0 K to the desired temperature followed by equilibration. Canonical ensemble, NVT (N: number of atoms in the system, V: system volume, T: absolute temperature), is usually applied during heating and initial equilibration. Harmonic restraint is gradually removed from the system before the isothermal, isobaric ensemble, NPT (N: number of atoms in the system, V: constant pressure, T: constant temperature), is applied. An early NVT ensemble is required during the initial small duration heating process because the calculation of pressure is inaccurate at low temperatures. However, as soon as the system is heated, the ensemble is shifted to NPT to correct the density. The rest of the simulation is generally continued at the NPT ensemble.
- *MD simulation:* Next, actual MD simulation is started at the NPT ensemble. The time scale varies between nanoseconds and milliseconds depending on the physiological property to be studied.
- *Posttrajectory analysis:* MD simulation trajectories are first checked for the stabilities in terms of RMSD, potential energy, etc. Stable trajectories are next analyzed as per the demand of the experiment. For example, Sharma et al. (Sharma and Wakode, 2020) confirmed the stability of their simulations using RMSD and potential energies and then performed root mean square fluctuation, dynamical cross-correlation matrix, PCA, and molecular mechanics Poisson–Boltzmann surface area analysis to compare the dynamics of a larger and smaller version of a therapeutic protein, PDE4B, both complexed with a small molecule inhibitor, NPV.

2.3.7.3 Tools

Table 2.12

Table 2.12 List of software widely used for MD simulations.

Software	Webpage
AMBER	https://ambermd.org
NAMD	https://www.ks.uiuc.edu/Research/namd/
Gromacs	http://www.gromacs.org
Desmond	https://www.schrodinger.com/desmond
LAMMPS	https://lammps.sandia.gov

It is imperative to mention that each package is capable of performing different types of simulations such as sampling configuration (replica exchange, accelerated molecular dynamics, steered molecular dynamics, nudged elastic band), free energy calculations (molecular mechanics/Poisson–Boltzmann (generalized born) surface area, nonequilibrium free energy, binding enthalpy measurements), etc.

2.3.7.4 Applications

- *Structure refinement:* MD simulations can be used to refine the homology model structure of a biomolecule. In addition, the algorithm can be combined with other software such as Phenix to refine X-ray crystal structures. Raval et al. (2012) performed >100 μs long MD simulation of 24 proteins and showed that the long simulations can indeed achieve the native-like conformations of the modeled proteins.
- *Ensemble docking:* Low-energy conformations from an MD simulation can be used as the ensemble of structures to dock small molecules. For example, Osguthorpe et al. (2012) performed replica-exchange MD to sample protein conformations and used the representative conformations to dock small molecules.
- *Identification of allosteric site:* It is likely for a dynamic biomolecule to undergo conformational rearrangements and such allosterisms are difficult to capture in experimentally solved static 3D structures. By combining flexible ligand docking and MD simulation, Schames et al. (2004) showed the existence of a new binding site in HIV integrase. The new site was previously unidentified in solved HIV integrase X-ray crystal structures. This new site was abundant in several frames of MD trajectories suggesting that the site was indeed energetically favored. A later study confirmed with the solved X-ray crystal structure that a cryptic trench site indeed existed in HIV integrase.
- *Induced-fit phenomenon:* Similar to allosterism, an induced-fit phenomenon is difficult to capture in static experimental 3D structures. Zhao et al. (2012) performed constrained MD simulation of an EphA3 structure in explicit solvent to create an induced-fit cavity. Using this induced-fit cavity, the authors carried out pharmacophore filtering and high-throughput docking, and identified 10 classes of novel molecules that could not be discovered using a primary X-ray crystal structure.
- *Advanced free energy calculations:* MD simulation has been advanced to calculate the binding free energies between two molecules using thermodynamic

integration, single-step perturbation, and free energy perturbation. Researchers have used different approaches to calculate binding free energies between molecules (Sharma et al., 2016; Moreau et al., 2017, 2020; Douglas et al., 2018; Sharma and Wakode, 2017; Sharma and Wakode, 2020).

- *Study of bond breakage/formation or transition-metal complex:* Although classical molecular mechanics-based MD simulation can predict several biological phenomena accurately, such simulations are not suitable to predict situations where the quantum effect is mandatory. Examples of such situations are (1) interaction with transition metals, or (2) breakage or formation of covalent bonds. To solve this issue, researchers have integrated classical MD simulations with quantum mechanical calculations. For example, Hong et al. (2011) complemented classical molecular mechanics-based MD simulation with quantum mechanics and studied the proton transfer mechanism of [Fe−Fe]H2ases.
- *Point mutations:* Padhi et al. (2012) utilized an all-atom MD simulation approach to understand the role of different point mutations on the function of human angiogenin. Using all-atom MD simulation, the authors studied the conformational switching of catalytic residue His114 and correlated with the mechanism causing loss of ribonucleolytic activity.

2.3.7.5 Challenges

- Large MD simulations of 1 μs or longer sometimes fail to predict the conformational rearrangements demanding implementation of sophisticated conformational sampling techniques.

References

Ajay, D., Sobhia, M.E., 2011. Simplified receptor based pharmacophore approach to retrieve potent PTP-LAR inhibitors using apoenzyme. Curr. Comput. Aided Drug Des. 7, 159−172.

Alogheli, H., Olanders, G., Schaal, W., Brandt, P., Karlén, A., 2017. Docking of macrocycles: comparing rigid and flexible docking in glide. J. Chem. Inf. Model. 57, 190−202.

Al-sha'er, M.A., Taha, M.O., 2010. Elaborate ligand-based modeling reveals new nanomolar heat shock protein 90α inhibitors. J. Chem. Inf. Model. 50, 1706−1723.

Anighoro, A., Pinzi, L., Marverti, G., Bajorath, J., Rastelli, G., 2017. Heat shock protein 90 and serine/threonine kinase B-Raf inhibitors have overlapping chemical space. RSC Adv. 7, 31069−31074.

Baldwin, J.J., Ponticello, G.S., Anderson, P.S., Christy, M.E., Murcko, M.A., Randall, W.C., Schwam, H., Sugrue, M.F., Springer, J.P., Gautheron, P., 1989. Thienothiopyran-2-sulfonamides: novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma. J. Med. Chem. 32.

Bernard, M., 2009. Lessons from 60 years of pharmaceutical innovation. Nat. Rev. Drug Discov. 8.

Boehm, H.-J., Boehringer, M., Bur, D., Gmuender, H., Huber, W., Klaus, W., Kostrewa, D., Kuehne, H., Luebbers, T., Meunier-keller, N., 2000. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. J. Med. Chem. 43, 2664−2674.

Daniel, F.V., Stephen, R.J., Hung-Yuan, C., Brain, R.S., Keith, W.W., Kenneth, D.K., 2002. Molecular properties that influence the oral bioavailability of drug candidates. J. Med. Chem. 45.

Douglas, R.G., Nandekar, P., Aktories, J.-E., Kumar, H., Weber, R., Sattler, J.M., Singer, M., Lepper, S., Sadiq, S.K., Wade, R.C., 2018. Inter-subunit interactions drive divergent dynamics in mammalian and Plasmodium actin filaments. PLoS Biol. 16, e2005345.

Egan, W.J., Merz, K.M., Baldwin, J.J., 2000. Prediction of drug absorption using multivariate statistics. J. Med. Chem. 43.

Ewing, T.J., Makino, S., Skillman, A.G., Kuntz, I.D., 2001. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput. Aid. Mol. Des. 15, 411−428.

Gasteiger, J., Marsili, M., 1980. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron 36, 3219−3228.

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1−17.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694−701.

Halgren, T.A., 1996. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J. Comput. Chem. 17, 520−552.

Hedia, M., Gilbert, L., Annai, K.Q., 2000. InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II. Microbiology 146 (Pt 2).

Hendrickson, J.B., 1991. Concepts and Applications of Molecular Similarity. Willey, New York.

Hong, G., Cornish, A., Hegg, E., Pachter, R., 2011. On understanding proton transfer to the biocatalytic [Fe–Fe] H sub-cluster in [Fe–Fe] H2ases: QM/MM MD simulations. Biochim. Biophys. Acta Bioenerg. 1807, 510−517.

Huang, S.-Y., Grinter, S.Z., Zou, X., 2010. Scoring functions and their evaluation methods for protein−ligand docking: recent advances and future directions. Phys. Chem. Chem. Phys. 12, 12899−12908.

Ji, Z.L., Wang, Y., Yu, L., Han, L.Y., Zheng, C.J., Chen, Y.Z., 2006. In silico search of putative adverse drug reaction related proteins as a potential tool for facilitating drug adverse effect prediction. Toxicol. Lett. 164, 104−112.

Ji-Xia, R., Lin-Li, L., Ren-Lin, Z., Huan-Zhang, X., Zhi-Xing, C., Shan, F., You-Li, P., Xin, C., Yu-Quan, W., Sheng-Yong, Y., 2011. Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on SVM model, pharmacophore, and molecular docking. J. Chem. Inf. Model. 51.

Jones, G., Willett, P., Glen, R.C., 1995. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. 245, 43−53.

Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., Shoichet, B.K., 2007. Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. 25, 197−206.

Kenji, M., Yoshiaki, M., Yutaka, U., Hiroshi, S., Ayumu, M., Osamu, T., Takane, Y., Toshio, F., Tadashi, O., Masami, O., Naohisa, O., Tadasi, A., Chie, O., Sachiko, T., Hidee, I., Yasuto, A., Akira, A., 2010. Identification of a new series of STAT3 inhibitors by virtual screening. ACS Med. Chem. Lett. 1.

Kumar, H., Kumar, R., Grewal, B.K., Sobhia, M.E., 2011. Insights into the structural requirements of PKCβII inhibitors based on HQSAR and CoMSIA analyses. Chem. Biol. Drug Des. 78, 283−288.

Kumar, H., Shah, A., Sobhia, M.E., 2012. Novel insights into the structural requirements for the design of selective and specific aldose reductase inhibitors. J. Mol. Model. 18, 1791−1799.

Kumar, H., Frischknecht, F., Mair, G.R., Gomes, J., 2015. In silico identification of genetically attenuated vaccine candidate genes for Plasmodium liver stage. Infect. Genet. Evol. 36, 72−81.

Kumar, H., Kehrer, J., Singer, M., Reinig, M., Santos, J.M., Mair, G.R., Frischknecht, F., 2019. Functional genetic evaluation of DNA house-cleaning enzymes in the malaria parasite: dUTPase and Ap4AH are essential in Plasmodium berghei but ITPase and NDH are dispensable. Expert Opin. Ther. Target. 23, 251−261.

Leach, A.R.G., Valerie, J., 2007. An Introduction to Chemoinformatics | SpringerLink, UK. Springer, Dordrecht.

Ling, W., Qiong, G., Xuehua, Z., Jiming, Y., Zhihong, L., Jiabo, L., Xiaopeng, H., Arnold, H., Jun, X., 2013. Discovery of new selective human aldose reductase inhibitors through virtual screening multiple binding pocket conformations. J. Chem. Inf. Model. 53.

Lipnski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. 46.

Liu, Z., Huang, C., Fan, K., Wei, P., Chen, H., Liu, S., Pei, J., Shi, L., Li, B., Yang, K., 2005. Virtual screening of novel noncovalent inhibitors for SARS-CoV 3C-like proteinase. J. Chem. Inf. Model. 45, 10−17.

Mcquarrie, D., 1976. Statistical Mechanics. Harper & Row, New York.

Miki, A., 2002. Current state and perspectives of 3D-QSAR. Curr. Top. Med. Chem. 2.

Moreau, C.A., Bhargav, S.P., Kumar, H., Quadt, K.A., Piirainen, H., Strauss, L., Kehrer, J., Streichfuss, M., Spatz, J.P., Wade, R.C., 2017. A unique profilin-actin interface is important for malaria parasite motility. PLoS Pathog. 13, e1006412.

Moreau, C.A., Quadt, K.A., Piirainen, H., Kumar, H., Bhargav, S.P., Strauss, L., Tolia, N.H., Wade, R.C., Spatz, J.P., Kursula, I., 2020. A function of profilin in force generation during malaria parasite motility that is independent of actin binding. J. Cell Sci. 134.

Morris, G.M., 2020. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function - Morris - 1998 - Journal of Computational Chemistry. Wiley Online Library.

Osguthorpe, D.J., Sherman, W., Hagler, A.T., 2012. Exploring protein flexibility: incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. J. Phys. Chem. B 116, 6952−6959.

Padhi, A.K., Kumar, H., Vasaikar, S.V., Jayaram, B., Gomes, J., 2012. Mechanisms of loss of functions of human angiogenin variants implicated in amyotrophic lateral sclerosis. PloS One 7, e32479.

Park, K., Cho, A.E., 2017. Using reverse docking to identify potential targets for ginsenosides. J. Ginseng Res. 41, 534−539.

Pirard, B., Brendel, J., Peukert, S., 2005. The discovery of Kv1. 5 blockers as a case study for the application of virtual screening approaches. J. Chem. Inf. Model. 45, 477−485.

Raghav, P.K., Verma, Y.K., Gangenahalli, G.U., 2012. Molecular dynamics simulations of the Bcl-2 protein to predict the structure of its unordered flexible loop domain. J. Mol. Model. 18, 1885−1906.

Raghav, P.K., Verma, Y.K., Gangenahalli, G.U., 2012. Peptide screening to knockdown Bcl-2's anti-apoptotic activity: implications in cancer treatment. Int. J. Biol. Macromol. 50, 796–814.

Raghav, P.K., Kumar, R., Kumar, V., Raghava, G.P., 2019. Docking-based approach for identification of mutations that disrupt binding between Bcl-2 and Bax proteins: inducing apoptosis in cancer cells. Mole. Gene. Genom. Med. 7, e910.

Ramachandran, G.T., Sasisekharan, V., 1968. Conformation of polypeptides and proteins. Adv. Protein Chem. Elsevier 23, 283–437. https://doi.org/10.1016/S0065-3233(08)60402-7.

Rarey, M., Kramer, B., Lengauer, T., Klebe, G., 1996. A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. 261, 470–489.

Raval, A., Piana, S., Eastwood, M.P., Dror, R.O., Shaw, D.E., 2012. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. Protein.: Struct., Funct., Bioinform. 80, 2071–2079.

Redshaw, S., 1993. Angiotensin-converting enzyme (ACE) inhibitors and the design of captopril. In: Ganellin cr, R.S. (Ed.), Medicinal Chemistry. The Role of Organic Chemistry in Drug Research 2ed. Academic Press, London: London.

Schames, J.R., Henchman, R.H., Siegel, J.S., Sotriffer, C.A., Ni, H., Mccammon, J.A., 2004. Discovery of a novel binding trench in HIV integrase. J. Med. Chem. 47, 1879–1881.

Sharma, V., Wakode, S., 2017. Structural insight into selective phosphodiesterase 4B inhibitors: pharmacophore-based virtual screening, docking, and molecular dynamics simulations. J. Biomol. Struct. Dyn. 35, 1339–1349.

Sharma, V., Kumar, H., Wakode, S., 2016. Pharmacophore generation and atom based 3D-QSAR of quinoline derivatives as selective phosphodiesterase 4B inhibitors. RSC Adv. 6, 75805–75819.

Sharma, V., Wakode, S., 2020. Investigating the role of N-terminal domain in phosphodiesterase 4B-inhibition by molecular dynamics simulation. J. Biomol. Struct. Dyn. 1–9. https://doi.org/10.1080/07391102.2020.1780154.32552529.

Simon, T., A.M., D., Paul David, L., Tudor, O., 1999. The design of leadlike combinatorial libraries. Angew. Chem. 38.

Steven, M.P., Daniel, S.M., Christopher, T.D., Charles, C.P., Bernard, H.M., Stacy, R.L., Aaron, L.S., 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat. Rev. Drug Discov. 9.

Thompson, N.D., Susan, L.M., Bryan, J.W., Thomas, P.K., Ravi, K., William, C.S., Daniel, T.C., Brain, K.S., 2002. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J. Med. Chem. 45.

Verlet, L., 1967. Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. Phys. Rev. 159, 98.

Wei, D., Jiang, X., Zhou, L., Chen, J., Chen, Z., He, C., Yang, K., Liu, Y., Pei, J., Lai, L., 2008. Discovery of multitarget inhibitors by combining molecular docking with common pharmacophore matching. J. Med. Chem. 51, 7882–7888.

Wermuth, C.-G., Ganellin, C., Lindberg, P., Mitscher, L., 1998. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Chimie Pure et Appliquee 70, 1129–1143.

Wlodawer, A., Vondrasek, J., 1998. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annu. Rev. Biophys. Biomol. Struct. 27.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2, 37−52.

Zachary, M., Keun-Sik, K., Do-Min, L., Vinod, K., Si Eun, B., Kwang, H.L., Yan-Yan, Z., Lin, A., Kimberly, C., Na-Ra, L., Shou, Z., Qingquan, Z., Yujin, J., Hyun-Young, J., Chang-Guo, Z., Wooin, L., Dong-Eun, K., Kyung Bo, K., 2015. Proteasome inhibitors with pyrazole scaffolds from structure-based virtual screening. J. Med. Chem. 58.

Zhao, H., Huang, D., Caflisch, A., 2012. Discovery of tyrosine kinase inhibitors by docking into an inactive kinase conformation generated by molecular dynamics. ChemMedChem 7, 1983−1990.

CHAPTER

Advances in structure-based drug design

3

Divya Jhinjharia[1], **Aman Chandra Kaushik**[2], **Shakti Sahi**[1]

[1]*School of Biotechnology, Gautam Buddha University, Greater Noida, India;* [2]*Wuxi School of Medicine, Jiangnan University, Wuxi, Jiangsu, China*

3.1 Introduction

Computer-aided drug design (CADD) includes highly effective techniques crucial in the discovery and development of a drug. The phases involved in drug discovery and development include lead identification and validation, target identification and validation, preclinical studies, and clinical trials. It takes approximately 13−15 years or more for a drug to come to market going through various stages and involves a cost of approximately 2.6 billion dollars for a single drug candidate to reach the market. Despite the considerable cost and time involved in the process, about 90% of drug candidates do not enter the Food and Drug Administration (FDA)-regulated clinical trials. They fail at various stages of the drug discovery and development process. About 75% of the cost involved is spent on stages before clinical trials. The escalating cost and extended time involved have led to the development of methods and strategies to reduce the time frame and cost involved in discovering and developing a drug candidate. CADD plays a pivotal role in screening out at early stages the molecules that are likely to fail in later stages as potential drug candidates. Through various in silico studies, a large number of ligands having the potential to lead are screened. Only the ligands showing promising results can then be experimentally tested, thereby reducing the cost and time. Most of the pharmaceutical and biotechnology industries are now using CADD as a vital part of the drug discovery and development pipeline. The CADD field started more than 40 years ago. CADD techniques and tools can, in no small way, give an accurate prediction of binding affinity, the effectiveness of ligands, probable side effects, mode of action, and pharmacokinetic profiling of candidate molecules. These results further help in designing and developing therapeutics having better efficacy and potency with minimal side effects. Many studies show the role of CADD in new drug development (Karthick et al., 2016; Clark et al., 2016; Chao et al., 2007; Tran et al., 2015).

The techniques involved in CADD are broadly classified into (1) structure-based drug design (SBDD) and (2) ligand-based or analog-based drug design. The basic concept behind SBDD is that the structures of both target and ligands are known. The 3D structure of the target molecule is obtained through experimental techniques like X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00009-9

In the absence of experimental structure, the 3D structure of the target is predicted using computational techniques like homology modeling, threading, and ab initio methods. The structure of the ligand is obtained through experimental methods or generated using 3D structure drawing tools. The 3D structures are used for designing potential drug candidates. Techniques like docking, high-throughput virtual screening, and de novo ligand designing are part of SBDD. However, ligand-based or analog-based drug design techniques are used when neither the experimental nor the predicted 3D structure of the target molecule is known. Methods like 3D and 4D quantitative structure−activity relationship (3D-QSAR and 4D-QSAR), pharmacophore search, and molecular similarity approach are analog-based drug design techniques. QSAR techniques generate the QSAR model correlating the biological activity of known ligand molecules with various descriptors and use the model to predict the activity for new ligand molecules.

This chapter focuses on computational SBDD methods, advances in techniques, and tools. SBDD has been crucial in the development of many FDA-approved drugs (Talele et al., 2010; Clark, 2006; Kitchen et al., 2004).

3.1.1 Structure-based drug design methods

The biological functions in living systems are a consequence of a network of biological interactions between proteins, nucleic acids, carbohydrates, lipids, substrates, and effectors. These interactions are crucial in various biological processes, like signal transduction and cellular regulation. An insight into how the physiological functions occur (particularly in areas of drug designing to find a cure for different diseases) is imperative to study how these interacting molecules influence each other in terms of structure, conformation, and functions. An insight into the mechanisms of biological processes is key to SBDD. The design and development of novel therapeutics or modification of existing drugs to create a highly potent, specific, and selective drug with minimal or no side effects require an understanding of the disease at the structure level of receptor and ligand.

3D structure information of the target molecule is required for SBDD. The 3D structure of the target molecules, usually proteins or RNA, is analyzed by SBDD methods. The objective is to identify critical residues and interactions responsible for biological activity. The Protein Data Bank (PDB) currently has 165,650 structures determined through X-ray crystallography or NMR. The significant techniques of SBDD are (1) docking, and (2) de novo ligand designing.

3.2 Molecular docking

Docking is a method to study the binding affinity of molecules (protein−protein, protein−ligand, protein−DNA) and their interaction mechanism. Docking involves two molecules: a target (receptor) and a ligand. The dynamic perturbations that range from small side-chain flexibility in the catalytic site to domain movements

or opening or closing of channels occur on the binding of the ligand to the receptor. These dynamic perturbations introduced are also known as the induced-fit effect. The biological activity results from the interactions between specific conformations of receptor and ligand. Not all the conformers (poses) of proteins and ligands result in biological activity. Docking predicts the preferred conformation (pose) of a molecule in its bound state to another molecule, which is energetically more stable and mimics the molecular recognition process. Docking techniques must accurately predict ligand poses and their binding affinity in agreement with experimental observation. Fig. 3.1 depicts a ligand bound in the active site of the receptor.

In a nutshell, for any two given biological molecules, docking aims to find:

Whether the two molecules interact.
If they interact with each other, then what is the orientation that maximizes the interaction and minimizes the complex's total energy?

3.2.1 Challenges in docking

The major factors that govern receptor–ligand interactions are:

i. Shape complementarity.
ii. Nonbonded interactions (H-bonding, electrostatic, and van der Waals interactions).
iii. Conformational flexibility of the receptor.

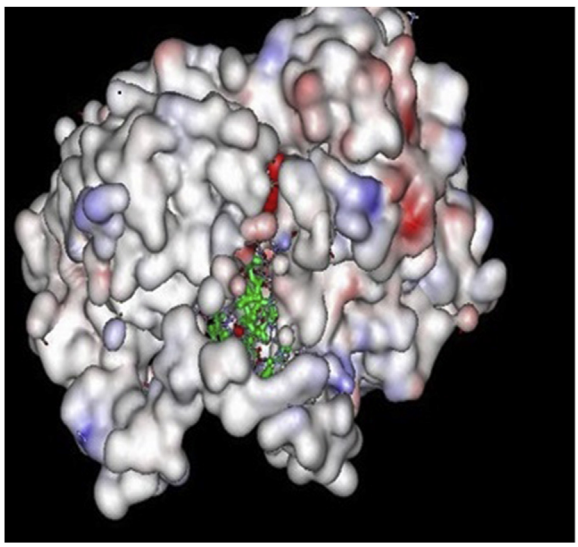


FIGURE 3.1

Surface view of COX-2 monomer bound with a peptide. The peptide is shown in green color in ball and stick mode.

iv. Conformational flexibility of the ligand.
v. Dynamic perturbations in the complex as a result of the binding of the ligand to the receptor.
vi. Presence or absence of solvent molecules.
vii. Presence or absence of other ions/molecules.
viii. Multiple binding sites.

Therefore any docking algorithm must deal with:

i. Six degrees of freedom of the ligand (three translational, three rotational).
ii. *N* degrees of freedom in case of full flexibility consideration of ligand and receptor in the native and bound states.
iii. Interaction of every atom in the ligand with each atom of the receptor.

3.2.2 Types of molecular docking

The receptor in most biological studies is usually a protein molecule, and the ligand may be another protein, small molecule (agonist, antagonist, substrate), DNA, lipid, or carbohydrate. Proteins are highly flexible molecules. Depending upon the type of receptor and ligand involved, docking may be of different types:

- Protein−ligand docking (small molecule docking).
- Protein−protein docking.
- Protein−DNA docking.

Table 3.1 contains a list of software in protein−protein and protein−ligand docking.

3.2.2.1 Rigid versus flexible docking

Docking methods are broadly categorized depending on the consideration of flexibility of receptor and ligand by the docking algorithms: rigid docking and flexible docking. Receptor−ligand docking is a dynamic process. It is a well-known fact that proteins and ligands can exist in different conformations depending on their flexibility. There are six degrees of rotational and translation freedom of the two molecules relative to each other and the conformational degree of freedom of each molecule. The more the number of rotatable bonds in a molecule, the more flexible it would be. The native state conformation of a molecule is different from its conformation in a bound state. Moreover, all the probable conformations of a molecule do not result in a biological effect. Only specific poses contribute to the biological effect.

Rigid docking: In this method, the protein and ligand are both considered as rigid bodies. Their conformational flexibility is not taken into consideration.

Flexible docking: In this method, the protein, as the receptor, is considered rigid and the ligand is treated as flexible. Fig. 3.2 depicts a ligand docked flexibly in the active site of the receptor. Some methods can consider the localized conformational

Table 3.1 Tools for protein–protein docking and protein–ligand docking and their accessibility.

S. no.	Program	Functionality/features	Web address	References
Protein–protein docking				
1.	pyDock	Protein–protein docking	https://omictools.com/pydock-tool	Cheng et al. (2007)
2.	EROS-DOCK	An approach that uses the physics-based ATTRACT function	https://erosdock.loria.fr/	Echartea et al. (2019)
3.	ClusPro	Direct docking of two interacting proteins	https://cluspro.bu.edu/home.php	Kozakov et al. (2017)
4.	ZDOCK	Performs a full rigid-body docking	https://zlab.umassmed.edu/zdock/index.shtml	Chen et al. (2003)
5.	HADDOCK	Flexible docking approach for biomolecular complexes	https://bianca.science.uu.nl/haddock2.4/	van Zundert et al. (2016)
6.	PatchDock	Docking algorithm based on the shape complementarity principle	http://bioinfo3d.cs.tau.ac.il/PatchDock/php.php	Schneidman-Duhovny et al. (2005)
7.	MemDock	Membrane–protein docking algorithm	http://bioinfo3d.cs.tau.ac.il/Memdock/php.php	Hurwitz et al. (2016)
8.	FiberDock	Method for flexible refinement of rigid-body docking	http://bioinfo3d.cs.tau.ac.il/FiberDock/php.php	Mashiach et al. (2010)
9.	LightDock	Open-source protein docking framework written in Python	https://lightdock.org/	Jiménez-García et al. (2018)
10.	FlexDock	Identifies hinge regions and rigid parts followed by docking	http://bioinfo3d.cs.tau.ac.il/FlexDock/php.php	Schneidman-Duhovny et al. (2007)
11.	ParaDock	Protein–DNA docking algorithm	http://bioinfo3d.cs.tau.ac.il/ParaDock/php.php	Banitt et al. (2011)
12.	Symmref	Reranking and refinement of symmetric docking solutions	http://bioinfo3d.cs.tau.ac.il/SymmRef/php.php	Mashiach-Farkash et al. (2011)

Continued

Table 3.1 Tools for protein–protein docking and protein–ligand docking and their accessibility.—*cont'd*

S. no.	Program	Functionality/features	Web address	References
13.	FireDock	Effective in rescoring of rigid-body protein –protein interaction	http://bioinfo3d.cs.tau.ac.il/FireDock/php.php	Mashiach et al. (2008)
14.	CombDock	Prediction of near-native assemblies	http://bioinfo3d.cs.tau.ac.il/CombDock/download/	Inbar et al. (2003)
15.	GRAMM-X	Search for rigid-body conformation	http://vakser.compbio.ku.edu/main/resources_gramm1.03.php	Vakser et al. (1999)
16.	3d GARDEN	Based on marching-cubes algorithm	http://www.sbg.bio.ic.ac.uk/~3dgarden/	Lesk et al. (2008)
17.	ATTRACT	Prediction of complex structures, supports two-body protein–protein docking protocol	http://www.attract.ph.tum.de/services/ATTRACT/ATTRACT.vdi.gz	de Vries et al. (2015)
18.	ICM-DOCK	Involves accurate individual docking sets	http://www.molsoft.com/docking.html	Abagyan et al. (1994)
19.	DOCK/PIERR	Protein–protein docking algorithm based on residue contact potential (PIE) and atomic potential for a given structure (PISA)	http://clsbweb.oden.utexas.edu/dock.html	Viswanath et al. (2014)
20.	LZerD software suite	Pairwise and multiple protein docking	http://kiharalab.org/proteindocking/index.php	Esquivel-Rodriguez et al. (2014)
21.	MEGADOCK	Protein–protein docking tool with ultrahigh performance	http://www.bi.cs.titech.ac.jp/megadock/	Ohue et al. (2014)
22.	HDOCK	Protein–DNA/RNA and protein–protein docking based on hybrid algorithm of template-based modeling and ab initio modeling	http://hdock.phys.hust.edu.cn/	Yan (2017)
Protein–ligand docking				
23.	Autodock	A suite of automated docking tools to predict the binding of small	http://autodock.scripps.edu/	Morris et al. (2009)

Table 3.1 Tools for protein–protein docking and protein–ligand docking and their accessibility.—*cont'd*

S. no.	Program	Functionality/features	Web address	References
		molecules to the 3D structure of a macromolecule		
24.	AutoDock Vina	Open-source program to perform molecular modeling	http://vina.scripps.edu/	Trott et al. (2010)
25.	Dockvision	A complete docking package with Monte Carlo and genetic algorithm and database screening docking algorithm	http://dockvision.sness.net/	Pagadala et al. (2017)
26.	Situs	Flexible refinement of protein structures against intermediate resolution density maps	https://situs.biomachina.org/index.html	Wriggers (2012)
27.	CaverDOCK	Analysis of transportation mechanism in proteins	https://loschmidt.chemi.muni.cz/caverdock/	Vavra et al. (2019)
28.	Flex X	Ensures accurate binding mode prediction	https://www.biosolveit.de/FlexX/	
29.	Glide	Offers high-throughput screening	https://www.schrodinger.com/glide	Friesner et al. (2004)
30.	Gold	Protein-ligand docking software having higher accuracy of binding modes of ligand with the protein	https://www.ch.cam.ac.uk/computing/software/gold-suite	Verdonk et al. (2003)

flexibility of the receptor. Therefore, based on the conformational flexibility of the receptor, flexible docking techniques can be further subclassified into:

Soft docking: This is one of the simplest methods, which takes into consideration of some amount of flexibility in the catalytic site of the receptor. In this method, the interatomic van der Waals interactions are relaxed to allow a small degree of overlap between the ligand and the receptor (Jiang and Kim, 1991; Ferrari et al., 2004). It is computationally less intensive compared to the other three methods.

Side-chain flexibility: This method considers the backbone of the receptor to be rigid and considers the flexibility of the protein's side chains using rotamer libraries (Schnecke and Kuhn, 2000).

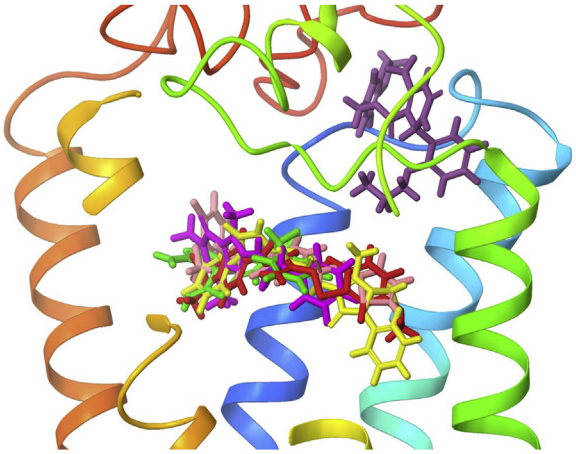


FIGURE 3.2

Different poses of a ligand are flexibly docked in the active site of the receptor. The ligand poses are shown in ball and stick model.

Molecular relaxation: The primary feature of this method is consideration of the backbone as well as the side-chain flexibility of the protein in the region around the ligand. It works in two steps. First, the ligand is docked in the active site of the target using rigid-body docking. In the second step, after rigid docking, the complex thus formed is relaxed or refined using techniques like molecular dynamics (MD) simulation and Monte Carlo (MC) simulations. The relaxation involves translation and rotation of the backbone and side chains of residues within the vicinity of the rigidly docked ligand to remove any steric hindrance and obtain an energetically favorable and stable complex (Apostolakis et al., 1998; Davis and Baker, 2009).

Ensemble based: The real challenge lies in considering the flexibility of both protein backbone and side chains. These methods can be very computationally intensive. Ensemble-based and induced-fit docking (IFD) consider the complete flexibility of the protein during docking.

Proteins are known to exist in different druggable conformations. The ensemble of protein conformations is extracted from the PDB. If the 3D structure of the protein is obtained through structural prediction algorithms, a conformational search is done to obtain all low-energy conformations. Therefore, in this method, a ligand is docked against the receptor's multiple conformations using MD simulations. Ensemble docking results in better pose predictions leading to virtual screening enrichment (Carlson and McCammon, 2000; Carlson, 2002; Teague, 2003; Cozzini et al., 2008; Totrov and Abagyan, 2008; Li et al., 2019). Various methods have been developed in ensemble-based docking (Amaro et al., 2018; De Paris et al., 2018; Fu and Meiler, 2018). The results of ensemble-based docking have found application in providing a structural basis

for the prediction of metabolism, toxicity, and off-target binding (Evangelista et al., 2016). Ensemble-based docking methods have led to a number of FDA-approved drugs (Schames et al., 2004; Hazuda et al., 2004; Summa et al., 2008).
Induced fit: IFD considers both the target and ligand molecules' flexibility. The target protein structures are treated as flexible. Schrodinger's tool, Glide (Sherman et al., 2006), incorporates the IFD method for an exhaustive search for possible binding modes of the ligand and the associated conformational changes within receptor active sites. It iteratively combines structure prediction methods with the docking of a rigid receptor. A Monte Carlo-based minimization algorithm is used by RosettaLigand (Meiler and Baker, 2006). RosettaLigand considers the residues' side-chain flexibility in the binding pocket of the target and flexibility of the backbone residues of the target. Adaptive BP-Dock uses perturbation response scanning in combination with the docking method of RosettaLigand. It has been tested on HIV reverse transcriptase and HIV protease (Boila and Ozkan, 2016). A mutually orthogonal Latin squares method has been developed for the conformational sampling of the flexible residues of the receptor and poses of the ligand (Paul and Gautham, 2017). IFD has been successfully implemented in designing novel lead molecules (Clark et al., 2016; Baumgartner and Evans, 2018).

3.2.2.2 Blind versus site-directed docking

The blind docking approach is used when there is no information available about the ligand-binding active site. The method can explore the probable binding pockets in a receptor, e.g., identification of a probable binding site in acetylcholine nicotinic receptors has been reported using a blind docking approach for allosteric modulators. The site-directed or guided docking approach refers to defining the putative site where ligand may bind. The putative site is defined either using information from site-directed mutagenesis about essential residues or through active site prediction methods.

3.2.3 Methodology

3.2.3.1 Generation of a 3D structure of receptor and ligand

The 3D structure of the receptor and ligand is essential for docking studies and used as input files. The 3D structure of both the receptor and the ligand is obtained from X-ray crystallography or NMR studies. The coordinates are downloaded from the PDB. In case the data from crystal structure/NMR is not available, then structural modeling is carried out. The 3D structure of the target is predicted using structure prediction methods like homology, threading, or ab initio predictions. The ligand 3D structure can be generated by drawing the 2D structure and converting it into a 3D structure followed by optimization using energy minimization techniques. Software like ChemkSketch, ISIS Draw, PubChem Sketcher, ChemDoodle, and Marvin is available (open source and commercial) for building 3D structures of ligands. For example, many of the modeling software, Maestro, MOE, has an in-built tool for building 3D structures.

3.2.3.2 Cleaning and refinement of structures

The accuracy of the docking study depends on input structures for the ligand as well as the target. The structural coordinate files obtained from the PDB consist of heavy atoms of protein of monomeric or heteromeric subunits of protein. It may also have water molecules or other solvent molecules, ligands, cofactors, and metal ions. The 3D structure, either experimentally determined or predicted, requires the following refinements: addition of H-atoms, addition of water (solvent) molecules, addition of any missing side chains or loop regions due to low resolution in a specific area, capping the termini residues, assignment of bond orders, creation of zero bond order to metals, topologies, or formal atomic charges, ionization, and tautomeric states.

Refinement is done using either molecular mechanics (MM) or quantum mechanics (QM). QM calculations solve approximations to Schrodinger's wave equation to determine the molecular properties such as electron density, free energy, transition moments, and others. In MM, a molecule is considered as a series of balls and springs. Hooke's law determines the energy of the molecule. Modules like Maestro of Schrodinger provide a pipeline for protein preparation. There is much standalone software available to carry out the different steps in the preparation of protein and ligand. The protonation states of the amino acids in the protein can be determined using PROPKA (Sondergaard et al., 2011; Olsson et al., 2011), H++ (Anandakrishnan et al., 2012), and SPORES (ten Brink and Exner, 2010). Several methods for the addition or removal of water molecules like 3D RISM (Kovalenko, 2003; Young et al., 2007; Abel et al., 2008), SZMAP (Rashin and Bukatin, 1991), JAWS (Michel et al., 2009), and WaterMap (Young et al., 2007; Abel et al., 2008; Schrodinger, 2020) are available.

3.2.3.3 Identification of active site

There can be multiple binding sites present in a target molecule. However, there is only one active site. The active site is the site where the ligand binding results in biological activity. Experimentally, mutagenesis studies, particularly site directive mutagenesis, can reveal information regarding residues affecting biological activity. Cocrystallization of protein with the ligand is also used to determine the binding site of a ligand in protein. Computationally, a comparison of the structure of a protein with known homologs or pocket detection algorithms is used to detect the active site in a target. All the available algorithms detect the active site based on two parameters: size and shape. The conventional algorithms available are divided into two categories: geometry-based and energy-based methods.

Geometry-based methods generally determine the molecular surface to identify the pockets. Solvent mapping is done with probes like hydrogen atoms or small organic molecules to identify binding sites on a 3D structure. For example, DOCK uses Connolly's algorithm for molecular surface determination. Some of the pocket-finding tools based on geometry-based methods include SURFNET-ConSurf (Glaser et al., 2006), CASTp (Tian et al., 2018), LIGSITE (Huang and Schroeder, 2006), PrankWeb (Jendele et al., 2019), SiteMap (Halgren, 2007,

2009; Schrodinger 2020), FTMap (Ngan et al., 2012), Fpocket (le Guilloux et al., 2009), MDpocket (Schmidtke et al., 2011), QsiteFinder (Laurie and Jackson, 2005), MED-SUMO (Doppelt-Azeroual et al., 2010), and SiteHound-web (Hernandez et al., 2009).

On the other hand, energy-based methods focus on calculating interaction energy between a probe and the protein. A probe molecule can be a hydroxyl, methyl, or any amino group. Q-SiteFinder is an example of an energy-based method where more than one probe may be used in one simulation to study the dynamic behavior. The free energy contributing between the protein and probes can be calculated and used for detecting binding sites. Tools using an energy-based method include MDMix (Seco et al., 2009), SILCS (Raman et al., 2011), and MixMD (Lexa and Carlson, 2011).

Water molecules play an essential role in the physiological system and impact ligands' interaction with the target molecules. The study of thermodynamic properties of water molecules solvating the binding sites and their interactions with the residues of the target and ligands provides significant insight into action mechanisms. Drug design must assess the binding affinity of a ligand. Thus water is also used as a probe for solvent mapping and identification of putative binding sites. Examples of tools include WaterFLAP (Baroni et al., 2007), WaterMap (Young et al., 2007; Abel et al., 2008; Schrodinger, 2020), 3D-RISM (Kovalenko, 2003; Young et al., 2007; Abel et al., 2008), SZMAP (Rashin and Bukatin, 1991), and AquaMMapS (Cuzzolin et al., 2018). A comparative study of water-mapping tools showed that considering water molecules enhances the efficacy of docking results (Bucher et al., 2018).

3.2.3.4 Conformational flexibility of ligand and receptor

In the case of rigid-body docking, the conformational analysis of ligand is carried out. The low-energy conformers are then selected and used for docking studies. Conformational analysis is done using a systematic search, random search, distance-based geometry approach, and genetic algorithm.

3.2.3.5 Docking

There are two essential aspects of any docking technique: algorithm and the scoring function used. These are discussed next. Fig. 3.3 shows a schematic representation of steps in the docking of a protein−ligand.

3.2.3.6 Analysis of docking results

The analysis of docking results is based on the geometrical and stereochemical considerations obtained from docking scores, specific hydrogen bonding, electrostatic interaction, and van der Waals interaction between the ligand and active site of the receptor, and molecular surface analysis of the complex. Fig. 3.4 shows the ligand docked in the active site and interacting with residues of the protein.

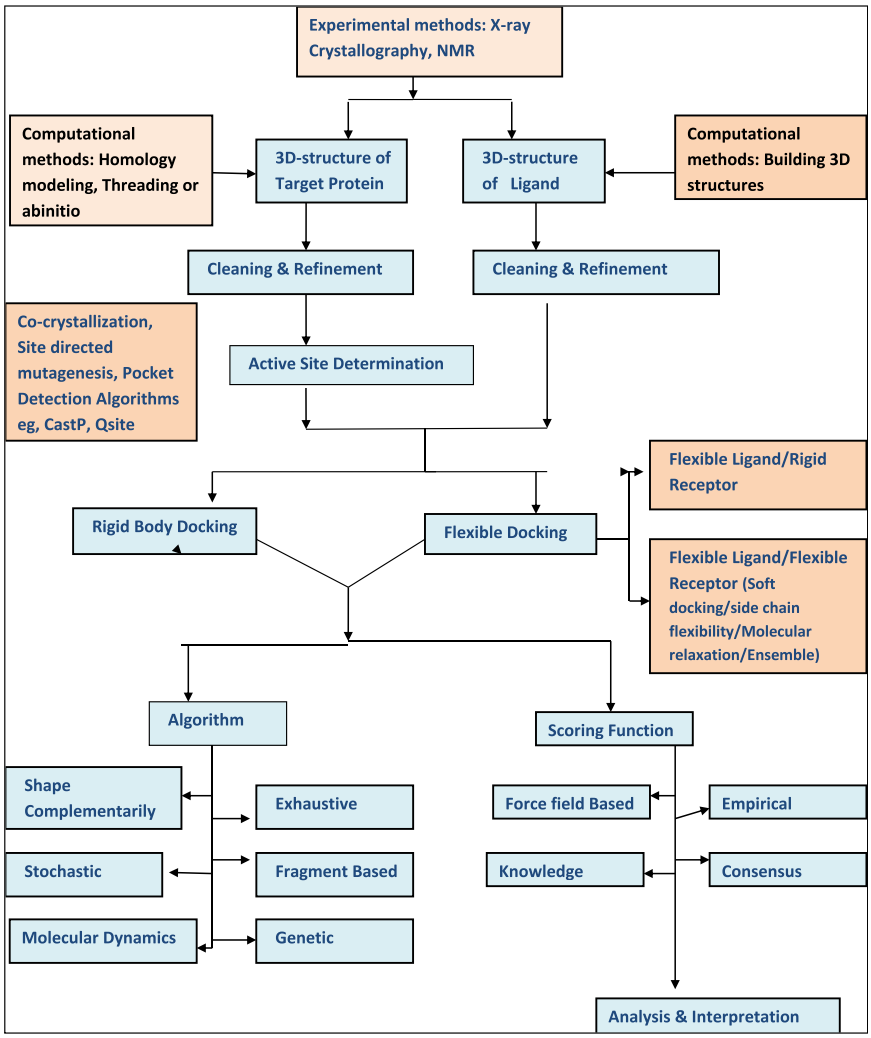


FIGURE 3.3

A flow chart depicting the various steps in molecular docking.

3.2.4 Docking algorithms

The algorithm deals with placing the ligand, within the receptor binding site, by conformational and orientation sampling of the ligand using different techniques. The primary docking algorithms used are:

i. Shape complementarity.
ii. Exhaustive systematic search.
iii. Fragment-based docking (FBD).

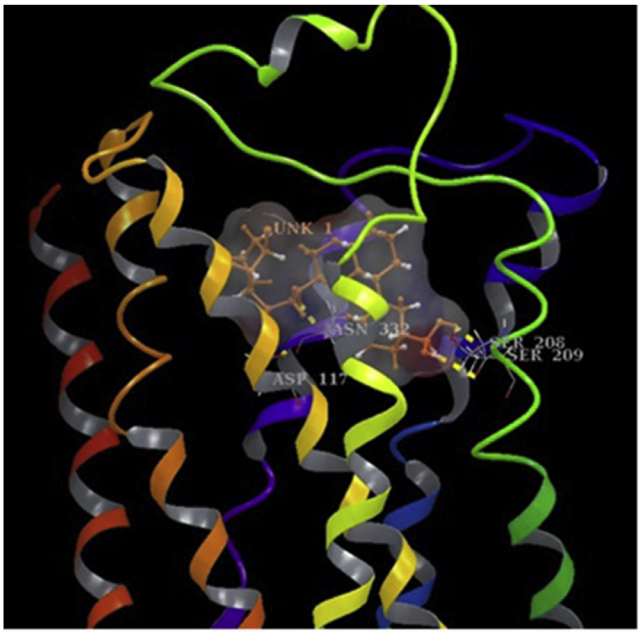


FIGURE 3.4

A docked complex of COX-2 with a ligand. COX-2 is shown as a ribbon and the ligand as a ball and stick model. Interacting residues are labeled.

iv. Stochastic search.
v. Genetic algorithm.
vi. MC simulation.
vii. MD simulation.

3.2.4.1 Shape complementarity algorithm

Shape complementarity is one of the simplest and crudest methods for initially placing the ligand in the binding pocket of the receptor. It is generally used as the first step in many advanced docking algorithms. The ligand is placed by matching the molecular surface of the ligand with that of the binding site in the receptor. Both ligand and receptor are considered as rigid bodies at this stage. The solvent-accessible surface area or the hydrophobic features of a receptor determine the receptor's molecular surface area. If the molecular surface of the ligand complements the molecular surface of the binding site, then the ligand is placed in the binding site.

Some of the docking programs based on shape complementarity are DOCK (Kuntz et al., 1982), FRED (McGann et al., 2003), EUDOC (Pang et al., 2001), LigandFit (Venkatachalam et al., 2003), Surflex (Jain, 2003), MS-DOCK (Sauton et al., 2008), MDock (Huang and Zou, 2007a, 2007b), LibDock (Diller and Merz, 2001), LIDAEUS (Taylor et al., 2008), Ph4DOCK (Goto et al., 2004), and Q-Fit (Jackson, 2002).

3.2.4.2 Exhaustive systematic search algorithm

Systematic search is a method for comprehensively sampling ligand conformations (using all six degrees of freedom) in the torsional space. It is used for flexible docking. It generates all putative low-energy conformers of the ligand. The number of rotatable bonds is identified, and then each rotatable bond is subjected to a torsional value between 0 and 360 using incremental values for the rotation step size. Each of the conformers thus generated is ranked according to the binding energy. With the increase in the number of rotatable bonds, the number of possible conformations also increases. Therefore the exhaustive search algorithm may result in a combinatorial explosion; that is, the number of conformations generated may be too large. With this method, some programs apply geometric and distance constraints to limit the number of conformations generated. MOLSDOCK and Glide (Friesner et al., 2004, 2006; Halgren et al., 2004) are examples of hierarchical sampling methods.

3.2.4.3 Fragment-based docking

The FBD method functions by breaking the ligand molecule into two or more low molecular weight fragments or substructures. These fragments or the poses of the fragments are then placed into the active site either by using the exhaustive systematic search or stochastic search.

In the incremental construction (IC) algorithm, the complete ligand is built within the active site by incrementally growing fragments one after the other in sequential order. All the fragments are placed in the active site and covalently connected using scaffolds or linkers. The scaffolds or linkers are used only for connecting the fragments, and they may have no interactions with the active site residues.

The advantage of using the FBD method is that it is computationally less exhaustive and thus more efficient. The substructures or fragments themselves have weak interactions with the receptor. However, these fragments are subjected to optimization within the active site and connected to build the complete ligand having a higher affinity for the receptor. This method is used in de novo drug design. Some tools based on FBD include DOCK (Ewing and Kuntz, 1997), LUDI (Bohm, 1992), FlexX (Rarey et al., 1996), and ADAM (Mizutani et al., 1994). Tools like DOCK 6.0 (Allen et al., 2015), Hammerhead (Welch et al., 1996), eHITS (Zsoldos et al., 2006), FLOG (Miller et al., 1994), PatchDock (Schneidman-Duhovny et al., 2005), and ProPOSE (Hogues et al., 2018) use the IC algorithm.

3.2.4.4 Stochastic search algorithm

The stochastic search algorithm places the ligand in the active site by random sampling of ligand conformation in both the Cartesian coordinate space and the torsional space. Random changes are made either in the Cartesian coordinates or torsional angles. The nature of the random movement and the criteria employed for accepting or rejecting the random moves vary depending on the method employed.

Stochastic search algorithms are further subclassified into:

1. MC search
   In this strategy, multiple runs are carried out from random starting positions and orientations. The probability of accepting or rejecting a random move is calculated using the Boltzmann probability function ($\exp(\Delta E/kT)$) through a comparison of probability against a random number. The move is accepted only if the Boltzmann probability of the move is larger than the random number. In case the value is smaller, the system returns to its original configuration.
   The ligand makes random moves around the receptor. At each step, random changes are made either in the Cartesian coordinates or torsional angles. The energy calculated for each change or step is compared to the previous energy value. The new step is only accepted if the energy value is lower than the previous step. In the case of higher energy, the step is accepted with a probability $\exp(\Delta E/kT)$.
   Programs that use the MC method include DockVision (Hart and Read, 1992), Prodock (Torsset and Scheraga, 1999), MCDOCK (Liu and Wang, 1999), ICM (Abgyan et al., 1994), RosettaLigand (Meiler and Baker, 2006), and AutoDock Vina (Trott and Olson, 2010).
2. Tabu search
   In this method, new states or conformations (N solutions) are randomly generated from an initial state. A record of the new conformations and the conformational search space explored is maintained so that the algorithm is forced to search into previously unexplored conformational spaces. The new conformations are scored and ranked in ascending order. Out of all the N solutions, only the best conformer (solution) is retained. The probability of acceptance of a new conformation is dependent on the previously explored conformational space. Tabu search makes use of the root mean square deviation (RMSD) to accept or reject the orientation or conformation of a ligand. The new conformation generated after a random search is accepted if the RMSD value between the new ligand conformation and any of the previous conformations is more than a threshold cut-off. In case the RMSD value of a new conformer and any previous conformers obtained is less than the cut-off, then the particular random move is rejected. Some examples of tools are PRO LEADS (Baxter et al., 1998) and PSI-DOCK (Pei et al., 2006).
3. Particle swarm optimization
   Particle swarm optimization (PSO) is a population-based stochastic optimization technique based on fish schooling's social behavior. In this method, initialization of the system is done using random solutions. The optimal solutions are obtained by updating generations. The potential solutions follow the current optimal solution to further move through the problem space. The best positions of the neighbors direct the movement of a ligand in the search space. Tools based on this method include SODOCK (Chen et al., 2007), Tribe-PSO (Chen et al., 2006), PSO@Autodock (Namasivayam and Gunther, 2007), and FIPS DOCK (Liu et al., 2013).

4. Genetic algorithm
   This comes under the category of evolutionary algorithms and involves stochastic search. Genetic algorithms are based on the theory of evolution and natural selection by Darwin. Different conformations and positions of ligands are generated. These initial states are considered to be at the lowest energy positions. These initial conformers are subjected to crossover and random mutations to generate new conformers. A set of values defines the conformation and orientation of a ligand and a protein called state variables (e.g., dihedral angles, ring geometry), which reflect the conformation and orientation of the ligand with respect to the receptor. Here, the genotype is the state of the ligand, and atomic coordinates define the phenotype. After successive steps of evolution, the best conformation of the ligand having the lowest energy is selected. The conformers having favorable genes are accepted and passed on to the next generation, and unfavorable conformers are eliminated. The selection of favorable/unfavorable conformer is through the fitness score. The interaction energy of the protein with the ligand defines the fitness score.
   GOLD (Jones et al, 1995, 1997), AutoDock (Morris et al., 1998), DIVALI (Clark, 1995), DARWIN (Taylor and Burnett, 2000), MolDock (Thomsen and Chirstensen, 2006), PSI-DOCK (Pei et al., 2006), FLIPDock (Zhao and Sanner, 2007), GAsDock (Li et al., 2004), Lead finder (Stroganov et al., 2008), and EADock (Grosdidier et al., 2007) are some of the tools that have implemented genetic algorithms for docking. A variant of genetic algorithm called Lamarckian genetic algorithm has also been developed to handle larger degrees of freedom.
5. MD simulation
   MD simulation is a technique to study the dynamic behavior of molecules. This method can help in understanding the dynamic perturbations occurring when a ligand binds to the receptor. It is based on Newton's second law of motion:

$$F = ma \text{ where } F \text{ is force, } m \text{ is mass, and } a \text{ is acceleration.} \qquad (3.1)$$

   This equation is integrated over a period of timeframe (usually in the order ranging from picoseconds to femtoseconds) to determine the new position, velocity, and acceleration of the molecules. The output is in terms of the new conformer of the complex. The protein is rigid, and the ligand is flexible. The conformations generated are docked into the protein in successive steps. MD simulation is carried out, followed by energy minimization of the system steps. Scoring is done based on energy values. This technique is useful in determining poses that are comparable with experimental structures.

3.2.5 Scoring functions in docking

The scoring function's objective is to analyze and rank the poses generated and select the best pose (conformer) for a given ligand based on binding affinity. The accuracy of the docking algorithm is determined by the scoring function used.

Depending on the method of derivation, the scoring functions have been categorized into the following types.

3.2.5.1 Forcefield-based scoring functions

Forcefield is a mathematical function defining the conformations based on energy terms. These scoring schemes approximate the free binding energy of protein–ligand complexes using forcefields. In other words, forcefields are sums of terms that correspond to bonded and nonbonded interaction, namely bond, angle, torsion, van der Waals, and electrostatic interaction energies as functions of conformation.

The forcefield parameters from AMBER (Weiner and Kollman, 1981), CHARMM (Brooks et al., 1983), OPLS (Jorgensen et al., 1996), OPLS3 (Harder et al., 2016), and MMF forcefields are used. The solvent effect is considered using (1) distance-dependent dielectric constant or (2) explicit solvents such as free energy pertubation and thermodynamic integration (Wang et al., 2001) or (3) implicit solvents such as Poisson–Boltzmann/surface area models (Rocchia et al., 2002; Grant et al., 2001) and the generalized-born/surface area models (Zou et al., 1999; Liu et al., 2004). The limitations of these methods are in the calculation of entropic effects and free energy calculations. Examples include AutoDock, G-Score, GOLD, DockScore, GoldScore, and HADDOCK scoring functions.

3.2.5.2 Empirical scoring functions

A set of protein–ligand complexes with known binding affinity is used for deriving the empirical scoring. The set of weighted empirical energy terms is used to calculate the binding energy score of a receptor–ligand complex. The empirical energy terms include van der Waals energy, electrostatic energy, hydrogen bonding energy, desolvation term, entropy term, and hydrophobicity term:

$$\Delta G = \sum Wi\ \Delta Gi \tag{3.2}$$

where $\{\Delta Gi\}$ shows individual empirical energy terms and $\{Wi\}$ represents the coefficients $\{Wi\}$ corresponding to it. The coefficients are determined using least square fitting by comparing the binding affinity data of a training set of receptor–ligand complexes with known structures (Eldridge et al., 1997; Krammer et al., 2005; Wang et al., 2002). The fit of the pose is evaluated according to this inferred potential. The empirical scoring functions are relatively more accurate than forcefield-based scoring functions. The use of binding constants of known protein–ligand complexes from the PDB can enhance the efficacy of empirical scoring functions.

Examples of an empirical scoring function using algorithms/programs are GlideScore (Friesner et al., 2004; Halgren et al., 2004), PLP (Gehlhaar et al., 1995), F-Score (Rarey et al., 1996), LigScore (Krammer et al., 2005), LUDI (Bohm 1994, 1998), SCORE (Wang et al., 1998), X-Score (Wang et al., 2002), ChemScore (Eldridge et al., 1997), Medusa Score (Yin et al., 2008), AIScore (Raub et al., 2008), and SFCscore (Sotriffer et al., 2008).

3.2.5.3 Knowledge-based scoring function

The knowledge-based scoring functions are based on the structural information available in known receptor–ligand complexes in the PDB. The energy of receptor–ligand binding is the sum of the interaction terms for all the receptor–ligand atom pairs in the complex. The probability distributions of interatomic distances are converted into distance-dependent interaction free energies of protein–ligand atom pairs. The "inverse Boltzmann law" is used to convert interatomic distance probability distributions into distance-dependent interaction free energies of protein–ligand atoms (Huang and Zou, 2010):

$$w(r) = -k\mathrm{B}T \ln[p(r) / p^*(r)] \tag{3.3}$$

Here, kB represents the Boltzmann constant, T the absolute temperature of the system, $p(r)$ the number density of the protein–ligand atom pair at distance r in the training set, and $p^*(r)$ the pair density in a reference state of no interatomic interactions. The reference states determine the weights between the various probability distributions. The knowledge-based scoring functions are efficient in terms of accuracy and speed as a large number of receptor–ligand complexes have been used to generate the potential term.

Examples include DrugScore (Gohlke et al., 2000; Velec et al., 2005), SMoG (DeWitte and Shakhnovich 1996; Ishchenko and Shakhnovich, 2002), BLEEP (Mitchell et al., 1999), GOLD/ASP (Mooji and Verdonk, 2005), MScore (Yang et al., 2006), and KScore (Zhao et al., 2008).

3.2.5.4 Consensus scoring

Consensus scoring is used to enhance the accuracy of a docking score by considering the scores from different scoring functions to minimize the errors in scoring functions. MultiScore and X-Cscore use consensus scoring (Wang et al., 2002; Terp et al., 2001).

3.3 High-throughput screening

High-throughput virtual screening (HTVS) is an extensively used method in SBDD. It is applied in the early stages of drug discovery and development. HTVS methods facilitate the fast screening of a large number of compounds against a specific biological target molecule to identify hits having a good affinity for binding to the target. The hits thus identified provide insight into the modulation mechanism of a particular biomolecular mechanism or pathway and interactions involved with the target in a particular physiological process at the cellular level. The screened hits can then be further modified and developed into lead compounds having the potential to be a drug molecule.

The methods, steps involved, and tools of HTVS are discussed here.

3.3.1 Methodology of virtual screening

3.3.1.1 Compound databases

There are many chemical compound databases available consisting of different chemical entities, either from synthetic sources or natural sources. These databases may be compiled from in-house data or compound databases having pretested or untested molecules (van Hilten et al., 2019;.Gong et al., 2017). The databases are either publicly available or commercially available. Some of the notable publicly available repositories for virtual screening include PubChem, ZINC, DrugBank, ChEMBL, ChemSpider, and NCI.

PubChem, a repository of NIH, contains millions of small molecules showing biological activity (Kim et al., 2019). It also contains macromolecules such as peptides, lipids, nucleotides, and carbohydrates. It provides details relating to structures, identifiers, physicochemical properties, activity, pharmacokinetic profile, and patent information for the molecules. Currently, it contains 102,768,482 molecules. ZINC database has more than 120 million compounds with drug-like properties and is readily available for purchase (Sterling and Irwin, 2015). The data for small molecules contain biological activity, physicochemical properties, structure, and commercial availability. The latest version is ZINC 15, which uses information from public databases such as ChEMBL, HMDB, DrugBank, and https://ClinicalTrials.gov for annotation of high activity compounds. So, the database is enriched with chemical compounds and biogenic molecules, natural products, metabolites, and approved drugs. DrugBank database (Wishart et al., 2018) is a comprehensive database of drugs and their targets. This database currently contains 13,579 drug entries, which include approved 2635 small molecule drugs, 1378 biologics, 131 nutraceuticals, and more than 6375 compounds under evaluation because of their probability of being drugs. It also has 5229 nonredundant protein sequences linked to the drug entries. ChEMBL (Gaulton et al., 2017) is a curated database of bioactive molecules. There are more than 1.6 million unique compounds in the current version, with 14 million activity values from 1.2 million assays mapped to about 11,000 targets.

Many commercial databases are also available from companies such as ChemBridge, ChemDiv, Maybridge, MedChemExpress, ChemNavigator, etc. These databases need refinement as the molecules are generally in 2D SDF format. ChemDiv has a large, diverse, and pharmacologically important collection of compounds, including more than 1,500,000 small drug-like molecules. MedChemExpress contains over 10,000 molecules having proven pharmacological activities. Schrodinger has a phase database of fragments and probable lead-like and drug-like compounds for which it has partnered with Enamine, MilliporeSigma, and MolPort.

3.3.1.2 Ligand preparation of the compound database

The next step involves the preprocessing and prefiltering of compounds. Most of the databases, except for the ZINC database, do not have compounds in a format that can be directly used for docking in virtual screening. Ligand preparation involves multiple preprocessing steps and format conversions. Generally, the compounds in the

libraries are stored in compact 1D formats like SMILES or 2D formats like SDF. The compounds in SMILES or SDF formats are converted to 3D format during preprocessing to assign the proper stereochemistry, tautomeric, protonation states, and energy minimization. Tools like OpenBabel (O'Boyle et al., 2011), an open-source software, convert into a 3D format. There are many software packages like LigPrep (Schrodinger, 2020), Epik (Greenwood et al., 2010; Schrodinger 2020), and SPORES (ten Brink and Exner, 2010), which are specifically for ligand preparation.

The screening of large libraries of compounds is computationally intensive and time consuming, and the screened compounds have redundant possibilities. Therefore many filters are applied to the compounds in the library to screen compounds having "drug-likeness" properties. One of the widely used filtering parameters is the "Lipinski rule of five." It states that drug-like compounds should have molecular weights lower than 500, lipophilicity lower than five, hydrogen bond donors less than five, and hydrogen bond acceptors less than 10. However, it has been reported that even many approved oral drugs do not precisely follow the Lipinski rule of five. There are many variations with flexibility like the rule of three, which states that molecular weight should be less than 300, logP value should be less than three, number of hydrogen bond donors and acceptors should be less than three, and number of rotatable bonds should be less than three.

Pfizer's rule of 3/75 correlates physicochemical properties to preclinical toxicity. The partition coefficient (ClogP) values are compared with the topological polar surface area (TPSA). According to this rule, compounds having a ClogP value lower than three and TPSA higher than 75 have a higher probability of passing in in vivo assays. The quantitative estimate of drug-likeness filter, which ranks chemical structures based on the properties of orally available drugs, is also used.

Absorption, distribution, metabolism, elimination, toxicity (ADMET) filtering may be applied at this stage or later to filter the compounds based on their bioavailability and toxicity.

3.3.1.2.1 Library design

During the preprocessing stage, customized libraries may be designed. Highly similar structures can be removed from a library by ligand similarity calculations. A customized library holds a wide range of chemically diverse molecules while reducing the size of the database. The compound libraries may also be customized to a specific target or have compounds with specific molecular property profiles based on different physicochemical properties like lipophilicity, partition coefficient, solubility, and fragment-likeness.

The open-source software for library design includes CLEVER (Song et al., 2009) and ChemT (Abreu et al., 2011). There are many other commercial tools for library design as per user-defined filters like Tripos, Diverse Solution, Accelrys Discovery Studio, and Medchem Studio. Table 3.2 contains details of the major HTVS tools available.

Table 3.2 Tools for high-throughput virtual screening workflow.

S. no.	Program	Functionality/ features	Web address	References
1.	PyRx	Screens library of compounds against potential drug targets	http://mgltools.scripps.edu/	Dallakyan et al. (2015)
2.	ProDy	An open-source Python package for protein structural dynamics analysis	http://prody.csb.pitt.edu/	Bakan et al. (2011)
3.	FragPELE	Hit-to-lead drug design tool	https://carlesperez94.github.io/frag_pele/first_steps.html	Perez et al. (2020)
4.	GpuSVMScreen	Ligand-based virtual screening	https://bio.tools/GpuSVMScreen	Jayaraj et al. (2019)
5.	ATT2	Designed for automatic lead optimization, also lead discovery	http://www.sioc-ccbg.ac.cn/software/att2/	Li et al. (2016)
6.	MolAr	Carries out the entire virtual screening process	http://www.drugdiscovery.com.br/software/	Maia et al. (2020)
7.	Octopus	Can perform fast and friendly docking simulation	http://www.drugdiscovery.com.br/software/	Maia et al. (2017)
8.	PRODIGY	Focuses on binding energy of biological complexes and identification of biological interfaces from crystallographic structures	https://bianca.science.uu.nl/prodigy/lig	Vangone et al. (2015)
9.	PSOVina	Fast docking tool optimization algorithm of particle swarm intelligence	https://cbbio.cis.um.edu.mo/software/psovina/	Tai et al. (2018)
10.	PoLi	Pipeline based on template pocket and ligand similarity	http://cssb.biology.gatech.edu/PoLi.	Roy A et al. (2015)
11.	Panther	Ultrahigh-throughput screening procedure	http://www.medchem.fi/panther/	Niinivehmas et al. (2015)
12.	GeauxDock	Binding of small ligands with	http://www.institute.loni.org/	Fang at al (2016)

Continued

Table 3.2 Tools for high-throughput virtual screening workflow.—*cont'd*

S. no.	Program	Functionality/ features	Web address	References
		pharmacologically relevant molecules	lasigma/package/dock/	
13.	LIDAEUS	High-throughput in silico screening program	http://opus.bch.ed.ac.uk/lidaeus/	Lim et al. (2011)
14.	GalaxySite	Ligand-binding site prediction	http://galaxy.seoklab.org/site	Heo et al. (2014)
15.	mRAISE	Descriptor-based bitmap search engine	http://www.zbh.uni-hamburg.de/raise	von Behren et al. (2016)
16.	GEMDOCK	Program to compute a ligand conformation and orientation relative to active site of the protein	http://gemdock.life.nctu.edu.tw/dock/	Yang et al. (2004)
17.	HomDock	Similarity based, used to improve efficiency and accuracy of complex binding of ligand and unknown protein	http://www.chil2.de/HomDock.html	Marialke et al. (2007)
18.	GlamDock	Based on Monte Carlo and minimization search in hybrid interaction	http://www.chil2.de/Glamdock.html	Tietze et al. (2007)
19.	eSimDock	Ligand docking and binding affinity prediction	http://www.brylinski.org/esimdock	Brylinski M (2003)
20.	Sliding Box Docking	Standalone tool for managing simulations of ligand docking at defined positions of 3D structures of DNA	https://sourceforge.net/projects/slidingboxdocki	Martins-José (2013)
21.	VSDocker	Uses AutoDock4 for optimized virtual screening	http://www.bio.nnov.ru/projects/vsdocker2	Prakhov et al. (2010)
22.	SwissDock	Small molecule docking screening	http://www.swissdock.ch/	Grosdidier et al. (2011)
23.	Exemplar	A map generated for the perfect ligand bound to the complement	https://rosie.rosettacommons.org/make_exemplar	Lyskov et al. (2013)
24.	PyPLIF	Method to interpret 3D interaction of		Radifar et al. (2013)

Table 3.2 Tools for high-throughput virtual screening workflow.—*cont'd*

S. no.	Program	Functionality/ features	Web address	References
		ligand and protein into bit array form (1D)	https://code.google.com/archive/p/pyplif/	
25.	BAPPL-Z	A server that predicts the binding affinity of a protein–ligand complex containing zinc	http://www.scfbio-iitd.res.in/software/drugdesign/bapplz.jsp	Jain et al. (2007)
26.	Fold X	Provides the importance of interactions contributing to the stability of protein –protein complexes	http://foldxsuite.crg.eu/	Schymkowitz et al. (2005)
27.	CRDOCK	Protein–ligand docking program similar to GLIDE	https://ub.cbm.uam.es/drug_design/crdock.php	Cortes Cabrera et al. (2012)
28.	VSDMIP	Virtual screening of chemical libraries integrated with MySQL relational database	https://ub.cbm.uam.es/drug_design/vsdmip.php	Gil-Redondo et al. (2009)
29.	AMMOS	Molecular mechanics optimization tool for high-throughput screening	http://drugmod.rpbs.univ-paris-diderot.fr/ammosHome.php.	Pencheva et al. (2008)

3.3.1.3 Target preparation

The receptor molecule is prepared. The protocol followed is the same as discussed earlier in the docking section (Fig. 3.3).

3.3.1.4 Docking

Each compound in the library is virtually docked into the active binding site. Details of docking algorithms and tools are given in the docking section 3.2.4.

3.3.1.5 Postprocessing

The screened compounds and their poses are ranked based on scores and analyzed for binding scores, poses, desirable chemical moieties, physicochemical properties, lead-likeness, chemical diversity, bonded and nonbonded interactions with the target, and ADMET profiling. The selected compounds are validated through experimental assays.

HTVS is an effective method to identify hits having diverse chemical structures, even in the absence of high-resolution crystallographic data or standard drug-binding data. Recently, a novel HTVS cascade protocol has been reported, combining both pharmacophore modeling and molecular docking to identify novel compounds for cancer immunotherapy (Serafini et al., 2020). HTVS has been used to successfully identify novel molecules for leukocyte antigen DR, Bruton's tyrosine kinase, and retinoic acid-related orphan receptors (Damm-Ganamet et al., 2019).

3.4 De novo ligand design

De novo ligand design uses knowledge about the 3D structure of receptor design novel lead molecules using molecular modeling tools. The high-resolution structure of the target or structure—activity relationship data of active modulators and well-defined binding site are required for de novo designing. The de novo design tools search the active site binding space for novel potent hit compounds. This technique provides the edge in designing leads with a defined selectivity profile and a unique molecular structure. The ligands designed may be similar to known inhibitors or novel scaffolds. These are then synthesized, and bioactivity assays carried out to validate the biological activity. The methods of de novo ligand design can be broadly classified into two categories: (1) whole molecule docking and (2) fragment-based techniques.

3.4.1 Whole molecule docking

Each of the proposed ligands is docked to position it in the receptor's active site or matched to a pharmacophore model representing the active site. The different conformations of the ligand are generated during docking to the active site to identify poses having good binding affinity. It makes use of properties like shape complementarity and electrostatic fitting for docking.

3.4.2 Fragment-based methods

The fragment-based methods are classified into four subcategories:

1. Site-point connection methods: Determine desirable locations of individual atoms ("site points") and then place suitable fragments.
2. Fragment connection methods: Start with previously positioned fragments in the active site and find "linkers" or "scaffolds" to connect those fragments without moving previously positioned fragments.
3. Sequential buildup methods: Construct ligands atom-by-atom or fragment-by-fragment within the active site. The set of building blocks is generally small, and the construction process may be random.
4. Random connection methods: Amalgamate different techniques incorporating randomness in designed ligands by incorporating specific features from various methods and bond-disconnection strategies.

There is much software available for de novo ligand design. Tools for ligand building such as biochemical and organic model builder (BOMB), SPROUTS, LigBuilder, SILCS, LigMerge, and ReLeaSE have been developed. BOMB (Barreiro et al., 2007; Jorgensen, 2009) builds molecules by fixing the core structure and adding substituents. SPROUTS has been successfully used to design inhibitors for *Escherichia coli* RNS polymerase (Gillet et al., 1994; Sova et al., 2009). LigMerge (Lindert et al., 2012) identifies the maximum common substructure by analyzing the known ligands. By systematically altering the distinct fragments attached to the common substructure at each complex, LigMerge produces multiple molecules with common features of the known ligands. SILCS uses MD simulations to find ligands with a high probability of binding to the receptor (Raman et al., 2012; Faller et al., 2015). The high probability binding areas of the target are analyzed from multiple simulation results. In LigBuilder (Yuan et al., 2011, 2020), the ligands are either grown or linked, and an empirical scoring function is used to estimate binding affinities. Tools like LUDI or Pocket identify the key interactions or hot spots at the binding site and convert these into 3D search queries and virtual screening. ReLeaSE (Popova et al., 2018) designs novel chemical compounds with desired properties by combining two deep neural networks namely-generative and predictive neural networks. Sequential graph generators have also been developed for de novo designing (Li, 2018). Designing of dual inhibitors of c-Jun N-terminal kinase 3 and glycogen synthase kinase-3 beta was accomplished using deep generative models. These compounds showed effective activity for both the targets.

The fragment-based de novo ligand design can assemble drug-like molecules in a highly reduced search space. They have also been used in de novo drug design, target selectivity, and receptor-based pharmacophore screening (Hartenfeller and Schneider, 2011; Schenider and Clark, 2019; Fischer et al., 2019; Amaravadhi et al., 2014). Table 3.3 shows details of the widely used de novo drug design programs.

3.5 Biomolecular simulations

The 3D structures of biomolecular complexes obtained from X-ray crystallography, NMR, and cryogenic electron microscopy reflect only certain aspects of molecular recognition. They only partly capture the dynamic behavior of biomolecules. However, biomolecules are dynamic. Both the ligand and the receptor may occur in multiple conformations. The native or unbound (apo) conformation of a receptor is different from its bound (holo) state conformation due to dynamic perturbations. Various factors such as solvent rearrangements and fluctuations, electrostatics—polarization, temperature, pH, ionic strength, presence of metal ions, and other molecules contribute to structural conformational transitions. All these factors are critical for structure-based and ligand-based drug design. There are multiple binding sites in a receptor molecule. The binding of ligands on one site may cause allosteric

Table 3.3 Software for de novo ligand design and their accessibility.

S. no.	Program	Functionality/ features	Web address	References
1.	REINVENT 2.0	Production-ready tool for de novo drug design	https://github.com/MolecularAI/Reinvent	Thomas et al. (2020)
2.	CzeekD	Fragment-based de novo drug design system	https://www.k-ct.jp/en/service/czeekd.html	Yoshikawa et al.
3.	DeepScaffold	A scaffold-based tool using deep learning	https://github.com/deep-scaffold	Li et al. (2019)
4.	SPROUT	Structure-based drug design	http://www.keymodule.co.uk/products/sprout/sprout-classic.html	Gillet et al. (1994)
5.	Glide	Offers high-throughput screening	https://www.schrodinger.com/glide	Friesner et al. (2006)
6.	AutoGrow4	Open-source genetic algorithm based	http://durrantlab.com/autogrow4	Spiegel et al. (2020)
7.	iSyn	WebGL-based interactive program, evolutionary-based algorithm that designs novel ligands	http://istar.cse.cuhk.edu.hk/iSyn.tgz	Li et al. (2014)
8.	Ludi	Designs candidate ligands for the active site of proteins	http://www.esi.umontreal.ca/accelrys/life/insight2000.1/ludi/1-Intro.doc.html	Böhm (1992)
9.	DOGS	Reaction-based program		Hartenfeller et al. (2012)
10.	LigBuilder V3	De novo multitarget approach and optimization	http://www.pkumdl.cn/ligbuilder3/	
11.	e-LEA3D	Performs CADD based on molecular fragments	https://chemoinfo.ipmc.cnrs.fr/LEA3D/index.html	Douguet D et al. (2005)

effects. Thus the finding of the biologically active conformation of a biomolecule is a challenging task.

Biomolecular simulation techniques are invaluable in understanding protein motion and conformational flexibility of the target molecule in apo form and holo form. The structural properties and the microscopic interactions between the assembly of molecules can be explored through simulations. Biomolecular simulation techniques can be broadly categorized in (1) MD and (2) MC simulations. There are also many hybrid techniques incorporating both MD and MC.

3.5.1 Molecular dynamics simulations

In MD, conformations of the system are generated by integrating Newton's laws of motion. The trajectories define positions and velocities of the particles over a period of time. MD methods provide an insight into the transient changes and dynamic perturbations. Forcefield parameters are used to mimic the dynamic behavior of real molecules (Durrant and McCammon, 2011). Realistic atomistic simulation of molecular systems is dependent on the accurate and reliable molecular mechanics forcefield. Several forcefields used in MD simulations include AMBER, CHARMM, OPLS, and GROMOS. They differ only in terms of parameterization. Some of the MD tools include AMBER, NAMD, GROMACS, and DESMOND.

Table 3.4 includes a list of important tools used for biomolecular simulation. With the advent of graphical processor unit architectures and increasing computational power, it is feasible to run long-range MD simulations with better accuracy. The estimation of thermodynamics and kinetics associated with drug−target recognition is enhanced by explicit structural flexibility and entropic effects. The range of timescale of a MD simulation is in nanoseconds to microseconds to milliseconds.

The purpose of MD simulation is to find all possible conformational states in which a molecule may exist. Individual states or conformations of the protein are often separated from others by extremely high energy barriers. The high computational demands limit conventional MD simulations to the order of microseconds, thereby resulting in inadequate sampling of conformational states. Enhanced sampling methods solve the issue of timescale in conventional MD and enable them to find biologically relevant conformational states. Fig. 3.5 shows a typical system setup in MD. The enhanced sampling algorithms solve the timescale problem and enhance the conformational sampling. Different techniques of enhanced sampling are reported: accelerated molecular dynamics, umbrella sampling, multicanonical algorithms, simulated tempering, transition path sampling, targeted molecular dynamics, and parallel tempering.

3.5.1.1 Accelerated molecular dynamics

Accelerated molecular dynamics simulation reduces energy barriers separating different states of a system. It improves the conformational space sampling. The potential energy landscape is modified by increasing energy wells that are below a certain threshold level. The energy wells above the threshold remain unaffected. Thus the energy barriers are reduced, and better conformational sampling is done (Hamelberg et al., 2004; Hamelberg and McCammon, 2005; Markwick et al., 2007; Bucher et al., 2011). It has been used in simulations of fast-folding proteins (Miao et al., 2015), G-protein coupled receptors (Miao et al., 2014), a silk-like polypeptide (Zhao et al., 2017), bovine pancreatic trypsin inhibitor (Pierce et al., 2012), streptavidin−biotin complex (Song et al., 2015), antitrypsin (Andersen et al., 2017), MSI-594 (Mukherjee et al., 2017), insulin (Nejad and Urbassek, 2018), and helical proteins in explicit water (Duan et al., 2019).

Table 3.4 Software for molecular dynamics and Monte Carlo simulations.

S. no.	Program	Functionality/ features	Web address	References
1.	LARMD	Based on conventional molecular dynamics	http://chemyang.ccnu.edu.cn/ccb/server/LARMD/index.php/home/index	Yang et al. (2019)
2.	Tinker-HP	Devoted to long-polarizable molecular dynamics simulations	http://tinker-hp.ip2ct.upmc.fr/	Lagardère et al. (2018)
3.	NAST	Generates RNA structures using knowledge-based forcefield	https://simtk.org/projects/nast	Jonikas et al. (2009)
4.	DelPhi Force	Calculates electrostatic force	http://compbio.clemson.edu/delphi-force/	Li et al. (2017)
5.	MDWeb	Runs standard molecular dynamics simulations	http://mmb.irbbarcelona.org/MDWeb/	Hospital et al. (2012)
6.	ProtPOS	Predicts preferred orientation of protein on the surface with initial absorption	https://cbbio.cis.um.edu.mo/software/protpos/	Jimmy et al. (2016)
7.	Desmond	High-speed molecular dynamic simulations	https://www.deshawresearch.com/resources_desmond.html	Robustelli et al. (2020)
8	Veinna-PTM	Molecular dynamics simulations for exploring posttranscriptional modifications	http://vienna-ptm.univie.ac.at/	Margreitter et al. (2013)
9.	LocalMove	Based on the Monte Carlo approach	http://bioinformatics.bc.edu/clotelab/localmove/	Meng et al. (2011)
10.	GROMACS	Fast and flexible program and freely accessible	http://www.gromacs.org/	Abraham et al. (2015)
11.	ProtoMol	An object-oriented component-based framework for molecular dynamics simulations	http://protomol.sourceforge.net/	Matthey et al. (2004)
12.	NAMD	Molecular dynamics code for high-yield simulation of macromolecules	http://www.ks.uiuc.edu/Research/namd/	Phillips et al. (2005)

Table 3.4 Software for molecular dynamics and Monte Carlo simulations.—*cont'd*

S. no.	Program	Functionality/ features	Web address	References
13.	Amber	Suite of programs to carry out molecular dynamics simulations	http://ambermd.org	Ferrer et al. (2013)
14.	MOIL	Suite of programs with a set of tools	http://clsbweb.oden.utexas.edu/moil.html	Ruymgaart et al. (2011)
15.	SQUEEZE	Method evolved to cut the cost of computational simulations	http://haddock.science.uu.nl/services/SQUEEZE/	Kastritis et al. (2014)
16.	OpenMM	Based on recent graphics processing units, open source	http://docs.openmm.org/7.4.0/userguide/index.html	Eastman et al. (2017)
17.	CHARMMing	Performs molecular dynamics simulations with in-built set of tools	https://www.charmming.org/charmming	Miller et al. (2008)
18.	H++	Calculation of pK values of ionizable groups and addition of H++ according to specific pH	http://biophysics.cs.vt.edu/	Anandakrishnan et al. (2012)
19.	DynOmics	Uses elastic network model	http://gnm.csb.pitt.edu/index.php	Li et al. (2017)

3.5.1.2 Umbrella sampling

Umbrella sampling (Torrie and Valleau, 1977) accelerates conformational sampling by flattening high-energy barriers. In umbrella sampling, artificial "umbrella" potentials mirroring the real barriers are added to flatten the energy landscape. Systematic sampling is done. It has been used for computing binding free energy and studying the dissociation process of ligand—receptor complexes. Umbrella sampling has been used in simulations of G-quadruplex DNA channels (Akhshi and Wu, 2017), troponin C isoforms (Bowman and Lindert, 2018), and the dissociation process of drugs with mitogen-activated protein kinase complex (You et al., 2019).

3.5.1.3 Metadynamics sampling

Metadynamics sampling uses a set of collective variables (CVs) describing the process (Laio, and Parrinello, 2002). The CVs are reaction coordinates accounting for relevant degrees of freedom in binding and unbinding states. Some of the CVs are

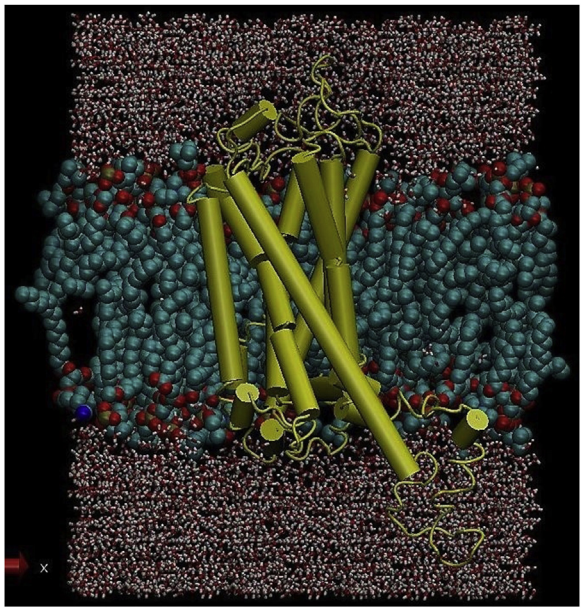


FIGURE 3.5

A typical setup of a molecular dynamics simulation system comprising protein, bilipid layer, and water molecules.

interatomic angles, dihedrals, and distances. After every dynamic step, a small Gaussian-shaped potential is applied to the reaction coordinate. The free energy profile is calculated by summing all the Gaussian's potentials (Sinko et al., 2013; Valsson et al., 2016; Yang et al., 2020,Yang et al., 2019). This technique has application in finding transition states, conformations, prediction of association, and dissociation and calculation of free energy profiles (Cavalli et al., 2015; De Vivo et al., 2016).

3.5.1.4 Targeted molecular dynamics

Targeted molecular dynamics (Schlitter et al., 1994; Ma et al., 2000) induces a conformational change in a target structure at normal temperature by applying a time-dependent geometrical constraint. It uses a moving distance constraint along reaction coordinates to find rare transitions. Targeted molecular dynamics are useful in searching stable intermediates during simulations (Wolf and Stock, 2018; Wolf et al., 2019). It has been used to study the transition between active and inactive structures in beta-adrenergic receptors (Xiao et al., 2015).

3.5.1.5 Parallel tempering method

In the parallel tempering method or replica exchange molecular dynamics (REMD) (Hukushima and Nemoto, 1996; Hansmann, 1997; Sugita and Okamoto, 1999;

García and Sanbonmatsu, 2002), replicas of the system can be parallel simulated at different temperatures to sample conformations. Thus high-energy barriers on the potential energy surface are overcome. REMD is a hybrid method coupling MD simulations with MC simulations. It is useful in calculating equilibrium properties and extracting kinetic information Stelzl and Hummer (2017). Parallel tempering has been used for simulation of human islet amyloid polypeptide (Qi et al., 2018) and folding of the G β-hairpin (Yu et al., 2016).

MD trajectories are analyzed to obtain free energy and kinetics measures. The results are compared with experimental data. Various applications of MD simulations in drug discovery have been reported (Leelananda and Lindert, 2016; Michel, 2014; Yu and MacKerell, 2017; De Vivo et al., 2016).

3.5.2 Monte Carlo simulations

MC methods rely on a random sampling technique to explore conformational space. Identifying drug-binding cavities at the protein−protein interaction (PPI) interface is challenging for designing inhibitors that can disrupt the PPIs (Da Silva et al., 2019). PPI cavities do not exhibit overlap in property with those of protein−drug complexes. Thus identifying a hit ligand is difficult. PPIs are dynamic. Therefore methods based on molecular simulations have been developed. Protein energy landscape exploration (PELE) is an excellent method to explore energy landscapes. PELE is made of structure prediction algorithms combined with MC techniques (Borrelli et al., 2005). It works in three steps: initial perturbation, side-chain sampling, and minimization. In the perturbation step, the protein is perturbed, and the ligand is randomly translated or rotated. In side-chain sampling, the structure is rebuilt using rotamer predictions. MC simulation has been used to identify protein−protein inhibitors in hemagglutinin found on influenza viruses (Diaz et al., 2020). MC techniques have been successfully applied in various drug design projects (Grebner et al., 2017; Kotev et al., 2018; Gilabert et al., 2019; Santiago et al., 2018).

3.6 ADMET profiling

Pharmacokinetics and pharmacodynamics study the effect of drugs on the human body. Pharmacokinetics deals with the absorption, distribution, metabolism, elimination, and toxicity of drugs (ADMET). Pharmacodynamics is the study of the drug's biochemical and physiological effects and helps to study the mechanism of action of the drug. In silico ADMET profiling of hits and leads is done in the early phases of drug discovery to filter out compounds that do not have drug-like properties or good oral bioavailability or that may be toxic. Thus the computational prediction of the ADMET profile of molecules is one of the essential steps in drug design.

A variety of filters for desired pharmacological features and ADMET profiling should be used. There are several criteria used for estimating the ADMET profile of compounds. A rule like the Lipinski rule of five and similar rules enable rapid

screening of compounds. Another extensively used rule,the "Jorgensen rule-of-three," implies that the logS (aqueous solubility) should be greater than −5.7, the cell permeability factor defined by Caco-2 should be faster than 22 nm/s, and the primary metabolites should be less than 7. These hold for the majority of oral drugs. Several online tools are available to efficiently filter compounds against such criteria, such as Qikprop and FAF2. Pan assay interference compounds (Baell and Holloway, 2010) and ALARMNMR (Metz et al., 2007) filters are useful in identifying compounds that are chemically reactive and assay interfering. Various statistical and mathematical models and machine learning algorithms such as neural networks, support vector machines, partial least squares discriminant analysis, and artificial neural networks have been used to develop prediction models.

Quantitative structure−property relationship (QSPR) models such as regression or classification models have also been developed. These QSPR models use the correlation of molecular descriptors to the receptor activity to predict various ADMET profiles. Tools like METEOR (Testa et al., 2005), MetabolExpert (Darvas, 1987) and META (Klopman et al., 1997) use the biotransformation reactions from biochemical and metabolic pathway databases to predict the probable metabolism of a compound. Software for toxicity prediction includes OncoLogic (Benigni et al., 2012), CASE (Saiakhov et al., 2013), TOPKAT (Venkatapathy et al., 2004), HazardExpert Pro (Dearden, 2003), ProTox (Drwal et al., 2014), and the open-source Toxtree (Mombelli and Deviller, 2010). Inverse screening approaches are used to predict any adverse effect resulting from off-target binding. Tools include idTarget (Wang et al., 2012), TarFisDock (Li et al., 2006), INVDOCK (Chen and Ung, 2001), ReverseScreen3D (Kinnings et al., 2011), PharmMapper (Liu et al., 2010), SEA (Keiser et al., 2007), and SwissTargetPrediction (Gfeller et al., 2014).

3.7 Conclusion

SBDD strategies help in the fast and cost-effective design of lead molecules and are an integral component of drug discovery and development projects. Computational methods like docking, HTVS, and de novo drug design are highly effective in screening, designing, and developing lead molecules. Using these rational drug design methods, the number of compounds to be screened in vitro for biological activities reduces considerably. Advances in algorithms have made it possible to efficiently screen in silico ligands having predicted activity comparable to experimentally determined biological activity. Experimental methods like X-ray crystallography usually represent proteins as static structures. Biomolecular simulations are highly useful in studying the dynamic behavior of proteins in native as well as complex forms. Molecular docking methods like IFD and ensemble-based approaches consider the target's flexibility and provide real results. HTVS methods can do a fast screening of sizable small molecule databases against a target to identify potential hits that can be developed into lead molecules. The pharmacokinetic profile of a compound is a major deciding factor in its development as a successful

drug molecule. In silico determination of the ADMET profile of a compound helps mostly in filtering out compounds likely to have an adverse effect on animal studies or preclinical trials.

Despite many advances and successes, CADD has many challenges in enhancing the efficacy of virtual screening methods and designing multitarget drugs and more efficient algorithms and tools to mimic the physiological system.

References

Abagyan, R., Totrov, M., Kuznetzov, D., 1994. ICM − a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J. Comput. Chem. 15, 488−506.

Abel, R., Young, T., Farid, R., Berne, B.J., Friesner, R.A., 2008. The role of the active-site solvent in the thermodynamics of factor Xa ligand binding. J. Am. Chem. Soc. 130, 2817−2831.

Abraham, M.J., Murtola, T., Schulz, R., Pall, S., Smith, J.C., Hess, B., Lindahl, E., 2015. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 19−25.

Abreu, R.M., Froufe, H.J., Daniel, P.O., Queiroz, M.J., Ferreira, I.C., 2011. ChemT, an open-source software for building template-based chemical libraries. SAR QSAR Environ. Res. 22 (5−6), 603−610.

Akhshi, P., Wu, G., 2017. Umbrella sampling molecular dynamics simulations reveal concerted ion movement through G-quadruplex DNA channels. Phys. Chem. Chem. Phys. 19 (18), 11017−11025.

Allen, W.J., Balius, T.E., Mukherjee, S., Brozell, S.R., Moustakas, D.T., Lang, P.T., Case, D.A., Kuntz, I.D., Rizzo, R.C., 2015. DOCK 6: impact of new features and current docking performance. J. Comput. Chem. 36 (15), 1132−1156.

Amaravadhi, H., Baek, K., Yoon, H.S., 2014. Revisiting de novo drug design: receptor based pharmacophore screening. Curr. Top. Med. Chem. 14 (16), 1890−1898.

Amaro, R.E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J.A., Miao, Y., Smith, J.C., 2018. Ensemble docking in drug discovery. Biophys. J. 114 (10), 2271−2278.

Anandakrishnan, R., Aguilar, B., Onufriev, A.V., 2012. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic Acids Res. 40 (Web Server issue), W537−W541.

Andersen, O.J., Risør, M.W., Poulsen, E.C., Nielsen, N.C., Miao, Y.L., Enghild, J.J., et al., 2017. Reactive center loop insertion in α-1-Antitrypsin captured by accelerated molecular dynamics simulation. Biochemistry 56, 634−646.

Apostolakis, J., Pluckthun, A., Caflisch, A., 1998. Docking small ligands in flexible binding sites. J. Comput. Chem. 19, 21−37.

Baell, J.B., Holloway, G.A., 2010. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J. Med. Chem. 53 (7), 2719−2740.

Bakan, A., Meireles, L.M., Bahar, 2011. I ProDy: protein dynamics inferred from theory and experiments. Bioinformatics 27 (11), 1575−1577.

Banitt, I., Wolfson, H.J., 2011. ParaDock: a flexible non-specific DNA–rigid protein docking algorithm. Nucleic Acids Res. 39 (20), e135.

Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., Mason, J.S., 2007. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. J. Chem. Inf. Model. 47 (2), 279−294.

Barreiro, G., Kim, J.T., Guimarães, C.R., Bailey, C.M., Domaoal, R.A., Wang, L., Anderson, K.S., Jorgensen, W.L., 2007. From docking false-positive to active anti-HIV agent. J. Med. Chem. 50 (22), 5324−5329.

Baumgartner, M.P., Evans, D.A., 2018. Lessons learned in induced fit docking and metadynamics in the drug design data resource grand challenge 2. J. Comput. Aided Mol. Des. 32 (1), 45−58.

Baxter, C.A., Murray, C.W., Clark, D.E., Westhead, D.R., Eldridge, M.D., 1998. Flexible docking using Tabu search and an empirical estimate of binding affinity. Proteins 33, 367−382.

Benigni, R., Bossa, C., Alivernini, S., Colafranceschi, M., 2012. Assessment and validation of US EPA's OncoLogicVR expert system and analysis of its modulating factors for structural alerts. J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev. 30, 152−173.

Böhm, H.J., 1992. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J. Comput. Aided Mol. Des. 6, 61−78.

Böhm, H.J., 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J. Comput. Aided Mol. Des. 8, 243−256.

Böhm, H.J., 1998. Prediction of binding constants of ptotein ligands: a fast method for the polarization of hits obtained from de novo design or 3D database search programs. J. Comput. Aided Mol. Des. 12, 309−323.

Bolia, A., Ozkan, S.B., 2016. Adaptive BP-dock: an induced fit docking approach for full receptor flexibility. J. Chem. Inf. Model. 56 (4), 734−746.

Borrelli, K.W., Vitalis, A., Alcantara, R., Guallar, V., 2005. PELE: protein energy landscape exploration. A novel monte carlo based technique. J. Chem. Theor. Comput. 1 (6), 1304−1311.

Bowman, J.D., Lindert, S., 2018. Molecular dynamics and umbrella sampling simulations elucidate differences in troponin C isoform and mutant hydrophobic patch exposure. J. Phys. Chem. B 122 (32), 7874−7883.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M., 1983. CHARMM − a programm for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4, 187−217.

Brylinski, M., 2003. Nonlinear scoring functions for similarity-based ligand docking and binding affinity prediction. J. Chem. Inf. Model. 53 (11), 3097−3112.

Bucher, D., Pierce, L.C., McCammon, J.A., Markwick, P.R., 2011. On the use of accelerated molecular dynamics to enhance configurational sampling in ab initio simulations. J. Chem. Theor. Comput. 7, 890−897.

Bucher, D., Stouten, P., Triballeau, N., 2018. Shedding light on important waters for drug design: simulations versus grid-based methods. J. Chem. Inf. Model. 58 (3), 692−699.

Carlson, H.A., 2002. Protein flexibility is an important component of structure-based drug discovery. Curr. Pharmaceut. Des. 8, 1571−1578, 2002.

Carlson, H.A., McCammon, J.A., 2000. Accommodating protein flexibility in computational drug design. Mol. Pharmacol. 57, 213−218.

Cavalli, A., Spitaleri, A., Saladino, G., Gervasio, F.L., 2015. Investigating drug−target association and dissociation mechanisms using metadynamics-based algorithms. Acc. Chem. Res. 48 (2), 277−285.

Chao, W.R., Yean, D., Amin, K., Green, C., Jong, L., 2007. Computer-aided rational drug design: a novel agent (SR13668) designed to mimic the unique anticancer mechanisms of dietary indole-3-carbinol to block Akt signaling. J. Med. Chem. 50 (15), 3412−3415.

Chen, Y.Z., Ung, C.Y., 2001. Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. J. Mol. Graph. Model. 20 (3), 199−218.

Chen, R., Tong, W., Mintseris, J., Li, L., Weng, Z., 2003. ZDOCK predictions for the CAPRI challenge. Proteins 52, 68−73.

Chen, K., Li, T., Cao, T., 2006. Tribe-PSO: a novel global optimization algorithm and its application in molecular docking. Chemometr. Intell. Lab. Syst. 82, 248−259.

Chen, H.M., Liu, B.F., Huang, H.L., Hwang, S.F., Ho, S.Y., 2007. SODOCK: swarm optimization for highly flexible protein-ligand docking. J. Comput. Chem. 28, 6 12−623.

Cheng, T.M., Blundell, T.L., Fernandez-Recio, J., 2007. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 68 (2), 503−515.

Clark, K.P., 1995. Flexible ligand docking without parameter adjust-ment across four ligand-receptor complexes. J. Comput. Chem. 16, 12 10−1226.

Clark, D.E., 2006. What has computer-aided molecular design ever done for drug discovery? Expet Opin. Drug Discov. 1 (2), 103−110.

Clark, A.J., Tiwary, P., Borrelli, K., Feng, S., Miller, E.B., Abel, R., Friesner, R.A., Berne, B.J., 2016. Prediction of protein-ligand binding poses via a combination of induced fit docking and metadynamics simulations. J. Chem. Theor. Comput. 12 (6), 2990−2998.

Cortes Cabrera, A., Klett, J., Dos Santos, H.G., Perona, A., Gil-Redondo, R., Francis, S.M., Priegos, E.M., Gago, F., Morreale, A., 2012. CRDOCK: an ultrafast multipurpose protein−ligand docking tool. J. Chem. Inf. Model. 52 (8), 2300−2309.

Cozzini, P., Kellogg, G.E., Spyrakis, F., Abraham, D.J., Costantino, G., Emerson, A., Fanelli, F., Gohlke, H., Kuhn, L.A., Morris, G.M., Orozco, M., Pertinhez, T.A., Rizzi, M., Sotriffer, C.A., 2008. Target flexibility: an emerging consideration in drug discovery and design. J. Med. Chem. 51 (20), 6237−6255.

Cuzzolin, A., Deganutti, G., Salmaso, V., Sturlese, M., Moro, S., 2018. AquaMMapS: an alternative tool to monitor the role of water molecules during protein-ligand association. ChemMedChem 13 (6), 522−531.

Da Silva, F., Bret, G., Teixeira, L., Gonzalez, C.F., Rognan, D., 2019. Exhaustive repertoire of druggable cavities at protein-protein interfaces of known three-dimensional structure. J. Med. Chem. 62 (21), 9732−9742.

Dallakyan, S., Olson, A.J., 2015. Small-molecule library screening by docking with PyRx. Methods Mol. Biol. 1263, 243−250.

Damm-Ganamet, K.L., Arora, N., Becart, S., Edwards, J.P., Lebsack, A.D., McAllister, H.M., Nelen, M.I., Rao, N.L., Westover, L., Wiener, J.J.M., Mirzadegan, T., 2019. Accelerating lead identification by high throughput virtual screening: prospective case studies from the pharmaceutical industry. J. Chem. Inf. Model. 59 (5), 2046−2062.

Darvas, F., 1987. Metabolexpert: an expert system for predicting metabolism of substances. QSAR Environ Toxicol 71−81.

Davis, I.W., Baker, D., 2009. RosettaLigand docking with full ligand and receptor flexibility. J. Mol. Biol. 385 (2), 381−392.

De Paris, R., Vahl Quevedo, C., Ruiz, D.D., Gargano, F., de Souza, O.N., 2018. A selective method for optimizing ensemble docking-based experiments on an InhA Fully-Flexible receptor model. BMC Bioinformatics 19 (1), 235.

De Vivo, M., Masetti, M., Bottegoni, G., Cavalli, A., 2016. Role of molecular dynamics and related methods in drug discovery. J. Med. Chem. 59 (9), 4035−4061.

de Vries, S.J., Schindler, C.E., Chauvot de Beauchêne, I., Zacharias, M.A., 2015. Web interface for easy flexible protein-protein docking with ATTRACT. Biophys. J. 108 (3), 462−465.

Dearden, J.C., 2003. In silico prediction of drug toxicity. J. Comput. Aided Mol. Des. 17, 119−127.

DeWitte, R.S., Shakhnovich, E.I., 1996. SMoG: de Novo design method based on simple, fast, and accutate free energy estimate. 1. Methodology and supporting evidence. J. Am. Chem. Soc. 118, 11733−11744.

Díaz, L., Soler, D., Tresadern, G., Buyck, C., Pérez-Benito, L., Saen-oon, S., Guallar, V., Soliva, R., 2020. Monte Carlo simulations using PELE to identify a protein−protein inhibitor binding site and pose. RSC Adv. 10, 7058−7064.

Diller, D.J., Merz Jr., K.M., 2001. High throughput docking for library design and library prioritization. Proteins 43 (2), 113−124.

Doppelt-Azeroual, O., Delfaud, F., Moriaud, F., de Brevern, A.G., 2010. Fast and automated functional classification with MED-SuMo: an application on purine-binding proteins. Protein Sci. 19 (4), 847−867.

Douguet, D., Munier-Lehmann, H., Labesse, G., Pochet, S., 2005. LEA3D: a computer-aided ligand design for structure-based drug design. J. Med. Chem. 48 (7), 2457−2468.

Drwal, M.N., Banerjee, P., Dunkel, M., Wettig, M.R., Preissner, R., 2014. ProTox: a web server for the in silico prediction of rodent oral toxicity. Nucleic Acids Res. 42. W53−8.

Duan, L., Guo, X., Cong, Y., Feng, G., Li, Y., Zhang, J., 2019. Accelerated molecular dynamics simulation for helical proteins folding in explicit water. Front. Chem. 7, 540.

Durrant, J.D., McCammon, J.A., 2011. Molecular dynamics simulations and drug discovery. BMC Biol. 9, 71.

Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.P., Simmonett, A.C., Harrigan, M.P., Stern, C.D., Wiewiora, R.P., Brooks, B.R., Pande, V.S., 2017. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. PLOS Comp. Biol. 13 (7), e1005659.

Echartea, M.E., Beauchêne, I.C., Ritchie, D.W., 2019. EROS-DOCK: protein−protein docking using exhaustive branch-and-bound rotational search. Bioinformatics 35 (23), 5003−5010.

Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V., Mee, R.P., 1997. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J. Comput. Aided Mol. Des. 11, 425−445.

Epik, Schrödinger, 2020. LLC, New York, NY.

Esquivel-Rodriguez, J., Filos-Gonzalez, V., Li, B., Kihara, D., 2014. Pairwise and multimeric protein-protein docking using the LZerD program suite. Methods Mol. Biol. 1137, 209−234.

Evangelista, W., Weir, R.L., Ellingson, S.R., Harris, J.B., Kapoor, K., Smith, J.C., Baudry, J., 2016. Ensemble-based docking: from hit discovery to metabolism and toxicity predictions. Bioorg. Med. Chem. 24 (20), 4928−4935.

Ewing, T.J.A., Kuntz, I.D., 1997. Critical evaluation of search algorithms for automated molecular docking and database screening. J. Comput. Chem. 18, 1175−1189.

Faller, C.E., Raman, E.P., MacKerell Jr., A.D., Guvench, O., 2015. Site identification by ligand competitive saturation (SILCS) simulations for fragment-based drug design. Methods Mol. Biol. 1289, 75−87.

Fang, Y., Ding, Y., Feinstein, W.P., Koppelman, D.M., Moreno, J., et al., 2016. GeauxDock: accelerating structure-based virtual screening with heterogeneous computing. PloS One 11 (7), e0158898.

Ferrari, A.M., Wei, B.Q., Costantino, L., Shoichet, B.K., 2004. Soft docking and multiple receptor conformations in virtual screening. J. Med. Chem. 47, 5076−5084.

Ferrer, R.S., Case, D.A., Walker, R.C., 2013. An overview of the Amber biomolecular simulation package. WIREs Comput. Mol. Sci. 3, 198−210.

Fischer, T., Gazzola, S., Riedl, R., 2019. Approaching target selectivity by de novo drug design. Expet Opin. Drug Discov. 14 (8), 791−803.

Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shaw, D.E., Shelley, M., Perry, J.K., Francis, P., Shenkin, P.S., 2004. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 47, 1739−1749.

Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., Greenwood, J.R., Halgren, T.A., Sanschagrin, P.C., Mainz, D.T., 2006. Extra precision Glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J. Med. Chem. 49, 6177−6196.

Fu, D.Y., Meiler, J., 2018. RosettaLigandEnsemble: a small-molecule ensemble-driven docking approach. ACS Omega 3 (4), 3655−3664.

García, A.E., Sanbonmatsu, K.Y., 2002. Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. Proc. Natl. Acad. Sci. U. S. A. 99, 2782−2787.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M.P., Overington, J.P., Papadatos, G., Smit, I., Leach, A.R., 2017. The ChEMBL database in 2017. Nucleic Acids Res. 45 (D1), D945−D954.

Gehlhaar, D.K., Verkhivker, G.M., Rejto, P.A., Sherman, C.J., Fogel, D.B., Freer, S.T., 1995. Molecular recognition of the inhibitor AG-1343 by HIV-1 Protease: conformationally flexible docking by evolutionary programming. Chem. Biol. 2, 3, 17-324.

Gfeller, D., Grosdidier, A., Wirth, M., Daina, A., Michielin, O., Zoete, V., 2014. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucleic Acids Res. 42 (Web Server issue), W32−W38.

Gil-Redondo, R., Estrada, J., Morreale, A., Herranz, F., Sancho, J., Ortiz, A.R., 2009. VSDMIP: virtual screening data management on an integrated platform. J. Comput. Aided Mol. Des. 23 (3), 171−184.

Gilabert, J.F., Grebner, C., Soler, D., Lecina, D., Municoy, M., Gracia Carmona, O., Soliva, R., Packer, M.J., Hughes, S.J., Tyrchan, C., Hogner, A., Guallar, V., 2019. PELE-MSM: a Monte Carlo based protocol for the estimation of absolute binding free energies. J. Chem. Theor. Comput. 15 (11), 6243−6253.

Gillet, V.J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z., Johnson, P., 1994. SPROUT: recent developments in the de novo design of molecules. J. Chem. Inf. Comput. Sci. 34 (1), 207−217.

Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A., Thornton, J.M., 2006. A method for localizing ligand binding pockets in protein structures. Proteins 62 (2), 479−488.

Gohlke, H., Hendlich, M., Klebe, G., 2000. Knowledge-based scoring function to predict protein-ligand interactions. J. Mol. Biol. 295, 337−356.

Gong, Z., Hu, G., Li, Q., Liu, Z., Wang, F., Zhang, X., Xiong, J., Li, P., Xu, Y., Ma, R., Chen, S., Li, J., 2017. Compound libraries: recent advances and their applications in drug discovery. Curr. Drug Discov. Technol. 14 (4), 216−228.

Goto, J., Kataoka, R., Hirayama, N., 2004. Ph4Dock: pharmacophore-based protein-ligand docking. J. Med. Chem. 47 (27), 6804−6811.

Grant, J.A., Pickup, B.T., Nicholls, A., 2001. A smooth permittivity function for Poisson-Boltzmann solvation methods. J. Comput. Chem. 22, 608−640.

Grebner, C., Lecina, D., Gil, V., Ulander, J., Hansson, P., Dellsen, A., Tyrchan, C., Edman, K., Hogner, A., Guallar, V., 2017. Exploring binding mechanisms in nuclear hormone receptors by Monte Carlo and X-ray-derived motions. Biophys. J. 112 (6), 1147−1156.

Greenwood, J.R., Calkins, D., Sullivan, A.P., Shelley, J.C., 2010. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. J. Comput. Aided Mol. Des. 24 (6−7), 591−604.

Grosdidier, A., Zoete, V., Michielin, O., 2007. EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. Proteins 67, 10 10−1025.

Grosdidier, A., Zoete, V., Michielin, O., 2011. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res. 39.

Halgren, T., 2007. New method for fast and accurate binding-site identification and analysis. Chem. Biol. Drug Des. 69, 146−148.

Halgren, T., 2009. Identifying and characterizing binding sites and assessing druggability. J. Chem. Inf. Model. 49, 377−389.

Halgren, T.A., Murphy, R.B., Friesner, R.A., Beard, H.S., Frye, L.L., Pollard, W.T., Banks, J.L., 2004. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J. Med. Chem. 47, 1750−1759.

Hamelberg, D., McCammon, J.A., 2005. Fast peptidyl cis-trans isomerization within the flexible gly-rich flaps of HIV-1 protease. J. Am. Chem. Soc. 127, 13778−13779. https://doi.org/10.1021/ja054338a.

Hamelberg, D., Mongan, J., McCammon, J.A., 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J. Chem. Phys. 120, 11919−11929. https://doi.org/10.1063/1.1755656.

Hansmann, U.H., 1997. Parallel tempering algorithm for conformational studies of biological molecules. Chem. Phys. Lett. 281, 140−150.

Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J.Y., Wang, L., Lupyan, D., Dahlgren, M.K., Knight, J.L., Kaus, J.W., Cerutti, D.S., Krilov, G., Jorgensen, W.L., Abel, R., Friesner, R.A., 2016. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. J. Chem. Theor. Comput. 12 (1), 281−296.

Hart, T.N., Read, R.J., 1992. A multiple-start Monte Carlo docking method. Proteins 13, 206−222.

Hartenfeller, M., Schneider, G., 2011. De novo drug design. Methods Mol. Biol. 672, 299−323.

Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., Schneider, G., 2012. DOGS: reaction-driven de novo design of bioactive compounds. PLoS Comput. Biol. 8 (2), e1002380.

Hazuda, D.J., Anthony, N.J., Gomez, R.P., Jolly, S.M., Wai, J.S., Zhuang, L., Fisher, T.E., Embrey, M., Guare Jr., J.P., Egbertson, M.S., Vacca, J.P., Huff, J.R., Felock, P.J., Witmer, M.V., Stillmock, K.A., Danovich, R., Grobler, J., Miller, M.D., Espeseth, A.S., Jin, L., Chen, I.W., Lin, J.H., Kassahun, K., Ellis, J.D., Wong, B.K., Xu, W., Pearson, P.G., Schleif, W.A., Cortese, R., Emini, E., Summa, V., Holloway, M.K., Young, S.D., 2004. A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. Proc. Natl. Acad. Sci. U. S. A 101 (31), 11233−11238.

Heo, L., Shin, W.H., Lee, M.S., Seok, C., 2014. GalaxySite: ligand-binding-site prediction by using molecular docking. Nucleic Acids Res. 42.

Hernandez, M., Ghersi, D., Sanchez, R., 2009. SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res. 17, W413−W416.

Hogues, H., Gaudreault, F., Corbeil, C.R., Deprez, C., Sulea, T., Purisima, E.O., 2018. PrоPOSE: direct exhaustive protein-protein docking with side chain flexibility. J. Chem. Theor. Comput. 14 (9), 4938−4947, 2018.

Hospital, A., Andrio, P., Fenollosa, C., Cicin-Sain, D., Orozco, M., Gelpí, J.L., 2012. MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. Bioinformatics 28 (9), 1278−1279.

Huang, B., Schroeder, M., 2006. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct. Biol. 6, 19.

Huang, S.Y., Zou, X., 2007. Efficient molecular docking of NMR structures: application to HIV-1 protease. Protein Sci. 16, 43−45, 2007, 1.

Huang, S.Y., Zou, X., 2007. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. Proteins 66, 399−421, 2007.

Huang, S.Y., Zou, X., 2010. Mean-force scoring functions for protein-ligand binding. Annu. Rep. Comput. Chem. 6, 281−296.

Hukushima, K., Nemoto, K., 1996. Exchange Monte Carlo method and application to spin glass simulations. J. Phys. Soc. Jpn. 65, 1604−1608.

Hurwitz, N., Duhovny, D.S., Wolfson, H.J., 2016. Memdock: an α-helical membrane protein docking algorithm. Bioinformatics 32 (16), 2444−2450.

Inbar, Y., Benyamini, H., Nussinov, R., Wolfson, H.J., 2003. Protein structure prediction via combinatorial assembly of sub-structural units. Bioinformatics 19.

Ishchenko, A.V., Shakhnovich, E.I., 2002. Small molecule growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions. J. Med. Chem. 45, 2770−2780.

Jackson, R.M., 2002. Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. J. Comput. Aided Mol. Des. 16 (1), 43−57.

Jain, A.N., 2003. Surflex: fully automatic molecular docking using a molecular similarity-based search engine. J. Med. Chem. 46, 499−511.

Jain, T., Jayaram, B., 2007. A computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes. Proteins: Struct. Funct. Bioinfo. 67, 1167−1178.

Jayaraj, P.B., Jain, S., 2019. Ligand based virtual screening using SVM on GPU. Comput. Biol. Chem. 83, 107143.

Jendele, L., Krivak, R., Skoda, P., Novotny, M., Hoksza, D., 2019. PrankWeb: a web server for ligand binding site prediction and visualization. Nucleic Acids Res. 47 (W1), W345−W349.

Jiang, F., Kim, S.H., 1991. Soft docking: matching of molecular surface cubes. J. Mol. Biol. 219, 79−102.

Jiménez-García, B., Roel-Touris, J., Romero-Durana, M., Vidal, M., Jiménez-González, Juan Fernández-Recio, D., 2018. LightDock: a new multi-scale approach to protein−protein docking. Bioinformatics 34 (1), 49−55.

Jimmy, C., Ngai, F., Mak, P.I., Siu, S.W., 2016. ProtPOS: a Python package for the prediction of protein preferred orientation on a surface. Bioinformatics 32, 2537−2538.

Jones, G., Willett, P., Glen, R.C., 1995. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. 245, 43−53.

Jones, G., Willett, P., Glen, R.C., Leach, A.R., Taylor, R., 1997. Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. 267, 727-7.

Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D., Altman, R.B., 2009. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA 15 (2), 189−199.

Jorgensen, W.L., 2009. Efficient drug lead discovery and optimization. Acc. Chem. Res. 42 (6), 724−733.

Jorgensen, W.L., Maxwell, D.S., Tirado-Rives, J., 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. 118, 11225−11236.

Karthick, V., Nagasundaram, N., Doss, C.G.P., Chakraborty, C., Siva, R., Lu, A., Zhang, G., Zhu, H., 2016. Infect. Dis. Poverty 5 (12).

Kastritis, P.L., Rodrigues, J.P., Bonvin, A.M., 2014. HADDOCK(2P2I): a biophysical model for predicting the binding affinity of protein-protein interaction inhibitors. J. Chem. Inf. Model. 54 (3), 826−836.

Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., Shoichet, B.K., 2007. Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. 25 (2), 197−206.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E., 2019. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 47 (D1), D1102−D1109.

Kinnings, S.L., Jackson, R.M., 2011. ReverseScreen3D: a structure based ligand matching method to identify protein targets. J. Chem. Inf. Model. 51, 624−634.

Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J., 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discov. 3 (11), 935−949.

Klopman, G., Tu, M., Talafous, J., 1997. META. 3. A genetic algorithm for metabolic transform priorities optimization. J. Chem. Inf. Comput. Sci. 37 (2), 329−334.

Kotev, M., Manuel-Manresa, P., Hernando, E., Soto-Cerrato, V., Orozco, M., Quesada, R., Pérez-Tomás, R., Guallar, V., 2018. Inhibition of human enhancer of zeste homolog 2 with tambjamine analogs. J. Chem. Inf. Model. 57 (8), 2089−2098.

Kovalenko, A., 2003. Three-dimensional RISM theory for molecular liquids and solid-liquid interfaces. In: Hirata, F. (Ed.), Molecular Theory of Solvation. In: Mezey, P.G. (Ed.), Series: Understanding Chemical Reactivity, vol. 24. Kluwer Academic Publishers, Dordrecht, pp. 169−275 vol. 360.

Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D., Yueh, C., Beglov, D., Vajda, S., 2017. The ClusPro web server for protein-protein docking. Nat. Protoc. 12 (2), 255−278.

Krammer, A., Kirchhoff, P.D., Jiang, X., Venkatachalam, C.M., Waldman, M., 2005. LigScore: a novel scoring function for predicting binding affinities. J. Mol. Graph. Model. 23, 395−407.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E., 1982. A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. 161 (2), 269−288.

Lagardère, L., Jolly, L.H., Lipparini, F., Aviat, F., Stamm, B., Jing, Z.F., Harger, M., Torabifard, H., Cisneros, G.A., Schnieders, M.J., Gresh, N., Maday, Y., Ren, P.Y., Ponder, J.W., Piquemal, J.P., 2018. Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced polarizable force fields. Chem. Sci. 9, 956−972.

Laio, A., Parrinello, M., 2002. Escaping free-energy minima. Proc. Natl. Acad. Sci. USA 99 (20), 12562−12566.

Laurie, A., Jackson, R., 2005. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21 (9), 1908−1916.

le Guilloux, V., Schmidtke, P., Tuffery, P., 2009. Fpocket: an open source platform for ligand pocket detection. BMC Bioinf. 10, 168.

Leelananda, S.P., Lindert, S., 2016. Computational methods in drug discovery. Beilstein J. Organic Chem. 12, 2694−2718.

Lesk, V.I., Sternberg, M.J.E., 2008. 3D-Garden: a system for modelling protein-protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm. Bioinformatics 24 (9), 1137−1144.

Lexa, K.W., Carlson, H.A., 2011. Full protein flexibility is essential for proper hot-spot mapping. J. Am. Chem. Soc. 133 (2), 200−202.

Li, H., Li, C., Gui, C., Luo, X., Chen, K., Shen, J., Wang, X., Jiang, H., 2004. GAsDock: a new approach for rapid flexible docking based on an improved multi-population genetic algorithm. Bioorg. Med. Chem. Lett 14 (18), 4671−4676.

Li, H., Gao, Z., Kang, L., Zhang, H., Yang, K., Yu, K., Luo, X., Zhu, W., Chen, K., Shen, J., Wang, X., Jiang, H., 2006. TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res. 34 (Web Server issue), W219−W224.

Li, H., Leung, K., Chan, C.H., Cheung, H.L., Wong, M., 2014. iSyn: WebGL-based interactive de novo drug design. In: 2014 18th International Conference on Information Visualisation, Paris, pp. 302−307.

Li, Y., Zhao, Z., Liu, Z., Wang, R., 2016. AutoT&T v2: an efficient and versatile tool for lead structure generation and optimization. J. Chem. Inf. Model. 56, 435−453.

Li, H., Chang, Y.Y., Lee, J.Y., Bahar, I., Yang, L.W., 2017a. DynOmics: dynamics of structural proteome and beyond. Nucleic Acids Res. 45, W374−W380.

Li, L., Jia, Z., Peng, Y., Chakravorty, A., Sun, L., Alexov, E., 2017b. DelPhiForce web server: electrostatic forces and energy calculations and visualization. Bioinformatics.15 33 (22), 3661−3663.

Li, Y., Zhang, L., Liu, Z., 2018. Multi-objective de novo drug design with conditional graph generative model. J. Cheminf. 10, 33.

Li, X., Zhang, X.X., Lin, Y.X., Xu, X.M., Li, L., Yang, J.B., 2019. Virtual screening based on ensemble docking targeting wild-type p53 for anticancer drug discovery. Chem. Biodivers. 16 (7), e1900170.

LigPrep, Schrödinger, 2020. LLC, New York, NY.

Lim, S.V., Rahman, M.B.A., Tejo, B.A., 2011. Structure-based and ligand-based virtual screening of novel methyltransferase inhibitors of the dengue virus. BMC Bioinf. 12, S24.

Lindert, S., Durrant, J.D., McCammon, J.A., 2012. LigMerge: a fast algorithm to generate models of novel potential ligands from sets of known binders. Chem. Biol. Drug Des. 80 (3), 358−365.

Liu, M., Wang, S., 1999. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. J. Comput. Aided Mol. Des. 13, 435−445, 1.

Liu, H.Y., Kuntz, I.D., Zou, X., 2004. Pairwise GB/SA scoring function for structure-based drug design. J. Phys. Chem. B 108, 5453−5462.

Liu, X., Ouyang, S., Yu, B., Liu, Y., Huang, K., Gong, J., Zheng, S., Li, Z., Li, H., Jiang, H., 2010. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res. 38 (Web Server issue), W609−W614.

Liu, Y., Zhao, L., Li, W., Zhao, D., Song, M., Yang, Y., 2013. FIPSDock: a new molecular docking technique driven by fully informed swarm optimization algorithm. J. Comput. Chem. 34 (1), 67−75.

Lyskov, S., Chou, F.C., Conchúir, S.Ó., Der, B.S., Drew, K., Kuroda, D., Xu, J., Weitzner, B.D., Renfrew, P.D., Sripakdeevong, P., Borgo, B., Havranek, J.J., Kuhlman, B., Kortemme, T., Bonneau, R., Gray, J.J., Das, R., 2013. Serverification of molecular modeling applications: the rosetta online server that includes everyone (ROSIE). PloS One 8 (5), e63906.

Ma, J.P., Sigler, P.B., Xu, Z.H., Karplus, M., 2000. A dynamic model for the allosteric mechanism of GroEL1. J. Mol. Biol. 302, 303−313.

Maia, E.H., Campos, V.A., Dos Reis Santos, B., Costa, M.S., Lima, I.G., Greco, S.J., Ribeiro, R.I., Munayer, F.M., da Silva, A.M., Taranto, A.G., 2017. Octopus: a platform for the virtual high-throughput screening of a pool of compounds against a set of molecular targets. J. Mol. Model. 23 (1), 26.

Maia, E.H.B., Medaglia, L.R., da Silva, A.M., Taranto, A.G., 2020. Molecular architect: a user-friendly workflow for virtual screening. ACS Omega 5 (12), 6628−6640.

Margreitter, C., Petrov, D., Zagrovic, B., 2013. Vienna-PTM webserver: a toolkit for MD simulations of protein post-translational modifications. Nucleic Acids Res. 41, W422−W442.

Marialke, J., Korner, R., Tietze, S., Apostolakis, J., 2007. Graph-based molecular alignment (GMA). J. Chem. Inform. Model. 47 (2), 591−601.

Markwick, P.R., Bouvignies, G., Blackledge, M., 2007. Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. J. Am. Chem. Soc. 129, 4724−4730.

Martins-José, A., 2013. Sliding Box Docking: a new stand-alone tool for managing docking-based virtual screening along the DNA helix axis. Bioinformation 9 (14), 750−751.

Mashiach, E., Schneidman-Duhovny, D., Andrusier, N., Nussinov, R., Wolfson, H.J., 2008. FireDock: a web server for fast interaction refinement in molecular docking. Nucleic Acids Res. 36.

Mashiach, E., Nussinov, R., Wolfson, H.J., 2010. FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. Nucleic Acids Res. 38 (Suppl. l), W457−W461.

Mashiach-Farkash, E., Nussinov, R., Wolfson, H.J., 2011. SymmRef: a flexible refinement method for symmetric multimers. Proteins 79 (9), 2607−2623.

Matthey, T., Cickovski, T., Hampton, S.S., Ko, A., Ma, Q., Nyerges, M., Raeder, T., Slabach, T., Izaguirre, J.A., 2004. ProtoMol: an object-oriented framework for prototyping novel algorithms for molecular dynamics. ACM Trans. Math Software 30 (3), 237−265.

McGann, M.R., Almond, H.R., Nicholls, A., Grant, J.A., Brown, F.K., 2003. Gaussian docking functions. Biopolymers 68 (1), 76−90.

Meiler, J., Baker, D., 2006. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins 65 (3), 538−548.

Meng, X.Y., Zhang, H.X., Mezei, M., Cui, M., 2011. Molecular docking: a powerful approach for structure-based drug discovery. Curr. Comput. Aided Drug Des. 7 (2), 146−157.

Metz, J.T., Huth, J.R., Hajduk, P.J., 2007. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. J. Comput. Aided Mol. Des. 21 (1−3), 139−144.

Miao, Y., Feixas, F., Eun, C., McCammon, J.A., 2015. Accelerated molecular dynamics simulations of protein folding. J. Comput. Chem. 36, 1536−1549. https://doi.org/10.1002/jcc.23964.

Miao, Y., Nichols, S.E., McCammon, J.A., 2014. Free energy landscape of G-protein coupled receptors, explored by accelerated molecular dynamics. Phys. Chem. Chem. Phys. 16 (14), 6398−6406.

Michel, J., 2014. Current and emerging opportunities for molecular simulations in structure-based drug design. Phys. Chem. Chem. Phys. 16 (10), 4465−4477.
Michel, J., Tirado-Rives, J., Jorgensen, W.L., 2009. Prediction of the water content in protein binding sites. J. Phys. Chem. B 113 (40), 13337−13346.
Miller, M.D., Kearsley, S.K., Underwood, D.J., Sheridan, R.P., 1994. FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. J. Comput. Aided Mol. Des. 8 (2), 153−174.
Miller, B.T., Singh, R.P., Klauda, J.B., Hodošček, M., Brooks, B.R., Woodcock, H.L., 2008. CHARMMing: a new, flexible web portal for CHARMM. J. Chem. Inf. Model. 48 (9), 1920−1929.
Mitchell, J.B.O., Laskowski, R.A., Alex, A., Thornton, J.M., 1999. Bleep − potential of mean force describing protein-ligand interactions: I. Generating potential. J. Comput. Chem. 20, 1165−1176.
Mizutani, M.Y., Tomioka, N., Itai, A., 1994. Rational automatic search method for stable docking models of protein and ligand. J. Mol. Biol. 243, 3 10−326.
Mombelli, E., Devillers, J., 2010. Evaluation of the OECD (Q)SAR Application Toolbox and Toxtree for predicting and profiling the carcinogenic potential of chemicals. SAR QSAR Environ. Res. 21, 731−752.
Mooij, W.T., Verdonk, M.L., 2005. General and targeted statistical potentials for protein-ligand interactions. Proteins 61, 272−287.
Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J., 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comput. Chem. 19, 1639−1662.
Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J., 2009. Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J. Comput. Chem. 16, 2785−2791.
Mukherjee, S., Kar, R.K., Nanga, R.P.R., Mroue, K.H., Ramamoorthy, A., Bhunia, A., 2017. Accelerated molecular dynamics simulation analysis of MSI-594 in a lipid bilayer. Phys. Chem. Chem. Phys. 19, 19289−19299.
Namasivayam, V., Gunther, R., 2007. PSO@Autodock: a fast flexible molecular docking program based on swarm intelligence. Chem. Biol. Drug Des. 70, 475−484.
Nejad, M.A., Urbassek, H.M., 2018. Insulin adsorption on functionalized silica surfaces: an accelerated molecular dynamics study. J. Mol. Model. 24, 89.
Ngan, C.H., Bohnuud, T., Mottarella, S.E., Beglov, D., Villar, E.A., Hall, D.R., Kozakov, D., Vajda, S., 2012. FTMAP: extended proteinmapping with user-selected probe molecules. Nucleic Acids Res. 40 (Web Server issue), W271−W275.
Niinivehmas, S.P., Salokas, K., Lätti, S., Raunio, H., Pentikäinen, O.T., 2015. Ultrafast protein structure-based virtual screening with Panther. J. Computer-aided Molecular Discovery 29 (10), 989−1006.
O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: an open chemical toolbox. J. Cheminf. 3, 33.
Ohue, M., Shimoda, T., Suzuki, S., Matsuzaki, Y., Ishida, T., Akiyama, Y., 2014. MEGADOCK 4.0: an ultra-high-performance protein-protein docking software for heterogeneous supercomputers. Bioinformatics 30 (22), 3281−3283.
Olsson, M.H.M., Sondergaard, C.R., Rostkowski, M., Jensen, J.H., 2011. PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. J. Chem. Theor. Comput. 7 (2), 525−537.

Pagadala, N.S., Syed, K., Tuszynski, J., 2017. Software for molecular docking: a review. Biophys. Rev. 9 (2), 91−102.

Pang, Y.P., Perola, E., Xu, K., Prendergast, F.G., 2001. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. J. Comput. Chem. 22 (15), 1750−1771.

Paul, D.S., Gautham, N., 2017. iMOLSDOCK: induced-fit docking using mutually orthogonal Latin squares (MOLS). J. Mol. Graph. Model. 74, 89−99.

Pei, J., Wang, Q., Liu, Z., Li, Q., Yang, K.L., Lai, L., 2006. PSI-DOCK: towards highly efficient and accurate flexible ligand docking. Proteins 62, 934−946.

Pencheva, T., Lagorce, D., Pajeva, I., Villoutreix, B.O., Miteva, M.A., 2008. AMMOS: automated molecular mechanics optimization tool for in silico screening. BMC Bioinf. 9, 438.

Perez, C., Soler, D., Soliva, R., Guallar, V., 2020. FragPELE: dynamic ligand growing within a binding site. A novel tool for hit-to-lead drug design. J. Chem. Inf. Model. 60 (3), 1728−1736.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., Schulten, K., 2005. J. Comput. Chem. 26, 1781−1802.

Pierce, L.C., Salomon-Ferrer, R., Augusto, F., de Oliveira, C., McCammon, J.A., Walker, R.C., 2012. Routine access to millisecond time scale events with accelerated molecular dynamics. J. Chem. Theor. Comput. 8, 2997−3002.

Popova, M., Isayev, O., Tropsha, A., 2018. Deep reinforcement learning for de novo drug design. Sci. Adv. 4 (7), eaap7885.

Prakhov, N.D., Chernorudskiy, A.L., Gainullin, M.R., 2010. VSDocker: a tool for parallel high-throughput virtual screening using AutoDock on Windows-based computer clusters. Bioinformatics 26 (10), 1374−1375.

Qi, R., Wei, G., Ma, B., Nussinov, R., 2018. Replica exchange molecular dynamics: a practical application protocol with solutions to common problems and a peptide aggregation and self-assembly example. Methods Mol. Biol. 1777, 101−119.

Radifar, M., Yuniarti, N., Istyastono, E.P., 2013. PyPLIF: python-based protein-ligand interaction fingerprinting. Bioinformation 9 (6), 325−328.

Raman, E.P., Yu, W., Guvench, O., Mackerell, A.D., 2011. Reproducing crystal binding modes of ligand functional groups using Site-Identification by Ligand Competitive Saturation (SILCS) simulations. J. Chem. Inf. Model. 51 (4), 877−896.

Raman, E.P., Vanommeslaeghe, K., Mackerell Jr., A.D., 2012. Site-specific fragment identification guided by single-step free energy perturbation calculations. J. Chem. Theor. Comput. 8 (10), 3513−3525.

Rarey, M., Kramer, B., Lengauer, T., Klebe, G., 1996. A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. 261 (3), 470−489.

Rashin, A.A., Bukatin, M.A., 1991. Continuum based calculations of hydration entropies and the hydrophobic effect. J. Phys. Chem. 95 (8), 2942−2944.

Raub, S., Steffen, A., Kamper, A., Marian, C.M., 2008. AIScore − chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes. J. Chem. Inf. Model. 48, 1492−1510.

Robustelli, P., Piana, S., Shaw, D.E., 2020. The mechanism of coupled folding-upon-binding of an intrinsically disordered protein. J. Am. Chem. Soc. 142 (25), 11092−11101.

Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., Honig, B., 2002. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J. Comput. Chem. 23, 128−137.

Roy, A., Srinivasan, B., Skolnick, J., 2015. PoLi: a virtual screening pipeline based on template pocket and ligand similarity. J. Chem. Inf. Model. 55 (8), 1757−1770.
Ruymgaart, A.P., Cardenas, A.E., Elber, R., 2011. MOIL-opt: energy-conserving molecular dynamics on a GPU/CPU system. J. Chem. Theor. Comput. 7 (10), 3072−3082.
Saiakhov, R., Chakravarti, S., Klopman, G., 2013. Effectiveness of CASE ultra expert system in evaluating adverse effects of drugs. Mol Inform. 32, 87−97.
Santiago, G., Martínez-Martínez, M., Alonso, S., Bargiela, R., Coscolín, C., Golyshin, P.N., Guallar, V., Ferrer, M., 2018. Rational engineering of multiple active sites in an ester hydrolase. Biochemistry 57 (15), 2245−2255.
Sauton, N., Lagorce, D., Villoutreix, B., Miteva, M., 2008. MS-DOCK: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. BMC Bioinf. 9, 184.
Schames, J.R., Henchman, R.H., Siegel, J.S., Sotriffer, C.A., Ni, H., McCammon, J.A., 2004. Discovery of a novel binding trench in HIV integrase. J. Med. Chem. 47 (8), 1879−1881.
Schlitter, J., Engels, M., Krüger, P., 1994. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. J. Mol. Graph. 12, 84−89.
Schmidtke, P., Bidon-Chanal, A., Luque, F.J., Barril, X., 2011. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics 27 (23), 3276−3285.
Schnecke, V., Kuhn, L.A., 2000. Virtual screening with solvation and ligand-induced complementarity. Perspect. Drug Discov. Des. 20, 17 1−190.
Schneider, G., Clark, D.E., 2019. Automated de novo drug design: are we nearly there yet? Angew Chem. Int. Ed. Engl. 58 (32), 10792−10803.
Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J., 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 33 (Web Server issue), W363−W367.
Schneidman-Duhovny, D., Nussinov, R., Wolfson, H.J., 2007. Automatic prediction of protein interactions with large scale motion. Proteins 69 (4), 764−773.
Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L., 2005. The FoldX web server: an online force field. Nucleic Acids Res. 33.
Seco, J., Luque, F.J., Barril, X., 2009. Binding site detection and druggability index from first principles. J. Med. Chem. 52 (8), 2363−2371.
Serafini, M., Torre, E., Aprile, S., Grosso, E.D., Gesù, A., Griglio, A., Colombo, G., Travelli, C., Paiella, S., Adamo, A., Orecchini, E., Coletti, A., Pallotta, M.T., Ugel, S., Massarotti, A., Pirali, T., Fallarini, S., 2020. Discovery of highly potent benzimidazole derivatives as indoleamine 2,3-Dioxygenase-1 (Ido1) inhibitors: from structure-based virtual screening to in vivo pharmacodynamic activity. J. Med. Chem. 63 (6), 3047−3065.
Sherman, W., Day, T., Jacobson, M.P., Friesner, R.A., Farid, R., 2006. Novel procedure for modeling ligand/receptor induced fit effects. J. Med. Chem. 49 (2), 534−553.
Sinko, W., Lindert, S., McCammon, J.A., 2013. Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design. Chem. Biol. Drug Des. 81 (1), 41−49.
SiteMap, Schrödinger, 2020. Schrödinger Release 2020-2. LLC, New York, NY.
Sondergaard, C.R., Mats, H.M.O., Rostkowski, M., Jensen, J.H., 2011. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. J. Chem. Theor. Comput. 7 (2), 2284−2295.
Song, C.M., Bernardo, P.H., Chai, C.L., Tong, J.C., 2009. CLEVER: pipeline for designing in silico chemical libraries. J. Mol. Graph. Model. 27 (5), 578−583.

Song, J., Li, Y., Ji, C., Zhang, J.Z., 2015. Functional loop dynamics of the streptavidin-biotin complex. Sci. Rep. 5, 7906.

Sotriffer, C.A., Sanschagrin, P., Matter, H., Klebe, G., 2008. SFCscore: scoring functions for affinity prediction of protein-ligand complexes. Proteins 73, 395−419.

Sova, M., Cadez, G., Turk, S., Majce, V., Polanc, S., Batson, S., Lloyd, A.J., Roper, D.I., Fishwick, C.W., Gobec, S., 2009. Design and synthesis of new hydroxyethylamines as inhibitors of D-alanyl-D-lactate ligase (VanA) and D-alanyl-D-alanine ligase (DdlB). Bioorg. Med. Chem. Lett 19 (5), 1376−1379.

Spiegel, J.O., Durrant, J.D., 2020. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. J. Cheminf. 12, 25.

Stelzl, L.S., Hummer, G., 2017. Kinetics from replica exchange molecular dynamics simulations. J. Chem. Theor. Comput. 13 (8), 3927−3935.

Sterling, T., Irwin, J.J., 2015. ZINC 15–ligand discovery for everyone. J. Chem. Inf. Model. 55 (11), 2324−2337.

Stroganov, O.V., Novikov, F.N., Stroylov, V.S., Kulkov, V., Chilov, G.G., 2008. Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening. J. Chem. Inf. Model. 48, 2371−2385.

Sugita, Y., Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. 314, 141−151.

Summa, V., Petrocchi, A., Bonelli, F., Crescenzi, B., Donghi, M., Ferrara, M., Fiore, F., Gardelli, C., Gonzalez Paz, O., Hazuda, D.J., Jones, P., Kinzel, O., Laufer, R., Monteagudo, E., Muraglia, E., Nizi, E., Orvieto, F., Pace, P., Pescatore, G., Scarpelli, R., Stillmock, K., Witmer, M.V., Rowley, M., 2008. Discovery of raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection. J. Med. Chem. 51 (18), 5843−5855.

Tai, H.K., Jusoh, S.A., Siu, S.W., 2018. Chaos-embedded particle swarm optimization approach for protein-ligand docking and virtual screening. J. Cheminf. 10, 62.

Talele, T.T., Khedkar, S.A., Rigby, A.C., 2010. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. Curr. Top. Med. Chem. 10 (1), 127−141.

Taylor, J.S., Burnett, R.M., 2000. Darwin: a program for docking flexible molecules. Proteins 41, 173−191.

Taylor, P., Blackburn, E., Sheng, Y.G., Harding, S., Hsin, K.Y., Kan, D., Shave, S., Walkinshaw, M.D., 2008. Ligand discovery and virtual screening using the program LIDAEUS. Br. J. Pharmacol. 153 (Suppl. 1), S55−S67. Suppl 1.

Teague, S.J., 2003. Implications of protein flexibility for drug discovery. Nat. Rev. Drug Discov. 2, 527−541.

ten Brink, T., Exner, T.E., 2010. pK(a) based protonation states and microspecies for protein-ligand docking. J. Comput. Aided Mol. Des. 24 (11), 935−942.

Terp, G.E., Johansen, B.E., Christensen, I.T., Jorgensen, F.S., 2001. A new concept for multidimensional selection of ligand conformations (MultiS elect) and multidimensional scoring (MultiS core) of protein-ligand binding affinities. J. Med. Chem. 44, 2333−2343.

Testa, B., Balmat, A.L., Long, A., Judson, P., 2005. Predicting drug metabolism–an evaluation of the expert system METEOR. Chem. Biodivers. 2 (7), 872−885.

Thomas, B., Josep, A.P., Hongming, C., Christian, M., Christian, T., Ola, E., et al., 2020. REINVENT 2.0 − an AI Tool for De Novo Drug Design. ChemRxiv.

Thomsen, R., Christensen, M.H., 2006. MolDock: a new technique for highaccuracy molecular docking. J. Med. Chem. 49, 33 15−3321.

Tian, W., Chen, C., Lei, X., Zhao, J., Liang, J., 2018. CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res. 46 (W1), W363−W367.

Tietze, S., Apostolakis, J., 2007. Glamdock: development and validation of a new docking tool on several thousand protein-ligand complexes. J. Chem. Inform. Model. 47 (4), 1657−1672.

Torrie, G.M., Valleau, J.P., 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J. Comput. Phys. 23, 187−199.

Totrov, M., Abagyan, R., 2008. Flexible ligand docking to multiple receptor conformations: a practical alternative. Curr. Opin. Struct. Biol. 18, 178−184.

Tran, N., Van, T., Nguyen, H., Le, L., 2015. Identification of novel compounds against an R294K substitution of influenza A (H7N9) virus using ensemble based drug virtual screening. Int. J. Med. Sci. 12 (2), 163−176.

Trosset, J.Y., Scheraga, H.A., 1999. Prodock: software package for protein modeling and docking. J. Comput. Chem. 20, 412−427.

Trott, O., Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J. Comput. Chem. 455−461.

Vakser, I.A., Matar, O.G., Lam, C.F., 1999. A systematic study of low-resolution recognition in protein-protein complexes. Proc. Natl. Acad. Sci. U.S.A. 96, 8477−8482.

Valsson, O., Tiwary, P., Parrinello, M., 2016. Enhancing important fluctuations: rare events and metadynamics from a conceptual viewpoint. Annu. Rev. Phys. Chem. 67, 159−184.

van Hilten, N., Chevillard, F., Kolb, P., 2019. Virtual compound libraries in computer-assisted drug discovery. J. Chem. Inf. Model. 59 (2), 644−651.

van Zundert, G.C.P., Rodrigues, J.G.L.M., Trellet, M., Schmitz, M., Kastritis, P.L., Karaca, E., Melquiond, A.S.J., Dijk, M.V., de Vries, S.J., Bonvin, A.M.J.J., 2016. The HADDOCK2.2 webserver: user-friendly integrative modeling of biomolecular complexes. J. Mol. Biol. 428, 720−725.

Vangone, A., Bonvin, A.M.J.J., 2015. Contact-based prediction of binding affinity in protein-protein complexes. eLife 4, e07454.

Vavra, O., Filipovic, J., Plhak, J., Bednar, D., Marques, S.M., Brezovsky, J., Stourac, J., Matyska, L., Damborsky, J., 2019. CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels. EEE/ACM Trans. Comput. Biol. Bioinform. Mar. 26 https://doi.org/10.1109/TCBB.2019.2907492.

Velec, H.F.G., Gohlke, H., Klebe, G., 2005. DrugScoreCSD-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J. Med. Chem. 48, 6296−6303.

Venkatachalam, C.M., Jiang, X., Oldfield, T., Waldman, M., 2003. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J. Mol. Graph. Model. 21 (4), 289−307.

Venkatapathy, R., Moudgal, C.J., Bruce, R.M., 2004. Assessment of the oral rat chronic lowest observed adverse effect level model in TOPKAT, a QSAR software package for toxicity prediction. J. Chem. Inf. Comput. Sci. 44, 1623−1629.

Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., Taylor, R.D., 2003. Improved protein-ligand docking using GOLD. Proteins 52 (4), 609−623.

Viswanath, S., Ravikant, D.V.S., Elber, R., 2014. DOCK/PIERR : web server for structure prediction of protein-protein complexes. Methods Mol. Biol. 1137, 199−207.

von Behren, M.M., Bietz, S., Nittinger, E., et al., 2016. mRAISE: an alternative algorithmic approach to ligand-based virtual screening. J. Comput. Aided Mol. Des. 30, 583−594.

Wang, R., Liu, L., Lai, L., Tang, Y., 1998. SCORE: a new empirical method for estimating the binding affinity of a protein-ligand complex. J. Mol. Model. 4, 379−394.

Wang, W., Donini, O., Reyes, C.M., Kollman, P.A., 2001. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annu. Rev. Biophys. Biomol. Struct. 30, 21 1−243.

Wang, R., Lai, L., Wang, S., 2002. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J. Comput. Aided Mol. Des. 16, 11−26.

Wang, J.C., Chu, P.Y., Chen, C.M., Lin, J.H., 2012. idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res. 40 (Web Server issue), W393−W399.

WaterMap, Schrödinger, 2020. Schrödinger Release 2020-2. LLC, New York, NY.

Weiner, P.K., Kollman, P.A., 1981. Amber − assisted model building with energy refinementła general program for modeling molecules and their interactions. J. Comput. Chem. 2, 287−303.

Welch, W., Ruppert, J., Jain, A.N., 1996. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chem. Biol. 3 (6), 449−462.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., Wilson, M., 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46 (D1), D1074−D1082.

Wolf, S., Stock, G., 2018. Targeted molecular dynamics calculations of free energy profiles using a nonequilibrium friction correction. J. Chem. Theor. Comput. 14 (12), 6175−6182.

Wolf, S., Amaral, M., Lowinski, M., Vallée, F., Musil, D., Güldenhaupt, J., Dreyer, M.K., Bomke, J., Frech, M., Schlitter, J., Gerwert, K., 2019. Estimation of protein-ligand unbinding kinetics using non-equilibrium targeted molecular dynamics simulations. J. Chem. Inf. Model. 59 (12), 5135−5147.

Wriggers, W., 2012. Conventions and workflows for using situs. Acta Crystallogr. D 68, 344−351.

Xiao, X., Zeng, X., Yuan, Y., Gao, N., Guo, Y., Pu, X., Li, M., 2015. Understanding the conformation transition in the activation pathway of β2 adrenergic receptor via a targeted molecular dynamics simulation. Phys. Chem. Chem. Phys. 17 (4), 2512−2522.

Yan, Y., Zhang, D., Zhou, P., Li, B., Huang, S.-Y., 2017. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res. 45.

Yang, J.M., Chen, C.C., 2004. GEMDOCK: a generic evolutionary method for molecular docking. Proteins: Struct., Funct. Bioinform. 55, 288−304.

Yang, C.Y., Wang, R.X., Wang, S.M., 2006. M-score: a knowledge-based potential scoring function accounting for protein atom mobility. J. Med. Chem. 49, 5903−5911.

Yang, J.F., Wang, F., Chen, Y.Z., Hao, G.F., Yang, G.F., 2020. LARMD: integration of bioinformatic resources to profile ligand-driven protein dynamics with a case on the activation of estrogen receptor. Brief Bioinform. 21 (6), 2206−2218.

Yang, Y.I., Shao, Q., Zhang, J., Yang, L., Gao, Y.Q., 2019. Enhanced sampling in molecular dynamics. J Chem Phys. 151 (7), 070902.

Yin, S., Biedermannova, L., Vondrasek, J., Dokholyan, N.V., 2008. MedusaScore: an accurate force-field based scoring function for virtual drug screening. J. Chem. Inf. Model. 48, 1656−1662.

Yoshikawa T , Kanai C, Yamamoto Y, Murakami R, Okuno Y, Czeek D., "De novo design system with PSO": ⟨https://www.insilico.jp/czeekd.html⟩.

You, W., Tang, Z., Chang, C.A., 2019. Potential mean force from umbrella sampling simulations: what can we learn and what is missed? J. Chem. Theor. Comput. 15 (4), 2433−2443.

Young, T., Abel, R., Kim, B., Berne, B.J., Friesner, R.A., 2007. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. Proc. Natl. Acad. Sci. U. S A. 104, 808−813.

Yu, W., MacKerell Jr., A.D., 2017. Computer-aided drug design methods. Methods Mol. Biol. 1520, 85−106.

Yu, T.Q., Lu, J., Abrams, C.F., Vanden-Eijnden, E., 2016. Multiscale implementation of infinite-swap replica exchange molecular dynamics. Proc. Natl. Acad. Sci. U. S. A 113 (42), 11744−11749.

Yuan, Y., Pei, J., Lai, L., 2011. LigBuilder 2: a practical de novo drug design approach. J. Chem. Inf. Model. 51 (5), 1083−1091.

Yuan, Y., Pei, J., Lai, L., 2020. LigBuilder V3: a Multi-Target de novo Drug Design Approach. Front Chem. 8, 142.

Zhao, Y., Sanner, M.F., 2007. FLIPDock: docking flexible ligands into flexible receptors. Proteins 68, 726−737.

Zhao, X., Liu, X., Wang, Y., Chen, Z., Kang, L., Zhang, H., Luo, X., Zhu, W., Chen, K., Li, H., Wang, X., Jiang, H., 2008. An improved PMF scoring function for universally predicting the interactions of a ligand with protein, DNA, and RNA. J. Chem. Inf. Model. 48, 1438−1447.

Zhao, B.W., Stuart, M.A.C., Hall, C.K., 2017. Navigating in foldonia: using accelerated molecular dynamics to explore stability, unfolding and self-healing of the β-solenoid structure formed by a silk-like polypeptide. PLoS Comput. Biol. 13, e1005446. https://doi.org/10.1371/journal.pcbi.1005446.

Zou, X., Sun, Y., Kuntz, I.D., 1999. Inclusion of solvation in ligand binding free energy calculations using the generalized-Born model. J. Am. Chem. Soc. 121, 8033−8043.

Zsoldos, Z., Reid, D., Simon, A., Sadjad, B.S., Johnson, A.P., 2006. eHiTS: an innovative approach to the docking and scoring function problems. Curr. Protein Pept. Sci. 7 (5), 421−435.

CHAPTER

Computational tools in cheminformatics

4

Rakhi Thareja[1], Jyoti Singh[2], Prerna Bansal[3]

[1]*Department of Chemistry, St. Stephen's College, University of Delhi, New Delhi, Delhi, India;*
[2]*Department of Chemistry, Hansraj College, University of Delhi, New Delhi, Delhi, India;*
[3]*Department of Chemistry, Rajdhani College, University of Delhi, New Delhi, Delhi, India*

4.1 Introduction

In the modern-day world, cheminformatics holds the key to the latest technology to carry out research effectively through computational tools in the areas of synthesis chemistry and medicinal chemistry. Researchers who have an extensive and elaborate knowledge of the combination of theoretical concepts of chemistry along with computational tools are essential for industries today. There are a number of detailed books available on cheminformatics but this chapter intends to provide a comprehensive outlook on computational problems in pharmaceutical sciences to achieve a higher success rate in the laboratory.

Cheminformatics is also known as an interface science as it combines physics, chemistry, biology, mathematics, biochemistry, statistics, and informatics (Arulmozhi and Rajesh, 2011; Engel, 2006; Bharati et al., 2009). Cheminformatics solves chemical and synthetic problems effectively by making use of information tools available on the web. Hence, we can say that it is recognized as a distinct discipline in computational molecular sciences. Cheminformatics has now made it easier to make or innovate newer designs of molecules of desirable pharmaceutical properties. Furthermore, it also helps in designing reactions and possible synthetic routes to obtain anticipated products. It helps with analysis and also aids in the structural elucidation of molecules isolated from various biological and environmental sources or from reaction pathways. Though modern-day research is based on an interdisciplinary approach, we must view cheminformatics in terms of understanding bioinformatics, which is primarily characterized by putting more emphasis on the processing of computational tools for using databases of biological information or sequences available in large amounts. In cheminformatics, we use databases of structures of chemical origin either in 2D or 3D forms with information on basic structural and physical properties of the respective molecules.

Cheminformatics has evolved from the basic representation of structures and the collection of different structures along with effective searching methodologies for the desired structure from vast databases. So, we may also summarize the basic definition of cheminformatics as a subject that deals with suitable applications of

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00012-9

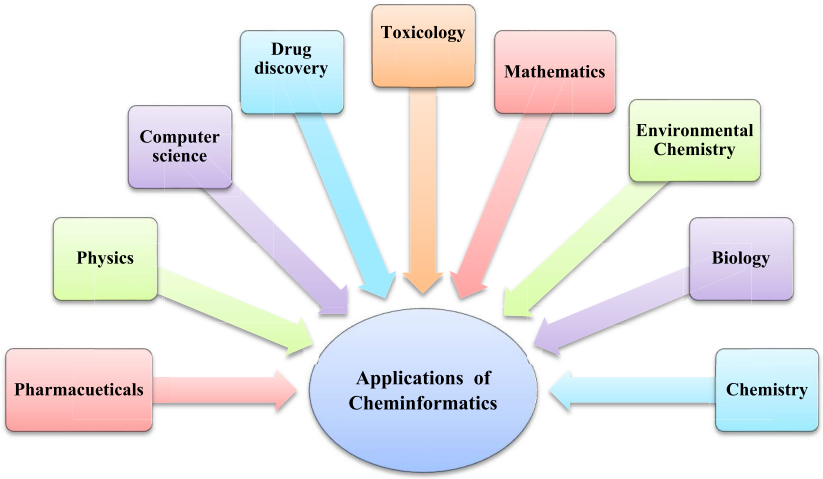


FIGURE 4.1

Applications of cheminformatics in different disciplines.

computational tools to molecules of chemical origin that possess a wide range of applications in other fields of importance for better and healthier living. Those fields may include biology, environmental chemistry, pharmaceutical chemistry, etc. as shown in Fig. 4.1. The methods involved in the representation of molecules and chemical reactions are discussed in the next section, which gives an understanding of the topic in depth. To explore the vast array of molecules of pharmaceutical importance, these methodologies will have to be adopted by practicing them simultaneously to gain confidence for the journey ahead. In the present work, a modest attempt has been made to provide a list of databases that are sufficient to start this quest.

A brief summary of cheminformatics covered in this chapter is given in Fig. 4.2. Fig. 5.2 shows how basic web cum computational tools help in setting up a link to understand molecules, reactions, spectra, structure, activities, and applications. It also gives an outline of the chapter.

It is pertinent to remember that the optimization tools mentioned in Section 4.6 are a part of molecular modeling, wherein computational tools for drawing and visualization of molecules of chemical origin are important to understand the basis of formation of the wide range of databases utilized for applications of cheminformatics and bioinformatics. Molecular modeling refers to that branch of chemistry that encompasses all computational methods that aid in drawing, visualizing, calculating, and interpreting results on chemical molecules. This branch of science includes not only drug designing in chemical biology but also designing and estimating the potential of novel materials in materials sciences without performing the experiments in a lab. It not only helps in the aforementioned projects but is also worthwhile when studying the reaction mechanism. Quantitative structure–activity relationship (QSAR) forms an interface between the structure and activity of various molecules of interest. It is generally available through multiple platforms

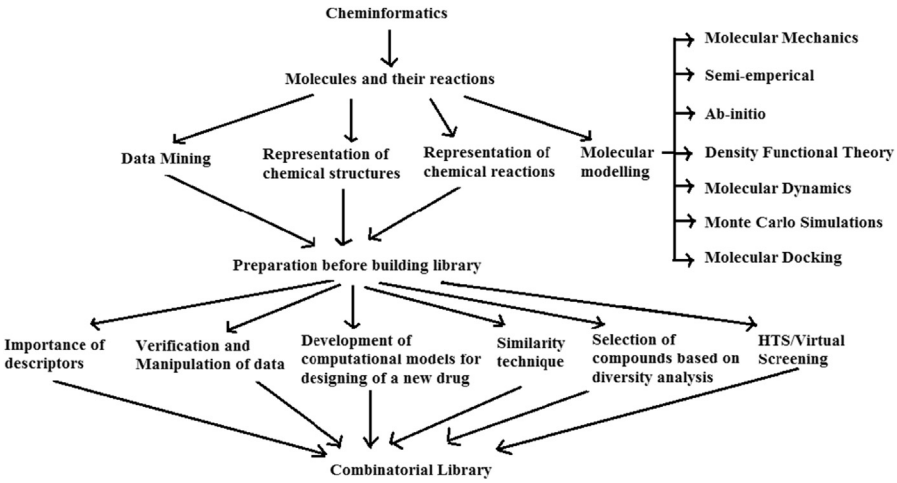


FIGURE 4.2

Summary of the chapter.
HTS, High-throughput screening.

inclusive of molecular modeling software like HyperChem (Froimowitz, 1993), Materials Studio (Accelrys, 2016), Arguslab (Thompson, 2004), AutoDoc 4.2 (Morris et al., 2009), Spartan'18 (Shao et al., 2006), MOE (Molecular Operating Environment (MOE), 2019), and many more.

Later, a general overview of the correlation of spectral analysis with structural aspects of molecules is discussed. It helps in effective structural elucidation and one can also appreciate how different spectroscopies are complementary. Using these techniques, confirmation of specific functional groups along with their orientations can be achieved successfully. Nowadays, there is software available (Logic for Structure Elucidation, LSD software) for elucidation of structure automatically that is good for beginners to understand how small molecules can be introduced by relating their structural properties to simple 1D and 2D spectra. It shows that effective collaboration of spectra such as HSQC, Cosy, and HMBC is effective for understanding interactions of nonhydrogen elements/atoms that are important for arriving at a solution (Nuzillard and Bertrand, 2017).

One will be able to appreciate the successful usage of information technology to help chemists and even biologists to investigate new-age problems of different types that keep evolving in newer forms. It not only enables investigation but also assists in organizing, analyzing, and understanding data available on the web for the development of newer drugs, materials, novel compounds, and reaction processes. The application sections cover examples from implementation of the knowledge of cheminformatics in modern-day science for better living. Now that the organization of the chapter has been discussed, it is now time to explore the role of cheminformatics in the science of discovery of newer and innovative drugs in a clearer and technologically self-sufficient way.

4.2 Molecules and their reactions: representation

It is imperative to represent a molecular structure in the correct way to develop a deeper understanding of its properties. Chemical structures are generally displayed as either 2D forms (formulae) or models of 3D forms. While 2D representations are based on graphs, 3D forms help in attaining better insights into conformational structures, which display important steric and structural properties with the clarity of electronic factors as well. 3D models are adequate to explain the spatial arrangement of bonds and atoms, but the availability of these in cyberspace does not ensure clarity of the additional properties possessed by them. To achieve the latter, the properties of interest must be transformed into suitable algorithms/methodological tools available with molecular modeling, which is described in Section 4.6 in detail. 3D representations or conformations are better and hence computational researchers show greater interest in the development of programs and databases for these. However, to reach the final outcome, the major challenge is to arrive at the most stable geometrically optimized structure, but, simultaneously, conformational flexibility cannot be ignored. In this section, we place emphasis on the widely used databases on 3D data information. Though information on representation is covered in later subsections, it is deemed fit to understand how one can use the discovery of knowledge in the already available large databases on various web portals. Mercury (Macrae et al., 2020), Chemcraft (https://www.chemcraftprog.com), VMD-Visual Molecular Dynamics (Humprey et al., 1996), Chemaxon (Marvin, 2014), and Chemspider (Swain, 2012) are some of the 2D and 3D visualization tools widely used.

One must begin by understanding the term "data mining," which refers to the discovery of arrangements and patterns in a vast number of datasets that involve techniques by using a suitable combination of computational tools, statistical tools, and databases. It paves the way for research into understanding chemistry in a better way.

4.2.1 Data mining

Data mining refers to a method that is a combination of statistical methods and computational tools (Fig. 4.3). The statistical methods include principal component analysis (PCA), principal component regression, multilinear regression analysis, and partial least squares regression. Other methods involved are factor analysis and correlation analysis. The computational tools are from the branch of machine learning that incorporates the study of the artificial neural networks like self-organizing, feedforward, counterpropagation, Bayesian, etc. Machine learning is also based on K-nearest neighbor analysis, decision learning trees like C5 and ID3, clustering algorithms, and genetic algorithms. The most comprehensively used databases for data mining include Cambridge Structural Database (CSD) and Protein Data Bank (PDB). The former, i.e., CSD, is a source that comprises crystal structures of metal organic molecules and simple organic molecules. These structures are mainly based on X-ray diffraction (XRD) or neutron diffraction studies. However, polypeptides and polysaccharides possessing more than 24 units, alloys, metals,

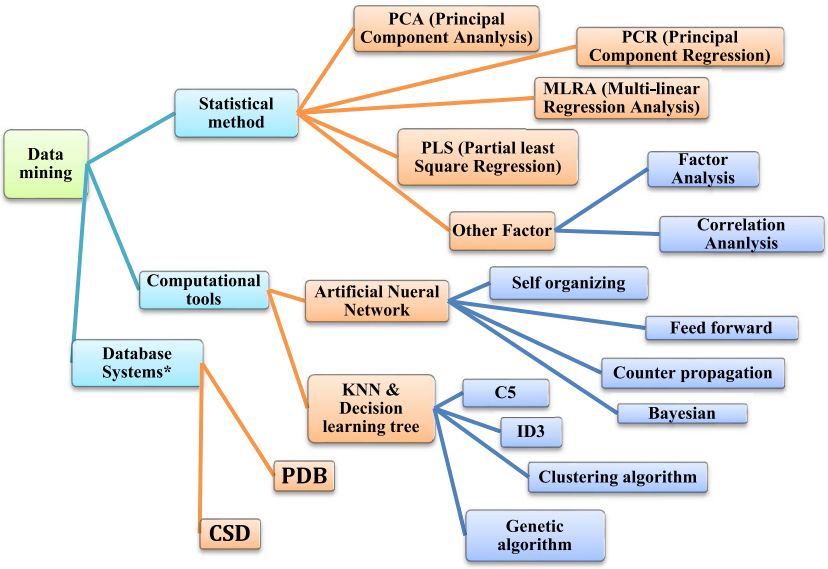


FIGURE 4.3

Methods and tools in data mining (*for details refer to Table 4.1).

and oligonucleotides are totally excluded from CSD. PDB is a database that deals with the structures of nucleic acids and proteins in 3D form. Here, the data available are derived from XRD studies or nuclear magnetic resonance (NMR) studies. More than 5000 structures are released per year. The list of databases available for chemical data is infinite in today's technological world. Descriptions of each one of them are beyond the scope of this chapter and hence they are described in Table 4.1 (Leach and Gillet, 2007).

These databases and many more based on similar compilations contain an infinite amount of data and information related to proteomics, phylogenetics, metalobomics, genomics, gene expression, nucleotide sequencing, chemical sciences, and more. The list is beyond the scope of this chapter but for a beginner, Table 4.1 provides enough information to successfully get started. This section discussed the discovery of information from large databases.

4.2.2 Representation of chemical structures

The efficacy of cheminformatics is directly related to the correct representation of molecules of chemical origin and their transformations through chemical reactions (Engel and Gasteiger, 2018). The previous section mentioned the wide range of databases in which information regarding the representation of chemical structures and other properties of interest for applications can be easily found. But correct representation of the chemical structure carries prime importance. Making use of

Table 4.1 List of databases that serve as an effective tool in cheminformatics.

Database	URL	Remarks
Cambridge Structural Database	www.ccdc.cam.ac.uk/products/csd/	Comprehensive library of chemical structures of organometallics and simple organic molecules.
Protein Data Bank	http://www.rcsb.org/pdb/home/home.do	Wide-ranging data of 3D representations of nucleic acids and proteins.
ChemIndustry	https://www.chemindustry.com/	Exhaustive directory for chemical and industrial researchers. It contains more than 50,000 entities with full texts.
ChemExper	https://www.chemexper.com/	Multidisciplinary in its approach, it combines chemical sciences with telecommunication and computational science.
PubChem	http://pubchem.ncbi.nlm.nih.gov/	2D structures of chemical compounds that are available for free (developed by the National Center for Biotechnology Information): PubChem Compounds, PubChem BioAssay, and PubChem Substances.
Beilstein database	http://info.crossfiredatabases.com/	Covers published information from 1771 to the present day.
CAS	www.cas.org/	CAS is a part of the American Chemical Society. It is effective for obtaining information on research related to biomedical sciences, materials, agriculture, chemical sciences, and many more.
NIOSHTIC-2	http://www2a.cdc.gov/nioshtic-2/	I A searchable database (in the form of a bibliography) of health-related research. Several other databases related to it are ACS, Science Direct, Elsevier, NLM, etc.
World of Molecular BioActivity	http://www.sunsetmolecular.com/	Aims at providing data related to clinical pharmacokinetics in addition to target information of drugs. It is ideal for computational drug discovery and relates to descriptors effectively for quantitative structure—activity relationship (QSAR) study.
ChemSpider	http://www.chemspider.com/	Chemical search engine that is centered around the information of the structure of the chemical molecule.

Table 4.1 List of databases that serve as an effective tool in cheminformatics.—*cont'd*

Database	URL	Remarks
ZINC	https://zinc.docking.org/	Fulfills the need virtual screening by providing a free database of thousands of commercially available compounds. The compounds available are put in a 3D ready-to-dock format.
Distributed Structure-searchable Toxicity	http://www.epa.gov/ncct/dsstox/index.html	A good tool for QSAR and quantitative structure–toxicity relationship
Registry of Toxic Effects of Chemical Substances	rtecsfile@symyx.com	Includes a database of a wide range of prescribed and nonprescribed drugs of biological, pesticidal, and chemical sciences.
ChemBank	http://chembank.broadinstitute.org/welcome.htm	Public database for screening of small molecules of biological and medical importance.
Unified Medical Language System	http://www.nlm.nih.gov/research/umls/	Library of biomedical compounds.
e-molecules	http://www.emolecules.com/	Focuses on a database of molecules of chemical origin.
Specs	https://www.specs.net/index.php	Founded in 1987 it provides databases of chemical compounds and molecules of importance for drug discovery in addition to some natural compounds obtained from nuclear magnetic resonance studies.
Biological databases	https://pdb101.rcsb.org/browse/biomolecules	Databases of molecules of biological origin that include information obtained from computational study as well as lab experiments.
PROSITE	http://www.expasy.ch/prosite/	Acts as an interface between cheminformatics and bioinformatics by relating the domains of various protein families.
MOLTABLE Web Portal	http://moltable.ncl.res.in/	Deserves a special mention because it has molecules of chemical origin with importance in pharmaceutical sciences.
Chemoinformatics.org	http://www.cheminformatics.org/menu.shtml	One of the best websites, which is noncommercial and provides information on all progams related to cheminformatics. Database sets for QSAR, quantitative structure–property relationship, and blood–brain barrier penetrations are also available.

Continued

Table 4.1 List of databases that serve as an effective tool in cheminformatics.—*cont'd*

Database	URL	Remarks
European Molecular Biology Laboratory	http://www.ebi.ac.uk/embl/	Nucleotide source that originated in Europe that also acts as a main source for providing sequences of RNA and DNA.
OMIM	http://www.ncbi.nlm.nih.gov/omim/	Database of all genetic diseases.
NCBI	http://www.ncbi.nlm.nih.gov/	Database of genome sequencing and hence carries importance because it is related to pharmaceutical sciences.
Medical Literature Analysis and Retrieval System Online	www.nlm.nih.gov/databases/databases_medline.html	Database based on bibliography.
MeSH	http://www.nlm.nih.gov/mesh/	Contains indexed articles of journals and books of biological sciences.

ChemDraw or ISISDraw (Science Museum Group. ISIS Draw chemical drawing software), ACD/Chemsketch (ACD/ChemSketch, 2020), ChemDoodle (Todsen, 2014), and JChemPaint (Krause et al., 2000) is an inherent part of earlier computational ways of representing chemical molecules. Nowadays, the ways this can be done are as follows:

1. **Linear representation:** This is done using Simplified Molecular Input Line Entry System (SMILES), International Chemical Identifier (InChI), and Single Line Notation. These are the codes that are used for effective representation of molecules in code language form.

The most extensively used linear notation is SMILES as it is much easier to comprehend. The set of rules for writing these strings is limited and comfortable for all to use. Several publications discuss SMILES in more detail, including Anderson et al. (1987), Weininger (1988), Weininger et al. (1989), and Hunter et al. (1987). The basic syntax rules for SMILES are limited to just five and if these rules are violated in a SMILES entry, then a warning is generated to the user, who is asked to reedit or reenter the structure. The codings support all elements of the periodic table. The extensions of SMILES are also very useful, e.g., SMART and SMIRKS, which is a line notation for generic reactions. One can read more about SMILES notations elsewhere (James et al., 2002).

Another widely used linear notation is InChI code, which was developed as a result of the requirement for a machine-readable standard nomenclature. It is developed by IUPAC (Stein et al., 2003). It is an open and freely available identifier that is based on a layer of hierarchy. The initial layers comprise information related to connection tables, and later, layers need to be additionally added that deal with

the complex nature of the structure of isomers and isotopes. These layers are flexible and can be extended. However, standard notation for InChI is when one notifies a predefined number for layers and if certain layers need to be added, in which case it is extended to nonstandard form of code that has additional layers for providing information related to the complexity of the structure of the chemical. It is unlike the earlier codes discussed as it refers to a canonical form of line/linear representation and hence constitutes a unique identifier that follows simple sets of rules. There is more than one SMILES string possible for a given structure, which is not the case in the present one. One must not confuse InChI with a registry system as it is a nomenclature describing the structural aspects of a chemical, and can even be generated for a chemical system that does not exist.

2. **Multidimensional representation:** There are a number of chemical table files available for different file formats. Multidimensional representation refers to a family of chemical file formats based on text that is able to correctly interpret molecules and the reactions they are involved in. It may be related to different types of coordinate system for representations. For multidimensional representation, there are several file formats available out of which mdl molfile extension is widely used. MDL molfile has a format that carries knowledge of the atoms, their connectivities, bonds between them along with coordinates of a molecule. Most of the cheminformatics software is compatible with the molfile format. The latter basically comprises the connection table having information on atoms, bonds, connections, and types in addition to other complex information. These days, molfile V2000 and the latest molfile V3000 are also being used widely. These are the extended connection tables.

Another file format widely used for structure representation is the .sdf file format, which stands for structure data file. Primarily, it puts emphasis on structural information. SDF files have a special feature of including data associated with the molecule while wrapping the molfile in it. It supports multiple line data representations and uses a high carriage return if there is extension of the text field beyond 200. It may be noted here that most codes violate the requirement of restriction within 200 character frames and hence this statement is often violated.

3. **File formats and visualization of chemical structures:** Apart from the aforementioned representations, SYBYL MOL/MOL2 can also be used to represent pdb structures in terms of Cartesian coordinates and CIF, the crystallographic file formats using the z-matrix.

4.2.3 Representation of chemical reactions

Apart from chemical structures being included in the core of cheminformatics, structural transformations also form an inherent part of it. This refers to the chemical reactions occurring between various molecules. It is more challenging to handle chemical reactions. On understanding the basic molecules that form part of a reaction either by being a reactant or a product, it is now appropriate to talk about information based on discoveries available for reactions of chemical origin.

The ways of representation for a reaction can be placed according to different standards like (1) description of reaction centers, (2) coding of electron and bond matrices, and (3) description of vectors or fingerprints of molecules (Faulon and Bender, 2010). Representation of certain types of reaction that undergo similar changes with respect to atoms and bonds irrespective of the intermediate states involves more complex approaches than merely just specifying a particular reaction involving the different reactants and products. In those cases, a set of chemicals involved in the entire reaction and products forms are used for representation (James et al., 2002).

4.3 Preparation before building libraries for databases in cheminformatics

4.3.1 Importance of descriptors

In recent times, drug designing and development of suitable and similar ligands for docking of macromolecular targets has become increasingly demanding especially because of the pandemic. Before we move on with the structures of files or libraries, it is pertinent to discuss the descriptors generally used to explain the similarities or dissimilarities in structures. These descriptors may be calculated from 2D or 3D representations. Some simple descriptors calculated from 2D representation are physicochemical properties, atomic pairs, fingerprints, kappa-shaped indices, topological indices, refractivity of the molar substance, BCUT descriptors, electrotopological state indices, simple counts, or other physicochemical properties. 3D descriptors always provide better in-depth knowledge about the compounds/structures under study. These include pharmacophore keys, 3D fragment screens, dipole moments, electrostatic potential, comparative molecular field analysis, and several additional descriptors, which are used for predicting absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties of ligands used as potential drugs for docking biomolecules such as proteins. The data obtained based on these descriptors are then verified and manipulated using simple methods such as scaling or a complicated technique such as PCA that further yields a whole new list of descriptors (Leach and Gillet, 2007).

For descriptors to be of importance to particular research, they must satisfy simple requirements such as: they must be able to interpret the structure, they should be able to correlate the structures with one or more properties, they should be able to differentiate between the isomers, they should be available for local application, they should be simple and easy to comprehend by the research society at large, they should not be deduced on the basis of experimental results or properties, they should be different from other descriptors and should not bear any relations with any other descriptors, and they should be based on similar concepts of structure. Descriptors must change with change in the structure of the molecule and if they are related to molecular size, they should be dependent on size to a high degree of correctness.

Therefore it would be ideal if one believes that molecular descriptors play a crucial role in pharmaceutical chemistry, wherein these structures are presumed to be real entities, which are later changed/modified to number representations on screen, and then mathematically treated for information relating to chemistry contained in the structure of the molecule. Therefore the molecular descriptors are described as logical and mathematical deductions that are transformed into information related to the chemical structures (Todeschini and Consonni, 2009). These descriptors are based on experiments or theory. Some descriptors based on the former, i.e., experimental descriptors, are log P, dipole moment, molar refractivity, and other physicochemical properties, and some descriptors based on the latter include a range of descriptors such as 0D, 1D, and 2D to 6D. To discuss each one of these is beyond the scope of this chapter but they are important for performing further research with molecules. QSARs are also based on these descriptors and these studies are important for the development and designing of newer compounds and drugs for pharmaceutical industries.

To understand the basis of these important properties, one must understand what descriptors are. Descriptors represent characteristics of molecules on the basis of their physicochemical properties. The most common descriptors are based on shape, size, geometry, interconnectivity of molecules, surface, electrostatic, hybrid, constitutional, and topological; all these are explicitly related to one another (Karelson et al., 2000). Constitutional descriptors are those that describe the chemical composition of compounds (http://www.vcclab.org/lab/indexhlp/consdes.html). Topological descriptors are those that are based on encoding the chemical constitution on the surface of the molecule that enables it to show permeability and solubility of a particular substance to understand its efficacy to prove itself as a drug. A descriptor that describes the extent of polarizability, electronic properties such as ionization energy, dipole moment, and electron density of a crystal is known as an electrostatic descriptor (http://www.codessa-ro.com/descriptors/electrostatic/index.htm). 3D descriptors are based on xyz coordinates that provide the orientation of the molecule in space and are known as geometrical descriptors. These descriptors are very useful in predicting biological activities (Balaban, 1997). There are quantum chemical descriptors too that incorporate in themselves all the characteristics related to geometrical and electronic parameters of a chemical system (Karelson et al., 1996). Quantum chemical descriptors include highest occupied molecular orbitals, lowest unoccupied molecular orbitals, delocalization of electrons in the system, and electronic density (Enoch, 2010). Apart from these, there are hybrid descriptors too, which are based on the diverse nature of chemical compounds and hence are helpful in predicting futuristic models (Stanton, 1999; Ma et al., 2012). In modern-day research, there many classes of descriptors. One of the most popular ones is the fingerprint descriptor, which is based on binary bit string for a similarity search in large database systems. There are several fingerprints known in the literature (http://rdkit.org/docs/api/rdkit.Chem.MACCSkeys-pysrc.html, Todeschini et al., 1996; Hinselmann et al., 2011; Rogers and Mathew, 2010; Bender, et al., 2010; Deursen et al., 2010).

These days, descriptors play a crucial role in the drug-discovery process too. Let us just explore some of the important ones here, such as the solubility parameter (Jorgenson and Duffy, 2002), which has great influence on the bioavailability of the drug. This constitutes a very important parameter of pharmacokinetic descriptors (Livingstone et al., 2009). It also plays a central role in pharmacy for lipid-based formulations (Persson et al., 2013). Another common and basic descriptor is log P, which is the ratio of the partition coefficient of water to the partition coefficient of octanol (Faller, 2007). Knowledge of these descriptors is important at the initial stages of the drug-discovery process. Knowledge of these descriptors of the biological target and the smaller-sized ligand are computed simultaneously to gain an insight into the side effects of drugs, and it also leads to further information about the upcoming area of research in medicine known as polypharmacology (Cortes-Cabrera et al., 2013).

Descriptors not only play a crucial role in the area of drug discovery but also are important in the field of materials science. In these areas, selection of the right descriptor has paved the way for the development of improved energetic substances (Rice and Byrd, 2013). Single adsorption isotherms have also been predicted by evaluation of important structural and molecular descriptors of both the adsorbate and the adsorbent (Garcia et al., 2013). The range of descriptors is huge but the importance of each and every descriptor can be exclusively studied under the topics of QSAR; however, those explained in this chapter will suffice at the initial level.

4.3.2 Verification and manipulation of data

In the previous subsection, the importance of descriptors was discussed. In this section, we will briefly discuss some of the tools that are employed for examining the characteristics of the structures or the representations under study. This is an important step before analysis of the compound. The distribution of values for a particular molecular descriptor must be evaluated to monitor how one descriptor is correlated to another. While doing this, researchers might feel the need to manipulate the data as well. As stated earlier, this can involve the use of methods ranging from simple to complex, or in other words ranging from scaling methods to PCA methods. Manipulation yields an altogether new set of descriptors with improved and desired features. Other methods are data spread and distribution of data, correlational analysis of descriptors, etc. PCA refers to reducing the dimensions, i.e., the number of variables, that are used to describe a particular molecule, object, or dataset regarding the number of overlaps in characteristic features of descriptors. This implies that if the number of correlations found is too many, then application of PCA is mandatory. PCA can be represented by a simple illustration as shown in Fig. 4.4.

In this, drug search is as important as the synthesis of newer compounds. Combinatorial libraries play an important role in the act of designing a drug.

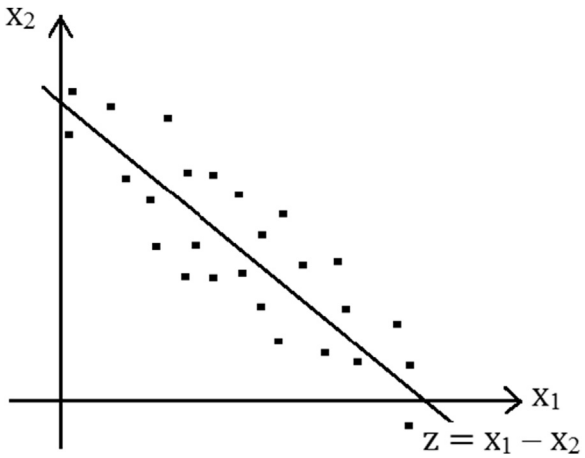


FIGURE 4.4

Illustration of principal component analysis.

4.3.3 Development of computational models for designing a new drug

The synthesis of drugs on the virtual mode is a vast topic in itself and consists of two broad categories: ligand-based drug design and structure-based drug design. Most of the designing protocols involve the traditional pathway that involves designing, synthesizing, and testing a new drug. Since 2019 drug testing has become very competitive because of the recent pandemic. The most common characteristic feature of this type of development using computational tools involves building a basic model that will enable the observation of desired activities or properties of the ligand or the drug so that it is docked efficiently.

The development of an efficient virtual model depends on a number of factors. One of the factors is whether the datasets are small or large in number. While the smaller ones can be analyzed using simple QSAR techniques, the larger datasets require high-throughput screening (HTS). QSAR models can be developed using simple or multiple linear regression analysis. The quality of these analyses then needs to be assessed, which can be done in a number of ways. The most common way is by making use of the square of the correlation coefficient or regression coefficient. This is generally followed by modification of the dataset where some values are removed from the dataset and a QSAR model is derived using the rest of the values, which is then applied to predict the results of the final model after the removal of those values from the dataset. One of the approaches generally discussed is the leave one out approach where just one value of data is removed and the process is carried out to cross-validate the squared correlation coefficient; hence, this method is known as cross-validation. The value of the square of the correlation coefficient thus found is lower than the normal value obtained. This process

when reiterated for best results proves to be the best method for predicting the efficiency of the QSAR equation developed for building the best model. Another method that can be employed for validating the predictability of regression function is the standard error of prediction that involves a mathematical expression dependence on the number of independent variables. This is further aided by having information on the number of degrees of freedom that are linked to each parameter.

QSAR in itself is a vast topic, which has now expanded to QSPR, i.e., quantitative structure property relationship, and QSTR, i.e., quantitative structure toxicity relationship. However, a general overview of QSAR is as follows.

The general protocol followed in QSAR can be summarized in five steps:

1. A set of chemical molecules is selected. These molecules should be capable of interacting with some macromolecular target and its activity should be known.
2. The descriptor should be calculated highlighting important features efficiently and easily.
3. Sets should be divided into training set and testing set.
4. A model is built wherein relations between properties and activities are established. In this step, machine learning, statistical methods, regression analysis, and more are employed.
5. The model is now tested on the testing data set.

QSAR is mainly based on a number of factors such as likeliness of drug and lead compounds and their prediction and calculation based on diversity; similarity in 2D and 3D representations; search of the subset and substructures; filtering of ligand or efficient hits on the basis of physicochemical properties, descriptors, pharmacophores, groups based on toxicity, reactivities, etc.; clustering of compounds, scaffold hopping, and linking of fragments; prediction of ADMET properties; designing of targeted and focused lead compounds; and development of focused libraries. Once this process is completed, the drugs that make it to the final stage are tested and tried. Additional methods that help in the process involve molecular dynamics simulation of biomolecules or target macromolecules, modeling of structures of proteins as well as ligands, and de novo designing or homology modeling. The latter are efficient tools of molecular modeling that form the basis for forming effective databases for support in the infinite world of cheminformatics.

4.3.4 Similarity techniques

There are several similarity techniques that are employed that may be based on several classification tools and methodologies. These searches that look for an effective substructure or 3D pharmacophore include the categorization of a specific problem or query, which is then input to look for a suitable database for identification of compounds that are potential hits. Each of these methods is associated with certain limitations such as restrictions on the number of active compounds shortlisted, which are very few in number. Searches may or may not match the query, and the researcher generally has no control over the number of compounds that will be obtained in the output.

This section, hence, introduces the reader to the similarity methods, which involve searching techniques that are supplementary alternatives to the searching of substructures and pharmacophores. In this method, a query compound is utilized to look for a database of potential hit compounds or leads that are most similar to the query by carrying out a comparison of queries with every shortlisted compound available in the database. The compounds are then listed in order of descending levels of similarity to the query in the database. This method has several advantages in comparison to other methods of searches such as eliminating the requirement for defining a specific substructure or pharmacophore because even one active molecule is sufficient to begin searching for other compounds to be listed in the database. The level of similarity can also be known and hence it is easier to form a list of those compounds that are more active than the others and which qualify to act as potential drugs. It is sometimes done or utilized along with a host of other methodologies such as virtual screening, which is discussed in the next section. The only limitation is that once the level of similarity between two molecules is known, it is difficult to give it a quantified justification. The two elementary steps are calculation of a set of descriptors for comparison of chemical structures and gaining knowledge about quantification of the level of similarity based on the descriptors calculated. The similarity searches commonly carried out are based on fingerprint knowledge, coefficients of similarity, distance coefficients, similarity searches involving subgraphs, reduced graphs, etc. Better pictures are provided by 3D similarity searches such as alignment-independent methods, field-based methods, gnomonic projection methods, and alignments looking for optimal similarities in orientations or conformations. These similarity searches are then compared and evaluated using simulated property prediction, enrichment factors, and hit rates. These methods have improved and evolved drastically in the last three decades, and today, data fusion is carried out, which uses the consensus scoring approach. The approach mentioned is generally used for the docking of ligands to proteins. This technique will now be discussed and explored in the following subsection, which is more concerned with analysis on the basis of the diverse nature of chemical molecules.

4.3.5 Selection of compounds based on diversity analysis

If we expect compounds with similarities in structure to show similar activities, then we look for maximum overlap in the activity space by selecting compounds with diversity in their structures. Such a library has more chances of containing molecules with a range of activities and the number of useless or waste compounds would also be less in that case. One needs such diverse combinatorial libraries for carrying out screening of a wide range of macromolecular targets such as proteins or when hardly anything is known about the active site of the target of interest in a particular research.

It has already been proved that when diversity in the number of compounds increases, then in a particular library the number of hits in assays of biology also increases. But just by increasing the number of different type of molecules, the condition of diversity will not be fulfilled; it is also important to take their

properties into account. Nowadays, these properties with advancement of molecular modeling can be easily studied and compared to experiments too if the data for the latter are available. The need of the hour is not only to achieve diversity but also to remain focused on the active compound that shows the highest potential to act as a drug. A collection of active compounds from a successful pharmaceutical research department may show about one million chemical structures at the research level, and these figures vary from 10^{20} to 10^{60} (Valler and Green, 2000). Diversity analysis offers a technique for exploring the most suitable chemical region so that identification of suitable subsets of chemical molecules for becoming drugs can be done for finally synthesizing, purchasing, and testing the lead compounds.

There are a number of methods for the selection of diverse sets and most of them incorporate the use of molecular descriptors to specify a chemical region. There are many approaches for the selection of diverse sets of chemical compounds but the most commonly used are cell-based methods, dissimilarity-based methods, cluster analysis, and well-known optimization techniques. The first three methods mentioned include a set of protocols of a basic algorithm that comprises generation of descriptors followed by calculation of similarity, which is succeeded by use of a clustering algorithm with the final step of the selection of the principal subset that is formed by selection of one or more chemical moieties from each available cluster. In dissimilarity-based methods, emphasis is placed on calculations of dissimilarities instead of similarities and a final representative subset is created. Cell-based methods are also known as partitioning methods that function with a predesignated lower-dimensional chemical region of space (Mason et al., 1994). The last discussed method is the optimization technique wherein the d-optimal design was one of the initial techniques used (Martin et al., 1995). Most of these techniques in the present-day world revolve around simulation techniques such as Monte Carlo and molecular dynamics with simulated annealing (Hassan et al., 1996; Agrafiotis, 1997). While each of these methods can be discussed extensively, they are beyond the scope of this chapter.

To summarize the selection of diverse compounds to form a representative set, the function must possess the following characteristics: if any useless molecule is added, which has properties similar to an existing molecule, then the diversity should not change. The addition of useful or nonredundant chemical molecules should result in a higher level of diversity. The function should be such that it is in favor of filling space to fill up voids in larger vacant spaces compared to already highly filled regions. For a finite descriptor space, a finite value for diversity function should be obtained when infinite chemical molecular structures are filled. The diversity should escalate when one molecule moves further away from others present in the set but must approach a finite and constant value. This approach matches the protocol followed by Gaussian functions too (Waldman et al., 2000). It is indeed a Herculean task to decide which is the best method suited for a selection of diverse subsets of chemical moieties for the development of a suitable drug in pharmaceutical sciences. Therefore the choice of descriptors, preplanning for choices of the subsets chosen, and finally computational demand must be taken into consideration.

4.4 High-throughput screening and virtual screening

After gaining knowledge of the optimized properties of molecules of importance in the pharmaceutical industry, it is important to screen a huge amount of data of compounds very effectively. HTS systems make it possible to screen thousands of compounds in a small number of days (Hertzberg and Pope, 2000). These runs generate large volumes of huge datasets for efficient analysis. It helps in measuring activities of each sample at a predesignated concentration. This information is then analyzed for identification of a subset of leads that will be processed for a later stage where an extensive protocol is followed for measurement of inhibition of the target protein or biomolecule. The most common method engages the formation of a dose–response curve where activities at varied concentrations are used for the determination of IC_{50} values (concentration of the drug needed to decrease the binding interactions of a small molecule or ligand or, in other words, the rate of change of concentration of a ligand with time by 50%). First, the predissolved liquid samples are analyzed and assays are performed using the samples. Finally, the dose–response curve is used for confirmation with the help of solid samples for which the purity can also be verified. HTS analysis yields innovative methodologies for designing newer compounds for screening in successive iterations that are carried out in the drug-discovery process. As stated earlier, the most potential compounds are selected for the next stage where determination of the number of selected molecules is done using throughput of the following assay.

Bioinformatics and chemoinformatics are crucial for the success of virtual screening of compound libraries, which is an alternative and complementary approach to HTS in the lead discovery process (Tropsha, 2008). A combination of drug-derived building blocks and a restricted set of reaction schemes are key for the automatic development of novel, synthetically feasible structures that can be docked into the active site of a drug target for lead identification using computers, which is the essence of virtual screening (Perola et al., 2000). The virtual screening of combinatorial libraries is used to rationally select compounds for biological in vitro testing from databases of hundreds of thousands of compounds. In addition to descriptors related to structural features such as fingerprints and pharmacophores, the application of relatively simple structural descriptors traditionally used in quantitative structure–activity studies offers speed and efficiency for rapidly measuring the molecular diversity of such collections capable of screening large datasets of organic compounds for potential ligands. Certain filters described in this section are used for computationally prioritizing a suitable candidate from molecular libraries for synthesis and screening of potential compound. In this regard, statistical methods are powerful because they provide a simple way to estimate the properties of the overall system.

Screening methods alleviate the drug development protocols by fastening the search for a true ligand that can efficiently dock the protein structure. HTS, discussed earlier, is a combination of several of the latest technological tools such as

controlling software, optical readers, robotics, and liquid handlers. The gist of the method is that small sets of compounds are tested against the target molecule or a bioassay. This is what we call screening in batches. Those compounds are shortlisted here as "hits," which bind to the target efficiently and are declared to be active. If shortlisted hits are agreeable to the research world of pharmaceutical and medical sciences, then they are tested for toxicity and if found nontoxic, they are developed further to act as potential leads or drugs against a particular target.

Another screening tool that complements the aforementioned process is known as virtual screening, which aids in choosing the right compounds for formation of a suitable subset for screening in HTS. It uses tools of computational chemistry depending on the information available regarding the target and the ligand molecules. A structure-based approach is employed when there is knowledge of the macromolecular target and then the computational tools involve molecular docking followed by the obtaining of subsequent scores of each of the ligand molecules bound to the target biomolecule (DiMasi et al., 2003). When there is knowledge of the activity of the ligand molecule or the smaller drug molecule, then in that case one prefers the ligand-based approach. Structure similarity tools may be used in case of the presence of fewer active compounds. On the other hand, if a lot of active compounds needed to dock the bioassay target are known, then discriminant analysis tools are used. This is done by selecting many ligands, whose activities are known for a specific target molecule, followed by developing appropriate models that predict and discriminate whether the ligand is active or inactive (Leach and Gillet, 2003). The main aim is to finally utilize these combination models in unscreened compounds to select active compounds that are then taken a step further in the research lab.

The main limitation of using computational tools and machine learning tools for such analyses is that they create an imbalance in the final output, in the sense that only one in 1000 inactive compounds proves to be active (Bradley, 2008). The tools that presume a balance are not efficient for predicting models for minority class ratio. Therefore one must keep in mind that the computational tools chosen must cope with this imbalance. This has paved the way for a cost-sensitive classifier. So, imbalanced pharmaceutical data are first subjected to virtual screening with two types of approaches: one with the use of classifiers without misclassification costs (Ehrman et al., 2007) and the other with the use of small datasets to reduce imbalance (Eitrich et al., 2007). A recent PubChem database was made available that made use of naive Bayes classifiers (Chan and Wild, 2009). All such databases and more are discussed in several research publications by putting more emphasis on drug-discovery processes, cost-effective classifiers, and bioassay databases available on cyberspace platforms. The present-day screening tools are primarily concerned with biological activities and other important descriptors of ligand molecules such as ADMET so that the drug-discovery process is faster, more efficient, and the best supplementary tool to serve humankind.

4.5 Combinatorial libraries

Diversity in molecular structure has proved to be a great tool in designing combinatorial libraries. These libraries are very useful for the identification of lead structures and they provide excellent information on the structure that is utilized for optimization of compounds via a pathway through specifically designed libraries with no bias for structure for a particular macromolecular or biological target. When nothing is known about the ligand or the biomolecular target, then extensive searching becomes mandatory. Hence, a library's diversity plays a crucial role because its size should be expanded with every new research. The search must always be extensive. The aim is to maximize the diversity of the library to put emphasis on specific compounds with, say, specific physicochemical properties or electronic or optical properties depending on the requirement of the drug. If the hit of the ligand gives us a promising compound, then optimization libraries could be utilized to bring about modifications in structure, stability, solubility, potency, etc. for yielding an improved drug. It all depends on the comparison of calculations of similarities and dissimilarities between the competing candidates with a potential to prove them as drugs. Some compounds also have the ability to do calculations for other additional binding properties and metabolic properties or toxicity, which proves to be beneficial. Additional data always serve as a tool for incorporation of intuitive compounds by experienced researchers in pharmaceutical and medical sciences.

Usually, a general multidimensional space, also known as a combinatorial building block, is created for each molecule, which is represented by a specific point. Here, similarity is reflected by proximity. Then, subsets are selected from a candidate's (ligand's) larger set that is used for filling the property space with as much efficiency as possible. This is done with the help of mathematical sampling methods. The ligands or candidates can also be differentiated on the basis of their docking abilities (docking into a 3D receptor). The additional categories into which these candidates can be put depend on the need of researchers. To choose selected substituents from different categories, stratified sampling is used. This type of calculation, which is diverse in nature, can be combined with information available in databases virtually to develop combinatorial libraries as per the need of the research world. There are a number of combinatorial library designs from which classification based on degrees of bias is quite popular. The first comprises design based on pure diversity, the second is based on bias of physicochemical properties of drugs that are orally available for use, the third is a library with efficient docking ligands that can be docked into the active site of the macromolecule, and the fourth is an optimization library of lead compounds designed especially to improve the molecular weight of the initial hit by decreasing it, thereby increasing potency.

Therefore it would not be wrong if it were stated that combinatorial library design is an important part of research where the aim is to develop or discover novel small drug molecules or ligands for a macromolecule or a target protein. De novo

design technology also carries significance for library selection and enumeration (Todorov et al., 2007). There are several examples cited in various research articles on the use of de novo design technology.

For more than a century, a combinatorial library has offered one of the best-suited methods to achieve great molecular diversity for the development of newer drugs. An ideal case for which these have been used is that of peptoids, which can be prepared easily with the help of a wide range of readily available primary amines. Screening methods with advancement have led to inexpensive and faster ways of developing specific macromolecular binding ligands from a plethora of available compounds. An example of a combinatorial library is SmiLab (combinatorial libraries using SMILES code) developed by Schuller, which is based on Java script and offers the construction of large combinatorial libraries. Combinatorial libraries can be created using various tools, e.g., ChemDraw (Science Museum Group. ISIS Draw chemical drawing software), ChemFinder (PerkinElmer Informatics), ChemACX (PerkinElmer Informatics), Double Dutch (Roehner et al., 2016), and many more.

4.6 Additional computational tools in cheminformatics: molecular modeling

The basis of all theoretical study comes from the knowledge of the tools of molecular modeling. This helps not only in the study of those compounds that cannot be tested experimentally but provides additional data on the experimental observations, if available. It solves the chemical problems and provides energy for the system along with details of structural, optical, and other physicochemical properties. Inputting the correct basis functions forms the basic framework of calculations in molecular modeling. To carry out molecular calculations successfully it is very important that the choice of basis sets is proper. Molecular modeling incorporates the number of methods. These include molecular mechanics and semiempirical, ab initio, and dynamics simulations (Monte Carlo simulations, molecular dynamics) (Fig. 4.5).

An effective overlap between molecular modeling and cheminformatics is occupied by the requirement to design and develop new chemical molecules that may have applicability in the areas of the pharmaceutical industry, medical sciences, environmental applications, etc. Most of the tools of molecular modeling belong to one of several optimization methods. The computational methodologies employed in molecular modeling bring together a plethora of tools and techniques to develop an altogether new approach. This new approach aids in forming many databases that constitute the study of cheminformatics effectively. Theory and experiments both contribute effectively to the formation of databases. If the databases are perfectly in order, they supplement suitable information to build a model pharmacophore, which gives important keynote features of a whole range of active molecules. In other cases, a 3D structure may be available in sources by virtue of XRD studies, comparative modeling, or NMR studies.

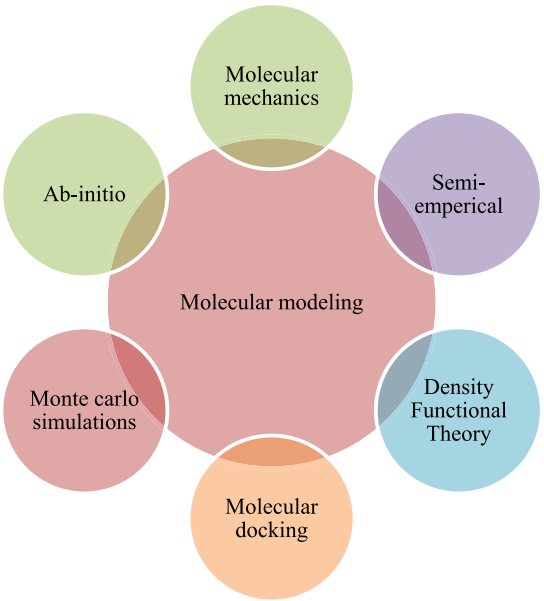


FIGURE 4.5

Methods incorporated in molecular modeling.

Active pharmacophore searching can be carried out in which a wide range of molecules is covered with the help of screening, as discussed earlier. Nowadays, these search methodologies come in a very handy form from the dropdown menu of the toolbars of several types of molecular modeling software. Once this model (pharmacophore) is developed, it gives direction to find newer analogs of active molecules that are capable of becoming efficient drugs. If the molecules are not completely rigid, then their conformational properties must also be taken into consideration. Making use of the concept of ensemble distance geometry, molecular dynamics, clique detection methods, or maximum similarity or likelihood methods can aid in taking conformational properties into account while building up newer molecules of interest. There are many tools that have been under development since the 1990s that make use of both energy minimization and an approach based on knowledge gained for the molecule: CORINA (Gasteiger et al., 1990) and CONCORD (Rusinko et al., 1988) are effective generators of structures and are used for the efficient prediction of structures.

The tools of molecular modeling that are used in combination with other simulation techniques to build up the subject of cheminformatics efficiently are discussed in the proceeding subsections to develop a deeper understanding of the world of optimization.

4.6.1 Molecular mechanics methods

The molecular mechanics method is used in both theoretical and computational works but also forms an important part of experimental studies, which include many X-ray and NMR-derived structures of proteins, nucleic acids, and their complexes, with other molecules deposited in the PDB (Berman et al., 2003) and Nucleic Acid Data Bank (Berman et al., 1992) that are the results of molecular mechanics refinements. Hence, this is a powerful tool with important applicability in drug discovery, as it is able to handle very large systems such as macromolecular targets, proteins, other biomolecules, etc. Use of the approach based on a forcefield tends to ignore the electronic calculations that are very time consuming for macrochemical moieties, and calculates the energy of the systems by taking them as a function of the position of their respective nucleus. In some cases, the forcefield concept has proved to be successful for providing accurate information on even the topmost calculations involving quantum mechanics. The only limitation of the method is that it does not give any information on the properties of the electronic nature of the molecule. All the calculations done in molecular mechanics are based on the presumption that atoms act as a ball and bonds act as a spring, and it is this behavior of ball and spring that is finally assessed.

Several valid assumptions set up the working of the molecular mechanics method involved in dynamic calculations. One such assumption is the famous Born–Oppenheimer approximation. This method is very simple in its approach for interactions occurring within the chemical moiety. The interactions range from simple stretching and bending of bonds to various nonbonded interactions such as the van der Waals interaction and Coulombic interactions. The attribution of transferability and correct parametrization of a forcefield mean that they yield nearly accurate results and take the research related to macromolecules a step further.

4.6.2 Semiempirical methods

These methods are a combination of partial theory, partial experimental verifications, and results obtained from the successful research carried out for the project under study. An easier and simplified version of Hartree–Fock methods of quantum mechanics semiempirical methods paves the way for performing calculations with great time efficiency and fairly accurate range of data because of quantum mechanics as well as empirical data, i.e., data derived from experiments. The methods would obviously be more time consuming in comparison to the molecular mechanics method but the time spent is worthwhile because the results obtained in many cases are comparable to the quantum mechanical data within the range of permissible limits of computational calculations.

The incorporation of values obtained from various experiments has made these approximate semiempirical methods equal to quantum mechanical methods such as ab initio. They are sometimes able to calculate certain properties in a more efficient way in comparison to ab initio methods. Several highly effective semiempirical methods were developed by Pople and Dewar such as NDDO, intermediate neglect

of differential overlap (INDO), and complete neglect of differential overlap (CNDO). The basic framework in this case is mainly based on approximation in the molecular orbital techniques that are again based on the popular Roothaan–Hall equations framework. The theoretical calculations are then combined with the experimental data and result in values that match the level of quantum mechanical calculations while saving a lot of time considering all electronic parameters. Some of the quantum mechanical (Hartree–Fock self-consistent field (SCF)) integrals are ignored during calculations by explicitly taking valence electrons of the system into account for calculations. The inner core electrons are presumed to be merged with the nucleus. But by considering all valence electrons, semiempirical methods turn out to be better than the Hückel molecular orbital approaches that take only pi electrons into consideration for calculating energies of multielectronic systems.

This technique invariable makes use of those orbitals that are orthogonal in nature such that they simplify the equations further for calculations and also make use of basis sets that comprise Slater-type orbitals such as s-, p-, and d-orbitals. Overlap and identity matrix are equated in this technique. Therefore all diagonal elements in the overlap matrix are one and off diagonal elements are simply equal to zero. Based on the type of approximation employed in the overlap between orbitals, semiempirical methods include ZDO, NDDO, CNDO, INDO, MINDO, MINDO/3, SAM1, PM3, AM1, and EHT. Each of these methods is discussed extensively in several books that deal with the study of molecular modeling (Leach, 2009; Cramers, 2013; Lewar, 2003; Hinchliffe, 2003).

4.6.3 Ab initio methods

The most expensive of all methods is ab initio because it is based entirely on the principles of quantum mechanics and validated approximations for calculations of energy and other structural, optical, and physicochemical properties of relatively smaller molecules that may act as a ligand or a drug. This method was highly time consuming until almost a decade ago, and with the advent of drastic improvements in efficiencies of hardware and software that are easy to use, the calculations that employ ab initio methodology seem easier and hence it has wider usage in the modern-day world. These calculations are also based on the Hartree–Fock SCF methodology that evaluates integrals correctly after manipulating them. These calculations are appropriate for performing calculations of the energy of the ground state of smaller/moderate-sized organic systems. The Roothaan–Hall equations are also employed with Hartree–Fock theory, which can be either spin restricted or spin unrestricted (Pople and Nesbet, 1954). These calculations can be done on closed shell systems (with no unpaired electrons) or open shell systems (with at least one unpaired electron). In the former, electron distribution is assumed to be zero throughout the system due to the pairing of all electrons, whereas in the latter, extra electron spin expressed in terms of spin density results in absorbing electronic density into calculations. Configuration interaction, electron correlation, and

perturbation theories are some of the approaches used in solving the energy of a chemical system using ab initio methodology.

When the nuclei are free to move for the calculation of energy, the process becomes tedious and hence highly time consuming. To make the process easier, different stages of theoretical calculations need to be employed at different levels of calculation. Tricks and tactics are to be employed judiciously to reduce the burden of calculations. One such technique involves carrying out optimization using ab initio lower basis set methodology and then carrying out further calculations of properties by giving a single point run at a higher basis set. Remember, here the choice of basis set for any ab initio calculation will matter a great deal as this is what decides the absolute accuracy of the results obtained. The SCF method used in molecular modeling in itself is direct or converged. As the value may be underestimated in earlier cases, in the case of ab initio, there are chances of the results being overestimated because of basis set superposition error. This error factor is attributed to the fall in energy of the overall system when two or more atoms come closer to each other due to favorable interactions between them, and the basis sets give more information about the electronic atmosphere around the molecule.

4.6.4 Density functional theory

This is one of the most extensively used tools for optimization, which is not only time efficient but has also proved its mettle in giving accurate results that are comparable to real-world data. It is based on the calculations of the electronic structure of chemical systems (Parr, 1983; Wimmer, 1997). It considers density of the electron as a functional for calculations of energy and other properties and hence is less time consuming than the elaborate ab initio method. It also takes into account single electron functions. Density functional theory (DFT) unlike the last method discussed does not calculate the entire "n" electron wave function but makes an attempt to evaluate the electronic energy of the entire system along with the density distribution of electrons. DFT is based on the relation between the former and the latter, i.e., electronic energy evaluated for the entire system and overall density distribution of electrons in the system. It became popular only after research proved that energy of the ground state and other important properties of a system could be defined on the basis of electron density completely (Hohenberg and Kohn, 1964). Hence, mathematically, energy is equated to the functional of electronic density dependent on the distance of the electron from the nucleus. It makes use of a number of mathematical and quantum mechanical concepts wherein contributions from interactions between different electrons, variational approach, Lagrangian multiplier, etc., are used. An equation is developed in DFT that is equivalent to the quantum mechanical Schrodinger wave equation. The popularity of this optimization methodology reached its zenith with the publishing of another pioneering paper by Kohn and Sham (Kohn and Sham, 1965). The equations developed by these authors for the calculation of energies and other properties gave proper definition to correlation and exchange functionals, where not only did the correlation and exchange

contributions matter but also contributions from the difference between the real kinetic energy of the chemical moiety and the entire energy of multielectronic systems containing "*n*" electrons are taken into account. To solve the equations developed in their paper, the approach used is based on the self-consistent method. A trial value is initially fed into the Kohn–Sham equations, which yields orbitals that give a better picture of the value of density. This is then followed by successive iterations until the calculations converge. DFT has now been extended to local spin density DFT where not only electron density but also spin density are considered to be fundamental quantities, with spin density being calculated as the difference between the up and down spin densities of electrons. One of the most important reasons for the extensive use of this method is the fact that even if a simple approximation is applied to the exchange correlation functional, the results obtained are favorable. Here, one must not forget that it owes its accuracy to quantum Monte Carlo methods (Ceperley and Alder, 1980) as these approaches are employed for calculation of all densities of interest in a particular research. The method has now been extended from local density approximations to gradient generalized approximations or gradient corrected functionals, and even hybrid Hartree–Fock methods of DFT.

DFT has paved the way for faster analogs to be more competitive with sometimes even superior values in comparison to ab initio techniques. The choice of basis sets is as important here as it was in the earlier case of full quantum mechanical calculations (Baboul et al., 1999). The gradients of energy calculated by also taking nuclear coordinates into account are one of the most important achievements of DFT in practical use.

4.6.5 Molecular dynamics

This is a simulation tool used for creating a real-time environment on screen for chemical systems. Integration of the equation of Newton's laws of motion is carried out to produce successive configurations of the chemical system under study. It yields a trajectory that shows how velocities and positions vary with time for particles in any chemical system. There are different cases that are often considered to apply Newton's laws of motion. One case is where the particles undergoing collision are under no force impact. Another case is where a constant force is believed to be acting on colliding particles. And a third case is where the force acts on the particle depending on its location relative to the location of the other particles in the system. Dynamics, when done with continuous potentials, includes finite difference techniques, predictor–corrector integrator models, multiple time step dynamics, etc. When it is finally set and made to run, constraint dynamics is also applied and temperature is calculated.

The models used in molecular dynamics range from simple to complex. Calculations done for properties that evolve with time include correlation functions,

various transport properties that refer to the flow of the substance from one point in space to another. Also, constant temperature and constant pressure dynamics may also be carried out while simultaneously incorporating the effect of solvents on these dynamics calculations. Due to varied aspects involved in these calculations, molecular dynamics tools not only help to obtain a better picture of energy of the chemical systems when put in a real environment, but also aid in carrying out conformational analysis. For smaller timescale calculations, atomistic simulations are done, and for systems demanding evaluations to be done for longer timespans, rather simpler models may be employed. For those that occur in intermediate timescales, mesoscale dynamics modeling may be employed in the form of dissipative particle dynamics. The latter refers to the super-quick movement of atoms, which is integrated and the left out basic units of beads interact with others through a suitable potential application (Koelman and Hoogerbrugge, 1993). The bead here refers to the smallest unit of the fluid. Force acting on each bead is a result of all dissipative forces and interactions of this basic unit with the rest in the system. Molecular dynamics simulations based on periodic box models overcome the barriers laid down by the primitive potentials, and their extension into Langevin and Brownian dynamics has proved significant in carrying out calculations of energy and other physicochemical properties easily by understanding of simple equations that are extensions of the equations of Newton's laws of motion.

4.6.6 Monte Carlo simulations

The birth of Monte Carlo simulations was marked by a serious note when it was established as a computer simulations tool in the form of a technique purely based on a statistical mechanical approach over the entire configuration space. Initially, it was done using the method of importance sampling and is now extended to metropolis Monte Carlo simulations involving random sampling where a correction is offered to the earlier model where the majority of the phase space is concerned with highly energetic configurations that are nonphysical. The Boltzmann factor has a substantial value only for a very small region of space considered. The box in this case is also considered to be periodic. These simulations have been successfully employed for rigid molecules as well as flexible molecules.

4.6.6.1 Importance of molecular dynamics simulations

Molecular dynamics simulations are used for the computation of balanced states at equilibrium and the transportation characteristics of multiple-body chemical systems. It is based on classical mechanical laws and is applicable to a wide range of chemical substances. It represents chemical systems in real-time environments wherein the observables are calculated over a chosen time range and interval. The longer the span is averaged out, the more accurate the measurements will be. To calculate properties of interest in these simulations, one must be able to express these properties as a function of its exact position in 3D space and momenta of the system.

4.6.6.2 Contrast between molecular dynamics simulations and Monte Carlo simulations

Monte Carlo simulations are based on random sampling techniques and also provide a real-time environment for molecular problems. The modeling of these is based on the principles of molecular dynamics simulations with the only difference being its approach, which depends on statistical mechanics at equilibrium that involves calculations using Boltzmann distribution. It is also popularly known as the metropolis Monte Carlo simulation technique. Another advantage of using Monte Carlo over molecular dynamics simulations is that the former can be used to model chemical systems that are to be defined on the basis of energy prescriptions. These simulations have further paved the way for developments of innovative methods for optimization such as simulated annealing.

Several software packages have been exclusively focused on use of the Monte Carlo metropolis algorithm, some of which are: Faunus (https://mlund.github.io/faunus/), ProtoMS (http://www.essexgroup.soton.ac.uk/ProtoMS/), Sire (https://siremol.org/), MCPro (Jorgenson and Tirado-Rives, 2005; Jorgenson, 1998), CP2K (https://www.cp2k.org/howto:gemc), and BOSS (Jorgenson and Tirado-Rives, 2005; Jorgenson, 1998).

4.6.7 Molecular docking

This is an important tool in molecular modeling, which helps in predicting the successful interaction of a ligand molecule with a biomolecular target. It also helps in predicting the correct orientation of the docking molecule with respect to the target so that the two can result in the formation of a stable complex. Macromolecules or biological targets include a wide range of chemical systems ranging from nucleic acids, proteins, carbohydrates, lipids, and peptides to supramolecules of the chemical world. This is the most important tool of molecular modeling, which aids in drug designing and hence marks its importance in the fields of pharmaceutical and medicinal sciences. The set of ligands shortlisted for docking through screening methodologies results in formation of prodigious databases for use by those who are into deep cheminformatics. Binding affinity has a significant role in the characterization of pairs of chemical and biological systems to expound basic processes involved in the field of biochemistry. It works on a principle analogous to the lock and key mechanism where the macromolecule is the lock and the drug in the form of a ligand molecule is the key. During interactions between the two, a best fit orientation results after appropriate conformational adjustments, also known as induced fit.

This technique is principally based on simulations of computational tools for the recognition of the correct molecules to achieve minimized conformation for the combination system of macromolecule and ligand in such a way that results in the minimization of the free energy of the entire system under research. Docking approaches also vary. One utilizes a match-making method wherein the ligand and its protein are considered as complementary surfaces to each other and another emphasizes simulation of the real docking method. The latter incorporates

calculation of the ligand−protein duo pairwise. But like all other methods, the advantages of these docking tools are accompanied by a few limitations as well. While the first approach is amenable to approaches based on featured pharmacophores as they utilize molecular descriptors of ligands to judge minimal binding to the target, the second focuses on the flexibility of the ligand, which makes it more real but highly time consuming because of the coverage of larger grid spaces. All these and more advancements have made docking a realistic tool for the formation of suitable databases for effective use in cheminformatics.

4.7 Conclusions

The chapter began with teaching methods suited for representations of structures and reactions in both 2D and 3D formats, which laid down the foundation stone of the basic purpose of this work. If representations are made properly, then poststorage search and recovery becomes easy. Identification of molecules with similar properties is done using database searching tools. Molecular descriptors barring a few are purely computational with values that can be predicted for new molecules that have yet to be discovered. There are several mathematical models that are also used for deriving QSAR or QSPR models effectively such as multiple linear regression or the partial least squares method. A complementary tool for pharmacophore searching or substructure searching is the similarity method. There were four major approaches discussed for selection tools for diverse subsets of chemical compounds, which were optimization tools, cell-based method, dissimilarity-based method, and cluster-based method. One must keep in mind that the descriptors chosen should have relevance to biological activity for the compound to be suitable for pharmaceutical application. The growing number of large-sized databases from HTS evaluates and tests the central ideology of similarity searches too. The number of chemical systems that can be calculated by making use of virtual screening keeps increasing with the development of new programs and advanced hardware. Accuracy and reliability are the keynote issues for making sure that the scores of molecular docking and models built for the prediction of ADMET properties are successful. Now with the rapidly increasing pace of experimental research with modernized real tools, computation tools for designing libraries have also been matched. Maintaining a balance between different features for selecting reagents for real lab synthesis is mandatory but one must not lose sight of the factors for “diverse” molecules and “focus” for the right hit.

The work presented in this chapter was by no means complete in this ever-diverging research world of converged analysis. The overview provided in the chapter placed emphasis on the important tools of cheminformatics that proved to be well organized for pharmaceutical data analysis and applications. Each of the tools described in the chapter has been extensively discussed in many research publications but this work gave a complete overview of the main tools and techniques that were able to aid in the study of many important applications such as

drug-discovery processes and other biochemical applications. The link between molecular modeling with cheminformatics is of great significance and cannot be ignored. Hence, the various optimization tools of molecular modeling act as "add-ons" to the main aim of discovering a potential drug or an active lead hit compound. Ranging from representations to searching databases, descriptors to QSAR analysis, similarity searching to deriving pharmacophores, and shortlisting datasets to forming combinatorial libraries, cheminformatics makes full use of the latest advancements of technologies available in the vast world of cyberspace. One must always be aware of the latest tools and techniques of cheminformatics that have emerged in recent years to make sure that the knowledge of cheminformatics when made to run with its analog in biology, i.e., bioinformatics, serves as a guiding principle for the development of newer molecules for widespread use as an application of pharmaceutical sciences and society at large.

References

Accelrys, 2016. Materials Studio. http://accelrys.com/products/collaborative-science/biovia-materials-studio/.

ACD/ChemSketch, version 2020.1.2, 2020. Advanced Chemistry Development, Inc., Toronto, ON, Canada. www.acdlabs.com.

Agrafiotis, D.K., 1997. Stochastic algorithms for maximising molecular diversity. J. Chem. Inf. Comput. Sci. 37, 841−851.

Anderson, E., Veith, G.D., Weininger, D., 1987. SMILES: A Line Notation and Computerized Interpreter for Chemical Structures. Report No. EPA/600/M-87/021. U.S. Environmental Protection Agency, Environmental Research Laboratory-Duluth, Duluth, MN 55804.

Arulmozhi, V., Rajesh, R., 2011. Chemoinformatics - A Quick Review. IEEE, 978-1-4244-8679-3/11.

Baboul, A.G., Curtiss, L.A., Redfern, P.C., Raghavachari, K., 1999. Gaussian-3 theory using density functional geometries and zero-point energies. J. Chem. Phys. 110, 7650−7657.

Balaban, A.T., 1997. From Chemical Topology to Three Dimensional Geometry. Plenum Press, New York, pp. 1−24.

Bender, A., Hamse, Y., Mussa, H.Y., Glen, C., 2010. Similarity searching of chemical databases using atom environment descriptors (Molprint 2D) evaluation of performance. J. Chem. Inf. Comput. Sci. 44, 1708−1718.

Berman, H.M., Olson, W.K., Beveridge, D.l., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A.R., Schneider, B., 1992. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J. 63, 751.

Berman, H.M., Henrick, K., Nakamura, H., 2003. Announcing the worldwide protein data bank. Nat. Struct. Mol. Biol. 10, 980.

Bharati, D., Jagtap, R.S., Kanase, K.G., Sonawame, S.A., Undale, V.R., Bhosale, A.,V., January - March, 2009. Chemoinformatics: newer approach for drug development. Asian J. Res. Chem. 2 (1). ISSN 0974-4169.

Bradley, D., 2008. Dealing with a data dilemma. Nat. Rev. Drug Discov. 7, 632−633.

Ceperley, D.M., Alder, B.J., 1980. Ground state of the electron gas by a stochastic method. Phys. Rev. Lett. 45, 566–569.
Chemcraft - Graphical Software for Visualization of Quantum Chemistry Computations. https://www.chemcraftprog.com.
Chen, B., Wild, D.J., 2009. PubChem BioAssays as a data source for predictive models. J. Mole. Graph. Model. 28 (5), 420–426.
Cortes-Cabrera, A., Morris, G.M., Finn, P.W., Morreale, A., Gago, F., 2013. Comparison of ultra fast 2D and 3D descriptors for side effect prediction and network analysis in polypharmacology. Br. J. Pharmacol. 170 (3), 557–567.
Cramer, C.J., 2013. Essentials of Computational Chemistry: Theories and Models, second ed. John Wiley and Sons Ltd, England.
Deursen, R., Blum-Lorenz, C.B., Reymond, J.L., 2010. A searchable map of PubChem. J. Chem. Inf. Model. 50 (11), 1924–1934.
DiMasi, J.A., Hansen, R.W., Grabowski, H.G., 2003. The price of innovations: new estimates of drug development costs. J. Health Econ. 22, 151–185.
Ehrman, T.M., Barlow, D.J., Hylands, J., 2007. Virtual screening of Chinese herbs with random Forest. J. Chem. Inf. Model. 47 (2), 264–278.
Eitrich, T., Kless, A., Druska, C., Meyer, W., Grotendorst, J., 2007. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. J. Chem. Inf. Model. 47, 92–103.
Engel, T., 2006. Basic overview of chemoinformatics. J. Chem. Inf. Model. 46, 2267–2277.
Engel, T., Gasteiger, J., 2018. Chemoinformatics: Basic Concepts and Methods. Wiley-VCH Verlag GmbH & Co, Germany.
Enoch, S.J., 2010. The use of quantum mechanics derived descriptors in computational toxicology. In: Puzyn, T. (Ed.), Challenges and Advances in Computational Chemistry and Physics, vol. 8. Springer Science, pp. 24–27.
Faller, B., Ertl, P., 2007. Computational approaches to determine drug solubility. Adv. Drug Deliv. Rev. 59, 533–545.
Faulon, J.L., Bender, A., 2010. Handbook of Chemoinformatics Algorithms. CRC Press, New York.
Froimowitz, M., 1993. HyperChem: a software package for computational chemistry and molecular modeling. Biotech. 14 (6), 1010–1013.
Garcia, E.J., Pellitero, P.J., Jallut, C., Pirngruber, G.D., 2013. Modeling adsorption properties on the basis of microscopic, molecular structural descriptors for non polar adsorbents. Langmuir 29 (30), 9398–9409.
Gasteiger, J., Rudolph, C., Sadowski, J., 1990. Automatic generation of 3D atomic coordinates for organic molecules. Tetrahedron Comput. Methodol. 3, 537–547.
Hassan, M., Bielawski, J.P., Hempel, J.C., Waldman, M., 1996. Optimisation and visualisation of molecular diversity of combinatorial libraries. Mol. Divers. 2, 64–74.
Hertzberg, R.P., Pope, A.J., 2000. High-throughput screening: new technology for the 21st century. Curr. Opin. Chem. Biol. 4, 445–451.
Hinchliffe, A., 2003. Molecular Modeling for Beginners. John Wiley and Sons, England.
Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., Zell, A.J., 2011. Compound Mapper: an open source JAVA library and command line tool for chemical fingerprints. J. Chemoinfor- m. 3, 3.
Hohenberg, P., Kohn, W., 1964. Inhomogeneous electron gas. Phys. Rev. B 136, B864–B871.
Humphrey, W., Dalke, A., Schulten, K., 1996. VMD - visual molecular dynamics. J. Mol. Graph. 14, 33–38. http://www.ks.uiuc.edu/Research/vmd/.

Hunter, R.S., Culver, F.D., Fitzgerald, A., 1987. SMILES User Manual. A Simplified Molecular Input Line Entry System. Includes Extended SMILES for Defining Fragments. Review Draft, Internal Report. Montana state university, institute for biological and chemical process control (IPA), Bozeman, MT.

ISISdraw/Biovia Draw, MDL Information Systems/dassault Systems.

James, C.A., Weininger, D., Delany, J., 2002. Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/theory.toc.html.

Jorgensen, W.L., 1998. In: Schleyer, P.v.R. (Ed.), BOSS - Biochemical and Organic Simulation System, the Encyclopedia of Computational Chemistry, vol. 5. John Wiley & Sons Ltd, Athens, USA, pp. 3281−3285.

Jorgensen, W.L., Tirado-Rives, J., 2005. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. J. Comput. Chem. 26, 1689−1700.

Jorgenson, W.L., Duffy, E.M., 2002. Prediction of drug solubility from structure. Adv. Drug. Deliv. Rev. 54, 355−366.

Karelson, M., 2000. Molecular Descriptors in QSAR/QSPR. Wiley.

Karelson, M., Lobanov, V., Katritzky, A.R., 1996. Quantum chemical descriptors in QSAR/QSPR studies. Chem. Rev. 96, 1027−1043.

Koelman, J.M.V.A., Hoogerbrugee, P.J., 1993. Dynamic simulations of hard-sphere suspensions under steady shear. Europhys. Lett. 21 (3), 363−368.

Kohn, W., Sham, J.L., 1965. Self-Consistent equations including exchange and correlation effect. Phys. Rev. A 140. A1133-113.

Krause, S., Willighagen, E., Steinbeck, C., 2000. JChemPaint - using the collaborative forces of the internet to develop a free editor for 2D chemical structures. Molecules 5, 93−98.

Leach, A.R., 2009. Molecular Modeling: Principles and Applications, second ed. Pearson education limited.

Leach, A.R., Gillet, V.J., 2003. An Introduction to Cheminformatics. Kluwer Academic Publishers, The Netherlands, Dordrecht.

Leach, A.R., Gillet, V.J., 2007. An Introduction to Cheminformatics, Revised Edition. Springer, Netherlands.

Lewars, E.G., 2003. Computational Chemistry: An Introduction to the Theory and Applications of Molecular and Quantum Mechanics, second ed. Springer.

Livingstone, D.J., Waterbeemd, V.D., Han, I., 2009. In silico prediction of human oral bioavail-
ability. Methods Princ Med. Chem. 40, 433−451.

Ma, S.L., Joung, J.Y., Lee, S., Cho, K.H., No, K.T., 2012. PXR ligand classification model with SFED weighted WHIM and CoMMA descriptors. SAR QSAR Environ. Res. 23 (5−6), 485−504.

Macrae, C.F., Sovago, I., Cottrell, S.J., Galek, P.T.A., McCabe, P., Pidcock, E., Platings, M., Shields, G.P., Stevens, J.S., Towler, M., Wood, P.A., 2020. Mercury 4.0: from visualization to analysis, design and prediction. J. Appl. Crystallogr. 53, 226−235.

Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., Moos, W.H., 1995. Measuring diversity: experimental design of combinatorial libraries for drug discovery. J. Med. Chem. 38, 1431−1436.

MarvinSketch (Version 6.2.2 , Calculation Module Developed by ChemAxon, 2014. http://www.chemaxon.com/products/marvin/marvinsketch/.

Mason, J.S., McLay, I.M., Lewis, R.A., 1994. Applications of computer-aided drug design techniques to lead generation. In: Dean, D.M., Jolles, G., Newton, C.G. (Eds.), New Perspectives in Drug Design. Academic Press, London, pp. 225−253.

Molecular Operating Environment (MOE), January, 2019. Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, p. 2020.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J., 2009. Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J. Comput. Chem. 16, 2785−2791.

Nuzillard, J.M., Bertrand, P., 2017. Tutorial for the structure elucidation of small molecules by means of the LSD software. Magn. Reson. Chem. 56 https://doi.org/10.1002/mrc.4612. MRC.

Parr, R.G., 1983. Density functional theory. Annu. Rev. Phys. Chem. 34, 631−656.

Perola, E., Xu, K., Kollmeyer, T.M., Kaufmann, H.S., Prendergast, F.G., Pang, Y.-P., 2000. Successful virtual screening of a chemical database of farnasyl transferase inhibitor leads. J. Med. Chem. 43 (3), 401−408.

Persson, L.C., Porter, C.J., Charman, W.N., Bergstrom, C.A., 2013. Computational prediction of drug solubility in lipid based formulation excipients. Pharm. Res. PMID:23771564.

Pople, J.A., Nesbet, R.K., 1954. Self-consistent orbitals for radicals. J. Chem. Phys. 22, 571−572.

Rice, B.M., Byrd, E.F., 2013. Evaluation of Electrostatic Descriptors for Crystalline Density. Langmuir.

Roehner, N., Youmg, E.M., Voigt, C.A., Gordon, D.B., Densmore, D., 2016. Double Dutch: a tool for designing combinatorial libraries of biologicalsystems. ACS Synth. Biol. 5, 507−517.

Rogers, D., Mathew, H., 2010. Extended connectivity fingerprints. J. Chem. Inf. Model. 50 (5), 742−754.

Rusinko III, A., Skell, J.M., Balducci, R., McGarity, C.M., Pearlman, R.S., 1988. CONCORD: A Program for the Rapid Generation of High Quality 3D Molecular Structures. The University of Texas at Austin and Tripos Associates, St Louis, MO.

Shao, Y., Molnar, L.F., Jung, Y., Kussmann, J., Ochsenfeld, C., Brown, S.T., 2006. Phys. Chem. Chem. Phys. 8, 3172.

Stanton, D., 1999. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. J. Chem. Inf. Comput. Sci. 39 (1), 11−20.

Stein, S.E., Heller, S.R., Tchekhovskoi, D., 2003. An open standard for chemical structure representation: the IUPAC chemical identifier. In: Proceedings of the 2003 International Chemical Information Conference. Nimes, France, October 19−22. Infonortics, Tetbury, UK, pp. 131−143.

Swain, M., 2012. Chemicalize.org. J. Chem. Inf. Model. 52 (2), 613−615.

Thompson, M.A., 2004. Molecular Docking Using ArgusLab, an Efficient Shape-Based Search Algorithm and the a Score Scoring Function. ACS Meeting, Philadelphia.

Todeschini, R., Consonni, V., 2009. Molecular Descriptors for Chemoinformatics (2 Volumes). Wiley-VCH Verlag GmbH & Co. KGaA, Germany.

Todeschini, R., Bettiol, C., Giurin, G., Gramatica, P., Miana, P., Argese, E., 1996. Modeling and prediction by using WHIM descriptors in QSAR studies: submitochondrial particles(SMP) as toxicity biosensors of chlorophenols. Chemosphere 33, 71−79.

Todorov, N.P., Alberts, I.L., Dean, P.M., 2007. Comprehensive medicinal chemistry II. Comp. Assist. Drug Design, De Novo Design 4 (13), 283−305.

Todsen, W.L., 2014. ChemDoodle 6.0. J. Chem. Inf. Model. 54 (8), 2391−2393.

Tropsha, A., 2008. Integrated chemo and bioinformatics approaches to virtual screening. In: Tropsha, A., Varnek, A. (Eds.), Cheminformatics Approaches to Virtual Screening. SC Publishing, pp. 295−325.

Valler, M.J., Green, D., 2000. Diversity screening versus focussed screening in drug discovery. Drug Discov. Today 5 (7), 286–293.
Waldman, M., Li, H., Hassan, M., 2000. Novel algorithms for the optimization of molecular diversity of combinatorial libraries. J. Mol. Graph. Model. 18, 412–426.
Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36.
Weininger, D., Weininger, A., Weininger, J.L., 1989. SMILES. 2. Algorithm for generation of unique SMILES notation. J. Chem. Inf. Comput. Sci. 29, 97–101.
Wimmer, E., 1997. Electronic structure methods. In: Catlow, C.R.A., Cheetham, A.K. (Eds.), New Trend in Material Chemistry, NATO ASI Series, C 498. Kluwer, Dordrecht.

Further reading

Gasteiger, J., Engel, T., 2003. Cheminformatics: A Textbook. Wiley-VCH Verlag GmbH & Co, Germany.
Kelder, J., Wagener, M., Timmer, M., 2004. Cheminformatics and Drug Design. Chapter 5. Cheminformatics I, Organon, N.V. The Netherland, pp. 111–127.
Olsson, T., Oprea, T., 2001. Cheminformatics: a tool for decision-makers in drug discovery. Curr. Opin. Drug Discov. Dev. 4 (3), 308–313.
Stefaniu, A., 2020. Cheminformatics and its Application. Intech Open Access Peer Reviewed Edited Volume, 978-1-83880-068-0.

CHAPTER

5 Structure-based drug designing strategy to inhibit protein-protein-interactions using *in silico* tools

Kailas D. Sonawane[1,2], V.G. Shanmuga Priya[3]

[1]*Structural Bioinformatics Unit, Department of Biochemistry, Shivaji University, Kolhapur-416 004, Maharashtra, India;* [2]*Department of Microbiology, Shivaji University, Kolhapur-416 004, Maharashtra, India;* [3]*Department of Biotechnology, KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belagavi, Karnataka, India*

5.1 Introduction

Most proteins interact directly with other protein(s) to carry out their functions, both inside and outside the cells. Such protein−protein interactions (PPIs) occur during various biological processes like metabolic pathways, antibody activity, cell-to-cell interactions, and cell developmental process control (Braun and Gingras, 2012). Experimental methods like the yeast two-hybrid system and affinity purification-coupled mass spectrometry are largely used to study interacting proteins (De Las Rivas and Fontanillo, 2010). Other techniques like co-immunoprecipitation, analytical ultracentrifugation, fluorescence spectroscopy, luminescence-based mammalian interactome mapping, protein-fragment complementation assay, surface plasmon resonance, and calorimetry are also engaged in identifying PPIs. Many computational methods based on genomic context, text mining, and machine learning are now extensively used for this purpose. The Rosetta Stone approach, conserved neighborhood method, and phylogenetic profiling are genome context-related methods (Marcotte, 1999; Raman, 2010). Text mining is a much less costly method, which generally detects binary relations between interacting proteins from individual sentences using rule/pattern-based information extraction methods from the literature (Badal et al., 2015). Machine learning methods like Random forest distinguish interacting protein pairs from others based on features such as cellular co-localization, gene co-expression, close location of genes on DNA, and so on (Qi et al., 2006). X-ray crystallography is presently the important technique to study the mode of interaction in a protein−protein complex at the atomic level. Also, nuclear magnetic resonance (NMR) spectroscopy, electron microscopy, and fluorescence resonance energy transfer are used for this purpose. Analysis of the structural

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00010-5

information derived from these techniques has been the inducing factor for the computational design of PPI modulators which are either inhibitors or stabilizing agents (Villoutreix et al., 2014).

Blocking PPIs using small molecules or peptides has been in practice for functional annotations of proteins. The idea of using PPIs as drug targets and modulating biochemical pathways for therapeutic significance has surfaced recently and today they are used as drug targets in various therapeutic fields like cancer (Li et al., 2016), heart failure and inflammation (Anand et al., 2013), neurological disorders (Hayes et al., 2017), tropical infectious diseases (Dawidowski et al., 2017), and oxidative stress (Lu et al., 2016; Ran and Gestwicki, 2018). One example of this success story is the ABT-199 compound derived from the fragment-based screening method for inhibition of Bcl-2 regulatory proteins (Oltersdorf et al., 2005) for the treatment of chronic lymphocytic leukemia.

PPI inhibitors can inhibit PPIs by binding either to the interface (orthosteric inhibition) or to a distal site (allosteric inhibition) (Cesa et al., 2015). The molecules are said to act either by hindering the formation of a protein complex or by destabilizing PPIs.

This chapter deals with a structure-based drug designing strategy for screening small molecules for orthosteric inhibition of PPIs and about the pharmacokinetic properties of these inhibitors. It reviews the databases that can be considered for exploring PPI interactions and their modulators. Here, transcription factors (TFs) are highlighted as one of the PPI drug targets and is dealt along with a case study. Finally, few *in silico* tools that can be used for studying PPI inhibition are discussed along with their algorithms, working procedures, and result analyses.

5.2 Methods to identify inhibitors of PPIs

The general means for identifying small-molecule inhibitors of PPIs include high-throughput screening (HTS) and computational techniques like fragment-based drug discovery, peptide-based drug discovery, and protein secondary structure mimetics (Meireles and Mustata, 2011). In HTS technologies, numerous compounds are screened in a relatively short period; however, HTS has been more effective for conventional targets like enzymes and receptors and less effective for PPI targets (Macarron, 2006). In fragment-based drug discovery, a small-molecular fragment library is screened initially against the given target. Fragments that bind at the required sites are made to have higher affinity using techniques like linking two fragments and growing fragments that are then developed into compounds (Winter et al., 2012; Coyne et al., 2010). Peptides are preferred for PPI inhibition and peptidomimetics are designed based on the knowledge of natural peptide ligands; Peptidomimetics are designed to circumvent some of the problems associated with a natural peptides like digestion by proteases and poor bioavailability (Eguchi et al., 2003). However, the problem of proteases is not completely resolved. Recently, using the peptidomimetics approach, small-molecule compounds resembling the

spatial arrangement of protein secondary structures which are involved in PPI, are designed or screened to act as competitive inhibitors (Marshall et al., 2009).

However, identification of small molecules that inhibit PPIs has many challenges to overcome. They include the absence of catalytic activity to screen and functional assays to monitor PPIs, the existence of various types of PPI modes like stable and transient, covalent, and noncovalent interactions and their different types of interfaces, and the size and character of typical small-molecule libraries. However, many of these problems are taken care of by advancements in molecular biology and computational modeling techniques (Arkin and Wells, 2004).

5.3 Nature of the PPI interface

The primary requirement in the phase of computational development of PPI modulators is to understand the PPI interface. Based on this knowledge, the inhibitors are either designed or screened. The structures of protein−protein complexes deposited in databases and the literature facilitate an understanding of the PPI interfaces. The interface between two proteins usually has an area of 1500−3000 $Å^2$ with ∼750−1500 $Å^2$ of surface area buried in each protein (Conte et al., 1999). The protein−protein complexes are stabilized by desolvation energy and van der Waals interactions (Fernandez and Scheraga, 2002). Besides hydrophobic interactions, electrostatic forces largely support complex formation and also determine the lifetime of protein complexes (Kundrotas and Alexov, 2006). In some interfaces, hydrogen bonding is responsible for interaction. It has been found that one hydrogen bond is present per 100−200 Å surface area in many complexes (Jones and Thornton, 1997). Moreover, all the residues on the interface between the proteins do not contribute equally to PPIs. A few residues donate more to the binding free energy and are called hot spots.

Regions having hot spots are seen more at the center of binding interfaces, which are small or medium in size (250−900 $Å^2$). A hot spot residue is identified as a residue that when substituted by an amino acid (e.g., alanine) shows a noteworthy increase in free energy of binding (generally >1.5 kcal/mol). Mutagenesis and structural studies show that small molecules targeted against these hot spots can modulate or inhibit PPIs (Kuenemann et al., 2015). To determine hot spots by experimental methods is a time-consuming and tiresome process and hence computational algorithms are developed for this. Alanine scanning analysis data indicate that hot spot residues are generally tryptophan (Trp), arginine (Arg), tyrosine (Try), leucine (Leu), isoleucine (Ile), and phenylalanine (Phe) (Bogan and Thorn, 1998). Energetic hot spots from alanine scanning correlate with structurally conserved residues in proteins. Of these, Trp and Tyr are involved in hydrophobic π-interactions and also form hydrogen bonds in the PPI region. The residue Arg generally forms hydrophobic interactions, salt bridges, and hydrogen bonds (Ma et al., 2003). In a protein-−protein complex, for hot spots that are spread over an area and are not

continuous, small molecules of larger size are preferred as PPI inhibitors and for PPIs with continuous hotspots, relatively smaller size molecules are preferred (Sable and Jois, 2015).

5.4 Computational drug designing

Today, *in silico* works are carried out to understand the molecular mechanism of diseases (Barale et al., 2019) and many *in vitro* and *in vivo* experiments are complemented with computational analysis and molecular simulations at different stages of the drug discovery process to save time and money. This in silico drug designing intends to select or design molecules that have complementary shape, charge, hydrogen bond donor−acceptor groups, along with hydrophobic, van der Waals, or electrostatic attraction towards the biomolecular target and thus can specifically bind to it with high affinity. Several molecular docking and molecular dynamics simulation studies have been carried out to understand the structure−function relationship of enzyme−ligand, protein−peptide, peptide−peptide, as well as PPIs (Dhanavade et al., 2013, 2016; Jalkute et al., 2015; Sonawane and Barage, 2014; Barage et al., 2014; Dhanavade and Sonawane, 2014).

Structure-based drug designing is the strategy undertaken for identifying inhibitors for PPIs when the protein−protein complex structure is available. This approach is favored when the 3D structure of the biological target is available preferably through experimental methods like X-ray crystallography, NMR, and CD or through *in silico* modeling by methods like *ab initio*, threading, and homology modeling. In this strategy, using the protein−protein complex comprising of the target and its interacting protein, the main residues and hot spots in the protein complex are identified and based on that information, small molecules or peptide inhibitors that bind with high affinity and selectivity to the target are designed or chosen by virtual screening and docking approaches. These selected ligands are tested for their binding stability by molecular dynamic simulations; their pharmacokinetic properties are evaluated and recommended as lead candidates for developing new drugs. In Fig. 5.1, the general strategy followed to identify small molecules for PPI inhibition is given in flow chart format.

Numerous work on PPI inhibitors has been carried out in the oncology field (Jin et al., 2014; Zinzalla and Thurston, 2009). In a study by Wang et al. using structure-based drug designing strategy, DDO-5936 was identified as a small-molecule inhibitor of the Hsp90−Cdc37 complex related to colorectal cancer and was tested to be active in *in vitro* conditions (Wang et al., 2015). The authors carried out molecular dynamics (MD) simulation with Gromacs to verify the hot spots, and docking against Hsp90 was performed with the Glide program. Binding free energy calculations were conducted using the molecular mechanics Poisson−Boltzmann surface area method with the AMBER program to predict the binding affinity.

A ligand-based drug designing strategy is undertaken to inhibit PPIs when natural or reduced affinity ligands to the target molecule are known. Here, pharmacophore

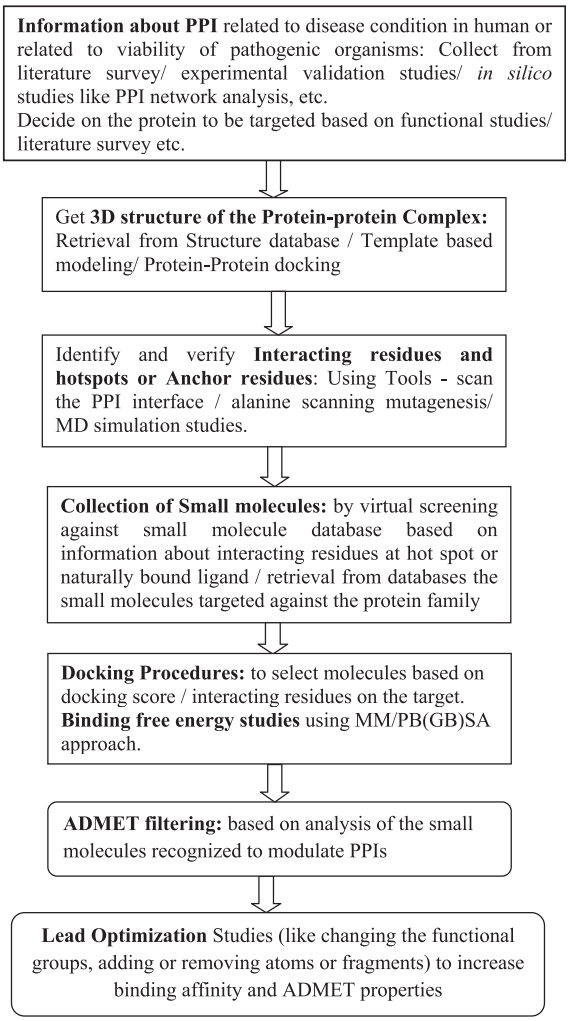


FIGURE 5.1

Structure-based drug designing strategy flow chart (one among many) to identify small-molecule orthosteric inhibitors of protein—protein interactions (PPIs).
ADME/T, Absorption, distribution, metabolism, excretion, and toxicity; *MD*, molecular dynamics; *MM/PB(GB)SA, Molecular Mechanics energies combined with Poisson—Boltzmann or Generalized Born and Surface Area continuum solvation*.

patterns are derived from known ligands and are screened against the small molecule databases to retrieve similar molecules. These molecules are then tested for their target binding affinity to select molecules having higher affinity towards the target.

Devi et al. performed studies to retrieve pharmacophoric patterns from known ligands of the BRD4 protein using PHASE and E-pharmacophore modules of the Schrodinger Maestro tool and screened against Zinc database clean leads subset (Devi et al., 2015). With this approach, three small molecules were identified, which could bind effectively with BRD4 and inhibit its interaction with histones for the treatment of BRD4-NUT midline carcinoma. The binding stability of the ligands with the target were verified with Gromacs simulation studies and binding free energy calculations.

5.5 Databases that play a significant role in the process of predicting PPI inhibitors: databases of PPIs, PPI modulators, and decoys

To treat a disease condition or to develop antimicrobials, knowledge of interacting proteins is needed. Among the numerous experimentally validated and predicted interactions, selecting PPIs of therapeutic significance is crucial. It depends on the information known from their interactions, the depth and width of the study made, and literature reviews. Databases and tools for PPI networks assist in this stage. Once a valid PPI is identified, *in silico* studies are to be carried out to smoothen the way to experimental studies to identify valid modulators of PPI. Structure databases provide the valuable information required for *in silico* studies. Also, databases of experimentally validated and predicted inhibitors or stabilizers or nonactive molecules can speed up drug development by providing knowledge on the mechanism and outcome of their binding, which can help in designing modulators of selected PPIs. Also, decoy databases aid in validating the docking process, which is an inevitable step in the drug designing process. Although many biological databases exist for the previously discussed subjects, an overview of a few important databases is discussed next.

5.5.1 Databases of PPIs

There are numerous databases that have information on PPIs, most of which are organized as interactive networks. Though most of them have collections of all types of PPIs, some are specific toward species, organisms, and protein families. A few important databases for PPI are discussed here. PPI data from experiments known from publications are collected in databases like the Database of Interacting Proteins (Xenarios, 2000), Human Protein Reference Database (Mishra, 2006), Biomolecular Interaction Network Database (Bader, 2003), Biological General Repository for Interaction Datasets (Oughtred et al., 2019), MIPS Protein Interaction Resource on Yeast, and MIPS Mammalian Protein−Protein Interaction Database (Pagel et al., 2004). Apart from publications, original PPI data are included in databases like the Agile Protein Interactomes Dataserver (Alonso-López et al., 2016),

Wiki-Pi (Orii and Ganapathiraju, 2012), and the Microbial Protein Interaction Database (Goll et al., 2008). Predicted information on PPIs are included in many databases like the Human Protein−Protein Interaction Prediction Database (PIPs) (McDowall et al., 2009) and STRING-db (Szklarczyk et al., 2016), a widely used database having both experimental and predicted data. Some databases like PiSITE have a collection of protein interaction sites (Higurashi et al., 2009). Some are huge databases that collect data from other databases. For example, IRefWeb has a large collection of data on PPIs in over 1000 organisms. This collection is consolidated from 14 major public databases (Turner et al., 2010).

To identify modulators of PPI by computational methods, their structural details are needed. The huge repository structure database Protein Data Bank (PDB) has structural data on PPI complexes, but the number is comparatively less compared to individual structures. Structures of domain−domain interactions are available in databases like 3did (Stein et al., 2010), iPfam (Finn et al., 2004), and KBDock (Ghoorah et al., 2013). The Interactome3D database (Mosca et al., 2013) provides 12,000 structural PPIs from eight organisms. Most of these databases have in-built tools to predict interaction between protein pairs or between a given set of proteins, while a few have only datasets to download.

Apart from these databases, there are many online tools that can predict whether two proteins can interact based on their sequences. Tools are also available for protein−protein docking, given a pair of structures, to understand the PPI interface. A list of these tools is found in the vls3d site (http://www.vls3d.com) under protein−protein docking and homology modeling of complexes. For example, one such tool is MEGADOCK.

MEGADOCK 4.0 is structural bioinformatics software for fast Fourier transform-based rigid docking to screen PPI pairs for an interactome prediction. It makes extensive use of recent heterogeneous supercomputers and shows powerful and scalable performance. For a user-submitted list of protein structures, the tool performs all-to-all docking to predict relevant PPI pairs and also performs interactome mapping. The procedure for PPI prediction consists of two sections called docking calculation and PPI decision. For a submitted set of protein structures, a docking calculation section performs all-to-all docking and generates high-scoring decoys for all possible combinations, based on shape complementarity and physicochemical properties. Later, for each pair of proteins, the PPI decision section analyzes the structural distributions of high-scoring decoys and concludes whether the two proteins can really interact. Also in the output, a possible PPI network is included that connects the positively predicted PPIs (Ohue et al., 2013; Ohue et al., 2016; http://www.bi.cs.titech.ac.jp/megadock).

5.5.2 Databases of PPI modulators

Though various studies both *in vitro* and *in silico* were carried out targeting PPIs for various purpose for past many years, separate databases for PPI modulators—iPPI-DB, 2P2I, and TIMBAL databases—have been developed over the past

10−12 years only. Also, small-molecule modulators for PPIs can be searched in ethnobotanical databases (Thakar et al., 2019; Thakar et al., 2015) and other natural sources (Barbosa and Roque, 2019).

iPPI database (iPPI-DB) (2012) is a database for orthosteric small-molecule inhibitors of PPIs and has no information on peptide inhibitors. Available details about structure, binding and activity, and pharmacological and pharmacokinetic properties are presented for 2054 PPI inhibitors as well as the profile of 35 families of PPI targets; these data are mostly retrieved from peer-reviewed scientific articles and world patents. Only compounds with activities below 30 μM are qualified for entry. The molecules in the database can be queried either by using physicochemical/pharmacological properties or with a user-defined structure.

Each compound has an individual ID card where all information are summarized under compound summary, physicochemistry, pharmacology, and drug similarity tabs. The compound summary has details on chemical structure, SMILES notation, and its IUPAC, along with external links to other databases such as ChEMBL, PubChem, and PubMed article, as well as patent information, if any. In the physicochemistry tab, the physicochemical profile of the compound is provided along with its compliance toward Lipinski's rule of 5 and Veber's and Pfizer's 3/75 rules. A principal component analysis map shows the position of the selected compound in the iPPI chemical space with respect to iPPI-DB compounds of the same family and iPPI-DB compounds on the same target. In the pharmacology tab, the available binding data are given together with assay type and activity type. Also, lipophilic and ligand efficiencies of the compound are compared with the same family of iPPI-DB compounds and all the modulators available for the same target are represented in a biplot. The drug similarity tab holds the chemical structure of the compounds together with the most similar drugs found in the MDL Drug Data Report database, along with links to DrugBank (Labbe et al., 2015; https://ippidb.pasteur.fr/).

2P2I database (2010) is a hand-curated database of only orthosteric inhibitors and has a collection of structural information about protein−protein and protein−ligand complexes and the small molecules involved, which are available in the PDB database and the literature. It encompasses 27 protein−protein complexes and 274 protein−inhibitor complexes related to 242 unique small-molecule modulators. The protein−protein complexes were subdivided into three classes: (1) protein−peptide complexes, (2) globular protein−protein complexes, and (3) bromodomains−histone complexes. For a given PPI family, it has 3D structures and data on geometry and the physicochemical nature of the interface, intermolecular nonbonded contacts, hydrogen bonds and salt bridges, and other binding parameters. In addition, the small molecules involved and its descriptors are portrayed and links to other webservers are present. It hosts a query tool to search for inhibitors within the database using standard molecular descriptors. The other in-built tool, 2P2I-inspector, calculates the physical and chemical nature of the PPI interface and bonds in the user-submitted complex structure (Basse et al., 2016; http://2p2idb.cnrs-mrs.fr).

TIMBAL database (2009) has a collection of approximately 8900 PPI orthosteric modulators of molecular weight <1200 Da. It was created initially by manually curating information extracted from relevant scientific publications and later was retrieved from the ChEMBL database (https://www.ebi.ac.uk/chembl/). As ChEMBL also has data on nonactive molecules, TIMBAL has a collection of inactive molecules against PPI targets (7%). Allosteric modulators are not included. Orthosteric small peptides having less than 10 peptide bonds are included in the collection and it contains more than 14,000 data points for nearly 7000 small molecules. The database also covers 50 known PPI drug targets, including PPIs that are stabilized by small molecules with therapeutic effect. Links to the PDB database and CREDO database are provided to retrieve experimental structures of protein−small molecules, protein−protein complexes, and unbound proteins, and to explore in detail the atomic interactions of these complexes, respectively. Other data entries are for integrins, the cell surface receptors that have been long recognized as therapeutic targets (Higueruelo et al., 2013; http://www-cryst.bioc.cam.ac.uk/timbal).

In a comparative view, the iPPI-DB derives its information from peer-reviewed scientific articles and world patents, 2PPI derives its information from both literature and PDB entries, and TIMBAL derives its information from the literature and the ChEMBL database. Both iPPI and 2P2I have collections of orthosteric inhibitors, whereas TIMBAL deals with both orthosteric inhibitors and stabilizers. Similarly, both iPPI and 2P2I have only small-molecule inhibitors but TIMBAL has both small molecules and peptide PPI modulators. Also, only TIMBAL has data on inactive molecules for PPI targets. Of the three, the iPPI-DB is still manually curated and can be queried with both PPI targets and small-molecule structure. The PPI target families in these three databases are listed in Table 5.1.

Currently, the 2P2I (active until March 14, 2019) and TIMBAL databases are not active. But some companies hold certain information on the molecules and targets from these databases and the molecules can be ordered for lab purposes. For example, Life Chemicals has PPI Focused Libraries, which includes 2400 compounds that were extracted from the Life Chemicals HTS Compound Collection (18,936 reference compounds) by a 2D fingerprint similarity search toward the TIMBAL database with Tanimoto 85% threshold. It also includes 4600 compounds obtained by filtering 2P2I and iPPI-DB compounds using the Rule of Four (https://lifechemicals.com).

5.5.3 Decoy databases for PPIs and modulators

Decoy sets of structures (false positive matches) are very important in developing intermolecular potentials and scoring functions. Servers like ZDOCK (http://zdock.umassmed.edu/) and RosettaDock (https://www.rosettacommons.org), which perform protein−protein docking, have their own decoy sets. Some decoy databases are developed from which the decoy sets can be downloaded and used for developing and validating protein−protein docking methodologies and results.

Table 5.1 List of protein–protein interaction (PPI) target families in PPI modulator databases.

iPPI	2P2I	TIMBAL
MDM2-like/P53	BRD2-1/H4	14-3-3/PMA
Bcl-2 like/BAX	BRD2-2/H4	ARF1/SEC7
LFA/ICAM	BRD3-1/H4	AuxinIAA-TIR1
XIAP/Smac	BRD3-2/H4	BIII/X11a
Bromodomain/ Histone	BRD4-1/H4	BRD2/Ack
CD4/gp120	BRD4-2/H4	BRD4/NUT
LEDGF/IN	BRDT-1/H4	BRDT/H4
CD80/CD28	Bcl2/Bax	CD40/CD154
TTR	BclXL/Bak	CD74/MIF
Beta-catenin/TCF-4	CIAP1_1/ CASPASE-9	c-Myc/Max
Survivin dimer	CIAP1_2/ SMAC	CRM1/Rev
MENIN/MLL	HDM2/P53	Cyclophilins
IL2/IL2R	HPV_E2/E1	E1/E2
VHL/HIF1alpha	HRAS/SOS1	HIF-1a/p300
E2/E1	IL-2/IL-2R	IL-2/IL-2Ra
PCNA trimer	Integrase/ LEDGF	FKBP1A/FK506
Myc/Max	KEAP1/NRF2	Integrins
NRP/VEGF	KRAS/SOS1	K-Ras/SOS1
14-3-3/TASK-like	MDM4/P53	Nrf2/Keap1
UPAR/UPA	Menin/MLL	Adenylyl cyclase dimer C1–C2 domains
14-3-3/PMA2	TNFR1A/TNFB	Annexin A2/S100-A10
Influenza NP	TNFalpha	Bcl-2 and Bcl-XL with BAX; BAK and BID
VEGF/VEGFR	VHL/HIF1A	Beta catenin/Tcf4 and Tcf3
MDM2-like dimer	XDM2/P53	CD80/CD28 (or CTLA-4)
Col1/Jaz1	XIAP/ CASPASE-9	Clathrin/adaptor and accessory proteins
CaM/CaMBD2	XIAP/SMAC	ESX/Sur-2 (DRIP130)
H-Ras/SOS1	ZIPA/FTSZ	LMO2/LDB1 or TAL1
Keap1/Nrf2		TNFa trimer or TNFa/TNFR
BRI1		Transthyretin tetramer
FAK/VEGFR3		UL30(Pol)/UL42 subunits of HSV T-1 DNAP
14-3-3/ER		XIAP/Caspase9 or SMAC
ZipA/ftsZTNF trimer		ZipA/FtsZ
CD40L-trimer		MLL/Menin

Table 5.1 List of protein–protein interaction (PPI) target families in PPI modulator databases.—*cont'd*

iPPI	2P2I	TIMBAL
WDRS/MLL		Neuropilin-1/VEGF-A
		PPAR-gamma/NRCoA1
		Plk1(PBD)/PBD substrate
		Rac1/GEFs
		Rad51/BRCA2
		RGS4/Galpha-o protein
		RRTF1/CBFb
		p53/S100B, p53/MDM2, p53/MDMX
		Tak1/Tab1
		Dimers of MAX, SOD1, STAT5, ToxT, STAT3 and tubulin

The **Dockground** project is developing docking software, including protein–protein docking and software for studying protein interfaces. The docking decoys were developed and designed for the unbound docking set, based on experimentally determined protein–protein complexes and their unbound structural forms from structure databases. The set consisting of 99 complexes was first gathered by selection, where sequence identity between bound and unbound structures is greater than 97%. Homomultimers and structures in the wrong formats were then deleted. GRAMM-X scan was engaged to build docking decoys from this set. The following characteristics were computed for 500,000 matches per complex: root mean square deviation (RMSD) of the backbone atoms of the interface residues, RMSD of the backbone atoms of the ligand (the smaller protein of the complex), the number of native residue–residue contacts in the predicted complex divided by the number of contacts in the native complex, and the number of non-native residue–residue contacts in the predicted complex divided by the total number of contacts in the complex. The set contains 61 decoy complexes, with each complex having 100 lowest energy non-native structures and at least one near-native structure (RMSD < 5.0 Å) and no native structures (Liu et al., 2008; http://dockground.bioinformatics.ku.edu). The decoys for protein–protein docking can be downloaded at http://dockground.bioinformatics.ku.edu/UNBOUND/decoy/decoy.php.

Also, ligand decoys can be engaged in the docking studies to identify valid small molecule inhibitors for PPIs. There are many databases dedicated to this, with the DUD•E database having a vast collection.

The **DUD•E** (Directory of Useful Decoys-Enhanced) database provides challenging decoys for molecular docking. It contains 22,886 ligands and their affinities against 102 targets, with an average of 224 ligands per target. For each ligand, 50 decoys are constructed having similar physicochemical properties but different 2D topology. DUD uses 2D similarity fingerprints to minimize the topological similarity

between decoys and ligands to minimize the likelihood of actual binding. To enable focused research on particular target classes, the following subsets are available: DUD38, the original DUD targets rebuilt; GPCR, seven transmembrane helix receptors; Kinase, protein kinases; Nuclear, nuclear hormone receptors; Protease, proteases; and Diverse, selected targets that are representative of the entire set. The actives and decoys can be downloaded in mol2 and SDF format. Also, decoys can be generated for user-entered active compounds by querying with SMILES, with or without an identifier. DUD decoys are matched to the physical chemistry of the queried ligands based on properties like molecular weight, number of rotatable bonds, calculated logP, and hydrogen bond donors and acceptors, and the decoy set is generated (Mysinger et al., 2012; http://dude.docking.org/).

5.6 Transcription factors as one of the PPI drug targets: importance, case study, and specific databases

Particular types of PPIs are frequently chosen as drug targets like transmembrane, cytoskeleton and mitotic proteins, and nuclear receptors, based on the ease of targeting them. TFs play a pivotal role in controlling cell signaling by regulating gene expression and hence are crucial for cell growth, cell division, embryonic development, etc. Dysfunction of specific TFs is known to be involved in a wide variety of diseases such as obesity, cancer, autoimmunity, diabetes, cardiovascular disease, and neurological disorders (Lee and Young, 2013). As they always function as dynamic complexes, they are held up as PPI targets for many therapeutic indications, even though there are known hurdles on the way (Fontaine et al., 2015).

Modulation of TF activity can be achieved by various approaches, like direct or indirect modulation of their own expression, altering their DNA binding activity and particularly affecting their ability to interact with partner proteins by PPI inhibition. The partner protein(s) can be the TF itself when they form homodimers (e.g., STAT), another TF (e.g., MYC/MAX), a cofactor/coactivator/mediator or repressor (e.g., Nrf2/Keap1), protein belonging to basal transcription machinery, RNA polymerase, and chaperones associated with nuclear translocation (Lambert et al., 2018). The first TF inhibited by modulating PPIs is the tumor suppressor transcription factor p53 (p53/mdm2). The first small-molecule AI-10-49 to target the fusion of TFs, CBFβ, and SMMHC was developed by a fragment building-based strategy (Illendula et al., 2015). This inhibition of PPI restored the transcriptional activity of RUNX1 and selectively induced cancer cell death *in vivo* in acute myeloid leukemia.

Modulation of transcriptional activity is carried out either to hinder, restore, enhance, or downregulate transcriptional activity of a single gene or set of genes related to the disease condition for therapeutic purposes. In humans, TF mutations are known to cause specific diseases like Rett syndrome, autoimmune diseases, multiple cancers, etc., and medications can be potentially targeted toward them.

Approximately 10% of genes in the human genome code for TFs, which makes this family the single largest family of human proteins. In fact, over 30 TFs have been identified as therapeutic targets of about 9% of the approved drugs in the DrugBank database. However, since 2001, though around 1700 human TFs and fewer than 200 coregulators have been predicted, only 62 TFs have been functionally validated (Vaquerizas et al., 2009; Schaefer et al., 2010); thus there is a huge space to study them for therapeutic purposes. Interestingly, most of the microbial TFs which are essential for their viability, are not highly conserved in eukaryotic cells, and hence are valid drug targets for antiinfective therapy. TFs like NusA, NusB/E, and NusG are well-recognized antibacterial targets (Ma et al., 2016).

Studies on microbial TFs are carried out to develop antibiotics. A bacterial TF, CarD, was known to interact with RNA polymerase (RNAP) to control rRNA transcription in *Mycobacterium tuberculosis* (*Mtb*) and their interaction was detected to be indispensable for the viability of the organism (Stallings et al., 2009); many experimental and in silico studies verified CarD as a valid target (Weiss et al., 2012; Priya et al., 2012). Based on this, studies by Priya et al. identified a small-molecule inhibitor to inhibit CarD−RNAP interaction (Priya et al., 2018). In their approach, initially, with the crystal structure of the CarD−RNAP complex (PDB ID:4KBM), Schrodinger's BioLuminate program panels and ANCHOR tool were utilized for identification of interacting residues and hot spot residues at the interface of this PPI. The structure of the CarD−RNAP complex is displayed in Fig. 5.2 along

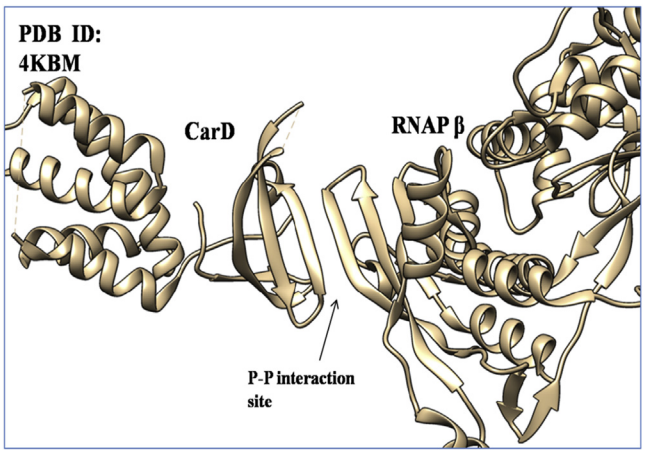


FIGURE 5.2

Crystal structure of CarD−RNA polymerase (RNAP) β-subunit complex in *Mtb* (PDB ID: 4KBM) (Gulten and Sacchettini, 2013). The RNAP binding site of CarD is located at its N-terminal domain and their interaction results in a 500 $Å^2$ buried surface area. There are eight hydrogen bonds and 69 nonbonded contacts at the interface, comprising electrostatic, hydrophobic, and van der Waals interactions, with the electrostatic force being the major contributor.

with details of the nature of their interface. It is a β−β interface, and based on the analysis, the residues ARG47, ARG25, THR45, Leu44, and VAL46 were identified as hot spot residues on CarD. Later, the pepMMsMIMIC tool was engaged for screening small molecules based on the RNAP residues which interact with hotspots on CarD. With the collection of these small molecules, docking analysis was carried out with Glide, and three small molecules were selected based on the CarD residues with which they interact and the binding score. The docking interaction between MMs0248919, one of the three small molecule hits and CarD is displayed in Fig. 5.3. Binding free energy calculations were done with the Prime program of Schrodinger using MM-GBSA calculations, and of the three hits, the small molecule MMs02420750 with least binding free energy was selected for a molecular dynamics simulation run in Gromacs to test the stability of the target−ligand complex. The absorption, distribution, metabolism, and excretion (ADME) properties of the molecule were predicted using QikProp from Schrodinger and was proclaimed as a valid inhibitor of CarD-RNAP complex.

Though the TF's sequence, structure, and other information like domain, family, and motifs are present in general biological databases, there are many specific databases dedicated to only TFs. They focus on the classification of TFs, their DNA binding sites, and regulatory interactions with other proteins. A few TF-specific databases are discussed next.

The **TRANSFAC** (TRANScription FACtor) database is a manually curated database of eukaryotic TFs and their DNA binding profiles. TFs are classified into families, classes, and superclasses based on the architecture of their DNA

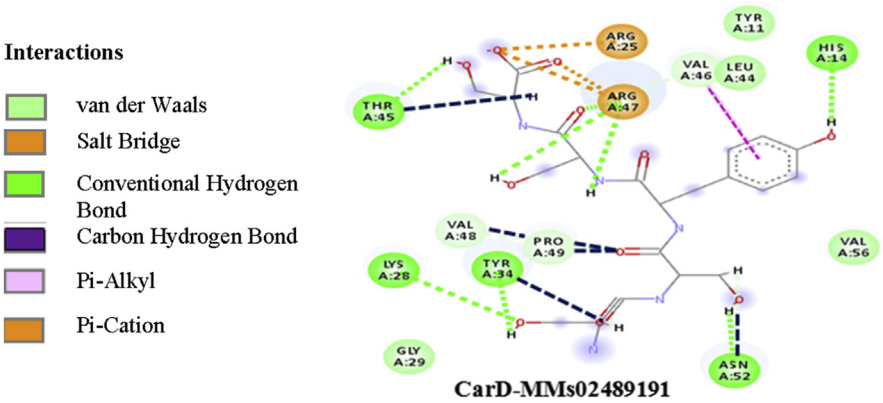


FIGURE 5.3

Interactions between the small molecule MMs0248919 with the hot spot residues ARG47, ARG25, THR45, Leu44, and VAL46 on the target CarD are represented in 2D. The nature of the interaction is indicated by different colored dotted lines, as given in color box (Schrodinger's visualization tool).

binding domains. Since July 2016, TRANSFAC has become a partially commercial database. It provides a list of classifications of TFs based on superclasses and classes along with the description, and also the families and subfamilies are provided along with links to databases like UniProt, Human Protein Atlas, TRANSFAC, and PDB. The previous version 2017−12 factor table is free to search, which can be carried out using various search terms relating to fields like gene, homolog, organism, species, reference authors, etc. This database has been widely used to predict TF binding sites as seen in many publications (Wingender et al., 1996; http://genexplain.com/).

The **TRRUST** (Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining) database consists of 8444 regulatory interactions for 800 TFs in humans and 6552 regulatory interactions for 828 TFs in mice, all derived from PubMed articles. If a TF gene is submitted as a query, it gives information about both the target genes it regulates along with the mode of regulation if known and other TFs involved in regulation of the submitted TF, along with a PubMed reference article. It also displays information about the other TFs that regulate the same target genes and also about the diseases and pathways with which the query TF is involved. If a non-TF gene is submitted, all the TFs involved in its regulation are displayed. For both searches, the protein interaction network is provided (Han et al., 2017; https://www.grnpedia.org/trrust/).

The **Animal Transcription Factor DataBase** (AnimalTFDB 3.0) has a vast collection of animal TFs and cofactors from 97 animal genomes. The TFs are further classified into 73 families based on their DNA-binding domain and the TF cofactors are classified into 83 families based on their function. For each entry, the database provides information about the gene model, protein sequence length, functional domain site, and orthologs and paralogs along with the similarity score, and has links to other databases like Ensembl and Pfam. It has a TF binding site prediction tool to identify potential binding TFs for nucleotide sequences and a TF prediction tool to identify whether the submitted sequence is a TF sequence. It also has a separate human TF database web interface (Hu et al., 2018; http://bioinfo.life.hust.edu.cn/AnimalTFDB/).

The **SM-TF database** holds 3D structures of TFs complexed with small molecules. There are presently 934 entries consisting of 176 TFs from various species. TFs are derived from bacteria, eukaryote, and archaeal lineages contributing 51%, 47%, and 2%, respectively. In the database, classification is done according to the organisms and species. TFs from *Homo sapiens* and *Mus musculus* are linked to TFClass, and TFs from *Escherichia coli* are linked to RegulonDB. In the list of TFs, the marking "DT" in front of their UniProt ID indicates that they are druggable targets. For each TF in the list, three PDB files are provided: (1) conformation of the small molecule in the binding mode, (2) binding site on the target protein, and (3) a clean binding site containing only the amino acid residues. It is suggested to use the file containing the small molecule for virtual screening studies with TFs and the files containing the binding site structure can be used in studies related to inverse docking (Xu et al., 2016; http://zoulab.dalton.missouri.edu/SM-TF).

5.7 Pharmacokinetic properties of small-molecule inhibitors of PPI

Many drug candidates fail during clinical studies due to poor pharmacokinetic properties, namely absorption, distribution, metabolism, excretion, and toxicity (ADME/T). These physicochemical and biochemical properties of drug molecules indicate what the body does to the drug. The ADME criteria of the drugs are discussed in brief next.

When a drug enters into the body it needs to reach the destined site at a required concentration to carry out its function to give the desired effect. This bioavailability of a drug depends on its absorption concentration. For oral drugs, absorption depends on the molecule's solubility, instability in the stomach, and intestinal transit time. Molecules that are less absorbed orally need to be administrated through other routes like nasal, parental, and dermal. Molecules that enter the blood stream must be delivered to the effector site and this process is called distribution. Distribution depends on the properties of the molecule like polarity, molecular size, and plasma protein binding nature. Drugs that have performed their function have to be excreted and hence need to be metabolized. Generally, from the time of entry into the body, metabolism of the molecule begins. Small-molecule drugs are mostly metabolized in the liver by a family of cytochrome P450 enzymes. Most of the current drugs are inactivated by metabolism and only a few drugs that are given as prodrugs are activated. Though the rate of metabolism depends on various physiological and pathological factors, the structure and properties of the molecules decide the rate of metabolism. Finally, drug excretion occurs through urine via the kidneys, through biliary excretion or fecal excretion, and through the lungs. The toxicity factor, on the other hand, arises due to many factors like inappropriate ADME properties, binding of the drug to other molecules (off-target binding), and the presence of harmful functional groups in the drug. The effect of toxicity can vary from mild to adverse and can even be fatal.

There are many in vitro assays carried out to predict ADME properties but predicting toxicity by in vitro assays is a tedious process. In the field of drug development, this information is used for prioritizing lead series, lead optimization, select compounds for in vivo studies, and the assessment of in vivo results (Balani et al., 2005). In silico prediction of ADME/T properties of the drug candidates prior to costly experimental procedures can eliminate unnecessary testing on compounds that will ultimately fail. Hence, during the drug designing process, the pharmacokinetic parameters like bioavailability, metabolic half-life, permeability, etc., of the small molecule ligands are predicted computationally using ADME/T tools.

The generalized chemical properties of the PPI inhibitors are predicted by analyzing PPI inhibitors discovered till now. PPI inhibitors are generally larger in size and molecular weight compared to traditional drugs (MW $>$ 400 Da). Their hydrophobicity values are high with an ALogP $>$ 4 and have more than four hydrogen bonds and four or more rings (Morelli et al., 2011; Villoutreix et al., 2014; Koes and Camacho, 2011) and hence they are not governed by Lipinski's

rule (Lipinski et al., 2012). Servers and tools existing at present have only databases of traditional drugs for reference. Hence, we need to predict the properties using these tools but analysis has to be done based on the knowledge of known PPI inhibitors.

There are two online servers, PPI-HitProfiler (http://www.cdithem.fr/) and $2P2I_{Hunter}$, which offer in silico filters to design libraries of PPI inhibitors from the huge set of conventional compound collections. PPI-HitProfiler uses the decision tree approach and $2P2I_{Hunter}$ uses support vector machine (SVM) algorithms to filter molecules based on the descriptors of known PPI inhibitors. $2P2I_{Hunter}$ is currently not available.

5.8 Strategies and tools to identify small-molecule inhibitors of PPIs

The general strategy followed in structure-based drug designing to inhibit PPIs is shown in Fig. 5.1 under Section 5.4. Various tools can be used in each step of the process. In this section, some important tools that are specifically used for studying PPI interfaces and to screen small molecules for the inhibition of PPIs are discussed in detail. Also, tools to study the pharmacokinetic profile of the molecules are reviewed.

5.8.1 Prediction of interacting residues and hot spots in protein–protein complexes

Given a protein–protein complex consisting of a target protein and an interacting protein, the main hot spot residues on the interacting protein can be used as a template to design or screen small molecules that can bind with the hot spot residues on the target protein at the PPI interface and hinder the interaction between the two proteins. Based on the structure of protein–protein complexes in the databases, computational tools are developed for the prediction of hot spots by alanine scanning mutagenesis. In this procedure, all the residues of the proteins in the complex or the residues at the interfaces are mutated to alanine and the residues whose mutation to alanine results in a decrease of at least 2.0 kcal/mol in binding free energy are identified as hot spots. The binding free energy ($\Delta G_{binding}$) is calculated as:

$$\Delta G_{binding} = \Delta G^{mut} - \Delta G^{wt}$$

where ΔG^{wt} and ΔG^{mut} are the binding free energies upon complex formation of the wild-type and alanine-mutated proteins, respectively (Moreira et al., 2007). Alanine is generally a first-choice residue for mutational scanning because it retains only the beta carbon but no other side chain chemistry, as beta carbon position depends on the backbone dihedral angles of the polypeptide and hence is really part of the main chain structure of the protein.

The tools follow various strategies to identify the interacting residues and hot spots, where some tools are based on dedicated energy functions, like FoldX (Schymkowitz et al., 2005) and PCRPi (Assi at al., 2009), and some tools rely on machine learning algorithms, like HSpred (Lise et al., 2011) and HotPoint (Tuncbag et al., 2010). Molecular dynamics simulation tools like CHARMM, GROMACS, NAMD, AMBER, and DESMOND are also used to manually verify the hot spots and of these, GROMACS (Spoel et al., 2005) is a free and extensively used software. Prediction reliability of these methods are high and hence combining them is recommended for accurate prediction. Commercial tools like the Flex_ddG method of Rosetta (Barlow et al., 2018) and the BioLuminate program of Schrodinger (Beard et al., 2013) are highly engaged software for the prediction of interacting residues and hot spots.

The **FoldX** server is used to predict the stability of protein−protein structures and calculate important interaction residues and hot spots between them using energy functions. This suite is freely available to academic and nonprofit research institutions for research purposes only. The binding free energy of a complex AB is calculated as:

$$\Delta G_{binding} = \Delta G_{AB} - (\Delta G_A + \Delta G_B)$$

where ΔG_{AB} is the Gibbs free energy of the complex and ΔG_A and ΔG_B are the individual free energies of A and B molecules. FoldX follows the given linear combination of empirical terms to calculate free energy (in kcal/mol):

$$\Delta G = a\Delta Gvdw + b\cdot\Delta GsolvH + c\cdot\Delta GsolvP + d\cdot\Delta Gwb + e\cdot\Delta Ghbond + f\cdot\Delta Gel + g\cdot\Delta Gkon + h\cdot T\Delta Smc + k\cdot T\Delta Ssc + l\cdot\Delta Gclash.$$

where ($a \ldots l$) are relative weights of the different energy terms used for free energy calculation, ΔG_{solvH} is the desolvation term related to hydrophobic groups, ΔG_{solvP} is the desolvation term related to polar groups, ΔG_{wb} is the explicit calculation of water molecules that makes more than two hydrogen bonds with the protein, ΔG_{vdw} is calculated in a similar fashion to desolvation but now taking into account experimental transfer energies from water to vapor, ΔG_{el} is calculated from a simple implementation of Coulomb's law, ΔG_{hbond} is for hydrogen bonds calculated on the basis of simple geometric considerations and their energy, ΔG_{kon} calculates the electrostatic contribution of interactions at interfaces, ΔS_{mc} is derived from a statistical analysis of the phi−psi distribution of a given amino acid as observed in a set of nonredundant high-resolution crystal structures (entropy penalty), ΔS_{sc} is the entropy cost of fixing a side chain in a particular conformation, and the ΔG_{clash} term provides a measure of the steric overlaps between atoms in the structure. The output contains the $\Delta G_{binding}$ for each pair of polypeptide chains in the .pdb file, decomposed into different energy terms (Schymkowitz et al., 2005; http://foldxsuite.crg.eu/).

The user needs to submit the PPI complex structure (.pdb file) and set the option for parameters like temperature, water, ionic strength, and van der Waals design.

The output displays the binding energies for each pair of polypeptide chains in the .pdb file, mentioning the contribution of each energy term like backbone hydrogen bond, side chain hydrogen bond, electrostatics, van der Waals, water bridge, entropy of side and main chain used by FoldX plus, and an additional term that reflects the intrachain clashes of residues forming part of the interface. To identify hot spots, an alanine scan is done, the effect of the truncation on the binding energy between the chains is calculated, and the list of changes in binding energy due to mutation for each residue is displayed.

HSPred, an SVM-based method, predicts hot spot residues given the structure of a complex. The basic energetic terms that contribute to hot spot interactions, i.e., van der Waals potentials, solvation energy, hydrogen bonds, and Coulomb electrostatics, are used as input features of an SVM classifier. Also, they have developed two additional SVM classifiers, specifically optimized for arginine and glutamic acid residues to improve the performance (Lise et al., 2011; http://bioinf.cs.ucl.ac.uk/psipred/).

The user needs to upload the PDB structure and specify the chains forming the interface by entering the chain identifiers for protein 1 and protein 2. HSPred mutates each amino acid at the interface to alanine and scores them. In the output, a score greater than zero represents predicted hot spot, and negative scores are predicted nonhot spots. In the resultant .pdb file, predicted hot spot residues are colored red, nonhot spot residues are colored white, and those that are not part of the interface are colored blue (specified in temperature column).

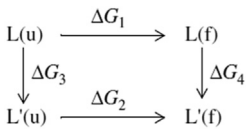
The **BioLuminate** program from Schrodinger through the **Protein Interaction Analysis panel** identifies the closest interacting residue neighbors in a submitted protein−protein complex. The following criteria are used to find the interacting residues: initially each residue is considered as the target residue and any residue that has an atom within the specified distance (default 4.0 Å) to the target residue is considered as its neighbor but interactions between backbone atoms are ignored. To detect hydrogen bonds, the four atoms involved in the hydrogen bond are designated as D−H … A−X, when the minimum acceptor angle H … A−X is 90 degrees, the minimum donor angle D−H … A is 120 degrees, and the maximum H … A distance is 2.5 Å. For detecting a salt bridge, the maximum distance between an ion and a protein atom is recommended to be 4.0 Å. For pi stacking, the maximum distance between the centroids of the two aromatic rings is expected to be 4.0 Å. Also, if $R_A + R_B - R_{AB}$, where R_A and R_B are the van der Waals radii and R_{AB} is the distance between atoms A and B, is greater than the allowable overlap of 0.4 Å, the atoms are considered to have van der Waals clash (Beard et al., 2013; BioLuminate, Schrödinger, LLC, New York, NY).

With a protein−protein complex .pdb file, the following steps are to be followed: open the BioLuminate interface of Schrodinger, import the Protein complex, and prepare the proteins using the Protein Preparation Wizard. Then, open the Protein Interaction Analysis panel from the task bar and in the Define Interacting Groups box, from the unassigned chain list displayed, select and assign the protein chains

for Group 1 and Group 2. In the Advanced option box, the default values for hydrogen bond, salt bridges, pi stacking, and van der Waals clashes displayed can be agreed or changed. Run by clicking Determine Protein Interactions.

The output of the program lists all the interacting residues of Group 1 protein with residue number and a three-letter code, and gives details of the residues in Group 2 protein with which interactions are seen along with particulars like distance in Angstrom, number of hydrogen bonds, salt bridges, pi−pi stacking interactions, disulfide bridges, and van der Waals clashes between them. It also includes van der Waals shape complementarity and percentage of buried solvent accessible surface area (SASA) of each interacting residue of Group 1 protein toward its interacting residues of Group 2 protein. Similar details are predicted for Group 2 protein and the analysis of both results gives a clear indication of the nature and mode of interactions at the PPI interface.

The **Residue Scanning panel** in the BioLuminate program of Schrodinger is engaged to identify hot spot residues at PPI sites in a submitted protein−protein complex by mutating residues of a protein (labeled as ligand) to any particular amino acid (e.g., alanine), or to an amino acid of the same physicochemical nature. Considering one protein in the complex as ligand and rest of the system as receptor, it then calculates the change in stability. Other changes like: change in total surface area due to the mutation, change in surface area of nonpolar atoms and polar atoms due to the mutation, change in p*K*a of the mutated residue, change in binding affinity of the mutated protein treated as the ligand (negative value—mutant binds better than the native protein), change in hydrophobicity or hydrophilicity of the mutated residue, and change in the van der Waals surface complementarity of residues at the interface due to the mutation are calculated only when opted by the user. Stability of the protein is estimated from a thermodynamic cycle to check the effect of mutation as shown here:

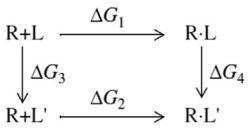


where $L(u)$ is the unfolded ligand, $L(f)$ is the folded ligand, $L'(u)$ is the unfolded mutated ligand, and $L'(f)$ is the folded mutated ligand. The change in stability is given as:

$$\Delta\Delta G(stability) = \Delta G2 - \Delta G1 = \Delta G4 - \Delta G3$$

Experimentation measures $\Delta G1$ and $\Delta G2$, but $\Delta G3$ and $\Delta G4$ are calculated to effectively cancel the error in the computational models. Prime MM-GBSA, which uses an implicit solvation model, is engaged for these calculations. Similarly, the

change in binding affinity of the protein due to the mutation is computed from a thermodynamic cycle, as given here:



where R is the receptor, L is the ligand, and L' is the mutated ligand. $R + L$ and $R + L'$ represent the separated receptor and ligand. $R \cdot L$ and $R \cdot L'$ represent the receptor bound to the ligand. The change in binding affinity is:

$$\Delta\Delta G(bind) = \Delta G2 - \Delta G1 = \Delta G4 - \Delta G3$$

where $\Delta G1$ and $\Delta G2$ are experimental values and $\Delta G3$ and $\Delta G4$ are calculated to avoid computational errors and these calculations are also done with Prime MM-GBSA (Beard et al., 2013; BioLuminate, Schrödinger, LLC, New York, NY).

The following steps are to be followed in the program: in the BioLuminate interface, the protein−protein complex (.pdb file) has to be imported, refined using the Protein Preparation Wizard, and displayed in the workspace. From the task bar, open the Residue Scanning Panel and choose one protein as ligand. Mutation in both the proteins can be done by labeling the other protein as ligand in the second run. Here, all the residues of the labeled protein will be displayed with chain name, number, and a three-letter code in the Residues column. The user can also choose to show Polar or Nonpolar residues only. In the Surface Complementarity column, the van der Waals surface complementarity for residues at the interface is provided, which assists the user to choose residues for mutation based on complementarity. All the residues can be mutated or only selected residues can be mutated by checking/selecting the box Mutate selected residues only. To carry out single residue mutation click the Mutation column, and against the selected residue, from the dropdown menu, select the amino acid to replace or select the residue group like polar, neutral, etc. and press Enter. Next, for a protein−protein complex, the property affinity needs to be selected and the program run.

In the results, the effects of mutation such as change in total surface area and change in surface area of polar atoms and nonpolar atoms, which are opted, are displayed. From change in binding affinity and change in the stability displayed for each mutated residue, the hot spot residues in the protein chains of the complex need to be predicted. For this, a negative value means that the mutant binds better than the native protein and vice versa. Residues that have a high positive value of more than 1 are to be considered as hot spots at the PPI interface.

Several research groups have used these tools to investigate the properties of PPIs. Mattapally et al. sequenced the NKX2.5 gene in 100 congenital heart disease patients and 200 controls, and identified seven mutations of which D16N was a novel

mutation that was associated with ventricular septal defect. Furthermore, they carried out computational analysis to verify the effect of mutation. Interaction of the NKX2.5 protein with GATA4, a TF, was studied using molecular modeling, docking, and MD simulation studies and the role of the mutation was analyzed. Then, using the BioLuminate residue scanning/affinity maturation panel and web servers DrugScorePPI and BeatMusic, alanine scanning of interface residues of the NKX2.5−GATA4 protein complex structure was carried out, which confirmed the importance of the mutation and also key residues of this PPI were identified (Mattapally et al., 2018).

ANCHOR is a web server program that identifies anchor residues in a protein−protein complex, which are amino acid side chains deeply buried at protein−protein interfaces, to discover possible druggable pockets to be targeted by small molecules. Also, the analogs of the side chain can aid in designing or screening PPI inhibitors. ANCHOR calculates the change in solvent accessible surface area (ΔSASA) upon binding for each residue, and it also approximates the contribution of each residue toward binding free energy. To characterize anchor residues in a given protein−protein complex structure, ANCHOR first adds missing atoms using the CHARMM19 force field and performs hydrogen minimization. It then substitutes each residue with alanine and calculates the ΔSASA for each residue's side chain by figuring the difference in SASA of the side chain in the unbound protein (isolated from complex) and in the bound protein complex. SASA is calculated with a computational program NACCESS. The binding free energy of each residue is calculated using FastContact, a fast empirical pairwise estimate that combines a standard distance-dependent dielectric "4r" electrostatic and a desolvation contact potential (Meireles et al., 2010; http://structure.pitt.edu/anchor/).

For this, open the tool on the web by typing "anchor tool for proteins" or use the web address. Upload the protein complex and specify chains for protein 1 and protein 2 and click Submit. In the second turn, interchange the protein chains and submit.

In the output, for protein 1 chain, the residues are listed in the decreasing order of ΔSASA or predicted binding energy as the user chooses. From the residues that contribute more to binding free energy (having least scores) and that show major percentage change in ΔSASA upon mutation, the anchor residues can be predicted. Analyze the results for both protein chains. In the tool through the Jmol visualizer, the user can visualize selected anchor residues in their pockets and the stereochemical nature of the surrounding region.

Dosztanyi et al. studied the prediction of the ANCHOR server for detecting binding regions in disordered proteins with the structure of human calcium/calmodulin-dependent protein kinase IV and verified its credibility (Dosztanyi et al., 2009).

5.8.2 Screening of small molecules

The information from interacting residues at hot spots from the protein−protein complex is used to screen small-molecule compounds using peptidomimetic techniques.

For this, residues of the interacting protein are submitted for small-molecule screening in tools like pepMMsMIMIC. Alternatively, the hot spot residues on the target can be submitted as targeting points and virtual docking of small-molecule datasets from databases can be carried out using docking tools. For example, DockBlaster is an online server that screens small molecules in the Zinc database against the submitted target (Irwin et al., 2009).

pepMMsMIMIC is a free web tool that carries out virtual screening of peptidomimetic compounds. When a peptide or protein 3D structure is given as a query, it does a multiconformers 3D similarity search based on pharmacophore and shape similarity against 17 million conformers stored in the MMsINC database, which are generated from 4.5 million commercially obtainable chemicals. With a submitted protein−protein/peptide complex structure, pepMMsMIMIC first identifies three key residues that are responsible for complex formation or recognizes the user-entered residues. Peptide complexity is reduced and the basic pharmacophore model is defined by its critical structural features in 3D space. All possible peptide pharmacophore feature arrangements are enumerated to form the basis of a peptide pharmacophore bitstring. pepMMsMIMIC performs pharmacophore screening against multiconformers in the MMsINC database. All possible conformer pharmacophore feature arrangements are estimated to form the basis of a conformer pharmacophore bitstring. Engaging two scoring approaches, pharmacophore fingerprint similarity (PFS) and ultrafast shape recognition (USR), and their consensus, peptidomimetic candidates are ranked according to similarity and the best top 200 are displayed as hits. The USR is a fast 3D similarity search method and in the encoding, the shape of the atomic ensemble is characterized by the distributions of atomic distances to four reference locations: the molecular centroid (*ctd*), the closest atom to ctd (*cst*), the farthest atom to ctd (*fct*), and the farthest atom to fct (*ftf*). Overall, each of these distributions is described through its first three vectors. In this way, each molecule is associated with a vector of 12 shape descriptors. PFS measure has been implemented based on a weighted similarity index (S_w), which is computed as:

$$S_w = c/(c + 2.5x\, m)$$

where c is the number of common bits between the peptide query fingerprint and the conformer's fingerprint, and m is the count of bits in the query fingerprint but not the conformer's fingerprint.

Four different scoring methods are actually implemented in the current version of pepMMsMIMIC: (1) shape score (ShS) based on the USR methods; (2) PFS based on weighted similarity coefficient S_w; (3) combined ShS and PFS filtering; and (4) hybrid scoring function, which is a weighted combination of the ShS and PFS approach (Floris et al., 2011; http://mms.dsfarm.unipd.it/pepMMsMIMIC).

For its use, using the web address, open the tool and upload the protein−protein complex .pdb file consisting of a target and an interacting protein. The 3D structure of the complex will be displayed on the Jmol visualization window. Based on the interactions in the complex analyzed previously, either directly from the 3D structure or from the display boxes below where all the residues are listed in the

pull-down menu, select three residues (linear or nonlinear) of the interacting protein that interacts with the hot spot residues on the target protein. Also, for these three residues, the user has a choice to select CO and/or NH interactors derived from the carbamide bonds of the backbone and/or the corresponding side chain moiety using the checker box to include the atoms of these for similarity search. Four different similarity searches are available: (1) only shape similarity, (2) only pharmacophoric similarity, (3) shape-based filtering of pharmacophoric similarity, and (4) hybrid search (60% pharmacophoric similarity, 40% shape), to search for small molecule conformers. Select a type of similarity search and run the program.

In the output of the program, for each search, the top 200 small molecule hits are displayed in their 2D representation, along with their MMsINC database ID and link to the database, where the known and predicted properties of the molecule can be discerned. The hits can be downloaded as a single .sdf file.

A host−pathogen interaction inhibition study was done by Alam, where PPI inhibition strategy was followed. Apical membrane antigen 1 protein of *Plasmodium falciparum* found on the surface of the organism, which interacts with rhoptry neck protein (PfRON2) on erythrocytes of humans, was targeted. Based on the interacting residues of PfRON2, a peptide similarity search to identify small molecules was done with the pepMMsMIMIC server and the top five peptidomimetics were taken for docking analysis with AutoDock Vina. The molecules MMs03919469, MMs03919369, MMs0391948, MMs03919367, and MMs02548719, which were found to bind at the expected hydrophobic groove of PfAMA1, were portrayed as potential lead compounds for designing antimalarials (Alam, 2014).

For entry into human cells, the human immunodeficiency virus needs its envelope glycoprotein gp120 to interact with the CD4 glycoprotein and a chemokine receptor on the human cell surface (Kwong et al., 1998). In a study by Andrianov et al. from crystal data of the gp120−CD4 complex, CD4 amino acid residues responsible for specific interactions with gp120 were used as the input data for the pepMMsMIMIC tool. The peptidomimetic candidates found were docked against gp120 and evaluated by MD simulations and binding free energy calculations to find the potential inhibitor (Andrianov et al., 2015).

5.8.3 Prediction of ADME/T properties

ADME/T properties are predicted to validate drug-likeness of compounds and to optimize their structure to enhance target binding affinities and drug-likeness qualities. It is suggested to carry out this study after the docking and simulation procedure to avoid filtering out valid inhibitors whose properties fall out of the recommended range. Online free tools like SwissADME (http://www.swissadme.ch/), admetSAR (http://lmmd.ecust.edu.cn/admetsar2/), and molinspiration (https://www.molinspiration.com) and commercial software like QikProp (Schrodinger), ADMET Descriptors/Collection (Accelrys), and MetaSite (Molecular Discovery) are widely used to predict ADME/T properties of the molecules.

The **QikProp** program from Schrodinger Maestro (commercial) is used to calculate the ADME properties of the small-molecule ligands. It evaluates numerically the pharmacokinetic properties of the molecules, which can be compared with the recommended values provided by the program. The ranges of the recommended values are calculated based on the analysis of 95% known drugs. In the development of QikProp, using BOSS program and OPLS-AA force field, Monte Carlo statistical mechanics simulations were performed on organic solutes in periodic boxes of explicit water molecules, which resulted in configurational averages for a number of descriptors. Correlations of these descriptors to experimentally determined properties were made, and then algorithms that mimic the full Monte Carlo simulations were developed. While performing an evaluation of the user-submitted molecules, QikProp rapidly analyses atom types and charges, rotor counts, volume, and surface area of the molecules. It then uses this information, along with the physical descriptors calculated using the QikProp developed algorithms, in the regression equations. The result is an accurate prediction of a molecule's pharmacologically relevant properties. QikProp is run in normal mode or fast mode, where in normal mode 44 properties for nearly 10,000 molecules are predicted in an hour and in fast mode 40 properties of approximately 300,000 compounds are predicted in an hour. Fast mode skips some calculations like dipole moment, ionization potential, etc. (QikProp, Schrödinger, LLC, New York, NY).

To run the program, from the Schrodinger Suite, select Maestro and run QikProp from the application. As the ligands to be submitted to QikProp should be in 3D structures and hydrogen atoms has to be explicit, the ligands are priorly prepared using LigPrep, the ligand preparation wizard in Maestro. In Maestro, on the Project toolbar, click Import and import the ligand molecules from the .mae file. The ligands are imported as individual entries. From the Project toolbar select the Table button to see the list of ligands and the entries that are selected. From Applications, open the QikProp panel. From Use structures from the option menu, select the Project Table (selected entries) or browse and select the file from system and click start. From the Start dialog box, from the Incorporate option, choose to replace existing entries. In the Name box, type any name (ligands) and click start. QikProp by default runs in normal mode; to run in fast mode, select Fast Mode. Also, the user can opt to search for a similar specified number of drug molecules before starting the process. When the job is over, .mae, .qpsa, .out, .log, and .csv files are found in the working directory. Choose Project and click Save as. In the box opening, name the project (.prj) and click Save. All the files will be saved as a project.

The .csv file contains all predicted properties of the molecules. In the output, the evaluation of descriptors and properties are numerically represented to be compared with the given recommended range values. A few of the descriptors and their recommended values are polarizability: from 13.0 to 70.0, hexadecane/gas partition coefficient: from 4.0 to 18.0, octanol/gas partition coefficient from 8.0 to 35.0, and octanol/water partition coefficient: from −2.0 to 6.5.

Falchi et al. reviewed studies on small-molecule modulators of PPIs obtained from different virtual screening strategies. Potential small-molecule inhibitors

from the studies were evaluated for their main pharmacokinetic properties using the QikProp program of Schrodinger (Falchi et al., 2014). This study gave an overview of the characteristics of small-molecule modulators used in PPI targeting studies.

Also, there are many tools to predict only the toxicity profile of small molecules, based on various strategies. Commercial tools like TOPKAT, DEREK, and MCASE and freely available tools like TEST, TOXTRE (http://toxtree.sourceforge.net/), and LAZAR (https://openrisknet.org/e-infrastructure/services/110/) are widely used. Also, toxicity prediction from the admetSAR tool (Cheng et al., 2012) is extensively used in many studies (Priya, 2017; Parulekar and Sonawane, 2017), and many prominent databases like DrugBank compute the pharmacokinetics profile of the drug compounds using this tool.

TEST (Toxicity Estimation Software Tool) estimates toxicity values for chemicals using quantitative structure−activity relationship (QSAR) methodologies, which calculate the toxicity profile based on physical characteristics of molecular structures called molecular descriptors. QSARs are mathematical models and simple QSAR models calculate the toxicity of chemicals using a simple linear function of molecular descriptors:

$$Toxicity = ax_1 + bx_2 + c$$

where x_1 and x_2 are the independent descriptor variables and a, b, and c are fitted parameters. For the user-submitted molecules, TEST calculates the required molecular descriptors and the toxicity is estimated using one of several advanced QSAR methodologies like the hierarchical method, functional data analysis (FDA) method, single-model method, group contribution method, and nearest neighbor method (Martin et al., 2012; https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test).

The program has to be downloaded and installed and the following steps are to be carried out: open the software window and either enter the ID of the small molecule or draw the structure in the chemical sketcher window. Also, the structure can be imported from the structure file or as a .mol file, or generated from SMILES. From the End Point pulldown menu, choose the toxicity dataset that has to be compared. The datasets available for comparison are 96 h fathead minnow LC50, 40 h Tetrahymena pyriformis IGC50, 48 h Daphnia magna LC50, oral rat LD50, developmental toxicity, bioaccumulation factor, and Ames mutagenicity. From the Method pulldown menu, select the methodology required. Hierarchical clustering, FDA, single model, and nearest neighbor are some of the methods that can be selected. From the options box, the user can allow relaxing fragment constraint and select the output folder. Click Calculate.

In the output, the numerical values for the predicted properties are displayed along with experimental values if available. Also, predicted values of the most similar compounds are displayed. Graphical representations of the predicted results are saved as a .png file.

Parulekar et al. carried out studies on aminoglycoside phosphotransferases (APHs) in the multidrug-resistant organism *Bacillus subtilis* strain RK, which is

responsible for aminoglycoside antibiotics resistance. In this study, through in vitro and in silico studies, the molecule ZINC71575479 was identified as a potential inhibitor for APH. The toxicity profile of this molecule was predicted with the TEST tool and its comparison with known inhibitor tyrphostin AG1478 identified it to be a valid molecule for drug development (Parulekar et al., 2019).

5.9 Conclusion

Present *in silico* tools and biological databases largely assist in the identification of small molecules for PPI inhibition. A deep understanding of the molecular basis of the disease and the investigation of the target will facilitate discovery of potent inhibitors to PPI using these tools. However, it should be noted that all PPIs are not easy to target because of their interaction complexity and limited number of experimental structures with high resolution. Also, the molecules in screening libraries are generated for traditional targets like enzymes, while molecules of a wider range of chemical diversities are required to inhibit PPIs. Hence, advancements in the molecular biology field, along with expansion of bioinformatics databases and screening libraries with more sophisticated tools and strategies, will make PPIs the most promising drug targets.

References

Alam, A., 2014. Bioinformatic identification of peptidomimetic-based inhibitors againstPlasmodium falciparumAntigen AMA1. Malar. Res. Treat. 2014, 1–8.

Alonso-López, D., Gutiérrez, M., Lopes, K., Prieto, C., Santamaría, R., De Las Rivas, J., 2016. APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Res. 44 (W1), W529–W535.

Anand, P., Brown, J., Lin, C., Qi, J., Zhang, R., Artero, P., et al., 2013. BET bromodomains mediate transcriptional pause release in heart failure. Cell 154 (3), 569–582.

Andrianov, A., Kashyn, I., Tuzikov, A., 2015. 185 Virtual screening of novel anti-HIV-1 agents targeting CD4-binding site of the envelope gp120 protein. J. Biomol. Struct. Dyn. 33 (Suppl. 1), 122–123.

Arkin, M., Wells, J., 2004. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. Nat. Rev. Drug Discov. 3 (4), 301–317.

Assi, S., Tanaka, T., Rabbitts, T., Fernandez-Fuentes, N., 2009. PCRPi: presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. Nucleic Acids Res. 38 (6) e86-e86.

Badal, V., Kundrotas, P., Vakser, I., 2015. Text mining for protein docking. PLoS Comput. Biol. 11 (12), e1004630.

Bader, G., 2003. BIND: the biomolecular interaction network database. Nucleic Acids Res. 31 (1), 248–250.

Balani, S., Miwa, G., Gan, L., Wu, J., Lee, F., 2005. Strategy of utilizing in vitro and in vivo ADME tools for lead optimization and drug candidate selection. Curr. Top. Med. Chem. 5 (11), 1033−1038.

Barage, S., Jalkute, C., Dhanavade, M., Sonawane, K., 2014. Simulated interactions between endothelin converting enzyme and Aβ peptide: insights into subsite recognition and cleavage mechanism. Int. J. Pept. Res. Therapeut. 20 (4), 409−420.

Barale, S., Parulekar, R., Fandilolu, P., Dhanavade, M., Sonawane, K., 2019. Molecular insights into destabilization of Alzheimer's Aβ protofibril by arginine containing short peptides: a molecular modeling approach. ACS Omega 4 (1), 892−903.

Barbosa, A., Roque, A., 2019. Free marine natural products databases for biotechnology and bioengineering. Biotechnol. J. 14 (11), 1800607.

Barlow, K., Conchúir, S., Thompson, S., Suresh, P., Lucas, J., Heinonen, M., et al., 2018. Flex ddG: Rosetta ensemble-based estimation of changes in protein−protein binding affinity upon mutation. J. Phys. Chem. B 122 (21), 5389−5399.

Basse, M., Betzi, S., Morelli, X., Roche, P., 2016. 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein−protein interactions. Database baw007, 2016.

Beard, H., Cholleti, A., Pearlman, D., Sherman, W., Loving, K., 2013. Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. PloS One 8 (12), e82849.

Bogan, A., Thorn, K., 1998. Anatomy of hot spots in protein interfaces. J. Mol. Biol. 280 (1), 1−9.

Braun, P., Gingras, A., 2012. History of protein-protein interactions: from egg-white to complex networks. Proteomics 12 (10), 1478−1498.

Cesa, L., Mapp, A., Gestwicki, J., 2015. Direct and propagated effects of small molecules on protein−protein interaction networks. Front. Bioeng. Biotechnol. 3, 119.

Cheng, F., Li, W., Zhou, Y., Shen, J., Wu, Z., Liu, G., et al., 2012. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. J. Chem. Inf. Model. 52 (11), 3099−3105.

Conte, L., Chothia, C., Janin, J., 1999. The atomic structure of protein-protein recognition sites 1 1Edited by A. R. Fersht. J. Mol. Biol. 285 (5), 2177−2198.

Coyne, A., Scott, D., Abell, C., 2010. Drugging challenging targets using fragment-based approaches. Curr. Opin. Chem. Biol. 14 (3), 299−307.

Dawidowski, M., Emmanouilidis, L., Kalel, V., Tripsianes, K., Schorpp, K., Hadian, K., et al., 2017. Inhibitors of PEX14 disrupt protein import into glycosomes and kill Trypanosoma parasites. Science 355 (6332), 1416−1420.

De Las Rivas, J., Fontanillo, C., 2010. Protein−protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput. Biol. 6 (6), e1000807.

Devi, S., Tandon, H., Ampasala, D., 2015. Identification of potent bromodomain4 (brd4) inhibitors by energy pharmacophore based virtual screening to target brd4-nut midline carcinoma. Int. J. Pharm. Pharm. Sci. 7 (4), 77−84.

Dhanavade, M., Jalkute, C., Barage, S., Sonawane, K., 2013. Homology modeling, molecular docking and MD simulation studies to investigate role of cysteine protease from Xanthomonas campestris in degradation of Aβ peptide. Comput. Biol. Med. 43 (12), 2063−2070.

Dhanavade, M., Parulekar, R., Kamble, S., Sonawane, K., 2016. Molecular modeling approach to explore the role of cathepsin B from Hordeum vulgare in the degradation of Aβ peptides. Mol. Biosyst. 12 (1), 162−168.

Dhanavade, M., Sonawane, K., 2014. Insights into the molecular interactions between amino-peptidase and amyloid beta peptide using molecular modeling techniques. Amino Acids 46 (8), 1853−1866.

Dosztanyi, Z., Meszaros, B., Simon, I., 2009. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform 11 (2), 225−243.

Eguchi, M., McMillan, M., Nguyen, C., Teo, J., Chi, E., Henderson Jr., W., 2003. Chemogenomics with peptide secondary structure mimetics. Comb. Chem. High Throughput Screen. 6 (7), 611−621.

Falchi, F., Caporuscio, F., Recanatini, M., 2014. Structure-based design of small-molecule protein−protein interaction modulators: the story so far. Future Med. Chem. 6 (3), 343−357.

Fernandez, A., Scheraga, H., 2002. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. Proc. Natl. Acad. Sci. U. S. A. 100 (1), 113−118.

Finn, R., Marshall, M., Bateman, A., 2004. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21 (3), 410−412.

Floris, M., Masciocchi, J., Fanton, M., Moro, S., 2011. Swimming into peptidomimetic chemical space using pepMMsMIMIC. Nucleic Acids Res. 39 (Suppl. l), W261−W269.

Fontaine, F., Overman, J., François, M., 2015. Pharmacological manipulation of transcription factor protein-protein interactions: opportunities and obstacles. Cell Regen. 4 (1), 4:2.

Ghoorah, A., Devignes, M., Smaïl-Tabbone, M., Ritchie, D., 2013. Kbdock 2013: a spatial classification of 3D protein domain family interactions. Nucleic Acids Res. 42 (D1), D389−D395.

Goll, J., Rajagopala, S., Shiau, S., Wu, H., Lamb, B., Uetz, P., 2008. MPIDB: the microbial protein interaction database. Bioinformatics 24 (15), 1743−1744.

Gulten, G., Sacchettini, J., 2013. Structure of the Mtb CarD/RNAP β-lobes complex reveals the molecular basis of interaction and presents a distinct DNA-binding domain for Mtb CarD. Structure 21 (10), 1859−1869.

Han, H., Cho, J., Lee, S., Yun, A., Kim, H., Bae, D., et al., 2017. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 46 (D1), D380−D386.

Hayes, M., Soto-Velasquez, M., Fowler, C., Watts, V., Roman, D., 2017. Identification of FDA-approved small molecules capable of disrupting the calmodulin−adenylyl cyclase 8 interaction through direct binding to calmodulin. ACS Chem. Neurosci. 9 (2), 346−357.

Higueruelo, A., Jubb, H., Blundell, T., 2013. TIMBAL v2: update of a database holding small molecules modulating protein−protein interactions. Database bat039, 2013.

Higurashi, M., Ishida, T., Kinoshita, K., 2009. PiSite: a database of protein interaction sites using multiple binding states in the PDB. Nucleic Acids Res. 37 (Database), D360−D364.

Hu, H., Miao, Y., Jia, L., Yu, Q., Zhang, Q., Guo, A., 2018. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 47 (D1), D33−D38.

Illendula, A., Pulikkan, J., Zong, H., Grembecka, J., Xue, L., Sen, S., et al., 2015. A small-molecule inhibitor of the aberrant transcription factor CBF -SMMHC delays leukemia in mice. Science 347 (6223), 779−784.

Irwin, J., Shoichet, B., Mysinger, M., Huang, N., Colizzi, F., Wassam, P., et al., 2009. Automated docking screens: a feasibility study. J. Med. Chem. 52 (18), 5712−5720.

Jalkute, C., Barage, S., Sonawane, K., 2015. Insight into molecular interactions of Aβ peptide and gelatinase from *Enterococcus faecalis*: a molecular modeling approach. RSC Adv. 5 (14), 10488−10496.

Jin, L., Wang, W., Fang, G., 2014. Targeting protein-protein interaction by small molecules. Annu. Rev. Pharmacol. 54 (1), 435−456.

Jones, S., Thornton, J., 1997. Analysis of protein-protein interaction sites using surface patches 1 1Edited by G.Von Heijne. J. Mol. Biol. 272 (1), 121−132.

Koes, D., Camacho, C., 2011. Small-molecule inhibitor starting points learned from protein−protein interaction inhibitor structure. Bioinformatics 28 (6), 784−791.

Kuenemann, M., Sperandio, O., Labbe, C., Lagorce, D., Miteva, M., Villoutreix, B., 2015. In silico design of low molecular weight protein−protein interaction inhibitors: overall concept and recent advances. Prog. Biophys. Mol. Biol. 119 (1), 20−32.

Kundrotas, P., Alexov, E., 2006. Electrostatic properties of protein-protein complexes. Biophys. J. 91 (5), 1724−1736.

Kwong, P., Wyatt, R., Robinson, J., Sweet, R., Sodroski, J., Hendrickson, W., 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393 (6686), 648−659.

Labbé, C., Kuenemann, M., Zarzycka, B., Vriend, G., Nicolaes, G., Lagorce, D., et al., 2015. iPPI-DB: an online database of modulators of protein−protein interactions. Nucleic Acids Res. 44 (D1), D542−D547.

Lambert, M., Jambon, S., Depauw, S., David-Cordonnier, M., 2018. Targeting transcription factors for cancer treatment. Molecules 23 (6), 1479.

Lee, T., Young, R., 2013. Transcriptional regulation and its misregulation in disease. Cell 152 (6), 1237−1251.

Li, Q., Quan, L., Lyu, J., He, Z., Wang, X., Meng, J., et al., 2016. Discovery of peptide inhibitors targeting human programmed death 1 (PD-1) receptor. Oncotarget 7 (40), 64967−64976.

Lipinski, C., Lombardo, F., Dominy, B., Feeney, P., 2012. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. 64 (1−3), 4−17.

Lise, S., Buchan, D., Pontil, M., Jones, D., 2011. Predictions of hot spot residues at protein-protein interfaces using support vector machines. PloS One 6 (2), e16774.

Liu, S., Gao, Y., Vakser, I., 2008. dockground protein-protein docking decoy set. Bioinformatics 24 (22), 2634−2635.

Lu, M., Tan, S., Ji, J., Chen, Z., Yuan, Z., You, Q., et al., 2016. Polar recognition group study of Keap1-Nrf2 protein−protein interaction inhibitors. ACS Med. Chem. Lett. 7 (9), 835−840.

Ma, B., Elkayam, T., Wolfson, H., Nussinov, R., 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc. Natl. Acad. Sci. U. S. A. 100 (10), 5772−5777.

Ma, C., Yang, X., Lewis, P., 2016. Bacterial transcription as a target for antibacterial drug development. Microbiol. Mol. Biol. Rev. 80 (1), 139−160.

Macarron, R., 2006. Critical review of the role of HTS in drug discovery. Drug Discov. Today 11 (7−8), 277−279.

Marcotte, E., 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285 (5428), 751−753.

Marshall, G., Kuster, D., Che, Y., 2009. Chemogenomics with protein secondary-structure mimetics. Methods Mol. Biol. 575, 123−158.

Martin, T., Harten, P., Young, D., 2012. TEST (Toxicity Estimation Software Tool) Ver 4.1. U.S. Environmental Protection Agency, Washington, DC. EPA/600/C-12/006.

Mattapally, S., Singh, M., Murthy, K., Asthana, S., Banerjee, S., 2018. Computational modeling suggests impaired interactions between NKX2.5 and GATA4 in individuals carrying a novel pathogenic D16N NKX2.5 mutation. Oncotarget 9 (17), 13713−13732.

McDowall, M., Scott, M., Barton, G., 2009. PIPs: human protein-protein interaction prediction database. Nucleic Acids Res. 37 (Database), D651−D656.

Meireles, L., Domling, A., Camacho, C., 2010. ANCHOR: a web server and database for analysis of protein-protein interaction binding pockets for drug discovery. Nucleic Acids Res. 38 (Web Server), W407−W411.

Meireles, L., Mustata, G., 2011. Discovery of modulators of protein-protein interactions: current approaches and limitations. Curr. Top. Med. Chem. 11 (3), 248−257.

Mishra, G., 2006. Human protein reference database–2006 update. Nucleic Acids Res. 34 (90001), D411−D414.

Moreira, I., Fernandes, P., Ramos, M., 2007. Hot spots-A review of the protein-protein interface determinant amino-acid residues. Proteins Struct. Funct. Bioinf. 68 (4), 803−812.

Morelli, X., Bourgeas, R., Roche, P., 2011. Chemical and structural lessons from recent successes in protein−protein interaction inhibition (2P2I). Curr. Opin. Chem. Biol. 15 (4), 475−481.

Mosca, R., Céol, A., Aloy, P., 2013. Interactome3D: adding structural details to protein networks. Nat. Methods 10 (1), 47−53.

Mysinger, M., Carchia, M., Irwin, J., Shoichet, B., 2012. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J. Med. Chem. 55 (14), 6582−6594.

Ohue, M., Matsuzaki, Y., Uchikoga, N., Ishida, T., Akiyama, Y., 2013. MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data. Protein Pept. Lett. 21 (8), 766−778.

Ohue, M., Matsuzaki, Y., Uchikoga, N., Ishida, T., Akiyama, Y., 2016. Megadock 4.0. An ultra-high-performance protein-protein docking software for heterogeneous supercomputers. Biophys. J. 110 (3), 327a.

Oltersdorf, T., Elmore, S., Shoemaker, A., Armstrong, R., Augeri, D., Belli, B., et al., 2005. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. Nature 435 (7042), 677−681.

Orii, N., Ganapathiraju, M., 2012. Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function. PloS One 7 (11), e49029.

Oughtred, R., Stark, C., Breitkreutz, B.J., Rust, J., Boucher, L., Chang, C., 2019. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 47 (D1), D529−D541.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., et al., 2004. The MIPS mammalian protein-protein interaction database. Bioinformatics 21 (6), 832−834.

Parulekar, R., Barale, S., Sonawane, K., 2019. Antibiotic resistance and inhibition mechanism of novel aminoglycoside phosphotransferase APH(5) from B. subtilis subsp. subtilis strain RK. Braz. J. Microbiol. 50 (4), 887−898.

Parulekar, R., Sonawane, K., 2017. Molecular modeling studies to explore the binding affinity of virtually screened inhibitor toward different aminoglycoside kinases from diverse MDR strains. J. Cell. Biochem. 119 (3), 2679−2695.

Priya, V.G., 2017. The ADME/T profiles of TB drugs - an in silico analysis. World J. Pharmaceut. Res. 6 (14), 1202−1211.

Priya, V.G., Muddapur, U., Mehta, M., 2012. Computational analysis of *M. tuberculosis* - CarD protein. ALST 6, 8−16.

Priya, V.G., Swaminathan, P., Muddapur, U., Fandilolu, P., Parulekar, R., Sonawane, K., 2018. Peptide similarity search based and virtual screening based strategies to identify small molecules to inhibit CarD−RNAP interaction in *M. tuberculosis*. Int. J. Pept. Res. Therapeut. 25 (2), 697−709.

Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J., 2006. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins Struct. Funct. Bioinf. 63 (3), 490−500.

Raman, K., 2010. Construction and analysis of protein−protein interaction networks. Autom. Exp. 2 (1), 2.

Ran, X., Gestwicki, J., 2018. Inhibitors of protein−protein interactions (PPIs): an analysis of scaffold choices and buried surface area. Curr. Opin. Chem. Biol. 44, 75−86.

Sable, R., Jois, S., 2015. Surfing the protein-protein interaction surface using docking methods: application to the design of PPI inhibitors. Molecules 20 (6), 11569−11603.

Schaefer, U., Schmeier, S., Bajic, V., 2010. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. Nucleic Acids Res. 39 (Database), D106−D110.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L., 2005. The FoldX web server: an online force field. Nucleic Acids Res. 33, W382−W388.

Sonawane, K., Barage, S., 2014. Structural analysis of membrane-bound hECE-1 dimer using molecular modeling techniques: insights into conformational changes and Aβ1−42 peptide binding. Amino Acids 47 (3), 543−559.

Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A., Berendsen, H., 2005. GROMACS: fast, flexible, and free. J. Comput. Chem. 26 (16), 1701−1718.

Stallings, C., Stephanou, N., Chu, L., Hochschild, A., Nickels, B., Glickman, M., 2009. CarD is an essential regulator of rRNA transcription required for *Mycobacterium tuberculosis* persistence. Cell 138 (1), 146−159.

Stein, A., Ceol, A., Aloy, P., 2010. 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 39 (Database), D718−D723.

Szklarczyk, D., Morris, J., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al., 2016. The STRING database in 2017: quality-controlled protein−protein association networks, made broadly accessible. Nucleic Acids Res. 45 (D1), D362−D368.

Thakar, S., Dhanavade, M., Sonawane, K., 2019. LegumeDB: development of legume medicinal plant database and comparative molecular evolutionary analysis of matK proteins of legumes and mangroves. Curr. Nutr. Food Sci. 15 (4), 353−362.

Thakar, S., Ghorpade, P., Kale, M., Sonawane, K., 2015. FERN Ethnomedicinal Plant Database: exploring fern ethnomedicinal plants knowledge for computational drug discovery. Curr. Comput. Aided Drug Des. 11 (3), 266−271.

Tuncbag, N., Keskin, O., Gursoy, A., 2010. HotPoint: hot spot prediction server for protein interfaces. Nucleic Acids Res. 38, W402−W406.

Turner, B., Razick, S., Turinsky, A., Vlasblom, J., Crowdy, E., Cho, E., et al., 2010. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database, 2010, baq023.

Vaquerizas, J., Kummerfeld, S., Teichmann, S., Luscombe, N., 2009. A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. 10 (4), 252−263.

Villoutreix, B., Kuenemann, M., Poyet, J., Bruzzoni-Giovanelli, H., Labbe, C., Lagorce, D., et al., 2014. Drug-like protein-protein interaction modulators: challenges and opportunities for drug discovery and chemical biology. Mol. Inf. 33 (6−7), 414−437.

Wang, L., Bao, Q., Xu, X., Jiang, F., Gu, K., Jiang, Z., et al., 2015. Discovery and identification of Cdc37-derived peptides targeting the Hsp90−Cdc37 protein−protein interaction. RSC Adv. 5 (116), 96138−96145.

Weiss, L., Harrison, P., Nickels, B., Glickman, M., Campbell, E., Darst, S., et al., 2012. Interaction of CarD with RNA polymerase mediates *Mycobacterium tuberculosis* viability, rifampin resistance, and pathogenesis. J. Bacteriol. 194 (20), 5621−5631.

Wingender, E., Dietze, P., Karas, H., Knüppel, R., 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 24 (1), 238−241.

Winter, A., Higueruelo, A., Marsh, M., Sigurdardottir, A., Pitt, W., Blundell, T., 2012. Biophysical and computational fragment-based approaches to targeting protein−protein interactions: applications in structure-guided drug discovery. Q. Rev. Biophys. 45 (4), 383−426.

Xenarios, I., 2000. DIP: the database of interacting proteins. Nucleic Acids Res. 28 (1), 289−291.

Xu, X., Ma, Z., Sun, H., Zou, X., 2016. SM-TF: a structural database of small molecule-transcription factor complexes. J. Comput. Chem. 37 (17), 1559−1564.

Zinzalla, G., Thurston, D., 2009. Targeting protein−protein interactions for therapeutic intervention: a challenge for the future. Future Med. Chem. 1 (1), 65−93.

CHAPTER

6 Advanced approaches and in silico tools of chemoinformatics in drug designing

Shweta Kulshrestha[1,a], Tanmay Arora[1], Manisha Sengar[2], Navneet Sharma[3], Raman Chawla[1], Shereen Bajaj[1], Pawan Kumar Raghav[4,a]

[1]*Division of CBRN Defence, Institute of Nuclear Medicine & Allied Sciences, DRDO, New Delhi, Delhi, India;* [2]*Department of Zoology, Deshbandhu College, University of Delhi, New Delhi, Delhi, India;* [3]*Department of Textile and Fiber Engineering, Indian Institute of Technology, New Delhi, Delhi, India;* [4]*Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, Delhi, India*

6.1 Introduction

Nowadays, to overcome the challenges of experimental drug designing, computer-aided (in silico) methods are widely accepted with the applications of chemoinformatics. The field of chemoinformatics has emerged in the past decades to aid in the conventional drug-discovery process by utilizing various computational tools and inductive learning procedures (Sliwoski et al., 2014). The term chemoinformatics was first introduced by Frank Brown to establish the use of chemical information for curating better strategies in developing drug–target interaction (DTI). It uses the power of computers, "informatics," and "chemistry" to design new drugs and predict complex activities such as toxicity, metabolism, carcinogenesis, drug–drug interactions, and chemical evaluation (Cooper, 2004). It has advanced chemical research and has substantially shaped the drug-discovery process.

Drug discovery is a broad field that encompasses the process of chemical identification, optimization, screening, and activity prediction (Sinha et al., 2017). The development of a new drug has always remained a quest in the current field of biomedicine. The process of drug discovery is divided into four steps, namely (1) data collection; (2) preprocessing; (3) high-throughput virtual screening; and (4) selectivity based on absorption, distribution, metabolism, excretion, and toxicity (ADMET) and chemical drug likeness (Lipinski rule of five) (Keiser et al., 2009) (Fig. 6.1).

The initial step of the drug-discovery process consists of potential target identification and compound collection from freely or commercially available databases. The preprocessing step involves the validation of the intended actual target and

[a] Equal contribution.

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00006-3

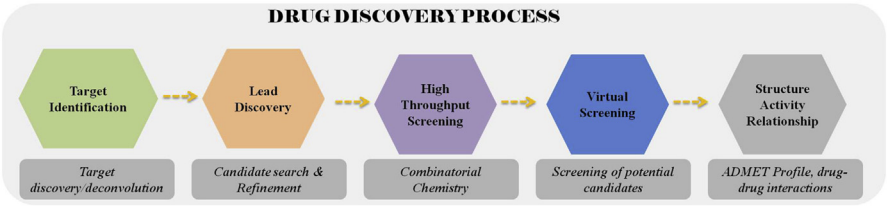


FIGURE 6.1

Schematic representation of the steps in in silico-based drug designing.
ADMET, Absorption, distribution, metabolism, excretion, and toxicity.

entity specifically interacting with the target (Lagarde et al., 2019). Furthermore, the computational method allows the screening of thousands of compounds that helps to discover potential drug candidates utilizing statistical and machine-learning (ML) models called "virtual screening (VS)" (Patrick Walters et al., 1998). VS is usually a knowledge-driven method that allows sequential filters to narrow down and select a set of "lead-like" hits. It is mainly divided into two broad categories, i.e., structure-based VS (SBVS) and ligand-based VS (LBVS). Databases of up to 10 million compounds can be handled for any category of VS experiment. SBVS is based on prior knowledge of the 3D structure of the biological target. It is the method of virtual high-throughput screening of compounds aimed at identifying whether a given dataset of compounds can interact with a prespecified target or not (Bohacek et al., 1996). It helps in the prediction of DTI in which unlikely drug−target combinations can be eliminated and high-affinity active combinations can be selected for clinical experimentation. SBVS can easily be explored using the molecular docking approach and tools working with different binding and scoring algorithms such as DOCK, AutoDock, Glide, GOLD, etc. (Bajusz et al., 2017).

On the other hand, LBVS is considered a nonstructure-based VS method. It relies on compound activity data derived from a set of known compound activities (Jahn et al., 2009). Pharmacophore designing is an important model in LBVS, and recognizes putative active compounds with diverse chemical features. It helps in the designing of new candidates and is easily explored using pharmacophore-mapping, shape-matching, and similarity search (Geppart et al., 2010; Lavecchia and Giovanni, 2013) tools such as PHASE, RAPID, Mo-inspiration, etc. Apart from screening, the pharmacophore model can also be used in quantitative structure−activity relationship (QSAR) studies for activity predictions. QSAR is the assumption with respect to the relationship of a compound activity with its biological potency. QSAR can be classified based on the dimensions of compound feature representations, namely 2D QSAR and 3D QSAR methods. 2D QSAR is based on topological features of the compound, while 3D QSAR utilizes the geometrical descriptors for model generation and activity prediction (Neves et al., 2018).

ML algorithms are gaining popularity in designing robust QSAR models. Different supervised and unsupervised learning algorithms such as *k*-nearest neighbor (*k*NN), support vector machine (SVM), artificial neural network (ANN),

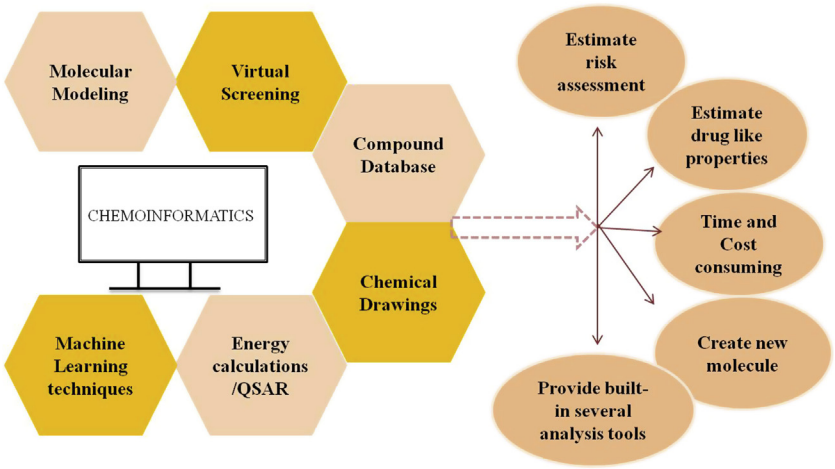


FIGURE 6.2

Graphical representation of scope of various activities of chemoinformatics.
QSAR, Quantitative structure–activity relationship.

and others have been applied for compound classification and regression scoring (Chen et al., 2018). ML is advantageous for chemoinformatics to train diversity of datasets and is capable of screening a large number of compounds with an impressive yield of activity prediction.

Understanding different applications and the need for chemoinformatics tools to follow different steps in the drug-discovery process are of utmost importance (Fig. 6.2). Therefore in the present framework, we explore various ligand resources that provide information that is vital to perform VS. Besides, we also report various tools to perform molecular docking such as AutoDock, AutoDock Vina, Glide, GOLD, FlexX, and Fred/Hybrid, and pharmacophore designing such as Molinspiration, MolSoft, Moka, and others. Furthermore, QSAR prediction tools and ML algorithms, mainly SVM, linear discriminant analysis (LDA), naïve Bayesian, random forest, *k*NN, ANN, and deep learning (DL), are included that help to outperform VS and QSAR tools.

6.2 Current chemoinformatics approaches and tools

6.2.1 Ligand databases/libraries

Compound databases are the platforms that allow users to implement efficient storage and retrieval of information about chemicals. To accomplish different VS, similarity search calculations, and QSAR tasks, various databases have been developed (Masoudi-Sobhanzadeh et al., 2020) that store vast amounts of information in a well-organized format. Table 6.1 provides information on the database tools, features, and available URLs.

Table 6.1 Public databases of compounds/chemicals/drug molecules and links used in chemoinformatics.

Databases	Description of tool (web link)
DrugBank	A chemical repository with physical and structural details. This database offers suites describing clinical-level information and free access to resources. It consists of more than 14,000 drug entries and Food and Drug Administration (FDA)-approved small molecules. Discovery phase molecules are also available in this database (Law et al., 2014). (http://www.drugbank.com)
PubChem	A chemically oriented public repository and information resource (Kim et al., 2016). It organizes the data into three linked databases, i.e., substance, compound, and bioassay database. It also holds the data of regulatory agencies such as the FDA, US Environmental Protection Agency, and substance registry (http://pubchem.ncbi.nlm.nih.gov)
ChemBL	A repository of bioactive molecules with drug-like properties determined using different assays such as binding affinity and absorption, distribution, metabolism, excretion, and toxicity assays. It shares its database with PubChem and complements information with shared portals. It also interlinks the chemical and biological data that aids in translational research (Gaulton et al., 2017) (https://www.ebi.ac.uk)
ZINC	A prepared library of comprehensive chemical compounds available with 3D structure features for virtual screening (Irwin and Shoichet, 2005). This library contains ~250,000 compounds and most of the compounds are drug-like or lead-like that are immediately usable by different virtual screening tools. The library is freely available in different formats such as SMILES, mol2, sdf, and DOCK files along with vendor and purchasing details (https://zinc.docking.org)
NCI	A directory of small molecules, therapeutic structures, and a depository for researchers mainly for cancer research (Bykov et al., 2002) (https://cactus.nci.nih.gov)
ChemDB	A database of stereochemical and geometrical information for small molecules (Chen et al., 2005). It handles stereochemical and geometrical information of molecules. It also provides a wide variety of flexible threshold functions and filters for determining molecular features (https://cdb.ics.uci.edu/)
Chemspider	An online chemical database consisting of more than 25 million chemical molecules and ~25,000 spectroscopic data. It aggregates the data with different platforms (nearly 400 different sources) and is known as the Google of chemicals (Pence and Williams, 2010) (www.chemspider.com)
BindingDB	Provides a library of compounds with measured binding affinities. It helps in the classification of new compounds by searching binding properties with similar compounds. It holds more than 25,000 binding affinity measurements utilized in virtual screening, and also provides a training set of ligands for quantitative structure–activity relationship (Gilson et al., 2016) (https://www.bindingdb.org)

Table 6.1 Public databases of compounds/chemicals/drug molecules and links used in chemoinformatics.—*cont'd*

Databases	Description of tool (web link)
PDB-Bind	A database that provides a collection of ligand–receptor complexes available in PDB format. The data from this database help in molecular recognition events where the information about structure and energy scoring functions are deposited. Currently, more than 25,000 binding complexes are deposited, including protein–ligand, protein–protein, protein–nucleic acid, and nucleic acid–ligand complexes (Liu et al., 2015) (https://www.pdbbind.org.cn)
PDBeChem	A chemical dictionary of structures, stereoisomers, isomers, and enantiomers of small molecules (Velankar et al., 2016) (https://www.ebi.ac.uk)
KEGG Compound	A collection of small compounds, biopolymers, and chemical substances applied to biological systems. This database has been employed with KEGG pathways and KEGG drug networking online portals (Kanehisa et al., 2017) (https://www.genome.jp)
HMDB	A human metabolome database consisting of a collection of small molecule metabolites found in biological systems. This database is interlined with KEGG, and Reactome collections to meet the requirements in metabolomics. It provides quantitative and analytical information about metabolites, associated enzymes, transporters, and disease-related pathology (Zhou et al., 2012) (https://hmdb.ca)
SMPDB	A visual dictionary specifically designed for small molecule pathways such as different metabolic and drug-action pathways. This database is hyperlinked to other databases such as DrugBank and HMDB, and is accompanied by detailed descriptions of all chemicals (Frolkis et al., 2009) (https://smpdb.ca)
HIT	Consists of herbal-derived compounds and protein target information. It covers a broad range of herbal ingredients that play roles as activators, inhibitors, agonists, and antagonists. It searches based on the keyword hit method and covers more than 5000 herbal entries in the database (Benesch et al., 2010) (https://omictools.com/hit-tool)
TTD	The first database that contains information about clinical therapeutic proteins and nucleic acid drug targets. Drugs with known specific target function are included in this with more than 10,000 data entries (Chen et al., 2002) (https://bidd.nus.edu.sg)
PharmGKB	A clinically oriented drug encyclopedia that displays genotype, molecular, and clinical information about drugs and candidate genes. It helps in building pharmacogenomics relationships with small compounds (Thorn et al., 2013) (https://www.pharmgkb.org)
SuperNatural	A database based on natural product descriptions. It mainly includes secondary plant metabolites along with physicochemical descriptions and toxicity profiles (Dunkel, 2006) (https://bioinformatics.charite.de)

In general, ligand libraries are mainly clinically oriented drug encyclopedias and databases. These are the resources that provide 2D and 3D structure, target activity measurement, physicochemical features, molecular descriptors, and clinically relevant literature-based data (Gupta et al., 2018). Prominent large-scale bioactivity data (IC_{50}, EC_{50}) and high-throughput screening experiment data have also been available in such databases, which adds complementary features. Furthermore, there are some resources that can be accessed to provide information about toxic metabolites and ligand–ligand interactions (Williams and Tkachenko, 2014; Richard et al., 2006).

6.2.2 In silico structure-based virtual screening

SBVS is also known as target-based VS, which predicts the interactions between ligand and molecular target. As a consequence, ligands are ranked according to affinity and the most potential ones are shown at the top. SBVS requires knowledge of the 3D structure of the target, so that the information can be predicted using in silico software. Among the approaches of SBVS, molecular docking is noteworthy in drug designing. The technique identifies drug-like binders from the extensive database of compounds. The main purpose is to devise specific electrostatic and stereochemical algorithms to search molecular recognition events, i.e., to link the interaction of the compound with biological targets (Kitchen et al., 2004). Shape and noncovalent interactions play essential roles in identifying the position, binding energetics, molecular interactions, and conformational changes between a compound (ligand) and target (receptor) for docking (Brooijmans and Kuntz, 2003). The docking protocols are composed of search algorithms and score functions to achieve accurate SBVS (Weng et al., 1996). The basic workflow of the molecular docking-based VS is shown in Fig. 6.3.

Search algorithms are used to search for the orientations and conformations of the ligand at the binding site. To predict the conformation, the algorithm considers three types of ligand flexibility: systemic, stochastic, and deterministic. Furthermore, scoring function in docking is used to estimate the noncovalent force in a ligand–target complex. Prediction of binding affinity is the primary factor that decides the failure or success of a molecule. Forcefield and empirical-based scoring functions are widely used in software; however, hybrid-based and ML-based functions such as SVM, decision tree, and ensemble methods are gaining attention for their reliable prediction (Schulz-Gasch and Stahl, 2004). Some evolutionary-based algorithms such as genetic algorithm, anticolony optimization, local search, linear programming, statistical search, Monte Carlo, conformational space annealing, or stimulated annealing and similarity-based approaches also represent the quality of the generated protein–ligand complexes (Maia et al., 2020).

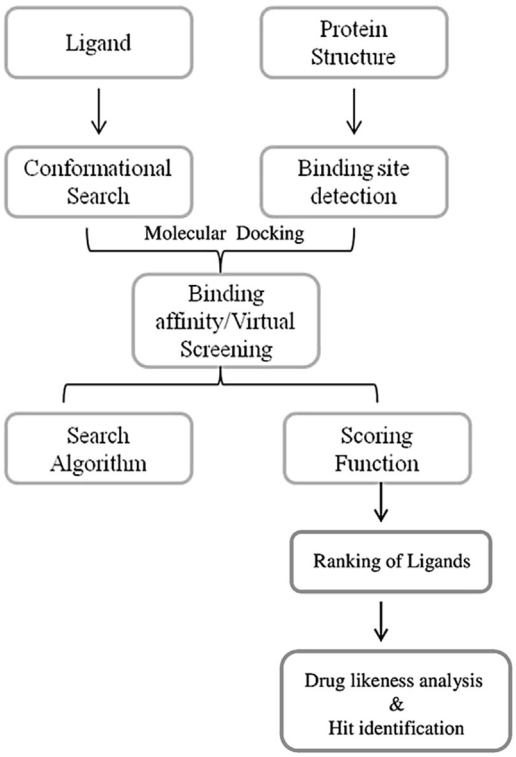


FIGURE 6.3

Molecular docking procedure for structure-based virtual screening of compounds.

6.2.2.1 Classes of molecular docking

1. Rigid docking
   In this type, the docking system does not allow movement in the conformations of receptor and ligand. It produces docked conformations with a favorable surface complementary method. The docking accuracy of a rigid system is suitable for protein−protein or protein−nucleic acid interaction (Zhao et al., 2015).
2. Semiflexible docking
   In this type, the docking system allows the alteration in the conformation of the ligand while keeping the conformation of the receptor fixed or unchanged. It is suitable for protein−ligand and nucleic acid−ligand interactions (Huanga and Caflischa, 2010).
3. Flexible docking
   In this type, the docking system allows movement in both receptor and ligand conformations. This type of docking is commonly applicable and provides

additional variables to increase conjugation affinity and make the process more reliable. It is suitable for ligand−ligand and ligand−receptor interactions (Rosenfeld et al., 1995).

6.2.2.2 Molecular docking tools

The performance of SBVS is based on the selection of a suitable docking program. A docking program provides the conformational search algorithm, best molecular complex poses, and high ranking of active compounds based on docking scoring function (Table 6.2).

Table 6.2 Commonly used docking programs used in chemoinformatics.

Program	Description (web link)
DOCK	Based on geometric molecular matching and forcefield scoring for semiflexible docking (https://dock.compbio.ucsf.edu)
AutoDock	Based on a Lamarckian genetic algorithm and forcefield scoring for semiflexible docking (https://autodock.scripps.edu)
GRAMM	Based on molecular matching exhaustive search and empirical free energy scoring for semiflexible docking (https://vakser.compbio.ku.edu)
FlexX	Based on an incremental construction algorithm and empirical free energy scoring for semiflexible docking (https://www.biosolveit.de/FlexX)
GOLD	Based on a genetic algorithm and molecular forcefield-based scoring for flexible docking (https://www.ccdc.cam.ac.uk/gold/)
Glide	Based on systemic search and empirical free energy-based scoring for semiflexible docking (https://www.schrodinger.com/glide/)
ICM	Based on stochastic random search, global minimization, and empirical free energy-based scoring for flexible docking (https://www.molsoft.com/docking.html)
CDOCKER (CHARMm)	Based on a molecular dynamics-simulated annealing algorithm and forcefield-based scoring for flexible docking
LigandFit	Based on a Monte Carlo or stochastic algorithm and forcefield-based scoring for rigid and flexible docking (https://accelrys.com/products/discovery-studio)
MolDOCK	Based on a hybrid dual algorithm-guided differential evolution and forcefield-based scoring for flexible docking (https://www.molsoft.com/docking)
AutoDock Vina	Based on a Broyden−Fletcher−Goldfarb−Shanno algorithm and empirical free energy-based scoring for semiflexible docking (https://vina.scripps.edu)
Surflex-DOCK	Based on an incremental construction algorithm, surface-based molecular algorithm, and forcefield-based scoring for flexible docking (https://www.tripos.com/index.php)
FRED	Based on an exhaustive search algorithm and Chemgauss scoring for rigid docking (https://www.eyesopen.com/oedocking)
HYBRID	Based on an exhaustive search algorithm and chemical Gaussian overlay for semiflexible docking (https://www.eyesopen.com/oedocking)
Affinity	Based on a Monte Carlo algorithm and forcefield-based scoring for flexible docking

1. DOCK
   DOCK is an open-source molecular docking tool. It is suitable for semiflexible-type docking using a geometric matching algorithm. It allows preorganization of the ligand into the molecular anchor, based on orientation; the geometric parameters are built for interactions. This method is usually called DOCK anchor docking. A forcefield energy-based scoring function is used, and optimization of the docked molecule is performed using the on-the-fly optimization method (Ewing et al., 2001).
2. AutoDock
   This is a freely available unbiased tool for semiflexible-type docking for proteins and ligands. It combines the simulated annealing and genetic algorithm to dock a complex rigid ligand with the target. The AMBER forcefield method is used for energy calculations and is generally used for substrate−enzyme complex evaluation (Morris et al., 1996).
3. AutoDock Vina
   AutoDock Vina is an open-source docking tool that uses the Broyden−Fletcher−Goldfarb−Shanno algorithm. The docking system is based on a gradient optimization method for manual selection of molecules using the AutoGrid program and allows cluster formations for scoring. It uses AutoDock PDBRT input file format and semiempirical free energy for evaluation purposes. This software is robust and fast for semiflexible docking with 80% accuracy results. Other AutoDock tools can be used with Vina to improve docking efficiencies such as LeDOCK, rDOCK, and UCSF DOCK (Docking et al., 1998; Gonczarek et al., 2018).
4. Glide
   Glide is a commercially available suite of semiflexible-type molecular docking, and utilizes the hierarchical filter to search for an optimized version of the native ligand with possible active sites. The shape and properties of the receptor are represented in the Glide grid to provide appropriate scoring of the bound ligand. Glide utilizes the systemic search algorithm for the conformational search function and an empirical-based scoring function algorithm to achieve acceptable results with ∼82% accuracy. The docking results are presented in the form of a Glide score. This program is suitable for ligand−receptor interactions and predicts binding affinities (Friesner et al., 2004).
5. GOLD
   GOLD (Genetic Optimization of Ligand Docking) is commercially applied software that performs automated flexible docking with full receptor−ligand flexibility and searches for space using a genetic algorithm. A simple scoring function is utilized in GOLD for hydrogen bonding, using a pairwise dispersion method to describe hydrophobic bonding. Furthermore, it uses a cavity detection method to define the active site in the receptor. The genetic algorithm's output is represented as the fittest conformations between ligand and receptor in the form of genetic algorithm fitness scores (Jones et al., 1997).

6. FlexX
   The FlexX (Fast Flexible Ligand Docking) program is a commercially available flexible-type docking tool based on the conformational flexibility model between receptor−ligand complexes predicted by geometric intermolecular constraints. FlexX requires the coordinates of the active site of the receptor for docking. Geometrically restricted interaction centers define the active site in the model. For docking, an incremental construction algorithm and complete linkage hierarchical cluster algorithm are used. These algorithms search for matching interaction groups between receptor and ligand and arrange them in the form of a tree-like structure. The output of FlexX is generally provided with ΔG values with the best prediction consisting the highest negative value (Rarey et al., 1996).
7. GRAMM
   GRAMM (Global Range Molecular Matching) is commercial software for semiflexible-type protein docking based on molecular function matching. It predicts high scoring best possible ligand conformations, which are further used for complex formation. It utilizes an exhaustive search algorithm through translation and molecular rotation to obtain a complex with high score steric fit. It allows 6D searches to predict the flexible interactions of molecular pairs, including protein−protein or protein−ligand docking. The program uses the empirical evaluation approach to produce the gross feature of the complex (Tovchigrechko and Vakser, 2005).
8. ICM
   ICM (Internal Coordinate Modeling) is a commercially flexible docking tool that involves internal coordinate variables such as bond length, bond angles, torsion angles, and phase dihedral angles to derive the algorithm for energy calculations. Systemic search or Monte Carlo simulation procedures are generally used to define essential docking components, i.e., energy function and search procedure. It predicts ligand−receptor interaction by global energy minimization (potential energy function) runs through the Cartesian coordinate space. It provides both a rigid and flexible type of docking with 52%−55% interaction accuracy (Abagyan et al., 1994).
9. CDOCKER
   CDOCKER (CHARMm) (CHARMm-based docking) is a molecular dynamics-simulated annealing-based algorithm for flexible protein−ligand docking. It is a combined grid-based docking tool that offers a forcefield energy scoring function. The docking strategy is based on the generation of several initial ligand conformations to target active sites followed by molecular dynamics-based simulation annealing and minimization for interactions. The output of CDOCKER is based on the flexibility of ligand, ligand size, and internal ligand geometry. Furthermore, this tool is best designed for medium-sized VS experiments with an accuracy of 50%−56% (Gagnon et al., 2016).

10. LigandFit
    LigandFit is a freely available shape-based docking tool that allows two essential procedures for interaction analysis: (1) identification of protein active site cavity using a flood fitting algorithm, and (2) docking that utilizes the Monte Carlo or stochastic algorithms for conformational search, selection of best ligand-compatible shape, and grid-based forcefield energy calculation to estimate the energy of receptor−ligand interaction. Being flexible and rigid, both types of docking are possible with a shape-directed ligand fit docking program with an accuracy of 45% (Venkatachalam et al., 2003).
11. Surflex
    Surflex is a commercially available automatic flexible docking algorithm. It reflects the incremental construction algorithm from the Hammerhead docking system with a search engine based on surface molecular similarity to generate the best possible poses. Results are presented by the Surflex utility screening tool using the molecular forcefield evaluation method. It is a fast method suitable for flexible docking with an 80% performance rate (Jain, 2003).
12. MolDOCK
    MolDOCK is a commercial flexible docking tool based on a new hybrid dual algorithm called guided differential evolution. It combines the heuristic search and cavity prediction algorithm for the conformational search process and accurate identification of molecule binding poses. The docking scoring function uses the piecewise linear potential along with hydrogen bond and electrostatic directionality. Furthermore, to improve docking accuracy, the rescoring function is introduced in the MolDOCK program (Thomsen and Christensen, 2006).
13. FRED and HYBRID
    FRED and HYBRID are commercially available tools that belong to OpenEye's OEDocking suite. FRED is a rigid docking program that uses only the structure of the protein to pose and score for docking, while HYBRID is a semiflexible program that uses the structure of the protein and ligand to pose and score. Both programs use an exhaustive search algorithm to dock the molecule. FRED allows only one protein to be docked with one ligand at a time and uses the Chemgauss scoring algorithm for scoring function. However, HYBRID allows multiple ligand docking and uses the chemical Gaussian overlay algorithm for a scoring of a docked molecule (McGann, 2012).

6.2.3 Pharmacophore development

The LBVS method applies because of the unavailability of drug−target structure. The most popular approaches for ligand-based VS are 3D pharmacophore development and QSAR modeling. A pharmacophore is an arrangement of molecular descriptors or elements related to biological activity (Yang, 2010). The pharmacophore comprises molecular descriptors such as hydrogen bond acceptor, hydrogen

bond donor, aromatic group, anion, cation, and hydrophobic group. These features are building blocks mainly extracted from the compounds that are known to be active (Wood et al., 2012). Subsequently, there are several commercial and free tools available that help in the prediction of molecular descriptors to design a pharmacophore. Generation of a pharmacophore requires the following steps: (1) a 3D structural database of ligands generated for feature extractions, (2) generation of multiple conformations from the compound database to produce new bioactive conformation, (3) identification and optimization of reciprocal properties, (4) alignment of features, and (5) generation of an active pharmacophore (Ghose et al., 2001).

6.2.3.1 Pharmacophore development tools

Pharmacophore designing can be performed using a suitable tool, either commercial or open source. A pharmacophore designing program provides the conformational space for chemical feature extraction and generation of new active conformations. Some important programs for pharmacophore designing are described next, while webserver and standalone software is listed in Table 6.3.

Table 6.3 List of pharmacophore designing tools.

Tool	Description (web link)
Mol-inspiration	Offers fragment-based ligand-based virtual screening (LBVS), molecular processing, and property calculations (http://molinspiration.com/egi-bin/proputus/)
QSIRIS Property Explorer	Helps in structure drawing and characterization of molecular descriptors (http://organic-chemistry.org/prog/peo/)
MolSoft	Helps in structure drawing, ligand editing, and clustering of large compound libraries (https://molsoft.com/mprop/)
MoKa	Helps in ligand editing using Grid molecular interaction fields (https://moldiscovery.com/software/moka)
Disco Tech	Generates low-energy conformations of pharmacophores.
GALAHAD	Helps in assigning macro definition features using genetic and rigid-body alignment algorithms
Ligand Scout	Allows extraction of molecular features and creates active conformations
PHASE	Identifies pharmacophore features and is used for overlapping and structure–activity data analysis (https://www.schrodinger.com/phase/)
GASP	Involves the superimposition of flexible molecules based on their proximity
PharmaGist	Free software for pharmacophore designing (https://bioinfo3d.cs.tau.ac.il/PharmaGist/)
ALADDIN	Helps in geometric, steric, and substructure searching
RAPID	Helps in identifying geometric invariants from the collection of small molecules

Table 6.3 List of pharmacophore designing tools.—*cont'd*

Tool	Description (web link)
MPHIL	Free software to derive a 3D pattern of pharmacophores utilizing feature and interfeature distances
SCAMPI	A program based on recursive partitioning and fast conformational search to design pharmacophores.
CoLibri	A standalone tool for LBVS (https://www.biosolveit.de/CoLibri/)
Decoy Finder	A standalone program that helps to find decoy molecules for active ligands (https://urvnutrigenomica-ctns.github.io/DecoyFinder/)
NNScore	A neural network-based program for LBVS (https://rocce-vm0.ucsd.edu/data/sw/hosted/nnscore/)
Epik	A standalone tool to develop ligand protonation states and tautomers (https://www.schrodinger.com/Epik)
SwissSimilarity	A webserver for complete LBVS and pharmacophore designing (https://www.swisssimilarity.ch/)
ZincPharmer	A web program that searches for pharmacophore descriptors from the ZINC compound library
ShaEP	A standalone tool used to align rigid molecular structures (https://users.abo.fi/mivainio/shaep/index.php)
BALLOON	Creates 3D atomic coordinates from distance geometry (https://web.abo.fi/fak/mnf/bkf/research/johnson/software.php)
React2D	Helps to combine fragmented libraries to form complete libraries
CATS	Performs chemical similarity search for small molecules
Autoclick Chem	A webserver and standalone software that performs chemical reactions
Shape-it	A tool for shape-based alignment using atomic Gaussians (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html)

1. Mol-inspiration
   Mol-inspiration is software that offers algorithms for molecular processing and property calculations. The program is written in Java language. It provides fragment-based VS of large databases, reproduces new conformations, and calculates various molecular properties such as pKa, ionic strength, and binding strength (Jarrahpour et al., 2012). The software is also available with molecular processing algorithms, including SMILES, SD file conversions, and high-quality molecule depiction.
2. QSIRIS Property Explorer
   This is free software written in Java language that computes different drug-relevant properties to form an active valid structure. It also helps to draw the structure, and calculates drug-relevant properties such as cLogP prediction, solubility prediction, and overall drug-likeness score (Martin et al., 1993).
3. MolSoft
   This is freely available software that helps in modeling structure and ligand editing. It allows the spatial organization of biological molecules, forecasts the

conformations of ligands, performs 2D to 3D conversion, allows clustering of large compound libraries, and predicts the compound descriptors and properties.

4. MoKa
   This is free downloading software for pharmacophore development. It implements ligand editing using the algorithm based on descriptors derived from the GRID molecular interaction field. It has an independent graphical user interface for system training and building of customized compound prediction models (Milletti et al., 2007).
5. Disco Tech
   Disco Tech (Distance Computing Technique) is a commercially available tool that helps to generate low-energy conformations. Using the location of atoms in the molecule and projections from the molecule to hydrogen bond acceptors and donors, it assists pharmacophore mapping. It uses a clique-specified number detection method to find the superpositions in the conformations. It is one of the fastest tools for mapping and is also able to compare the predicted pharmacophore model with alternative pharmacophore maps (Spitzer et al., 2010).
6. GALAHAD
   GALAHAD (genetic algorithm with linear assignment for the hypermolecular alignment of the database) is a commercially available pharmacophore developer tool. It helps in the iterative construction of hypermolecules that retain the individual attribute to identify target pharmacophores. It works on two algorithms: (1) genetic algorithm and (2) rigid-body alignment algorithm. This software uses macro definition fillers (such as acidity, basicity, and tautomerization) of encountering compounds to assign features in pharmacophores and allow molecular features to overlap. It is also available as a default suite with SYBYL-2 software for pharmacophore mapping (Richmond et al., 2006).
7. Ligand Scout
   Ligand Scout is a commercially available tool for the LBVS of pharmacophores. It allows extraction of active ligand features, creation of active conformations, validation, and interpretation of models within its graphical user interface (Wolber and Langer, 2005).
8. PHASE
   PHASE is an outstanding pharmacophore development suite available with the Schrodinger package. It is used to identify target pharmacophore features, mapping, overlapping, and rationalizing structure−activity data to develop a pharmacophore. It utilizes the Monte Carlo multiple minimum algorithm and low mode conformational searching algorithm for mapping and the rapid torsion sampling algorithm to generate a pharmacophore with six built-in descriptors: hydrogen bond donor, acceptor, hydrophobe, negative ionizable, positive ionizable, and aromatic ring (Dixon et al., 2006).

9. GASP program
   GASP is utilized in commercial programs. It involves the superimposition of flexible molecules by minimizing the distance between known pharmacophore points in the two molecules that are being compared. It also encodes information about the intermolecular mapping between the structural features. The collection of data in GASP is called chromosomes, and two algorithms, CROSSOVER and MUTATION, are applied by the program to remove the least fit model and select the active base molecule. To run the program in GASP, the input structure is normally created using the SYBYL BUILD module. Input is decoded on the basis of fitness function using a least square fitting technique (Jones et al., 1995). Mapping is performed by superimposing each molecule on the base molecule obtained from the least square technique. Furthermore, a similarity score is generated to filter overlay molecules, and finally a fitness score is generated to develop an active new molecule (Hou and Xu, 2004).
10. PharmaGist
    PharmaGist is the first academic webserver for pharmacophore development. It handles a set of drug-like molecules to find the highest scoring 3D pattern of molecular features. It uses the pairwise alignment and multiple alignment (pivot iteration) algorithm to generate pharmacophores with the hydrogen bond acceptor, donor, cation, anion, and hydrophobe (Schneidman-Duhovny et al., 2008).
11. ALADDIN
    ALADDIN is an integrated computational tool for the design and recognition of pharmacophores from geometric, steric, and substructure searching. It is mainly used to design analogs to probe a flexible bioactive conformation with more subtle variations in shape of the structure. It utilizes the geometric description language and includes the provision to test the molecule in the actual coordinate system to generate and store 3D structures. Implementation of ALADDIN is based on GENIE. It is a language that incorporates chemical searches and helps in substructure specification, recognition, and enumeration (Van Drie et al., 1989).
12. RAPID
    RAPID is a randomized pharmacophore identification for drug designing academic-based tools. It helps in the generation of pharmacophores by identifying geometric invariants among the collection of small molecule datasets. This tool is based on finding the largest common point sets in the input data, which are system tractable and without noise. The RAPID algorithm generates a large number of conformations at random (Finn et al., 1997). Then, the information is partitioned into sets that reflect geometric similarities, followed by the clustering of possible conformations of the molecule. These clusters are used as inputs to identify the invariants. This tool

uses pairwise matching and multiple alignments to determine the invariants. Finally, the possible invariants in the cluster are overlaid to produce new active conformations (Humblet and Marshall, 1980).

13. CLEW
    CLEW is freely available software that utilizes the combination of two different algorithms, i.e., correlation theorem and ML classification method, to generate a complete set of new conformations. First, it designs pharmacologically important features such as hydrogen bond acceptor, donor, anion, cation, and hydrophobe using a correlation theorem from active analogs. Second, it classifies the designed pharmacophore utilizing the ML algorithm (Dolata et al., 1998).
14. MPHIL
    MPHIL is an academically available tool that identifies the smallest 3D pattern of pharmacophore points (*k*-point) by measuring feature and interfeature distances within the input ligands. It also employs genetic algorithms to build a pharmacophore feature such as hydrogen bond donor, hydrogen bond acceptor, extensions of hydrogen bond, and electrostatic interactions of binding sites (Holliday and Willett, 1997).
15. SCAMPI
    SCAMPI (Statistical Classification of Activities of Molecules for Pharmacophore Identifications) uses integrated recursive partitioning and fast conformational search followed by clustering analysis. It creates a correspondence space in which all possible chemical features and configurations are indicated. It uses an ensemble distance geometry and active analogs approach to present correspondence search features among different compound datasets. Furthermore, the student *t*-test is used for recursive partitioning of datasets to form a binary molecular descriptor matrix. This matrix helps to guide the presence and absence of particular molecular descriptors in the compound (Chen et al., 1999). This tool also performs energy minimization of all input datasets to restrict the conformations within low-energy regions. The random search and minimum accessible distance calculation algorithms are used in the new version of SCAMPI.
16. LigBuilder
    LigBuilder builds ligands by using an organic fragments approach used in structure-based drug design. This tool uses programs for binding pocket analysis, building up methods, scoring methods, and genetic algorithms. It uses growing strategy and linking strategy, GROW and LINK, respectively, for building pharmacophore sites with prominent positions for interactions (Wang et al., 2000).
17. Other tools
    There are other software packages available such as CoLibri, DecoyFinder, MOLA, NNScore, Epik, SwissSimilarity, ZincPharmer, ShaEP, BALLOON, React2D, ChemCom, CATS, Autoclick Chem, GMA, Shape-it, Li-SiCA, DANTE, APOLLO, GAMMA, and Apex 3D tools for pharmacophore development (Agrawal et al., 2018).

6.2.4 Quantitative structure–activity relationship prediction

QSAR is one of the most important models employed in chemoinformatics. QSAR is a mathematical model that has been developed to relate the structural features of a pharmacophore to its biological and physiochemical activity. QSAR modeling has the potential to provide information by reducing time, cost, and animal model testing (Gramatica, 2007). QSAR is available with various variants such as quantitative structural toxicity relationship and quantitative structure–pharmacokinetics relationship to model toxicological and pharmacological activities, respectively (Tropsha et al., 2003; Devillers, 2004).

The principle to search for QSAR works on the assumption that similar structures have similar activities. Therefore these methods are often called predictive methods (Gadaleta et al., 2016). They can predict different activities such as biological activity (IC_{50}), class of compounds to which a compound belongs (inhibitor or activator), developmental toxicity, mutagenicity, and can give rise to the property of interest.

Developing QSAR models for the prediction process involves various modeling methods such as linear regression and logistic regression (Cao et al., 2010). However, knowledge-based ML algorithms are gaining importance and describe the empirical relationship between structure and compound. ML algorithms involves the preparation of a training compound set of represented similar compounds through the scanning of optimal molecular descriptors. Furthermore, training sets are run prior to testing, providing program learning. Finally, the test is run with newly developed pharmacophores and different algorithms are applied to predict the QSAR of new molecules (Martins and Ferreira, 2013) (Fig. 6.4).

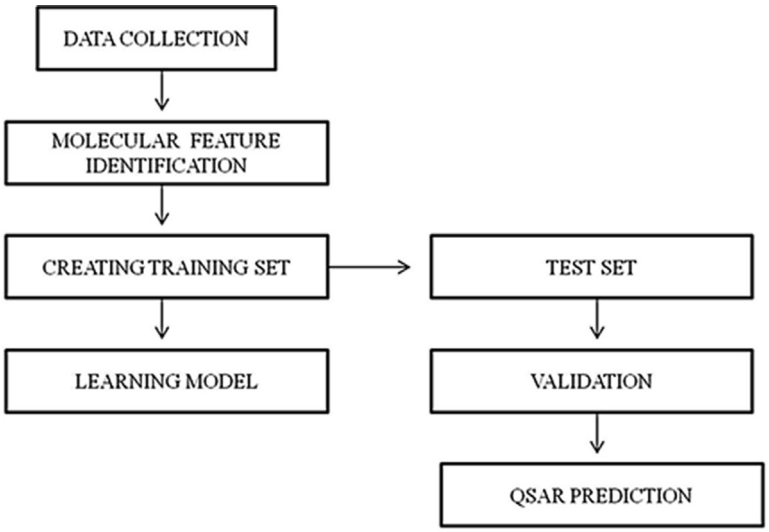


FIGURE 6.4

Schematic representation of quantitative structure–activity relationship (QSAR).

6.2.4.1 Types of QSAR

Various dimension-based QSAR approaches cater for the selection of desirable features and energy calculations of compounds, mainly 1D, 2D, 3D, 4D, 5D, and 6D QSARs. 1D and 2D QSARs are classic forms corresponding to pKa and logP properties for activity predictions. Besides, 3D QSAR allows spatial arrangement of molecules to set the 3D dynamic lattice algorithms for biological activity prediction. Comparative molecular field analysis (COMFA) and comparative molecular similarity indices analysis (CoMSIA) are the most frequent methods employed for 3D QSAR modeling. COMFA and CoMSIA are alignment-dependent and ligand descriptor-based methods for building QSAR. The concept focuses on placing ligands on energy grids and at each lattice point in the grid the energy is calculated. COMFA correlates energy fields in terms of electrostatic (Coulombic) and steric properties (van der Waals), while CoMSIA is capable of providing more stable information correlates with steric, electrostatic, hydrogen bond donor, and hydrogen bond acceptor interactions with the aim of increasing the potential of the compounds (Sharma et al., 2016).

4D QSAR is an advanced version that allows conformational flexibility and freedom of alignment features in 3D QSAR analysis. Moreover, for virtual model building for advanced pharmacokinetics properties predictions, multiple representations of chemicals generally provide new dimensions in 4D QSAR. Representation of a 4D prediction model for multiple induced fit is called a 5D QSAR model. The new dimension "solvation function" addition to 5D QSAR to study the noncovalent interactions in activity prediction is called a 6D QSAR model. The advantage of each approach of the QSAR model is efficiently shown by scoring functions of internal validation such as cross-validation (q^2), least square fit, and external validation (Damale et al., 2014).

Cross-validation and least square fit are internal validation methods to determine how large a model can be used for the dataset. They have promising predictive abilities that represent the relationship between predictors and response or experimental activity. Besides this, external validation of a QSAR model is based on predicted and observed activities of external test sets and search for the correlation coefficient and coefficient of determination. Internal validation is a popular method that defines robustness and assesses the model fit (Veerasamy et al., 2011).

6.2.4.2 QSAR modeling tools

1. SYBYL
   SYBYL is a commercial tool providing a wide range of structure-building, optimization, and basic comparison models to relate to the structure. The tool comes with a selection of broad forcefields that can be used in compound activity prediction. COMFA is generally employed for affinity representation in the SYBYL tool.
2. CODESSA
   CODESSA is a commercially available tool for QSAR prediction. It is able to calculate a range of molecular descriptors based on 3D structure of the

compound on the basis of constitutional, topological, geometrical, electrostatic, charged surface area, quantum chemical, molecular orbital-related, and thermodynamic features. The use of CODESSA can be integrated with AMPAC for compound property prediction of the chemical structure.

3. Auto-QSAR
   Auto-QSAR is an application of the commercially available software Schrodinger. It is automated, high-quality guesswork applied to generate models and predict structural property. It uses the practice included in the OECD-QSAR guidelines. It estimates the property using structural similarity among the training set and returns a yes or no indication for a particular prediction. The results for QSAR prediction can be analyzed via Maestro visualization application (Karnik et al., 2020).
4. OECD-QSAR toolbox
   To facilitate the QSAR approaches in regulatory government-based industries and to improve regulatory acceptance, the OECD developed a QSAR toolbox application to access the hazards of any chemical. The tool consists of several programs for identification of potential mechanisms of action of compounds, identification of other similar compounds, and prediction of activities (Oecd, 2004).
5. Vega QSAR
   Vega QSAR is a freely available tool providing accessibility to various applications for toxicity prediction of a compound. It allows users to develop their own model of prediction to predict any property of the compound scripted in the Java language (Benfenati et al., 2013).
6. TEST
   The toxicity estimation software tool is freely available software developed to estimate toxicity and physical properties of a compound using various QSAR algorithms. It predicts toxicity based on the physical characteristics of the compound. The best part of the software is that it does not require any external program to run it. The input chemical structure can draw on a chemical sketcher window of the software, or a structural text file (SMILES) can be used or directly imported from the list of databases. The software includes models for estimation of IC_{50}, LC_{50}, LD_{50}, developmental toxicity, and mutagenicity (US EPA, 2010).
7. Caesar 2.0
   Caesar version 2.0 software is a freely available tool that has been integrated with the QSAR model to predict developmental toxicity and mutagenicity. It is a Java application and is easy to use for analysis (Cassano et al., 2010).
8. PASS Prediction
   PASS (prediction of activity spectra for substances) is an online application available with Way2Drug predictive services. PASS is employed to predict biological activity of new compounds, including pharmacological activity, mechanism of action, toxicity, and adverse effects (Rudik et al., 2019).

Table 6.4 List of quantitative structure–activity relationship (QSAR) prediction tools.

Tools	Description (web links)
SYBYL	Commercial-based molecular field QSAR analysis-based tool (https://www.tripos.com)
CODESSA	Commercial-based 1D and 2D QSAR modeling tool (https://www.codessa-pro.com/index.htm)
Auto-QSAR	Commercial QSAR tool in Schrodinger (https://schrodinger.com/autoqsar)
Vega-QSAR	Freely available 2D QSAR tool (https://vegahub.eu/download/vega-qsar-download/)
TEST	Free tool for absorption, distribution, metabolism, excretion, and toxicity prediction (http://epa.gov/chemical-research/toxicity-estimation-software-tool-test/)
PASS	Webserver activity prediction spectra tool (http://pharmaexpert.ru/passonline/)
OECD QSAR toolbox	Webserver tools for activity prediction (http://oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm)

9. DEMETRA
   DEMETRA is freely available software that is useful for predicting the toxicity of molecules particularly for pesticide chemicals.
10. Other tools
    E-Dragon, GRIN/GRID, Danish QSAR, TOX match, Toxtree, OCHEM, AMBIT, ChemAxon Marvin, MetaDrug, PreADME, TerraQSAR, Derek, Hazardexpert, TIES, and QSAR Pro are other tools available to predict structure–activity relationships based on different prediction models or algorithms (Yousefinejad and Hemmateenejad, 2015) (Table 6.4).

6.3 Machine learning approaches and tools for chemoinformatics

The area of ML is currently one of the most rapidly evolving approaches in the field of chemoinformatics. ML methods are primarily employed for pattern recognition algorithms to construct a model for structure–activity prediction. It mainly works on the "modes of statistical interference" and "predictive modeling levels" to develop a prediction model (Simeone, 2018). These methods in chemoinformatics provide knowledge-based relationships between the structure and property of interest. The optimal learning parameters are mainly used to perform two important tasks (Shalev-Shwartz and Ben-Daid, 2013):

1. Retrieving chemical information (feature description) to extract domain knowledge with desired behavior, called "ENCODING."
2. Learning by building a hypothesis class for chemoinformatics models called "MAPPING" to illustrate the structure–activity relationship.

This section highlights the introduction of newer ML methods and advanced QSAR tools that are commonly used to build and validate prediction models in chemoinformatics.

6.3.1 Techniques of ML

ML algorithms are broadly classified into two main techniques: supervised learning and unsupervised learning.

1. Unsupervised learning
   Unsupervised learning is defined as "learning of the pattern from the unlabeled data, i.e. training sets consist of input without the labeled desired output for activity prediction." There is no indication of desirable output; however, it could predict the properties of the mechanism generating the data. Data clustering is the main learning task of unsupervised learning (Odziomek et al., 2017).
2. Supervised learning
   Supervised learning is defined as "external reinforcement of information to produce a learning hypothesis, i.e. training set consists of labeled input and output for activity prediction." Generation of pattern or hypothesis to predict chemical activity via supervised learning involves dataset features collections, construction of new features, and selection of learning algorithms to test compound activity (Lavecchia, 2015). Different platforms utilizing ML techniques are listed in Table 6.5.

Table 6.5 Platforms for machine learning (ML)-based quantitative structure–activity relationship (QSAR) modeling.

Tool	Description (web link)
LibSVM	Commercial program based on ML support vector machine (SVM) and super vector regression (SVR) algorithms. This tool performs priority-wise extraction of influential datasets and cross-validates to search for the best molecular feature. This technique in SVM is called Information Gain or InfoGain. The suite contains SVM-scale tools that comprise certain parameters on which the data are classified such as S SVM_type, t kernel_type, d degree, g gamma, wi weight, v-n validation mode, and q quiet mode (https://www.csie.ntu.edu.tw/libsvm/)
DPubChem	This tool is a freely available MATLAB-based program that performs different classification and regression data visualizations, simulations, image processing, and computational modeling (Soufan et al., 2018). (https://www.cbrc.kaust.edu.sa/dpubchem)
MOE	MOE is commercial drug discovery software with ML algorithms to perform QSAR, pharmacophore discovery, protein modeling, molecular modeling, and simulation medicinal chemistry application and method development. Given a set of known training sets, an MOE-designed QSAR model can also help to correlate the activities (Vilar et al., 2008) (https://www.chemcomp.com/software.htm/)

Continued

Table 6.5 Platforms for machine learning (ML)-based quantitative structure–activity relationship (QSAR) modeling.—*cont'd*

Tool	Description (web link)
Ezqsar	A free R-program-based QSAR tool. This is open-source tool performs a variety of linear or nonlinear statistical modeling, time series analysis, classification, regression, and clustering tools (Tsiliki, 2015) (https://www.omicstools.com)
CORAL	A free R-program-based QSAR and nano-QSAR tool. (https://www.insilico.eu/coral/)
QSARINS	A free R-program-based QSAR tool. It performs a variety of linear or nonlinear statistical modeling and time series regression analyses (Tsiliki, 2015) (https://www.vegatools.com)
RRegrs	A free R-program-based QSAR tool (https://www.r-project.org/)
WeKa	Open-source program for ML-based QSAR modeling. This suite of ML is a popular open source for performing feature selection, clustering, classification, association rule mining, and regression (Pyka et al., 2012) (https://www.cs.waikato.ac.nz/ml/weka/)
KNIME	Konstanz Information Miner is a free standalone R-program-based tool for QSAR. It is an academic platform for data integration, processing, analysis, and modeling. The given module of a dataset is organized in the form of nodes that provide data modeling, visualization, and data flow (Berthold et al., 2006) (https://www.knime.org)
Rapid Miner	Open-source Weka-based tool for QSAR. The system operates with the Java project for implementing ML and data mining algorithms. It also integrates the Weka attributes for library modeling. It has an "optimize parameter" operator to allow the semiautomation of different tools (Choudhary et al., 2018) (https://rapid-i.com)
Tanagra	Free tools that allow support data visualization, one-way ANOVA, Welch ANOVA, paired *t*-test, normality test, feature selection (remove constant, define feature status), regression (regression tree, SVR, multiple linear regression algorithm), factorial analysis (principal component analysis, principal factor analysis), clustering (*k*-nearest neighbor), and classification (SVM, random forest, decision tree, and naïve Bayes classifiers) (http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html)
Keel	Open-source tool for finding evolutionary relationships (https://sci2s.ugr.es/keel/)
AZOrange	AZOrange is an open-source tool for ML, developed for absorption, distribution, metabolism, excretion, and toxicity, in particular for QSAR models in drug discovery. It provides fundamental scientific principles of reproducibility, which are guided by OECD for QSAR modeling. It also consists of some open-source codes such as OpenCV package, PLearn interface, and APPSPACK especially for QSAR modeling (Stålring et al., 2011) (https://www.orange.biolab.si/)

6.3.2 Types of supervised learning

1. Classification
   Classification methods classify data into specific classes based on the finite and discrete similarity output.
2. Regression
   Regression methods refer to the regression algorithm to improve the accuracy of predicted classification models by providing quantitative true values. The objective of regression methods is to provide the mathematical equation to predict the outcome.

6.3.3 Algorithms for classification and regression problems in drug designing

1. Support vector machine
   The SVM learning algorithm maps the data within the high-dimensional space along with demarcation of the separating hyperplane. A hyperplane is defined by a linear discriminant function comprising a linear combination of molecular descriptors. The SVM approach is based on the linear kernel function; it means that when the 2D data are not separated using a straight line, they are projected in the form of an SVM hyperplane that allows data to be linearly separated (Geppert et al., 2008). Each test instance is classified depending on which side of the hyperplane boundary they lie. Ranking could be achieved by measuring the distance between a hyperplane and the instance. SVM is one of the most common classifier approaches used in chemoinformatics adopted for binary or multiple classification (Cortes and Vapnik, 1995). SVM learning algorithms used in bioactivity prediction for repurposing drugs, inhibitors, and receptor compounds are based on the structural risk minimization principle (Zhao et al., 2006). Furthermore, it is also used to predict toxicity-related properties and physicochemical property prediction that include solubility, pKa, logP, and melting point. LibSVM, Weka, and MOE are efficient platforms utilizing the SVM algorithm.
2. Linear discriminant analysis
   An LDA classifier works on data that has categorical target properties and molecular descriptors in continuous variables. It also finds the separating hyperplane that can separate different classes. The hyperplane is generated using the LDA of molecular features. The LDA approach classifies the unknown compound on the basis of L-discriminant scores either below or above the hyperplane margin score. Applications of LDA in modeling are used to predict mutagenicity, antiparasitic drugs, toxicity of pesticides, and antitrypanosomal and trichomonacidal symptoms (Vert and Jacob, 2008). SPSS, SAS, R-program, and Tanagra are commonly developed tools utilizing LDA for prediction.

3. Naïve Bayesian algorithm
   Naïve Bayes classification algorithms are probabilistic approaches for estimating the probabilities of class membership. They are based on Bayes theorem of conditional probability in which the test instance is correctly assigned to the highest estimated probability class, and has the benefit of conceptual simplicity (Klon, 2009). The use of naïve Bayesian analysis has been investigated extensively to predict biological activities of multiple drugs with newly proposed features; a well-known example of this approach was recently reported by Bai et al. (2018). This study performed the systemic analysis of large random drug pairs using the naïve Bayesian algorithm to predict effective drugs combinations for cancer and metabolic disorder based on two new features: metabolic enzymes of drugs and transporters of drugs. The method demonstrated better performance and constructed a more stable and accurate predictive model compared to other ML classifiers, indicating the role of the naïve Bayes approach in predicting drug combinations. In addition, the naïve Bayes approach is often used for performance enhancement, protein target prediction, bioassay classification for drug-like molecules, and toxicity.
4. Random forest algorithm
   The decision tree is a hierarchical arrangement of nodes and branches. Decision tree structure mainly consists of three main nodes: root nodes, middle nodes, and terminal or leaf nodes, respectively. Two nodes, i.e., root and middle node, form the test condition, which is assigned with molecular descriptors; the terminal nodes are assigned with target properties to classify unknowns. The classification of unknown molecules is based on the terminal or leaf node. The information undergoes a series of inquiries through root and internal middle nodes, with rules and regulations. A compound will be classified on the basis of matching of properties from the given set of descriptors. Most commonly, Hunt's algorithm runs the deciding tree. It specifies the threshold of molecular descriptors that specify the best splitting of the unknowns. Applications of the decision tree have been applied in QSAR to cytochrome P450, catalyst prediction, peptide−protein binding affinity, inhibitors, and substrate predictions (Sela and Simonoff, 2012).
   In contrast, random forest classifiers work on the development of the consensus of large numbers of decision tress, thus forming a forest. The majority predicted score from each tree in the forest forms the final prediction. Random forest is easy to use if provided with the number of trees in the forest and a number of molecular descriptors (Segal, 2004). Therefore a large number of trees or relationships can be utilized to classify the unknown molecule in random forest.
5. *k*-nearest neighbor algorithm
   *k*NN is a nonparametric and lazy ML algorithm for classifying test sets. It allows simple implementation to classify the instances and does not need any training dataset for model development. Each test feature/instance is classified based on a class common to its closest neighbor "*k*" present in a high-dimensional feature space. The algorithm finds the distance between the two instances using Euclidean distance, Hamming distance, Manhattan distance,

and Minkowski distance, followed by finding and voting to search for the closest neighbor (Kauffman and Jurs, 2001). Each instance is defined as a position vector in the feature space that is often chosen to be a small integer, and the neighbor is chosen to be closest to the point for which the instances need to be predicted. Let the classes of a feature be 2, then $k = 1$ is selected as the nearest neighbor in the feature space to vote. The number of neighbors in *k*NN has to be decided by the requirement of the dataset. The small number of neighbor fits flexible with high variance and low biasness into the network. *k*NN can easily identify the class of the dataset; however, regression can also be performed by assessing the attribute of each test instance. Furthermore, it is also helpful to assess the contribution of the properties of the neighbors (Mitchell, 2014). *k*NN is commonly used in ligand-based drug designing, which classifies based on the assumption of compound similarity with nearest neighbor, predicting binding affinity of receptor ligand complexes such as activity of anti-HIV isatin analogs, T-helper cell antagonists, and others.

6. Least square algorithms

The linear regression model in ML fits the classified data according to the least square method to acquire the accurate sampling data and reduce the square of the error. This method predicts the linear function scores of one variable from the given set of training data points. The predicting group is called the criterion variable and another known group is called the predictor variable. The values of criterion and predictor variables are the model parameters. The regression plot utilizes the values of both of criterion and predictor variables to construct the best fitted straight "regression line" to the data points (Marill, 2004). The fitted line minimizes the distance between the data nodes along the dimensions to give outcome variables. Hansch and Free-Wilson analysis of QSAR models makes extensive use of linear regression algorithms. It is also applied to the prediction of luteinizing hormone-releasing factor and interleukin-1 antagonists (Frank and Friedman, 1993).

Several techniques of the least square method such as principal component regression and super vector regression are available to combat model complexity. L2 regularization methods like ridge regression and Gaussian process decrease the number of predictor variables and select the small subsets that are being predicted by any chemoinformatics model such as QSAR or QSPR (Seeger, 2004). Principal component regression (PCR) is a type of multiple regression method under unsupervised learning that converts the large variable dataset into the smaller set of variables with unrelated features. This type of regression method is commonly used to identify mechanistic features among the variables. Another popular method of regression is partial least squares, which mainly couples the PCR algorithm with multivariate regression. This method transforms the given set of predictor variables into noncorrelated variables and is the first choice for QSAR and 3D QSAR modeling, predicting compound inhibitors for *Chlorella vulgaris* and human immunodeficiency virus (HIV) (Lo et al., 2018).

7. Artificial neural networks algorithms
   ANNs are a family of ML neural network algorithms. They are inspired by the operations of brain networks; unlikely a mathematical model used for pattern recognition. The network is based on a consecutive connected layered architecture, an input layer for recognition (analogous to dendrites) input signal, some intermediate or hidden layers that generate an activation response (analogous to cell body), and an output layer that passes the output signals to subsequent connected nodes (analogous to axons) (Niculescu, 2003). Each connection between the neurons carries the signal in the form of desired patterns or scores. The network learns how to connect the input and output data during the training phase; this approach is called pretraining. The ANN has adjustable internal parameters called weights or knobs to fine tune the function. It is utilized for speed recognition and prevent overfitting for better generalization of the functions. ANN algorithms have been widely applied to all branches of chemoinformatics, including modeling QSAR/QSPR properties of drug-like molecules, pharmacokinetic and pharmacodynamic analysis, and toxicological and physicochemical property analysis (Patel and Goyal, 2008). R-program, MATLAB, and Neuralware are some tools used for ANN development.
8. Deep learning neural network algorithm
   Deep learning network (DLN) is closely associated with ANN-type architecture, with multiple (hundreds to millions) hidden layers. Each hidden layer has its own weights, activation functions, and biases that help to modify the internal function for better output. The recent success of DL provides an opportunity to develop tools for molecular graph generation to set descriptors of chemical structures (Ekins, 2016). DL was first introduced in QSAR evaluation to predict complex statistical patterns among thousands of molecular descriptors. It helps in numerous chemoinformatics applications of drug−target identification, de novo molecular design, small molecule optimization, and QSAR prediction. One advantage of DLN algorithms is that they have different flexible algorithms to fine tune the method.
   A convolutional neural network (CNN) is a type of DL algorithm that is designed to predict performance using local filter scans and different hidden layers. Each hidden layer in ConvNets acts with independent function. It takes advantages of feature extractions by pooling data, sharing weights, and using different hidden layers in the network. It mainly helps in image recognition by employing simple local features to complex models. CNN does not have any sort of dependency in the sequential input data. The output from the CNN is self-dependent based on the training model (Duvenaud et al., 2015). For example, if 200 different inputs are run, the results do n'ot have any correlation between previous inputs and the next input (nonbiased data classification).
   In contrast, when the identification of a new model is dependent on previous data generated, then results biasness based on previous output is required. In this scenario, the second type of DL, i.e., recurrent neural network (RNN), helps to channel information through a series of operations with some sense of memory

of previous input or output in the sequence of data. In the RNN, all hidden layers are merged into a single recurrent layer that helps to store all the previous functional input, and merges that information with the current input (Olivecrona et al., 2017). This kind of output information indicates the correlation between the current data step and previous input. This complex algorithm is useful in building models for QSAR, de novo synthesis, QSPR activities, and long arrays of data analysis, such as gene expression data.

6.4 Conclusion

The effectiveness of the chemoinformatics approach to drug discovery is associated with various tools used in conjugation. In spite of advances, a number of drug candidates fail to reach the clinical phase, so there is a need to adopt techniques that will be easy to use and show minimal loss in the designing process. Chemoinformatics applications enclosed within software-based platforms have enabled researchers from across the world to collaborate with their respective areas of drug discovery and vaccine development. The in silico chemoinformatics tools discussed in this chapter provide a different insight into pharmaceutical studies so that various facets of modern biomedicine may be understood to expand their horizons without the use of textbooks . The previously existing attitudes and conceptions of drug designing mean that software-based computational approaches are now redundant. These disciplines are no longer separate and the advancements in each of these fields inevitably impact each other. It is therefore imperative for students, scientists, learners, and researchers to understand this nexus and to make concerted efforts to develop new strategies for accelerating the drug-discovery process for countless diseases that impact humanity.

References

Abagyan, R., Totrov, M., Kuznetsov, D., 1994. ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J. Comput. Chem. 15, 488−506.

Agrawal, P., Raghav, P.K., Bhalla, S., Sharma, N., Raghava, G.P.S., 2018. Overview of free software developed for designing drugs based on protein-small molecules interaction. Curr. Top. Med. Chem. 18, 1146−1167.

Bai, L.Y., Dai, H., Xu, Q., Junaid, M., Peng, S.L., Zhu, X., Xiong, Y., Wei, D.Q., 2018. Prediction of effective drug combinations by an improved naïve bayesian algorithm. Int. J. Mol. Sci. 19.

Bajusz, D., Ferenczy, G.G., Keseru, G.M., 2017. Structure-based virtual screening approaches in kinase-directed drug discovery. Curr. Top. Med. Chem. 17 (20), 2235−2259.

Benesch, M., Weber-Mzell, D., Gerber, N.U., Von Hoff, K., Deinlein, F., Krauss, J., et al., 2010. Ependymoma of the spinal cord in children and adolescents: a retrospective series from the HIT database: clinical article. J. Neurosurg. Pediatr. 6, 137−144.

Benfenati, E., Manganaro, A., Gini, G., 2013. VEGA-QSAR: AI inside a platform for predictive toxicology. In: CEUR Workshop Proc. 1107, 21−28.

Berthold, M.R., Cebron, N., Dill, F., Di Fatta, G., Gabriel, T.R., Georg, F., Meinl, T., Ohl, P., Sieb, C., Wiswedel, B., 2006. 2006. KNIME: the konstanz information miner. In: 4th International Industrial Simulation Conference 2006. ISC, pp. 58−61.

Bohacek, R.S., McMartin, C., Guida, W.C., 1996. The art and practice of structure-based drug design: a molecular modeling perspective. Med. Res. Rev. 16, 3−50.

Brooijmans, N., Kuntz, I.D., 2003. Molecular recognition and docking algorithms. Annu. Rev. Biophys. Biomol. Struct. 32, 335−373.

Bykov, V.J.N., Issaeva, N., Selivanova, G., Wiman, K.G., 2002. Mutant p53-dependent growth suppression distinguishes PRIMA-1 from known anticancer drugs: a statistical analysis of information in the National Cancer Institute database. Carcinogenesis 23 (12), 2011−2018.

Cao, D.S., Liang, Y.Z., Xu, Q.S., Li, H.D., Chen, X., 2010. A new strategy of outlier detection for QSAR/QSPR. J. Comput. Chem. 31, 592−602.

Cassano, A., Manganaro, A., Martin, T., Young, D., Piclin, N., Pintore, M., Bigoni, D., Benfenati, E., 2010. CAESAR models for developmental toxicity. Chem. Cent. J. 4, S1−S4.

Chen, X., Rusinko, A., Tropsha, A., Young, S.S., 1999. Automated pharmacophore identification for large chemical data sets. J. Chem. Inf. Comput. Sci. 39, 887−896.

Chen, X., Ji, Z.L., Chen, Y.Z., 2002. TTD: therapeutic target database. Nucleic Acids Res. 30, 412−415.

Chen, J., Swamidass, S.J., Dou, Y., Bruand, J., Baldi, P., 2005. ChemDB: a public database of small molecules and related chemoinformatics resources. Bioinformatics 21, 4133−4139.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., Blaschke, T., 2018. The rise of deep learning in drug discovery. Drug Discov. Today 23 (6), 1241−1250.

Choudhary, S., Namdeo, V., Dwivedi, A., 2018. Performance comparison of machine learning techniques in intrusion detection using rapid. Miner. Int. J. Comput. Sci. Eng. 6, 1001−1005.

Cooper, M.E., 2004. Chemoinformatics: concepts, methods and tools for drug discovery. Drug Discov. Today 9, 957−959.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273−297.

Damale, M., Harke, S., Kalam Khan, F., Shinde, D., Sangshetti, J., 2014. Recent advances in multidimensional QSAR (4D-6D): a critical review. Mini. Rev. Med. Chem. 14, 35−55.

Devillers, J., 2004. Prediction of mammalian toxicity of organophosphorus pesticides from QSTR modeling. SAR QSAR Environ. Res. 15, 501−510.

Dixon, S.L., Smondyrev, A.M., Knoll, E.H., Rao, S.N., Shaw, D.E., Friesner, R.A., 2006. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J. Comput. Aid. Mol. Des. 20, 647−671.

Docking, A.J.A., Autodock, A.J., Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J., 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comput. Chem. 19, 1639−1662.

Dolata, D.P., Parrill, A.L., Walters, W.P., 1998. CLEW: the generation of pharmacophore hypotheses through machine learning. SAR QSAR Environ. Res. 9, 53−81.

Dunkel, M., 2006. SuperNatural: a searchable database of available natural compounds. Nucleic Acids Res. 34, D678−D683.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., Adams, R.P., 2015. Convolutional networks on graphs for learning molecular fingerprints. In: Advances in Neural Information Processing Systems, pp. 2224−2232.

Ekins, S., 2016. The next era: deep learning in pharmaceutical research. Pharm. Res. 33, 2594−2603.

Ewing, T.J.A., Makino, S., Skillman, A.G., Kuntz, I.D., 2001. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput. Aided Mol. Des. 15, 411−428.

Finn, P.W., Kavraki, L.E., Latombe, J.C., Motwani, R., Shelton, C., Venkatasubramanian, S., Yao, A., 1997. RAPID: randomized pharmacophore identification for drug design. In: Proceedings of the Annual Symposium on Computational Geometry, vol. 10, pp. 324−333.

Frank, L.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35, 109−135.

Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., et al., 2004. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 47, 1739−1749.

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D.D., Liu, P., Gautam, B., Ly, S., Guo, A.C., Xia, J., Liang, Y., Shrivastava, S., Wishart, D.S., 2009. SMPDB: the small molecule pathway database. Nucleic Acids Res. 38, 480−487.

Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A., Nicolotti, O., 2016. Applicability domain for QSAR models. Int. J. Quant. Struct. Relationship. 1, 45−63.

Gagnon, J.K., Law, S.M., Brooks, C.L., 2016. Flexible CDOCKER: development and application of a pseudo-explicit structure-based docking method within CHARMM. J. Comput. Chem. 37, 753−762.

Gaulton, A., Hersey, A., Nowotka, M.L., Patricia Bento, A., Chambers, J., Mendez, D., et al., 2017. The ChEMBL database in 2017. Nucleic Acids Res. 45, D945−D954.

Geppert, H., Horváth, T., Gärtner, T., Wrobel, S., Bajorath, J., 2008. Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. J. Chem. Inf. Model. 48, 742−746.

Geppert, H., Vogt, M., Bajorath, J., 2010. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J. Chem. Inf. Model. 50, 205−216.

Ghose, A.K., Viswanadhan, V.N., Wendoloski, J.J., 2001. The fundamentals of pharmacophore modeling in combinatorial chemistry. J. Recept. Signal Transduct. 21, 357−375.

Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J., 2016. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045−D1053.

Gonczarek, A., Tomczak, J.M., Zareba, S., Kaczmar, J., Dąbrowski, P., Walczak, M.J., 2018. Interaction prediction in structure-based virtual screening using deep learning. Comput. Biol. Med. 100, 253−258.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694−701.

Gupta, P.P., Bastikar, V.A., Chhajed, S.S., 2018. Chemical Structure Databases in Drug Discovery, vol. 3, pp. 47–61.

Holliday, J.D., Willett, P., 1997. Using a genetic algorithm to identify common structural features in sets of ligands. J. Mol. Graph. Model. 15, 221–232.

Hou, T., Xu, X., 2004. Applications of genetic algorithms to computer-aided drug design. Prog. Chem. 16, 35–38.

Huanga, D., Caflischa, A., 2010. Library screening by fragment-based docking. J. Mol. Recogn. 23, 183–193.

Humblet, C., Marshall, G.R., 1980. Pharmacophore identification and receptor mapping. Annu. Rep. Med. Chem. 15, 267–276.

Irwin, J.J., Shoichet, B.K., 2005. Zinc - a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. 45, 177–182.

Jahn, A., Hinselmann, G., Fechner, N., Zell, A., 2009. Optimal assignment methods for ligand-based virtual screening. J. Cheminf. 1, 14.

Jain, A.N., 2003. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J. Med. Chem. 46, 499–511.

Jarrahpour, A., Fathi, J., Mimouni, M., Hadda, T.B., Sheikh, J., Chohan, Z., Parvez, A., 2012. Petra, Osiris and Molinspiration (POM) together as a successful support in drug design: antibacterial activity and biopharmaceutical characterization of some azo Schiff bases. Med. Chem. Res. 21, 1984–1990.

Jones, G., Willett, P., Glen, R.C., 1995. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. J. Comput. Aided Mol. Des. 9, 532–549.

Jones, G., Willett, P., Glen, R.C., Leach, A.R., Taylor, R., 1997. Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. 267, 727–748.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K., 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353–D361.

Karnik, K.S., Narula, I.S., Sarkate, A.P., Wakte, P.S., 2020. Auto QSAR- A fast approach for creation and application of QSAR models through automation. Chem. Select 5, 5756–5762.

Kauffman, G.W., Jurs, P.C., 2001. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. J. Chem. Inf. Comput. Sci. 41, 1553–1560.

Keiser, M.J., Keiser, M.J., Setola, V., Setola, V., Irwin, J.J., Irwin, J.J., et al., 2009. Predicting new molecular targets for known drugs. Nature 462, 175–181.

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., et al., 2016. PubChem substance and compound databases. Nucleic Acids Res. 44, D1202–D1213.

Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J., 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discov. 3, 935–949.

Klon, A., 2009. Bayesian modeling in virtual high throughput screening. Comb. Chem. High Throughput Screen. 12, 469–483.

Lagarde, N., Goldwaser, E., Pencheva, T., Jereva, D., Pajeva, I., Rey, J., Tuffery, P., Villoutreix, B.O., Miteva, M.A., 2019. A free web-based protocol to assist structure-based virtual screening experiments. Int. J. Mol. Sci. 20 (18), 4648.

Lavecchia, A., Giovanni, C., 2013. Virtual screening strategies in drug discovery: a critical review. Curr. Med. Chem. 20, 2839–2860.

Lavecchia, A., 2015. Machine-learning approaches in drug discovery: methods and applications. Drug Discov. Today 20, 318–331.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., et al., 2014. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 42, D1091−D1097.

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., Wang, R., 2015. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics 31, 405−412.

Lo, Y.C., Rensi, S.E., Torng, W., Altman, R.B., 2018. Machine learning in chemoinformatics and drug discovery. Drug Discov. Today 23, 1538−1546.

Maia, E.H.B., Assis, L.C., de Oliveira, T.A., da Silva, A.M., Taranto, A.G., 2020. Structure-based virtual screening: from classical to artificial intelligence. Front. Chem. 8, 343.

Marill, K.A., 2004. Advanced statistics: linear regression, Part II: multiple linear regression. Acad. Emerg. Med. 11, 94−102.

Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I., Pavlik, P.A., 1993. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. J. Comput. Aided Mol. Des. 7, 83−102.

Martins, J.P., Ferreira, M.M.C., 2013. Qsar modeling: a new open source computational package to generate and validate qsar models. Quim. Nova 36. 554-U250.

Masoudi-Sobhanzadeh, Y., Omidi, Y., Amanlou, M., Masoudi-Nejad, A., 2020. Drug databases and their contributions to drug repurposing. Genomics 112 (2), 1087−1095.

McGann, M., 2012. FRED and HYBRID docking performance on standardized datasets. J. Comput. Aid. Mol. Des. 26, 897−906.

Milletti, F., Storchi, L., Sforna, G., Cruciani, G., 2007. New and original pKa prediction method using grid molecular interaction fields. J. Chem. Inf. Model. 47, 2172−2181.

Mitchell, B.O., 2014. Machine learning methods in chemoinformatics. Wiley Interdiscip. Rev. Comput. 4, 468−481. J.B.O.

Morris, G.M., Goodsell, D.S., Huey, R., Olson, A.J., 1996. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J. Comput. Aided Mol. Des. 10, 293−304.

Neves, B.J., Braga, R.C., Melo-Filho, C.C., Moreira-Filho, J.T., Muratov, E.N., Andrade, C.H., 2018. QSAR-based virtual screening: advances and applications in drug discovery. Front. Pharmacol. 9, 1275.

Niculescu, S.P., 2003. Artificial neural networks and genetic algorithms in QSAR. J. Mol. Struct. 622, 71−83.

Odziomek, K., Rybinska, A., Puzyn, T., 2017. Unsupervised Learning Methods and Similarity Analysis in Chemoinformatics: Handbook of Computational Chemistry, pp. s2095−2132.

Oecd, 2004. OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationships models. Biotechnology 1−2.

Olivecrona, M., Blaschke, T., Engkvist, O., Chen, H., 2017. Molecular de-novo design through deep reinforcement learning. J. Cheminf. 9.

Patel, J., Goyal, R., 2008. Applications of artificial neural networks in medical science. Curr. Clin. Pharmacol. 2, 217−226.

Patrick Walters, W., Stahl, M.T., Murcko, M.A., 1998. Virtual screening - an overview. Drug Discov. Today 3, 160−178.

Pence, H.E., Williams, A., 2010. Chemspider: an online chemical information resource. J. Chem. Educ. 87 (11).

Pyka, M., Balz, A., Jansen, A., Krug, A., Hüllermeier, E., 2012. A WEKA interface for fMRI data. Neuroinformatics 10, 409−413.

Rarey, M., Kramer, B., Lengauer, T., Klebe, G., 1996. A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. 261, 470−489.

Richard, A.M., Gold, L.S., Nicklaus, M.C., 2006. Chemical structure indexing of toxicity data on the Internet: moving toward a flat world. Curr. Opin. Drug Discov. Dev 9 (3), 315−325.

Richmond, N.J., Abrams, C.A., Wolohan, P.R.N., Abrahamian, E., Willett, P., Clark, R.D., 2006. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. J. Comput. Aided Mol. Des. 20, 567−587.

Rosenfeld, R., Vajda, S., DeLisi, C., 1995. Flexible docking and design. Annu. Rev. Biophys. Biomol. Struct. 24, 677−700.

Rudik, A.V., Dmitriev, A.V., Lagunin, A.A., Filimonov, D.A., Poroikov, V.V., 2019. PASS-based prediction of metabolites detection in biological systems. SAR QSAR Environ. Res. 30, 751−758.

Schneidman-Duhovny, D., Dror, O., Inbar, Y., Nussinov, R., Wolfson, H.J., 2008. PharmaGist: a webserver for ligand-based pharmacophore detection. Nucleic Acids Res. 36, 223−228.

Schulz-Gasch, T., Stahl, M., 2004. Scoring functions for protein-ligand interactions: a critical perspective. Drug Discov. Today Technol. 1 (3), 231−239.

Seeger, M., 2004. Gaussian processes for machine learning. Int. J. Neural Syst. 14, 69−106.

Segal, M.R., 2004. Machine learning benchmarks and random forest regression. Biostatistics 18, 1−14.

Sela, R.J., Simonoff, J.S., 2012. RE-EM trees: a data mining approach for longitudinal and clustered data. Mach. Learn. 86, 169−207.

Shalev-Shwartz, S., Ben-David, S., 2013. Understanding Machine Learning: From Theory to Algorithms, Understanding Machine Learning: From Theory to Algorithms.

Sharma, R., Dhingra, N., Patil, S., 2016. CoMFA, CoMSIA, HQSAR and molecular docking analysis of ionone-based chalcone derivatives as antiprostate cancer activity. Indian J. Pharmaceut. Sci. 78, 54−64.

Simeone, O., 2018. A very brief introduction to machine learning with applications to communication systems. IEEE Trans. Cogn. Commun. Netw. 4, 648−664.

Sinha, S., Vohora, D., 2017. Drug Discovery and Development: An Overview in Pharmaceutical Medicine and Translational Clinical Research, pp. 19−32.

Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E.W., 2014. Computational methods in drug discovery. Pharmacol. Rev. 66 (1), 334−395.

Soufan, O., Ba-Alawi, W., Magana-Mora, A., Essack, M., Bajic, V.B., 2018. DPubChem: a web tool for QSAR modeling and high-throughput virtual screening. Sci. Rep. 8, 9110.

Spitzer, G.M., Heiss, M., Mangold, M., Markt, P., Kirchmair, J., Wolber, G., Liedl, K.R., 2010. One concept, three implementations of 3D pharmacophore-based virtual screening: distinct coverage of chemical search space. J. Chem. Inf. Model. 50, 1241−1247.

Stålring, J.C., Carlsson, L.A., Almeida, P., Boyer, S., 2011. AZOrange - high performance Open Source machine learning for QSAR modeling in a graphical programming environment. J. Cheminf. 3, 28.

Thomsen, R., Christensen, M.H., 2006. MolDock: a new technique for high-accuracy molecular docking. J. Med. Chem. 49, 3315−3321.

Thorn, C.F., Klein, T.E., Altman, R.B., 2013. PharmGKB: the pharmacogenomics knowledge base. Methods Mol. Biol. 1015, 311−320.

Tovchigrechko, A., Vakser, I.A., 2005. Development and testing of an automated approach to protein docking. Protein Struct. Funct. Genet. 60, 296−301.

Tropsha, A., Gramatica, P., Gombar, V.K., 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. In: QSAR Comb. Sci. 22, 69–77.

Tsiliki, G., Munteanu, C.R., Seoane, J.A., Fernandez-Lozano, C., Sarimveis, H., Willighagen, E.L., 2015. RRegrs: an R package for computer-aided model selection with multiple regression models. J. Cheminf. 7, 46.

US EPA, 2010. User's Guide for T.E.S.T. (Version 4.1) (Toxicity Estimation Software Tool) [WWW Document]. http://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test.

Van Drie, J.H., Weininger, D., Martin, Y.C., 1989. ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. J. Comput. Aided Mol. Des. 3, 225–251.

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C.P., Agrawal, R.K., 2011. Validation of QSAR models - strategies and importance. Int. J. Drug Des. Disocovery 2, 511–519.

Velankar, S., Van Ginkel, G., Alhroub, Y., Battle, G.M., Berrisford, J.M., G.J., et al., 2016. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic Acids Res. 44, D385–D395.

Venkatachalam, C.M., Jiang, X., Oldfield, T., Waldman, M., 2003. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J. Mol. Graph. Model. 21, 289–307.

Vert, J.-P., Jacob, L., 2008. Machine learning for in silico virtual screening and chemical genomics: new strategies. Comb. Chem. High Throughput Screen. 11, 677–685.

Vilar, S., Cozza, G., Moro, S., 2008. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr. Top. Med. Chem. 8, 1555–1572.

Wang, R., Gao, Y., Lai, L., 2000. LigBuilder: a multi-purpose program for structure-based drug design. J. Mol. Model. 6, 498–516.

Weng, Z., Vajda, S., Delisi, C., 1996. Prediction of protein complexes using empirical free energy functions. Protein Sci. 5, 614–626.

Williams, A., Tkachenko, V., 2014. The Royal Society of Chemistry and the delivery of chemistry data repositories for the community. J. Comput. Aid. Mol. Des. 28, 1023–1030.

Wolber, G., Langer, T., 2005. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J. Chem. Inf. Model. 45, 160–169.

Wood, D.J., Vlieg, J., De, Wagener, M., Ritschel, T., 2012. Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. J. Chem. Inf. Model. 52, 2031–2043.

Yang, S.Y., 2010. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discov. Today 15, 444–450.

Yousefinejad, S., Hemmateenejad, B., 2015. Chemometrics tools in QSAR/QSPR studies: a historical perspective. Chemometr. Intell. Lab. Syst. 149, 177–204.

Zhao, C.Y., Zhang, H.X., Zhang, X.Y., Liu, M.C., Hu, Z.D., Fan, B.T., 2006. Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology 217, 105–119.

Zhao, C., Xia, C.G., Yu, M., Pan, Y., Wang, L., 2015. Application of molecule docking software in drug design. Chin. J. Antibiot. 40, 234−240.
Zhou, B., Wang, J., Ressom, H.W., 2012. Metabosearch: tool for mass-based metabolite identification using multiple databases. PloS One 7.

CHAPTER

7 Chem-bioinformatic approach for drug discovery: in silico screening of potential antimalarial compounds

Himanshu Ojha[1], Mamta Sethi[2], Rita Kakkar[3], Malti Sharma[2], Manisha Saini[1], Mallika Pathak[2]

[1]*CBRN Protection and Decontamination Research Group, Division of CBRN Defence, Institute of Nuclear Medicine and Allied Sciences, New Delhi, Delhi, India;* [2]*Department of Chemistry, Miranda House, University of Delhi, New Delhi, Delhi, India;* [3]*Computational Chemistry Laboratory, Department of Chemistry, University of Delhi, New Delhi, Delhi, India*

7.1 Importance of technology in medical science

Computational approaches involving both cheminformatics and bioinformatics have made remarkable progress in drug discovery. Since the beginning of the 20th century, theoretical chemists have developed various chem-bioinformatic approaches to identify and validate targets, design new drugs, perform bimolecular interaction of new potential targets with chosen targets, and realize virtual screening of potential drugs for a disease (Augen, 2002). Conventional drug discovery involves huge investment in terms of money, time, and labor. Therefore drug discovery is a challenging task for the entire scientific fraternity. With the advent of high-speed computers and user-friendly computational software, pharma companies are employing these tools on a regular basis to save precious resources and shorten the drug discovery cycle.

7.2 Origin of cheminformatics

With the intent of reducing the cost and time of discovering a drug, pharmaceutical companies joined hands with information technologists to expedite the process. As a result of the amalgamation of these two fields, the term "cheminformatics" originated. With the foundation of the *Journal of Chemical Information and Modelling*, chemists embraced the importance of information technology in the field of chemistry in 1961 and it still remains the core journal for cheminformatics (Willett, 2008). Cheminformatics primarily deals with the processing of chemical data extracted

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00005-1

from the molecular structure, which is then analyzed to establish a meaningful relation between the structure features of a compound and its activity. Strictly speaking, cheminformatics extracts, assimilates, manipulates, stores, visualizes, and interprets the chemical information stored in molecules (Gasteiger, 2006). In the beginning, cheminformatics was introduced as a tool to facilitate drug discovery; nevertheless, it has extended its applications to other areas of chemistry and biology (Prakash and Gareja, 2010).

7.2.1 Role of cheminformatics in drug designing

The dawn of cheminformatics has brought a breakthrough in pharmaceutical research. It has become an inseparable approach to discover new drugs as the entire process is expedited leading to remarkable reductions in the time and cost of developing a drug. The approach to design a new drug has been given a completely new outlook from hit-and-trial to tailored-made designing. Traditional drug designing methods relied on synthesizing a random library of chemical compounds that contained many nondrug-like molecules, whereas computational techniques are based on a rational approach for searching large databases and modeling their physicochemical and biological properties to design potent and commercially feasible drugs. What has made cheminformatics so instrumental in revolutionizing pharmaceutical research? It has facilitated advances in high-throughput screening and combinatorial sciences. As a result of these advances, structural and bioactivity data have really burgeoned, adding millions of chemical compounds with each passing year. This necessitates the use of more sophisticated informatics techniques that can efficiently handle this avalanche of data. For many decades, cheminformatics has successfully been employed in the discovery of various drugs. Although it is hard to tabulate all the data applications, an attempt has been made to include the major ones in this chapter.

Normally, the process of discovering a novel drug commences with the shortlisting of an appropriate target that is involved in a particular disease (Turk and Cantley, 2003). This is then followed by searching, designing, and screening the potential compounds, which can behave like a drug (usually inhibitors of the target molecule). At this stage, various in silico techniques are employed to systematically identify drug-like candidates, which undergo various screening criteria to prove their efficacy. The following are the cheminformatics approaches that are usually adopted in a drug-discovery process.

7.2.1.1 Selection of a compound library

The selection of appropriate compound and generating virtual libraries is one of the first and foremost requirements in drug discovery as it is practically very difficult to screen huge numbers of compounds that combinatorial chemistry has added to the pool; only a few of these compounds turn into potential hits (Hall et al., 2001). Various computational methods have been developed to generate chemically diverse libraries comprising molecules similar to existing drugs. Some of these methods are

library enumerations, structural similarity algorithms, structural descriptor-based calculations, and other classical algorithms. Nevertheless, these approaches added little to the comfort of researchers. The screened molecules using these methods are not necessarily potential drug candidates. Moreover, other related issues like lead optimization, target validation, etc. cropped up during the process. However, improving the screening of designed compounds using these cheminformatics approaches remained a gigantic task. Fortunately, the virtual screening approach offered a better solution to the previous approaches. A computational approach mediated the generation of a virtual library of designed compounds with diversity, pharmacokinetics, and synthetic accessibility being the central criteria for the screening of drug-like compounds (Downs and Barnard, 1997; Walters et al., 1998; Bajorath, 2002; Lobanov and Agrafiotis, 2002).

7.2.1.2 Virtual screening

Virtual screening is basically a process of filtering out the less competent candidates that would not have resulted in potent drugs at a preliminary stage by imposing certain constraints that can be either ligand based or structure based (Oprea et al., 2005). In other words, virtual screening employs computational techniques to pick potential hits from virtual fragment libraries. This can be done by introducing filters that sort the compounds based on certain factors such as bioavailability (amount of medication after entering the body that is actually circulated), solubility, chemical reactivity/toxicity of chemical compounds, and absorption, distribution, metabolism, and excretion (ADME) (Lipinski et al., 1997; Huuskonen et al., 2000; Zuegge et al., 2001). Structure-based virtual screening is adopted when the target structure is well characterized and involves the docking of compounds with the targeted structure, whereas ligand-based virtual screening is adopted when the target structure is not known and involves a comparison of new compounds with drug-like compounds and their similarity index is analyzed for identification of structural features which are primarily responsible for pharmacological action. (Abagyan and Totrov, 2001; Diller and Merz, 2001; Duca and Hopfinger, 2001; Makara, 2001; Willett, 2000). However, there may be a case when the structures of the target and ligand are not known; then, structure−activity relationship paradigms may be determined by screening the experimental data using statistical tools (Hopfinger and Duca, 2000; Roberts et al., 2000; Gedeck and Willett, 2001). Besides, virtual screening is a useful technique for designing a combinatorial library for a given target. ZINC has been identified as one of largest databases, which provides access to the compounds that are available commercially for repurposing.

7.2.1.3 High-throughput screening

High-throughput screening is an excellent technique that allows swift testing of millions of compounds for their biological activities via an automated screening process. The process is so efficient that it can screen 10^3−10^6 molecules in parallel fashion (Attene-Ramos et al., 2014). Conventionally, the entire library of compounds is tested at a single concentration, whereas a more advanced version

of high-throughput screening called quantitative high-throughput screening provides a means to carry out testing of the compounds at various concentrations. Their response curves are generated immediately after performing screening and this output can be analyzed to predict the potential hits. High-throughput screening basically aims at identifying the active molecules that can alter a target in a favorable way (Camp et al., 2012). The basic information for high-throughput screening can be borrowed from virtual screening. High-throughput screening yields results that are more accurate and are comparable to online available libraries.

7.2.1.4 Structure–activity relationship on high-throughput screening data and sequential screening

As we move forward, each step works at narrowing down the cost of production and eliminating undesired candidates. The data retrieved from high-throughput screening can be further subjected to sequential high-throughput screening (Hawkins et al., 1997) where compounds are screened in an iterative manner, their activities are analyzed until desired, and nanomolar and novel leads are identified. This technique of further shortlisting the compounds is driven by structure–activity relationship analysis (Tropsha and Zheng, 2002).

7.2.1.5 In silico ADMET

ADMET expands ADME to adsorption, distribution, metabolism, elimination, and toxicity. Therefore ADMET determines the safety, uptake, elimination, metabolic behavior, and effectiveness of a drug. The emergence of predicting ADMET properties arose from that fact that almost 60% of the candidates fail in clinical trials. After identifying a lead compound, it is essential to evaluate the properties related to ADMET so that the effects of these compounds on the human body can be assessed. Modern-day in silico ADMET studies rely on computational methods (Paul Gleeson et al., 2011) to select molecules with reasonable values of these properties so that less competent candidates can be eliminated at an early stage and only the desired ones can be carried forward for synthesis and biological testing. This would further lead to cost reduction by avoiding in vivo and in vitro testing. The famous Lipinski's "rule of five" is also often employed to determine the drug-likeliness of a potential drug candidate. It is a set of five rules introduced by Christopher A. Lipinski in 1997 to underline the importance of physical and chemical properties of a molecule, i.e., ADME properties that could be a determining factor whether a given molecule is likely to be orally bioavailable or not. The rule was based purely on the observation that relatively small molecules with moderate lipophilicity have more chances of being orally active and their use is limited to only orally administered drugs. He then formulated it in the form of rules:

- Not more than five hydrogen bond donors.
- Not more than 10 hydrogen bond acceptors.
- The molar mass should be less than 500 Da.
- Maximum value of log P (partition coefficient) can be five.

A molecule is unlikely to behave like a drug if it violates two or more of the aforementioned conditions. Under all the mentioned conditions, the multiple of five lies at the core to determine the drug likeliness, hence this rule is known as the rule of five (Lipinski et al., 1997).

7.3 Role of bioinformatics in drug discovery

Earlier, we learnt how techniques of cheminformatics help in handling the data stored in chemical structures. A close observation, however, suggests that it only deals with small molecules. Informatics techniques are also helpful in treating complicated biomolecules, which gives rise to the birth of "bioinformatics." Both these fields complement each other for exploring the human physiological processes. A number of computational methods are available to design and develop novel lead inhibitors, which are called computer-aided drug designing (CADD) techniques. CADD is primarily of two types: structure-based and ligand-based in silico drug designing. There are other computational techniques too that assist drug designing: 3D pharmacophore modeling and the molecular dynamics simulation approach.

7.3.1 In silico designing of a drug using the structure-based approach

Structure-based drug designing basically focuses on determining and analyzing the 3D structures of target molecules. Drug targets are proteins and enzymes intrinsically involved in a specific metabolic pathway linked with a disease. Drug molecules are generally inhibitors of the target that are capable of altering the disease-related pathway to produce desired therapeutic results (Kaushik et al., 2018). The human genome project, after its successful completion, led to an exponential increase in information on various targets that created ample opportunities for drug discoverers. Structure-based designing facilitates the swift screening of potential drug candidates, which may subsequently be validated by employing simulation and visualization techniques. The key steps for structure-based drug designing are briefly explained next.

7.3.1.1 Selection of the target

A drug target is usually selected depending on the requirements of the problem disease; hence, this step is primarily biologically or biochemically driven. Ideally, the target must be associated with a disease and it must have a suitable binding-pocket/active site into which a drug or drug-like molecule can bind. Generally, proteins are good targets but sometimes RNA can also serve the purpose. Enzymes usually contain small grooves or pockets into which substrate (a small ligand) can easily bind and inhibit it; therefore, enzymes are excellent drug targets.

7.3.1.2 Evaluation of the drug target

After identifying a target, it is essential to have a well-characterized 3D structure of the enzyme/protein. The well-characterized 3D structures provide key information on the binding site that will assist in developing novel drugs for a particular disease. X-ray crystallographic and nuclear magnetic resonance techniques can be used to obtain the 3D structure of protein targets. Imagine the level of precision at which both these techniques are operating! Working at such a high resolution (atomic level) will allow us to explore the intermolecular interactions that are present in a protein–drug complex very precisely. This is what makes structure-based drug designing an indispensable tool in drug designing. In case an experimental-derived 3D structure of the target is not available, then the alternative approach is homology modeling. This is actually predictive modeling that works on the principle that "3D structures of two proteins will be similar if there is similarity in their sequencing." In other words, if the protein sequence of a protein is known, then the same sequence can be copied for other similar proteins. Accuracy of the derived protein structure through homology modeling depends on its probability (Enyedy et al., 2001a,b; Schapira et al., 2001).

7.3.1.3 Refining the target structure

This step is also called protein preparation. The protein obtained earlier must be refined before its interactions are studied with a drug molecule by adding hydrogen (because hydrogen cannot be detected by X-ray crystallography). Various tautomeric and protonation states are generated for the residues wherever required. Water molecules close to the binding sites are only retained and the rest of the water molecules and unnecessary heteroatoms are removed.

7.3.1.4 Locating the binding site

Locating the binding site or active site for the target protein is a crucial step. The binding site is a pocket in the target structure formed by protein folding, which allows a small drug-like molecule to bind into it. It is usually characterized by the presence of molecular surfaces, number of hydrogen bond donors and acceptors, and number of hydrophobic functionalities that help it to hold a substrate (Filikov et al., 2000; Lind et al., 2002).

7.3.1.5 Docking ligands into the binding site

After successfully identifying the structure of the target and its active site, it is important to model the interactions between inhibitors and target protein. It has been observed that molecules having similar binding affinities will show similar biological effects and hence careful analysis of binding site of a target will help in predicting novel ligands. At this stage, molecular docking, which is a structure-based drug design technique, is usually employed. A binding model clearly depicts the behavior of an inhibitor in the binding site of the enzyme and gives an insight into the interactions between ligand and receptor at the atomic level, which facilitates understanding of biochemical processes (McConkey et al., 2002). Molecular docking involves predicting the preferred conformation of a ligand in

the active site of the target and assessing its binding affinity. The earliest theory of molecular docking suggested that a lock-and-key type mechanism operates between a ligand and a receptor (Fischer, 1894). There was a shortcoming of this method in that it allowed refining of the ligand–receptor complex by fixing its relative orientations. This type of docking is called rigid docking. A more advanced version of this theory is induced-fit theory. As per induced-fit theory, the target receptor and ligand can orient themselves to achieve the best fit conformation (Koshland, 1963; Hammes, 2002). Contrary to rigid docking, flexible docking gives more accurate information about the binding events but of course it is computationally very expensive. Therefore docking is usually performed by keeping the receptor fixed and allowing the ligand to change its orientation continuously (Moitessier et al., 2008).

7.3.2 In silico drug designing using the ligand-based approach

The foremost condition for structure-based drug designing is the availability of a 3D structure of the protein/enzyme, but if the structure is not available, then it is not possible to apply the structure-based approach. Alternatively, information about ligands can play an important role in developing pharmacologically active ligands. This is called the ligand-based approach. There are ligand-based approaches such as quantitative structure–activity relationship (QSAR) and pharmacophore modeling, which are used commonly for drug designing (Loew et al., 1993; Mason et al., 2001).

7.3.2.1 Pharmacophore modeling

The presence of pharmacophoric features associated with the 3D spatial arrangement of structural properties such as hydrophobic surfaces, aromatic rings, hydrogen bond donating and accepting groups, etc. helps in binding a ligand inside the active site of the target (Alvarez, 2004; Verma et al., 2010). The term pharmacophore was first introduced to the world of drug designing by Ehrlich. It is essentially a framework that contains key structural features primarily driving biological activity of a ligand (Ehrlich, 1909). Normally, atom and functional group type, their position and orientation, and stereochemistry of a ligand are encoded into a pharmacophore model (Van Drie, 2003).

To create a pharmacophore, a dataset of biologically active ligands with a diverse scaffold (substituents can also be diverse) is selected. Conformational analysis of each ligand must be performed to construct a 3D model, whereas information about atoms and their connectivity is sufficient to construct a 2D model. These conformations are then superimposed on each other to find the common pharmacophoric features. Models are then generated by employing an algorithm. More than one model may be constructed using the same set of ligands and the one with highest score is finally selected for further procedure. These models are ranked using various scoring functions incorporated in the pharmacophore-building software (Güner, 2002). The top models are validated by exploring the receptor–ligand binding. A valid model is the one that gives a clear insight into the active site of the target and does not contradict the established mechanism.

7.3.2.2 Quantitative structure–activity relationship

QSAR methods are reliable predictive methods that can intuit the biological activities of untested compounds. QSAR essentially depends on a simple assumption that the pharmacological activity of compounds and their properties has a direct correlation with the structure of the molecule. Mathematically, it can be expressed as:

$$Biological\ activity = f\ (Physiochemical\ property)$$

Mathematically, it is possible to devise a relation to express biological activity in terms of physicochemical properties of a compound. It has been observed that molecules having similar physical and chemical properties tend to show similar biological activities (Akamatsu, 2002; Verma and Hansch, 2009). These relationships are prepared using the statistical equations that are further employed for designing better inhibitors. The following steps are generally followed to build and test a QSAR model:

- Ligands with desired experimental biological activities are identified. To ensure a large variation in activities, selected compounds must be structurally similar yet diverse.
- Molecular descriptors derived from the structural and physicochemical properties of these compounds are determined.
- Quantitative relations between these descriptors and activity are formulated to explain the correlation between structure and activity by employing various statistical models.
- Developed QSAR models are tested for their predictability and robustness.

7.3.3 Another exquisite tool: molecular dynamics

Molecular dynamics (MD) is yet another powerful simulation method to investigate the dynamics of conformational space where ligand and receptor are both allowed to move over a definite time period. In other words, it can be said that MD visualizes temporal motion of the receptor–ligand complex to generate a trajectory for the whole system. The algorithms used in MD are much more advanced than any method applied for flexible docking. Newton's laws of motion are used for energy minimization. Although local optimization is amazingly effective in MD simulations, the process occurs in small steps; thus overcoming energy conformational barriers can be a challenge. Ideally, simulations must be performed at a remarkably high temperature because it will allow the system to overcome high-energy barriers that can lead to insufficient sampling. Structures are picked up from the course at a regular period for further energy minimization (Leach et al., 2007). Hence, a viable strategy may be used for random search and to identify preferred conformation of the ligand followed by more precise MD simulations. MD can be used for simulations of protein shapes and refinement of X-ray structures (Polanski, 2009).

7.4 Applications of cheminformatics and bioinformatics in the development of antimalarial drugs

7.4.1 Background of the disease

The scientific fraternity is working assiduously to discover new drugs for various diseases, yet a significantly high death rate due to malaria remains one of the major tragedies of this century. It is the foremost cause of mortality and morbidity in hot and humid places worldwide and remains an unresolved disease control priority (Ngoungou et al., 2006; Solomon et al., 2007; White et al., 2014). As per the WHO (2020) report, approximately 229 million cases of malaria were reported, which is a clear indication that the conquest of this ancient disease is still a long way off. As per the WHO malaria report 2020, India witnessed a sharp fall in total number of cases from 10 million in 2000 to 5.6 mllion cases in 2019. Although, India accounted for 84% of malaria related deaths in south east Asia. In humans, there are five species of parasites that cause malaria; these are *P. falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*, and *Plasmodium knowlesi* (Wilson et al., 2011). Drug resistance is a major challenge to clear the parasites from the bloodstream of the human host (Derbyshire et al., 2011). Since 78% of total malaria cases are caused by *P. falciparum*, under this study suitable antimalarials are screened with respect to *P. falciparum*.

7.4.2 Antimalarials commercially available

There are quite a large number of antimalarial drugs in use belonging to particular classes of chemical compounds having different modes of action and specificity for various biochemical targets of antimalarial therapy. The currently used antimalarial drugs are classified as per targets and mode of action in Table 7.1.

7.4.3 Hybrid molecules: an alternative to conventional antimalarial drugs

A popular strategy that has been adopted by various research groups is to synthesize hybrid compounds carrying more than one functional group to block or inhibit more than one pathway essential for the survival of the parasite. Quinoline-containing molecules have been in use for some time but the popularity of renowned drugs like chloroquine has fallen due to increased drug resistance (Vangapandu et al., 2007; Kouznetsov and Gómez-Barrio, 2009). Development of alternative drugs has prompted synthetic chemists to develop novel drugs as a first line of defense. It was proposed that hybrid molecules [an example shown in Fig. 7.1] carry pharmacophoric features that lead to parasite killing without any drug resistance over a long period (Dechy-Cabaret et al., 2000; Wenzel et al., 2010). This multitarget strategy gave rise to various hybrid molecules, for example, 4-aminoquinoline-trioxane, 4-aminoquinoline-based Mannich bases (Jarrahpour et al., 2007), 4-aminoquinoline-isatin (Agarwal et al., 2005a,b), and 4-aminoquinoline-ferrocene (Biot et al., 1997).

Table 7.1 Principally available antimalarial drugs.

Drugs	Target location	Pathway/ mechanism	Target molecule
Chloroquine, amodiaquine, piperaquine, quinine, quinidine, mefloquine, halofantrine, lumefantrine, primaquine	Food vacuole	Heme polymerization	Inhibit hematin detoxification
Artemisinin, arteether, artemether, artesunate, dihydroartemisinin	Food vacuole	Unknown	Free radical generations
Pyrimethamine, chloroproguanil	Cytosol	Folate metabolism	DHFR inhibitor
Sulfadoxine, sulfalene, dapsone	Cytosol	Folate metabolism	DHPS inhibitor
Atovaquone	Mitochondria	Electron transport	Cytochrome-C oxidoreductase
Azythromycin, clindamycin,	Apicoplast	Protein synthesis	Apicoplast ribosome

DHFR, *Dihydrofolate reductase;* DHPS, *dihydropteroate synthase.*

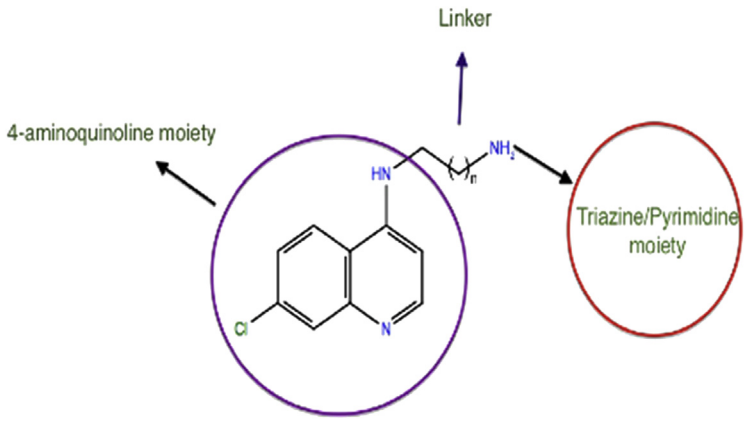


FIGURE 7.1

Hybrid molecule depicting two separate chemical entities connected by a linker.

4-Aminoquinoline moiety is known to block or inhibit hematin polymerization and convert it into a nontoxic product, while triazines and pyrimidines are well-established dihydrofolate reductase (DHFR) enzymes (Falco et al., 1951; Rosowsky et al., 1973; Agarwal et al., 2005a,b; Katiyar et al., 2005; Ojha et al., 2011; Müller and Hyde, 2013). DHFR of *P. falciparum* is an enzyme that mediates the conversion of nicotinamide adenine dinucleotide phosphate-dependent reduction of 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate in the folate metabolism (Osborne et al., 2001; Yuthavong et al., 2012; Rao and Tapale, 2013). Hence, it is an essential pathway for the survival of *P. falciparum* as it will lead to the

synthesis of purines and thymidine as well as the remethylation cycle of homocysteine to methionine. Inhibition of *P. falciparum* (pf)-DHFR by compounds like cycloguanil or pyrimethamine, etc. for long periods leads to mutations on amino acids like Ala16, Ile51, Cys59, Ser108, and Ile164 in the binding pocket of DHFR of *P. falciparum* (Liao et al., 2011).

In the present work, hybrid molecules were designed by combining 4-aminoquinolines and triazines/pyrimidines. 4-Aminoquinolines are known for their antimalarial efficacy against cycloguanil-sensitive and -resistant strains of *P. falciparum* (Manohar et al., 2012; Manohar et al., 2013), whereas s-triazines are capable of inhibiting the folate mechanism by targeting dihydrofolate reductase. One chemical entity is linked with another through a spacer arm, which is a linear-chained diaminoalkane termed a linker (Tables 7.3–7.5).

7.4.4 Computational details

7.4.4.1 Collection of dataset

A dataset of IC_{50} values for a total of 47 compounds based on functional moiety; 4-aminoquinoline, were used for pharmacophore modelling and subsequent 3D QSAR. The reported IC_{50} values were obtained during testing of these 47 compounds both against the sensitive and resistant strains of *P. falciparum* for the drug chloroquine.

7.4.4.2 Steps involved in pharmacophore and 3D QSAR model building

The steps incloved in pharmacophore modelling and 3D QSAR buidling were shown as in Figs. 7.2–7.6.

7.4.4.3 Preparation of ligands

A pharmacophore alignment and scoring engine module of Schrödinger 19.2 was performed to predict common pharmacophores. The reported IC_{50} values of dataset comprising analogs of 4-aminoquinolines, tabulated in Table 7.2 were transformed as a negative log of IC_{50} (pIC_{50}) to achieve uniform distribution. To discern the active molecules from inactive molecules, a threshold limit was set in pIC_{50} as 0.68 and 0.45 for active and inactive ligands, respectively. The inactive set can be used in scoring to screen out hypotheses that match both active and inactive ligands on the basis of Bayes classification. LigPrep was used to refine the geometries of ligands. The ionization, tautomeric, and stereoisomeric states at physiological pH (7 ± 2) for each of these molecules were generated. To search the flexibility and sophisticated conformational analysis of ligands, a mixed large-scale low-mode search was employed along with a mixed Monte Carlo low-mode search (Rosipal and Krämer, 2005). The search was performed using a dielectric solvation model along with optimized potential for liquid simulation (OPLS) version 2005 for tautomer generation. The limit value of 10 kcal/mol relative to the global energy minimum conformer was fixed for this iterative process.

Table 7.2 List of ligands selected based on 4-aminoquinoline derivatives. [Fig. 7.2]

Ligand Name	*n*	*X*	R_1	IC_{50} in μM (D6 clone) *Plasmodium falciparum*	IC_{50} in μM (W2 clone) *P. falciparum*	Ratio of activity (D6/W2)
1.	1	O	Aniline	0.29	0.17	1.705
2.	1	O	4-Ethylaniline	0.47	0.67	0.701
3.	1	O	4-fluoroaniline	0.24	0.48	0.500
4.	1	O	4-methoxyaniline	0.25	0.59	0.423
5.	1	O	$HNCH_2CH_2OH$	0.49	1.70	0.288
6.	1	O	$HNCH_2CH_2CH_2OH$	0.67	1.56	0.429
7.	1	O	$HNCH_2CH_2CH_2CH_2OH$	0.48	1.52	0.315
8.	1	CH_3	$HNCH_2CH_2CH_2CH_2OH$	0.14	0.40	0.350
9.	1	CH_3	$HNCH_2CH_2CH_2OH$	0.21	0.83	0.253
10.	1	CH_3	$HNCH_2CH_2OH$	0.19	0.49	0.387
11.	2	O	Aniline	0.44	1.05	0.419
12.	2	O	4-Ethylaniline	0.40	0.42	0.952
13.	2	O	4-fluoroaniline	0.43	0.33	1.303
14.	2	O	4-methoxyaniline	0.16	0.15	1.066
15.	2	O	$HNCH_2CH_2OH$	0.50	2.06	0.242
16.	2	O	$HNCH_2CH_2CH_2OH$	0.63	1.52	0.414
17.	2	O	$HNCH_2CH_2CH_2CH_2OH$	0.51	1.12	0.455
18.	3	O	Aniline	0.29	0.25	1.160
19.	3	O	4-Ethylaniline	0.39	0.18	2.166
20.	3	O	4-fluoroaniline	0.24	0.11	2.181
21.	3	O	4-methoxyaniline	0.07	0.71	0.090
22.	3	O	$HNCH_2CH_2OH$	0.19	0.59	0.322
23.	3	O	$HNCH_2CH_2CH_2OH$	0.22	0.57	0.385
24.	3	O	$HNCH_2CH_2CH_2CH_2OH$	0.09	0.49	0.183
25.	3	CH_3	$HNCH_2CH_2CH_2CH_2OH$	0.10	0.28	0.357
26.	3	CH_3	$HNCH_2CH_2CH_2OH$	0.06	1.17	0.051
27.	3	CH_3	$HNCH_2CH_2OH$	0.19	0.42	0.452

Table 7.3 List of 4-aminoquinoline-pyrimidine hybrids (6a–6d) [Fig. 7.3].

Ligand Name	*n*	IC_{50} in μM (D6 clone) *Plasmodium falciparum* (resistant)	IC_{50} in μM (W2 clone) *P. falciparum* (sensitive)	Ratio of activity (D6/W2)
6a	1	0.16	0.5	0.320
6b	2	0.33	0.70	0.471
6c	3	0.12	0.68	0.176
6d	5	0.44	0.54	0.814

Table 7.4 List of 4-aminoquinoline-pyrimidine hybrids (7a–7d) [Fig. 7.4].

Ligand Name	*n*	IC_{50} in μM (D6) *Plasmodium falciparum*	IC_{50} in μM (W2) *P. falciparum*	Ratio of activity (D6/W2)
7a	1	0.21	0.81	0.259
7b	2	0.24	1.17	0.205
7c	3	0.17	0.64	0.265
7d	5	0.14	0.58	0.241

Table 7.5 List of 4-aminoquinoline-pyrimidine hybrids (8b–8n) [Fig. 7.5].

Ligand Name	*n*	R_2	IC_{50} in μm (D6) *Plasmodium falciparum*	IC_{50} in μm (W2) *P. falciparum*	Ratio of activity (D6/W2)
8b	2	piperidine	0.02	0.21	0.095
8c	3	piperidine	0.02	0.09	0.222
8d	5	piperidine	0.06	0.10	0.600
8e	1	morpholine	0.02	0.14	0.142
8f	2	morpholine	0.02	0.05	0.400
8h	5	morpholine	0.03	0.14	0.214
8i	1	*N*-Methylpiperazine	0.005	0.03	0.166
8j	2	*N*-Methylpiperazine	0.007	0.016	0.116
8k	3	*N*-Methylpiperazine	0.021	0.023	0.913
8l	5	*N*-Methylpiperazine	0.007	0.016	0.437
8m	2	*N*-Methylpiperazine	0.006	0.06	0.100
8n	3	*N*-Methylpiperazine	0.02	0.023	0.869

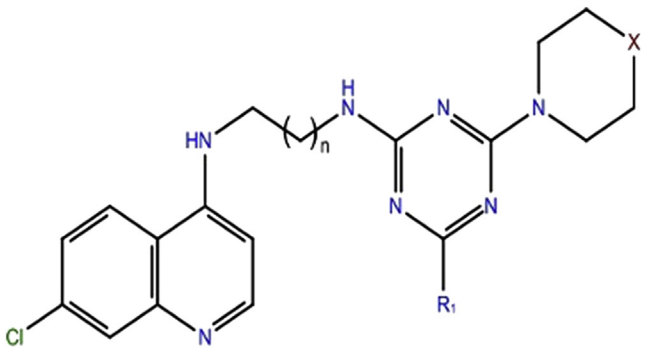


FIGURE 7.2

4-Aminoquinoline-triazine hybrids (1–27).

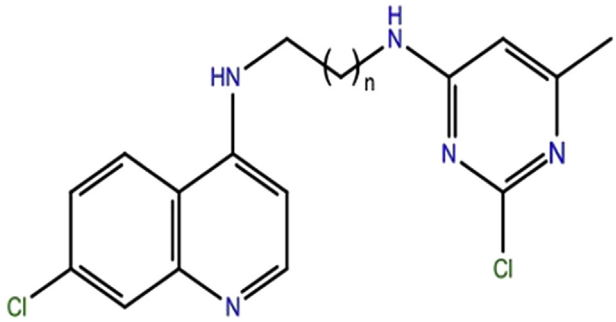


FIGURE 7.3

4-Aminoquinoline-pyrimidine hybrids (6a–6d).

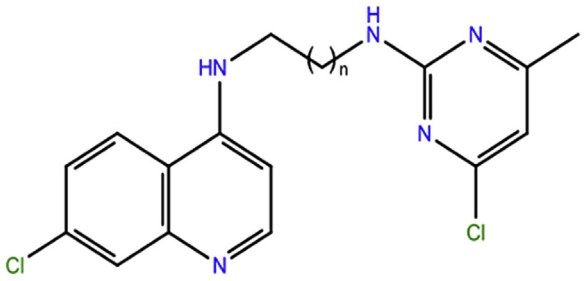


FIGURE 7.4

4-Aminoquinoline-pyrimidine hybrids (7a–7d).

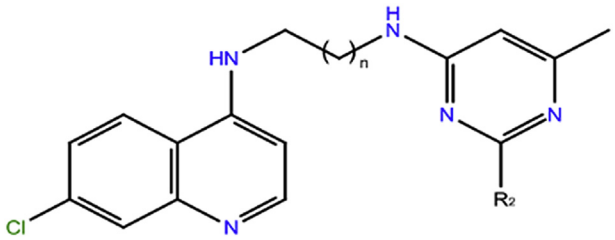


FIGURE 7.5

4-Aminoquinoline-pyrimidine hybrids (8b–8n).

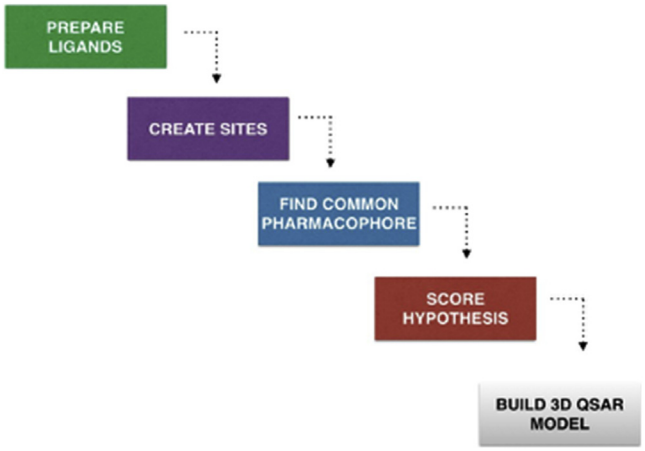


FIGURE 7.6

Fundamental steps required for pharmacophore modeling and 3D quantitative structure–activity relationship (QSAR) using phase-incorporated Schrödinger Inc. The schematic diagram portrays the steps followed to generate common pharmacophore models and to build an atom-based 3D QSAR.

7.4.4.4 Site creation and finding pharmacophores

The pharmacophore models were generated using all states of the compounds as discussed in Section 7.4.4.3. The chemical features of all the ligands were mapped to identify common pharmacophoric features (CPHs). A common tree-based portioning approach used active ligand conformations and identified pharmacophore models with 2 Å set distance. Table 7.6 shows the numbers of hypotheses that can be generated for individual variants in 10 active analogs. A hypothesis can be classified as either good or bad depending on its active and inactive features.

Each of the 123,262 hypotheses contained a number of similar pharmacophores, any of which could be considered to be a common pharmacophore. However, with the help of a variety of user-adjustable criteria, a single pharmacophore was selected from each variant, and these were deemed to be the common pharmacophore hypotheses.

Table 7.6 Identified pharmacophore hypotheses.

Variant	Maximum hypothesis
ADHRR	12,255
ADDHR	8,530
AHHRR	10,515
AHHHR	8,176
AAADR	1,714
ADDRR	1,018
AAAHR	4,069
AAHRR	5,670
ADHHR	30,883
AADHR	18,385
AADRR	2,703
AADDR	1,502
AAHHR	17,842
	Total = 123,262

7.4.4.5 Scoring of pharmacophores

First, the scores with respect to active analogs were calculated. The "survival score" is a product of root mean square deviation (RMSD) values and other score values like volume, site, etc. with certain weightages. The scoring was performed on all 13 variants selected in Table 7.6. Furthermore, the hypotheses were filtered based on their match to the inactive ligands. The resulting scoring by the hypothesis with highest survival from that of the inactive score was used subsequently for further analysis. The best perceived pharmacophore hypothesis was AADDR, which was picked for further validation.

7.4.4.6 Model validation: 3D QSAR

To ensure accuracy of CPHs, an atom-based 3D QSAR model based on partial least square (PLS) regression was used. The atom-based 3D QSAR approach basically considered all atoms as van der Waals spheres. All atoms were classified into the following six categories:

W: represents atoms that are nonionic oxygen and nitrogen and are treated as withdrawing atoms.
D: represents hydrogen atoms bonded to a polar atom and essentially function as donors.
H: atoms that are represented as hydrophobic atoms and these atoms are carbon and halogens.
N: represents atoms that carry a negative charge.
P: represents atoms that carry a positive charge.
X: represents the rest of the atoms classified as miscellaneous.

In this study, the whole dataset was divided into a training set and a test set. The training set ligands were used to develop the QSAR models and the test set was used for externally validating the generated QSAR. PLS regression was employed to generate the QSAR models. The PLS factor was kept at 3 with a grid spacing of 1 Å. In the current study, 70% of the dataset was assumed as a training set and the rest were treated as a test set. For the purpose of QSAR development, the best generated hypothesis, i.e., AADDR, was chosen.

7.4.4.7 Creating a virtual library

After the generation of CPHs and 3D QSAR validation, the foremost step was to create the 3D ZINC database for carrying out the screening process. ZINC is a freely available commercial library of chemical compounds for the screening of potential drugs (Wold et al., 2001). Site creation and the common pharmacophore protocol were repeated to obtain the various pharmacophore models for each of the 7781 molecules in the ZINC database, which were then stored with the pdb extension. The best generated hypothesis AADDR was used as a 3D structural input file to search the database for compounds that contained the pharmacophore features required of active ligands.

7.4.4.8 Molecular docking

The prepared ligands and the protein were then docked using the Glide module from the Schrödinger suite. Grid-based ligand docking with energetics was employed to screen out potential candidates on the basis of their binding affinity with the targeted protein. Furthermore, an attempt was made to dock pf-DHFR with ligands. The 3D target receptors *P. falciparum* dihydrofolate reductase-thymidylate synthase (pf-DHFR-TS) (PDB ID: 4DPD) and a hematin (CCD ID: 162267) structure were obtained from the Protein Data Bank and the Cambridge Crystallographic Database, respectively. The hematin structure was subjected to geometry optimization with the PM3 semiempirical method, whereas the protein target DHFR-TS was subjected to "protein preparation" Under this step, the missing bond orders were corrected along with addition of missing loops. Subsequently, cocrystallized water molecules were kept intact up to 5 Å distance from the binding site. Finally, hydrogen atoms were added because X-ray diffraction structure lacks them. The prepared protein structure was subjected to energy minimization using an OPLS forcefield. Minimization was processed in an iterative manner until the RMSD value of the heteroatoms reached 0.3 Å.

Side chain refinement was performed using the Prime module with default settings. To perform Prime GBSA for calculating free energy changes, the ligands were minimized using the density functional theory method and active residues of the binding pockets were subjected to a B3LYP/LACVP* basis set and minimized using the OPLS forcefield. LACVP* employs the 6-31G* basis set for nontransition elements in the active site. In MM/GBSA, energy values obtained using OPLS minimization (*EMM*) were combined with solvation contribution for polar solvents

(*GSB*) and nonpolar (*GNP*) solvents. The non-solar component consists of accessible surfaces and van der Waals contacts.

$$\Delta G_{binding} = G_{protein} - \left(G_{protein} + G_{ligand}\right)$$

$$G = EMM + GSB + GNP$$

7.4.4.9 In silico rapid ADME prognosis

The 457 hits that were lifted from the ZINC database were run on a QikProp module of Schrödinger software. The module was used to track down the candidates that possess drug-like properties. It is a speedy, precise, easy-to-use program designed by Professor William L. Jorgensen, which is used to predict ADME of a drug-like molecule by comparing its properties with those of known drugs. The 10 drug molecules that are most similar to the input molecules were identified for running the job. The module predicted 44 properties for the hits and these are tabulated in S1.

7.4.5 Results and discussion

The pharmacophore models were generated using 3D structural attributes of listed 4-aminoquinoline-based hybrids that were keys for hematin {Fe(III)(PPIX)} inhibition. The pharmacophore hypothesis generated gave the impression of the relative binding of the ligands to ferri(III)protoporphyrin IX, and for the predictability of the hypothesis, a 3D QSAR model was generated to identify overall aspects of the molecular structure that governed the activity. A pharmacophore essentially embodies all key binding aspects that have been collected from an experimental dataset of ligands bound to a receptor. Subsequently, variants of pharmacophore models were prepared and scored, and it was found in the present study that five featured pharmacophore models scored better than models containing three and four pharmacophoric features. These five featured pharmacophore models' CPHs were scored in terms of their alignment of active ligands with an RMSD value limit of 1.2−2 A with the default values as set for distances tolerance in the module of the software. The alignment values were expressed in terms of survival score as:

$$S = W_{site}S_{site} + W_{vec}S_{vec} + W_{vol}S_{vol} + W_{sel}S_{sel} + W^{m}_{rew}$$

where W and S represent weights and scores, S_{site} stands for alignment score, S_{vec} represents vector score, S_{vol} represents volume score, and S_{sei} represents selectivity score. The default values of W_{site} and W_{sei} were kept as 1 and 0 Å, respectively. $W^{m}{}_{\text{rew}}$ stands for reward weights defined as m^{-1}. The value of m is defined as a hypothesis of how many active sites can be matched. The CPHs so developed were again tested in terms of a score value by considering all 25 inactive ligands with a weightage of 1 Å. Later, CPHs were evaluated by subtracting their survival scores from inactive scores. Out of 10 variants, AADDR CPH was the best variant selected based on scoring. AADDR has the maximum predictive potential out of the 10

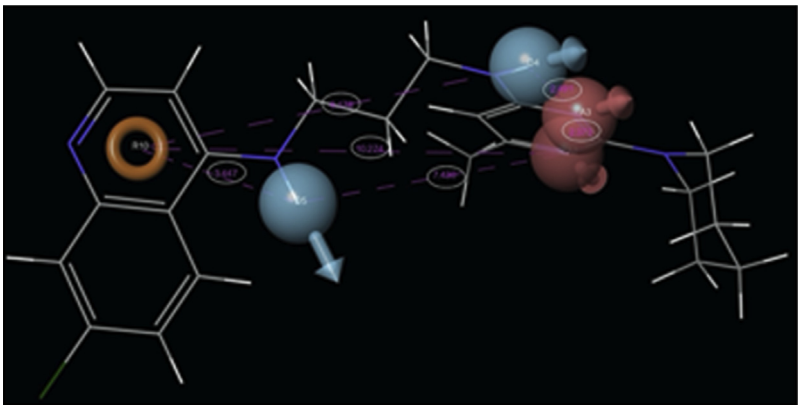


FIGURE 7.7

Fitting of the most active ligand (8b) in the created common pharmacophore with fit value 3 Å. (The orange, pink and light blue spheres represent aromatic ring, O_2 hydrogen bond acceptors, and O_2 hydrogen bond donors.)

variants of CPHs. Fig. 7.7 represents the spatially positioned features with their site distance AADDR.

To generate a QSAR model, out of 47 inhibitors, 33 (hematin detoxification) were used as testing compounds for internal validation and 14 were used for external validation of the QSAR model. A PLS approach was utilized to design the model. It built a linear model based on numerous dependent variables *Y*, and a set of independent factors that were used for the prediction of biological activity (Wold et al., 2001). Currently, "*x*" was the experimental biological activity and "*y*" was the predicted biological activity (both in PIC_{50}). Table 7.7 shows the 3D QSAR models for PLS factors 1, 2, and 3, and it was observed that for PLS factor 3, the statistical factor had significantly improved ($R^2 = 0.854$ and $Q^2 = 0.44$). Furthermore, increasing the PLS factor did not improved the statistics.

When all the ligands were superimposed over the created pharmacophore model, fitness was observed for the ligands with R^2 value 0.854 and Q^2 value of 0.449. Hence, the obtained pharmacophore model AADDR can be used satisfactorily for the screening of inhibitors out of a commercially available library. Fig. 7.8 shows the correlation of experimental activity with predicted activities for both the training and test sets of ligands and all necessary information in Table S2 (supplementary information).

After the generation of 12,362 compounds and screening of the library, exactly 1263 compounds were identified for further study with maximum fitness value of 2.96 (77% match) and lower fitness of 0.13 (4% match). All the necessary information is tabulated in Table S2.

Table 7.7 Partial least square (PLS) statistical parameters of the selected 3D quantitative structure—activity relationship model of hypothesis AADDR.

ID	PLS factor	SD	R^2	*F*	*P*	RMSD	Q^2	Pearson-*R*
AADDR	1	0.222	0.426	23.1	3.754e-05	0.292	0.074	0.314
	2	0.169	0.679	31.8	3.884e-08	0.277	0.169	0.452
	3	0.116	0.854	56.7	3.056e-12	0.225	0.449	0.675

RMSD, *Root mean square deviation;* SD, *standard deviation.*

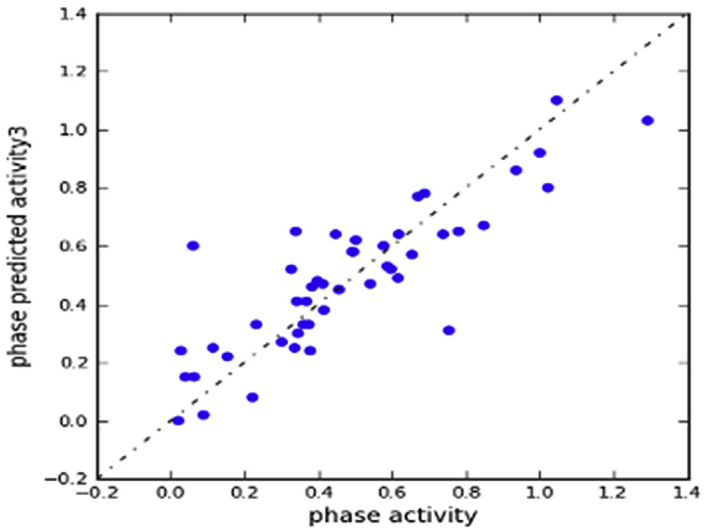


FIGURE 7.8

Fitness graph between observed activity versus phase-predicted activity for training and test set compounds.

7.4.5.1 3D QSAR visualization

A notable conjecture of the predicted model can only be achieved after the visualization of atoms in 3D space. A specialized pattern of cubes was created that illustrated the structural features responsible for intensifying as well as diminishing the inhibitory effect against hematin detoxification. Fig. 7.9A–C shows the conceptual representation of the generated contours. In these pictorial representations, the yellow cubes signify the favorable regions, i.e., if the functional groups are added to this position it may enhance the biological activity, whereas the purple cubes indicate unfavorable regions where the placement of the functionality may not lead to any enhancement of the activity.

7.4.5.2 Virtual database screening

Database searching can competently be used to spot unique and potential inhibitors. This enables us to search a database to identify potential inhibitors by filtering the inactive ones. AADDR was considered as a 3D structural query to search the entire ZINC database, which comprises seven thousand seven compounds. An iterative search was carried out, which involves screening of those compounds, which encompasses the same model with similar pharmacophoric features as in AADDR hypothesis. The search retrieved 457 potent candidates.

7.4.5.3 Drug resemblance analysis

The 457 hits were made to pass through ADME screening. All the four parameters had a profound effect on the plasma kinetics of drugs in the systemic circulation.

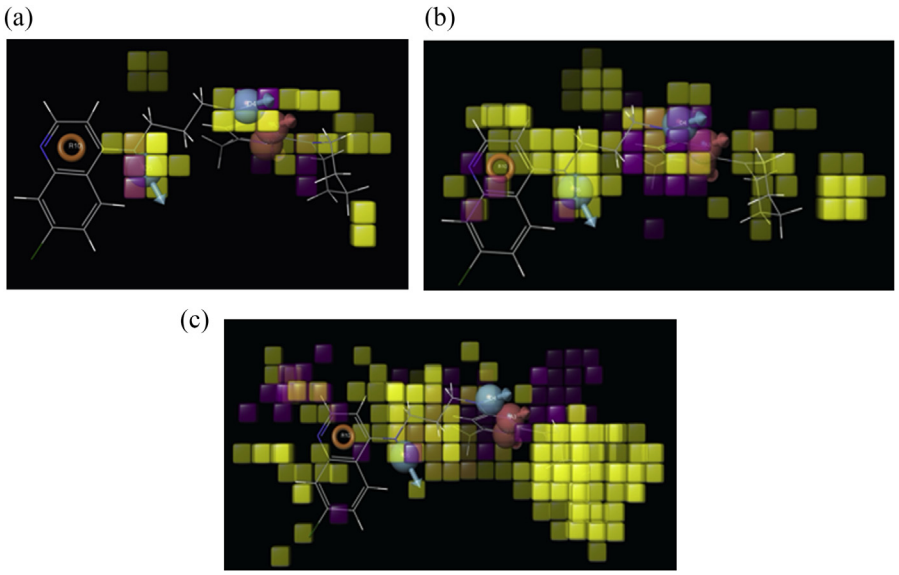


FIGURE 7.9

(A) 3D visualization of hydrogen bond donor in a quantitative structure–activity relationship (QSAR) model (yellow cubes depict positive potential and purple cubes exhibit negative potential for H bond substitution). (B) 3D visualization of electron withdrawing group in the QSAR model (yellow cubes depict positive potential and purple cubes exhibit negative potential for electron withdrawing substitution). (C) 3D visualization of hydrophobic/nonpolar groups in the QSAR model (yellow cubes depict positive potential and purple cubes exhibit negative potential for hydrophobic/nonpolar substitutions).

A total of 44 parameters such as solubility, partition coefficient, blood–brain barriers, gut–blood barriers, log HAS for serum binding, and Lipinski's five rules were obtained for all the hits. All the values are summarized in Table S3. Forty-nine compounds could successfully pass ADME screening. ADME properties of the selected four lead candidates are summarized in Table S3 (supplementary information).

To gain an insight into receptor–ligand binding, the 49 successful compounds obtained from ADME were docked into a ferri(III)protoporphyrin IX ring, which is a structural unit of hematin and the active site of pf DHFR-TS, respectively, using a nondeterministic sampling method. The 49 compounds showed diverse Glide scores with both targets. The molecular docking results are summarized in Table 7.8. In agreement with our experimental results, the Glide scores of all the compounds are better than already known drugs, which validates that they inhibit pf-DHFR-TS more effectively and are highly active compared to the standard drugs cycloguanil, pyrimethamine, and chloroquine. We considered only the top four scoring poses for further analysis (Table 7.8). The pose that is best for the docking of ligand with the receptor was selected based on various parameters such as Glide score, model energy score (E_{model}), nonbonded contribution, and internal energy of generated conformation.

Table 7.8 Glide scores of top four hit molecules.

Lead molecules	Glide score for inhibition of Fe(III)PPIX	Glide score for inhibition of pf-DHFR-TS enzyme
Lead 1	−5.49	−5.38
Lead 2	−5.46	−5.17
Lead 3	−4.47	−4.22
Lead 4	−5.30	−5.24
Cycloguanil	NA	−3.44
Pyrimethamine	NA	−4.0
Chloroquine	−3.42	NA

pf-DHFR, *Plasmodium falciparum dihydrofolate reductase.*

7.4.5.4 Docking of lead molecules with Fe(III)PPIX ring

Fe(III)PPIX is a complex consisting of a protoporphyrin ring bonded to one Fe(III) atom at the center. Singh et al., (2014) suggested that the iron atom in the protoporphyrin ring forms a bond with the heteroatoms in inhibitors. Fig. 7.10 displays the complex structures formed between the Fe(III)PPIX and the top lead molecule. All the docked structures illustrated that Fe(III) atoms form the center of the ring and connect with the atoms of the 1,2,4-triazole ring of the lead molecules, which is in agreement with the conclusion deduced by Singh et al. in their pioneering study (Singh et al., 2014). Therefore our attempt indicated that our lead molecules substantially bind to Fe(III) atoms of the protoporphyrin ring.

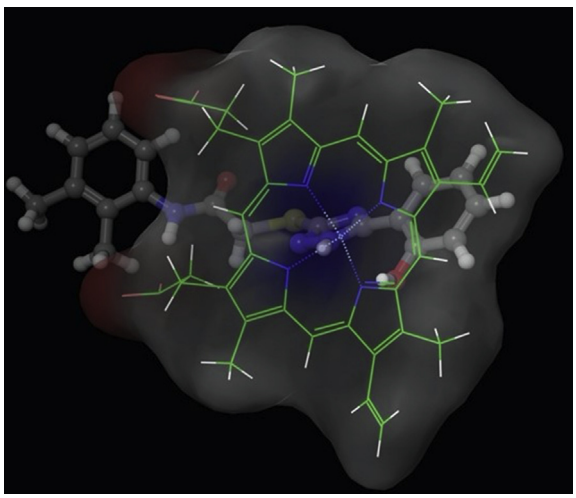


FIGURE 7.10

Interaction between Fe(III)PPIX and the lead molecule 1.

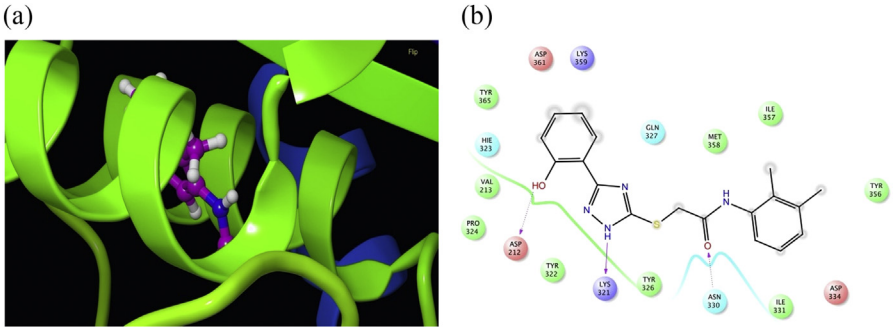


FIGURE 7.11

(A) A still image of docking between lead molecule 1 and the pf-DHFR, *Plasmodium falciparum* dihydrofolate reductase (pf-DHFR-TS) enzyme. (B) 2D interaction of lead molecule 1 with amino acids of the active site of the pf -DHFR-TS enzyme.

7.4.5.5 Docking of lead molecules with pf-DHFR

To validate further, our speculation, i.e., the presence of triazine/pyrimidine motif in the hybrid molecules, inhibits pf-DHFR. Docking was performed between the 49 database hits and the enzyme pf DHFR-TS. Fig. 7.11 shows a docking pose of the ligand, which scored the maximum for docking with the receptor.

It was deduced from the figure that lead molecule 1 forms the hydrogen bond with ASP 212, LYS 321, and ASN 330 inside the binding pockets of the dihydrofolate in the enzyme pf-DHFR-TS. LYS 321 interacts with the 1,2,4-triazole ring via ASP 212 connecting with the phenolic group and the ASN 330 amino acid forming a bond with the carbonyl group of the amide linkage.

3D and 2D representations of all shortlisted compounds are given in the supplementary information (Fig. S1).

7.5 Conclusions

The approach of drug designers has changed from "how to make" to "what to make." As a result, it has become mandatory to eliminate nondrug-like candidates as early as possible. With the advent of cheminformatics and bioinformatics, it is now possible to design novel drugs at a faster pace, saving time, money, effort, chemicals, and use of animals for in vivo testing. The case study of searching for potential antimalarials as discussed here concluded that for inhibition of Fe(III)PPIX and pf-DHFR, 1,2,4-triazole is the key functionality for the binding. After successfully applying the techniques embedded in CADD workflow, namely pharmacophore model creation, 3D QSAR modeling and database search, in silico ADME, and molecular docking, four lead compounds were shortlisted, which were found to inhibit Fe(III)PPIX and pf-DHFR more effectively than well-known antimalarial drugs such as chloroquine, cycloguanil, and pyrimethamine. These results clearly suggest that the choice of hybrid molecules as antimalarials can be a promising alternative to conventional aminoquinoline-based drugs. It becomes mandatory here to reveal that some of these hybrid compounds have also made their way to clinical trials (Meunier, 2008).

Electronic Supplementary information

Chem-bioinformatic approach for drug discovery: in silico screening of potential antimalarial compounds

(a-d) 3D and 2D representations of the top four lead molecules.

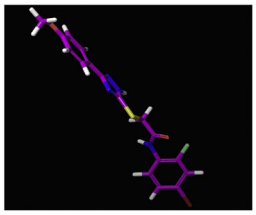
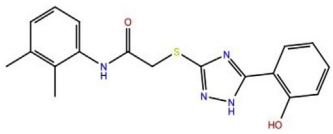
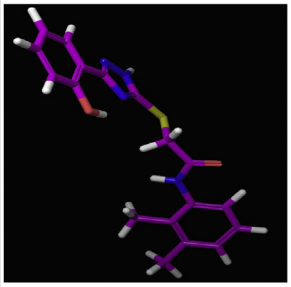
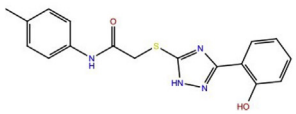
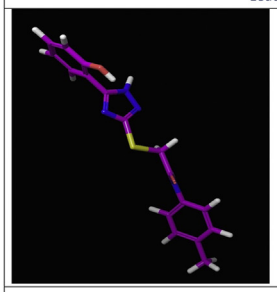
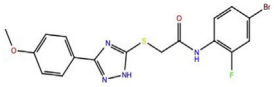
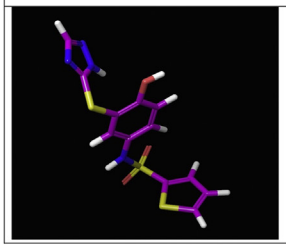
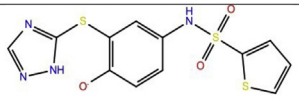
3D ilustration	2D ilustration
Lead 1: ZINC05520252	
	
Lead 2: ZINC06556763	
	
Lead 3: ZINC00633984	
	
Lead 4: ZINC06270540	
	

Table S1 QikProp determine the following parameters.

Property or descriptor	Description	Acceptable range/ recommended values for 95% of known drugs
#stars	Number of property or descriptor values that fall outside the 95% range of similar values for known drugs. Outlying descriptors and predicted properties are denoted with asterisks (*) in the .out file. A large number of stars suggests that a molecule is less drug-like than molecules with few stars. The following properties and descriptors are included in the deter- mination of #stars: MW, dipole, IP, EA, SASA, FOSA, FISA, PISA, WPSA, PSA, volume, #rotor, donorHB,accptHB, glob, QPpolrz, QPlogPC16, QPlogPoct, QPlogPw, QPlogPo/w, logS, QPLogKhsa, QPlogBB,#metabol	0 −5
#amine	Number of non-conjugated amine groups	0 −1
#amidine	Number of amidine and guanidine groups	0
#acid	Number of carboxylic acid groups	0-1
#amide	Number of non-conjugated amide groups	0 −1
#rotor	Number of non-trivial (not CX3), non-hindered (not alkene, amide, small ring) rotatable bonds	0 − 15
#rtvFG	Number of reactive functional groups; the specific groups are listed in the output file. The presence of these groups can lead to false positives in HTS assays and to decomposition, reactivity, or toxicity problems *invivo*	0 −2
CNS	Predicted central nervous system activity on a −2 (inactive) to +2 (active) scale	−2 (inactive), +2 (active)

Table S1 QikProp determine the following parameters.—*cont'd*

Property or descriptor	Description	Acceptable range/ recommended values for 95% of known drugs
mol_MW	Molecular weight of the molecule	130.0 – 725.0
dipole†	Computed dipole moment of the molecule	1.0 – 12.5
SASA	Total solvent accessible surface area (SASA) in square angstroms using a probe with a 1.4 A radius	300.0 – 1000.0
FOSA	Hydrophobic component of the SASA (saturated carbon and attached hydrogen)	0.0 – 750.0
FISA	Hydrophilic component of the SASA (SASA on N, O, H on heteroatoms, carbonyl C)	7.0 – 330.0
PISA	π (carbon and attached hydrogen) component of the SASA	0.0 – 450.0
WPSA	Weakly polar component of the SASA (halogens, P, and S)	0.0 – 175.0
volume	Total solvent-accessible volume in cubic angstroms using a probe with a 1.4 A radius	500.0 – 2000.0
donorHB	Estimated number of hydrogen bonds that would be donated by the solute to water molecules in an aqueous solution. Values are averages taken over a number of configurations, so they can be non-integer	0.0 – 6.0
accptHB	Estimated number of hydrogen bonds that would be accepted by the solutefrom water molecules in an aqueous solution. Values are averages taken over a number of configurations, so they can be non-integer	2.0 – 20.0
dip^2/V†	Square of the dipole moment divided by the molecular volume	0.0 – 0.13
ACxDN^.5/SA	Index of cohesive interaction in solids. This term represents the relationship (accptHB(donorHB))/(SA)	0.0 – 0.05

Table S1 QikProp determine the following parameters.—*cont'd*

Property or descriptor	Description	Acceptable range/ recommended values for 95% of known drugs
glob	Globularity descriptor, $(4\pi r2)/(SASA)$, where *r* is the radius of a sphere with a volume equal to the molecular volume. Globularity is 1.0 for a spherical molecule	0.75 − 0.95
polrz	Predicted polarizability in cubic angstroms	13.0 − 70.0
logPC16	Predicted hexadecane/gas partition coefficient	4.0 − 18.0
logPoct‡	Predicted octanol/gas partition coefficient	8.0 − 35.0
logPw/g	Predicted water/gas partition coefficient	4.0 − 45.0
logPo/w	Predicted octanol/water partition coefficient	−2.0 − 6.5
logS	Predicted aqueous solubility, log S. S in mol dm−3 is the con- centration of the solute in a saturated solution that is in equi- librium with the crystalline solid	−6.5 − 0.5
CI-logS	Conformation-independent predicted aqueous solubility, log S S in mol dm-3 is the concentration of the solute in a sat- urated solution that is in equilibrium with the crystalline solid	−6.5 − 0.5
logHERG	Predicted IC50 value for blockage of HERG K+ channels	concern below −5
PCaco	Predicted apparent Caco-2 cell permeability in nm/sec. Caco- 2 cells are a model for the gut-blood barrier. QikProp predictions are for non-active transport	<25 is poor, >500 is great
logBB	Predicted brain/blood partition coefficient. QikProp predictions are for orally delivered drugs	−3.0 − 1.2
PMDCK	Predicted apparent MDCK cell permeability in nm/sec. MDCK cells are considered to be a good mimic for the blood- brain barrier. QikProp predictions are for non-active transport	<25 is poor, >500is great

Table S1 QikProp determine the following parameters.—*cont'd*

Property or descriptor	Description	Acceptable range/ recommended values for 95% of known drugs
logKp	Predicted skin permeability, log *Kp*	−8.0 − −1.0
IP(ev)†	PM3 calculated ionization potential	7.9 − 10.5
EA(eV)†	PM3 calculated electron affinity	−0.9 − 1.7
#metab‡	Number of likely metabolic reactions	1 − 8
logKhsa	Prediction of binding to human serum albumin	−1.5 − 1.5
Human Oral Absorption	Predicted qualitative human oral absorption	1, 2, or 3 for low, medium, or high.
% Human Oral Absorption	Predicted human oral absorption on 0 to 100% scale. The pre- diction is basedon a quantitative multiple linear regression model. This property usually correlates well with HumanOral- Absorption, as both measure the sameproperty	>80% is high <25% is poor
SAFluorine	Solvent-accessible surface area of fluorine atoms	0.0 − 100.0
SAamideO	Solvent-accessible surface area of amide oxygen atoms	0.0 − 35.0
PSA	Van der Waals surface area of polar nitrogen and oxygen atoms and carbonyl carbon atoms	7.0 − 200.0
#NandO	Number of nitrogen and oxygen atoms	2 − 15
Rule of Five	Number of violations of Lipinski's rule of five. The rules are: mol_MW<500, logPo/w < 5, donorHB ≤ 5, accptHB≤ 10. Compounds that satisfy these rules are considered drug- like	maximum is 4
Rule of Three	Number of violations of Jorgensen's rule of three. The three rules are: QPlogS> -5.7, QP PCaco> 22 nm/s, # Primary Metabolites < 7. Compounds with fewer (and preferably no) violations of these rules are more likely to be orally available	maximum is 3
#ringatoms	Number of atoms in a ring	
#in34	Number of atoms in 3- or 4-membered rings	

Table S1 QikProp determine the following parameters.—*cont'd*

Property or descriptor	Description	Acceptable range/ recommended values for 95% of known drugs
#in56	Number of atoms in 5- or 6-membered rings	
#noncon	number of ring atoms not able to form conjugated aromatic systems .	
#nonHatm	Number of heavy atoms (nonhydrogen atoms)	
Jm	Predicted maximum transdermal transport rate, *K*p × MW × S (μg cm−2 hr−1). *K*p = predicted skin permeability MW=molecular weight S= solubility	

Table S2 Fitness and predicted activity of training and test set for 3D QSAR studies.

Ligand name	QSAR set	PIC_{50} = -(log IC_{50}) ratio of activity (D6/W2)	PLS factors	Predicted activity	Pharm set	Fitness
1.	Training	0.232	3	0.33	inactive	1.86
2.	Training	0.154	3	0.22	inactive	1.82
3.	Test	0.301	3	0.27	inactive	1.82
4.	Training	0.374	3	0.33	inactive	1.82
5.	Training	0.541	3	0.47		1.87
6.	Training	0.368	3	0.41	inactive	1.86
7.	Test	0.502	3	0.62		1.95
8.	Training	0.456	3	0.45		1.91
9.	Training	0.597	3	0.52		1.89
10.	Training	0.412	3	0.47	inactive	1.88
11.	Training	0.378	3	0.24	inactive	2.68
12.	Training	0.021	3	0.00	inactive	2.64
13.	Test	0.115	3	0.25	inactive	2.66
14.	Test	0.028	3	0.24	inactive	2.64
15.	Test	0.616	3	0.49		2.86

Table S2 Fitness and predicted activity of training and test set for 3D QSAR studies.—*cont'd*

Ligand name	QSAR set	PIC_{50} = -(log IC_{50}) ratio of activity (D6/W2)	PLS factors	Predicted activity	Pharm set	Fitness
16.	Test	0.383	3	0.46	inactive	2.84
17.	Training	0.342	3	0.41	inactive	2.82
18.	Training	0.064	3	0.15	inactive	2.06
19.	Training	0.336	3	0.25	inactive	1.7
20.	Training	0.339	3	0.65	inactive	1.81
21.	Training	1.046	3	1.10	active	1.81
22.	Test	0.492	3	0.58		1.77
23.	Training	0.415	3	0.38	inactive	1.89
24.	Training	0.738	3	0.64	active	1.77
25.	Test	0.447	3	0.64	inactive	1.74
26.	Training	1.292	3	1.03	active	1.83
27.	Training	0.345	3	0.30	inactive	1.96
6a	Training	0.495	3	0.58		1.93
6b	Training	0.327	3	0.52	inactive	2.86
6c	Test	0.754	3	0.31	active	2.14
6d	Training	0.089	3	0.02	inactive	1.1
7a	Training	0.587	3	0.53		1.78
7b	Training	0.688	3	0.78	active	1.48
7c	Training	0.577	3	0.60		1.34
7d	Test	0.618	3	0.64		1.17
8b	Test	1.022	3	0.80	active	3
8c	Training	0.654	3	0.57		2.23
8d	Training	0.222	3	0.08	inactive	1.31
8e	Test	0.848	3	0.67	active	1.91
8f	Training	0.398	3	0.48	inactive	2.82
8h	Training	0.67	3	0.77		1.47
8i	Training	0.78	3	0.65	active	1.9
8j	Test	0.936	3	0.86	active	2.96
8k	Training	0.04	3	0.15	inactive	2.11
8l	Training	0.36	3	0.33	inactive	1.09
8m	Training	1	3	0.92	active	2.94
8n	Test	0.061	3	0.60	inactive	2.21

Table S3 The ADME properties of the selected four lead candidates.

Lead	logPo/W	logS	PCaco	logKhsa	logBB	log HERG	% human oral
molecules	(-2.0 to 6.5)	(-6.5 to 0.5)	(<25 is poor &>500 is great)	(-1.5 to 1.5)	(-3.0 to 1.2)	(below -5.0)	absorption (<25% is poor & >80% is high)
Lead 1	3.214	-5.255	522.453	0.251	-1.036	-6.207	94.410
Lead 2	2.874	-5.316	283.807	0.175	-1.403	-6.584	87.678
Lead 3	4.129	-6.306	1000.186	0.363	-0.523	-6.491	100.000
Lead 4	1.153	-3.332	114.601	-0.442	-1.532	-5.437	70.550

Acknowledgments

Dr. Mallika Pathak, Dr. Mamta Sethi, and Dr. Malti Sharma are thankful to Principal Miranda House for her encouragement for this chapter. Dr. Himanshu Ojha is thankful to Director INMAS and Mr. Vinod Kumar Kaushik, Head of Division of CBRN Defence for their support. The authors have not received any funding in this regards.

References

Abagyan, R., Totrov, M., 2001. High-throughput docking for lead generation. Curr. Opin. Chem. Biol. 5 (4), 375–382.

Agarwal, A., Srivastava, K., Puri, S.K., Chauhan, P.M., 2005. Antimalarial activity and synthesis of new trisubstituted pyrimidines. Bioorg. Med. Chem. Lett 15 (12), 3130–3132.

Agarwal, A., Srivastava, K., Puri, S.K., Chauhan, P.M., 2005. Syntheses of 2, 4, 6-trisubstituted triazines as antimalarial agents. Bioorg. Med. Chem. Lett. 15 (3), 531–533.

Akamatsu, M., 2002. Current state and perspectives of 3D-QSAR. Curr. Top. Med. Chem. 2 (12), 1381–1394.

Alvarez, J.C., 2004. High-throughput docking as a source of novel drug leads. Curr. Opin. Chem. Biol. 8 (4), 365–370.

Attene-Ramos, M.S., Austin, C.P., Xia, M., 2014. High throughput screening. In: Wexler, P. (Ed.), Encyclopedia of Toxicology, third ed. Academic, Oxford, pp. 916–917.

Augen, J., 2002. The evolving role of information technology in the drug discovery process. Drug Discov. Today 7 (5), 315–323.

Bajorath, J., 2002. Virtual screening in drug discovery: methods, expectations and reality. Curr. Drug Discov. 2, 24–28.

Biot, C., Glorian, G., Maciejewski, L.A., Brocard, J.S., Domarle, O., Blampain, G., Millet, P., Georges, A.J., Abessolo, H., Dive, D., Lebibi, J., 1997. Synthesis and antimalarial activity in vitro and in vivo of a new ferrocene– chloroquine analogue. J. Med. Chem. 40 (23), 3715–3718.

Camp, D., Davis, R.A., Campitelli, M., Ebdon, J., Quinn, R.J., 2012. Drug-like properties: guiding principles for the design of natural product libraries. J. Nat. Prod. 75 (1), 72–81.

Dechy-Cabaret, O., Benoit-Vical, F., Robert, A., Meunier, B., 2000. Preparation and antimalarial activities of "trioxaquines", new modular molecules with a trioxane skeleton linked to a 4-aminoquinoline. Chembiochem 1 (4), 281–283.

Derbyshire, E.R., Mota, M.M., Clardy, J., 2011. The next opportunity in anti-malaria drug discovery: the liver stage. PLoS Pathogen. 7 (9), e1002178.

Diller, D.J., Merz Jr., K.M., 2001. High throughput docking for library design and library prioritization. Protein.: Struct., Funct., Bioinform. 43 (2), 113–124.

Downs, G.M., Barnard, J.M., 1997. Techniques for generating descriptive fingerprints in combinatorial libraries. J. Chem. Inf. Comput. Sci. 37 (1), 59–61.

Duca, J.S., Hopfinger, A.J., 2001. Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. J. Chem. Inf. Comput. Sci. 41 (5), 1367–1387.

Ehrlich, P., 1909. Über den jetzigen Stand der Chemotherapie. Ber. Dtsch. Chem. Ges. 42 (1), 17–47.

Enyedy, I.J., Lee, S.L., Kuo, A.H., Dickson, R.B., Lin, C.Y., Wang, S., 2001. Structure-based approach for the discovery of bis-benzamidines as novel inhibitors of matriptase. J. Med. Chem. 44 (9), 1349−1355.

Enyedy, I.J., Ling, Y., Nacro, K., Tomita, Y., Wu, X., Cao, Y., Long, Y.Q., 2001. Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. J. Med. Chem. 44 (25), 4313−4324.

Falco, E.A., Goodwin, L.G., Hitchings, G.H., Rollo, I.M., Russell, P.B., 1951. 2: 4 Diaminopyrimidines—a new series of antimalarials. Br. J. Pharmacol. Chemother. 6 (2), 185.

Filikov, A.V., Mohan, V., Vickers, T.A., Griffey, R.H., Cook, P.D., Abagyan, R.A., James, T.L., 2000. Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. J. Comput. Aided Mol. Des. 14 (6), 593−610.

Fischer, E., 1894. Einfluss der Configuration auf die Wirkung der Enzyme. Ber. Dtsch. Chem. Ges. 27 (3), 2985−2993.

Gasteiger, J., 2006. The central role of chemoinformatics. Chemometr. Intell. Lab. Syst. 82 (1−2), 200−209.

Gedeck, P., Willett, P., 2001. Visual and computational analysis of structure−activity relationships in high-throughput screening data. Curr. Opin. Chem. Biol. 5 (4), 389−395.

Guner, O.F., 2002. History and evolution of the pharmacophore concept in computer-aided drug design. Curr. Top. Med. Chem. 2 (12), 1321−1332.

Hall, D.G., Manku, S., Wang, F., 2001. Solution-and solid-phase strategies for the design, synthesis, and screening of libraries based on natural product templates: a comprehensive survey. J. Comb. Chem. 3 (2), 125−150.

Hammes, G.G., 2002. Multiple conformational changes in enzyme catalysis. Biochemistry 41 (26), 8221−8228.

Hawkins, D.M., Young, S.S., Rusinko III, A., 1997. Analysis of a large structure-activity data set using recursive partitioning. Quant. Struct.-Act. Relat. 16 (4), 296−302.

Hopfinger, A.J., Duca, J.S., 2000. Extraction of pharmacophore information from high-throughput screens. Curr. Opin. Biotechnol. 11 (1), 97−103.

Huuskonen, J., Rantanen, J., Livingstone, D., 2000. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. Eur. J. Med. Chem. 35 (12), 1081−1088.

Jarrahpour, A., Khalili, D., De Clercq, E., Salmi, C., Brunel, J.M., 2007. Synthesis, antibacterial, antifungal and antiviral activity evaluation of some new bis-Schiff bases of isatin and their derivatives. Molecules 12 (8), 1720−1730.

Katiyar, S.B., Srivastava, K., Puri, S.K., Chauhan, P.M., 2005. Synthesis of 2-[3, 5-substituted pyrazol-1-yl]-4, 6-trisubstituted triazine derivatives as antimalarial agents. Bioorg. Med. Chem. Lett 15 (22), 4957−4960.

Kaushik, A.C., Kumar, A., Bharadwaj, S., Chaudhary, R., Sahi, S., 2018. Bioinformatics Techniques for Drug Discovery: Applications for Complex Diseases. Springer International Publishing.

Koshland, D.E., 1963. Correlation of structure and function in enzyme action. Science 142 (3599), 1533−1541.

Kouznetsov, V.V., Gómez-Barrio, A., 2009. Recent developments in the design and synthesis of hybrid molecules basedon aminoquinoline ring and their antiplasmodial evaluation. Eur. J. Med. Chem. 44 (8), 3091−3113.

Leach, A.R., John, B.T., David, J.T., 2007. Comprehensive Medicinal Chemistry II. Elsevier, Oxford, pp. 87−118.

Lind, K.E., Du, Z., Fujinaga, K., Peterlin, B.M., James, T.L., 2002. Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. Chem. Biol. 9 (2), 185–193.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. 23 (1–3), 3–25.

Loew, G.H., Villar, H.O., Alkorta, I., 1993. Strategies for indirect computer-aided drug design. Pharmaceut. Res. 10 (4), 475–486.

Lobanov, V.S., Agrafiotis, D.K., 2002. Scalable methods for the construction and analysis of virtual combinatorial libraries. Comb. Chem. High Throughput Screen. 5 (2), 167–178.

Makara, G.M., 2001. Measuring molecular similarity and diversity: total pharmacophore diversity. J. Med. Chem. 44 (22), 3563–3571.

Manohar, S., Rajesh, U.C., Khan, S.I., Tekwani, B.L., Rawat, D.S., 2012. Novel 4-aminoquinoline-pyrimidine based hybrids with improved in vitro and in vivo antimalarial activity. ACS Med. Chem. Lett. 3 (7), 555–559.

Manohar, S., Khan, S.I., Rawat, D.S., 2013. 4-Aminoquinoline-Triazine-Based hybrids with improved in vitro antimalarial activity against CQ-sensitive and CQ-resistant strains of plasmodium falciparum. Chem. Biol. Drug Des. 81 (5), 625–630.

Mason, J.S., Good, A.C., Martin, E.J., 2001. 3-D pharmacophores in drug discovery. Curr. Pharmaceut. Des. 7 (7), 567–597.

McConkey, B.J., Sobolev, V., Edelman, M., 2002. The performance of current methods in ligand–protein docking. Curr. Sci. 845–856.

Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., Corbeil, A.C., 2008. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br. J. Pharmacol. 153 (S1), S7–S26.

Müller, I.B., Hyde, J.E., 2013. Folate metabolism in human malaria parasites—75 years on. Mol. Biochem. Parasitol. 188 (1), 63–77.

Ngoungou, E.B., Quet, F., Dubreuil, C.M., Marin, B., Houinato, D., Nubukpo, P., Dalmay, F., Millogo, A., Nsengiyumva, G., Kouna-Ndouongo, P., Diagana, M., 2006. Epidemiology of epilepsy in sub-Saharan Africa: a review. Cahiers d'études et de recherches francophones/Santé 16 (4), 225–238.

Osborne, M.J., Schnell, J., Benkovic, S.J., Dyson, H.J., Wright, P.E., 2001. Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. Biochemistry 40 (33), 9846–9859.

Oprea, T.I., Bologa, C., Olah, M., Alvarez, J., Shoichet, B., 2005. Compound selection for virtual screening. Virtual Screen. Drug Discov. 89–106.

Paul Gleeson, M., Hersey, A., Hannongbua, S., 2011. In-silico ADME models: a general assessment of their utility in drug discovery applications. Curr. Top. Med. Chem. 11 (4), 358–381.

Prakash, N., Gareja, D.A., 2010. Cheminformatics. J. Proteonomics Bioinf. 3, 249–252.

Polanski, J., 2009. Receptor dependent multidimensional QSAR for modeling drug-receptor interactions. Curr. Med. Chem. 16 (25), 3243–3257.

Rao, A.S., Tapale, S.R., 2013. A study on dihydrofolate reductase and its inhibitors: a review. Int. J. Pharmaceut. Sci. Res. 4 (7), 2535.

Roberts, G., Myatt, G.J., Johnson, W.P., Cross, K.P., Blower, P.E., 2000. LeadScope: software for exploring large sets of screening data. J. Chem. Inf. Comput. Sci. 40 (6), 1302–1314.

Rosipal, R., Krämer, N., 2005. February. Overview and recent advances in partial least squares. In: International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection. Springer, Berlin, Heidelberg, pp. 34−51.

Rosowsky, A., Chen, K.K.N., Lin, M., 1973. 2, 4-Diaminothieno [2, 3-d] pyrimidines as antifolates and antimalarials. 3. Synthesis of 5, 6-disubstituted derivatives and related tetracyclic analogs. J. Med. Chem. 16 (3), 191−194.

Schapira, M., Raaka, B.M., Samuels, H.H., Abagyan, R., 2001. In silico discovery of novel retinoic acid receptor agonist structures. BMC Struct. Biol. 1 (1), 1−7.

Singh, K., Kaur, H., Smith, P., de Kock, C., Chibale, K., Balzarini, J., 2014. Quinoline−pyrimidine hybrids: synthesis, antiplasmodial activity, SAR, and mode of action studies. J. Med. Chem. 57 (2), 435−448.

Solomon, V.R., Haq, W., Srivastava, K., Puri, S.K., Katti, S.B., 2007. Synthesis and antimalarial activity of side chain modified 4-aminoquinoline derivatives. J. Med. Chem. 50 (2), 394−398.

Tropsha, A., Zheng, W., 2002. Rational principles of compound selection for combinatorial library design. Comb. Chem. High Throughput Screen. 5 (2), 111−123.

Turk, B.E., Cantley, L.C., 2003. Peptide libraries: at the crossroads of proteomics and bioinformatics. Curr. Opin. Chem. Biol. 7 (1), 84−90.

Van Drie, J.H., 2003. Pharmacophore discovery-lessons learned. Curr. Pharmaceut. Des. 9 (20), 1649−1664.

Vangapandu, S., Jain, M., Kaur, K., Patil, P., Patel, S.R., Jain, R., 2007. Recent advances in antimalarial drug development. Med. Res. Rev. 27 (1), 65−107.

Verma, R.P., Hansch, C., 2009. Camptothecins: a SAR/QSAR study. Chem. Rev. 109 (1), 213−235.

Verma, J., Khedkar, V.M., Coutinho, E.C., 2010. 3D-QSAR in drug design-a review. Curr. Top. Med. Chem. 10 (1), 95−115.

Walters, W.P., Stahl, M.T., Murcko, M.A., 1998. Virtual screening—an overview. Drug Discov. Today 3 (4), 160−178.

Wenzel, N.I., Chavain, N., Wang, Y., Friebolin, W., Maes, L., Pradines, B., Lanzer, M., Yardley, V., Brun, R., Herold-Mende, C., Biot, C., 2010. Antimalarial versus cytotoxic properties of dual drugs derived from 4-aminoquinolines and Mannich bases: interaction with DNA. J. Med. Chem. 53 (8), 3214−3226.

White, N.J., Pukrittayakamee, S., Hien, T.T., Faiz, M.A., Mokuolu, O.A., Dondorp, A.M., 2014. Malaria. Lancet 383 (9918), 723−735.

Willett, P., 2000. Chemoinformatics−similarity and diversity in chemical libraries. Curr. Opin. Biotechnol. 11 (1), 85−88.

Willett, P., 2008. A bibliometric analysis of the literature of chemoinformatics. In: Aslib Proceedings. Emerald Group Publishing Limited.

Wilson, M.E., Kantele, A., Jokiranta, T.S., 2011. Review of cases with the emerging fifth human malaria parasite, Plasmodium knowlesi. Clin. Infect. Dis. 52 (11), 1356−1362.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. 58 (2), 109−130.

Yuthavong, Y., Tarnchompoo, B., Vilaivan, T., Chitnumsub, P., Kamchonwongpaisan, S., Charman, S.A., McLennan, D.N., White, K.L., Vivas, L., Bongard, E., Thongphanchang, C., 2012. Malarial dihydrofolate reductase as a paradigm for drug development against a resistance-compromised target. Proc. Natl. Acad. Sci. U. S. A. 109 (42), 16823−16828.

Zuegge, J., Schneider, G., Coassolo, P., Lavé, T., 2001. Prediction of hepatic metabolic clearance. Clin. Pharmacokinet. 40 (7), 553−563.

Ojha, H., Gahlot, P., Tiwari, A.K., Pathak, M., Kakkar, R., 2011. Quantitative structure activity relationship study of 2, 4, 6-Trisubstituted-s-triazine derivatives as antimalarial inhibitors of plasmodium falciparum dihydrofolate reductase. Chem. Biol. Drug Des. 77 (1), 57−62.

Liao, C., Sitzmann, M., Pugliese, A., Nicklaus, M.C., 2011. Software and resources for computational medicinal chemistry. Fut. Med. Chem. 3 (8), 1057−1085.

World malaria report, WHO, 2020 ISBN 978-92-4-001579-1, https://www.who.int/malaria.

CHAPTER

Mapping genomes by using bioinformatics data and tools

8

Md Shoaib[1], **Anju Singh**[1,2], **Srishty Gulati**[1], **Shrikant Kukreti**[1]

[1]*Nucleic Acid Research Lab, Department of Chemistry, University of Delhi, North Campus, New Delhi, Delhi, India;* [2]*Department of Chemistry, Ramjas College, University of Delhi, New Delhi, Delhi, India*

8.1 Background

Information technology has paved the way for generating an exceptional amount of knowledge and data for various research fields. Bioinformatics represents a comprehensive field that plays a central role in combining other research fields like statistics, computation, molecular biology, and mathematics. It has been defined in various ways, but it simply deals with the collection, storage, and interpretation of biological information. The bioinformatics tools are computer algorithms and programs that can analyze the biological data more efficiently. We have witnessed (especially the last two decades) a revolution in the progress of computer-based technologies that have helped in revealing fundamental mechanisms involved in biological pathways, disease processes, and evolution in molecular biology. The Human Genome Project and microarray expression profiling have proved to be the revolution that established bioinformatics as a well-recognized discipline.

Any function of a species is the outcome of genetic, epigenetic, and treatment options. There have been several studies that have shown the risk factors and involvement of macromolecules that are DNA, RNA, protein, and metabolites in disease progression. Molecular biology deals with huge loads of data of the biological processes involved, which can be very hard to interpret manually. Then, there is the role of bioinformatics tools. Bioinformatics tools help to develop programs and methodologies to systemically collect and store large volumes of biological data so that we may better understand them. Fig. 8.1 represents a relationship between biological systems and databases.

Moreover, bioinformatics involves various duties like the study of the genome, data mining, structure prediction, molecular dynamics simulation, molecular docking, and designing. A few general bioinformatics tasks are shown in Fig. 8.2.

8.1.1 Emergence and evolution of bioinformatics

The origin of bioinformatics was associated with the evolution of protein sequencing back in the 1960s. However, it all started with the determination of the first protein

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00002-6

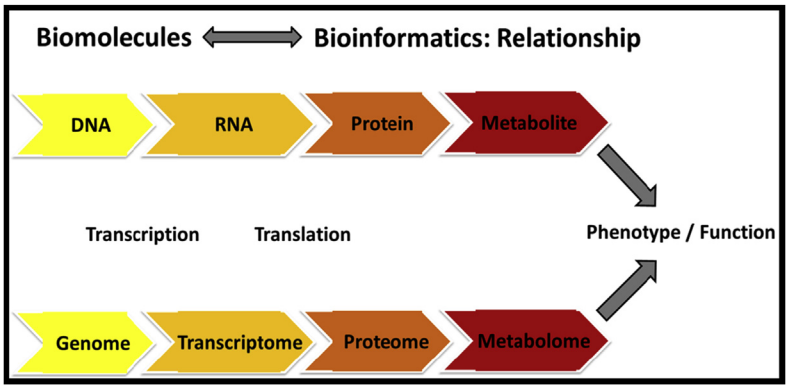


FIGURE 8.1

Genotype-to-phenotype relationships.

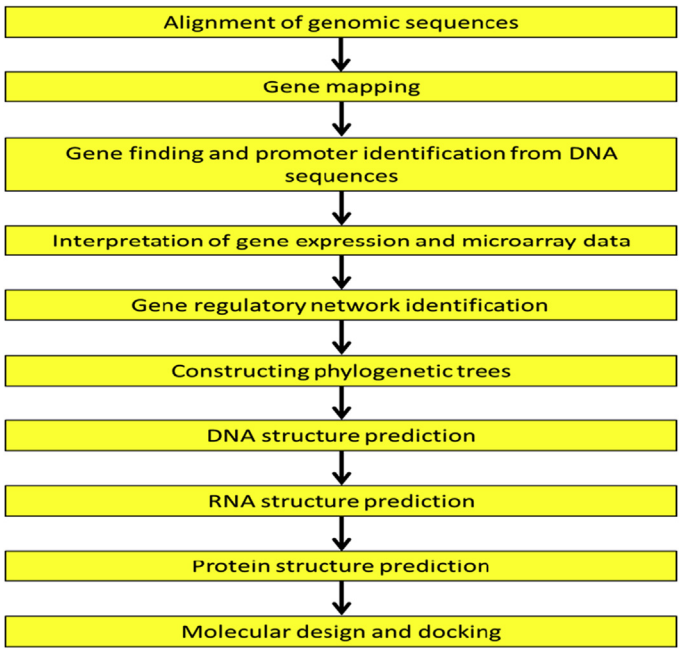


FIGURE 8.2

General tasks of bioinformatics tools.

sequence, insulin, in the early 1950s (Sanger and Thompson, 1953a,b). This breakthrough encouraged researchers to work in this field, but sequencing really came into existence with the development of Edman's degradation method of protein

sequencing (Edman, 1949). The simplicity of this method kept it very popular and useful for over a decade. With the increasing number of protein sequences determined, it started to become difficult to store the data, and also for large sequences, assembling a large number of small fragments of residues to get a complete sequence became problematic. This issue brought the emergence of the first computer-based program, named COMPROTEIN, which compiled Edman sequencing data and helped in the preparation of the Protein Information Resource (PIR) (Dayhoff and Ledley, 1962). This collection of sequences of proteins was published in 1965 as an atlas. The term "bioinformatics," however, was first coined in 1970 as information technology by the Dutch biologist, Hogeweg (Hesper and Hogeweg, 1970). This paradigm of protein sequencing began to shift to DNA sequencing when Francis Crick's hypothesis of central dogma was put forward. By that time, it became clear that it was DNA that governed the essential biological processes, thus regulating the synthesis of protein (Crick, 1958). In the 1970s, the field progressed rapidly with the growth of DNA sequencing techniques such as Sanger sequencing (Sanger and Coulson, 1975) and Maxam−Gilbert sequencing (Maxam and Gilbert, 1977). Thereafter, in the early 1980s with the advancement in computation, various DNA sequence databases were created by compiling DNA sequence data. The popular ones are Genbank (http://www.ncbi.nlm.nih.gov/genbank), EMBL (the European Molecular Biology Laboratory) (http://www.embl.org,), and DDBJ (DNA Data Bank of Japan) (http://www.ddbj.nig.ac.jp). The SWISS-MODEL server became the first widely used web-based automated modeling platform, which brought a massive change with its user-friendly interface in the 1990s (Brooks et al., 1993). Then, one of the most popular and useful web-based platforms from NCBI, the "BLAST tool" (Stephen et al., 1990), became available, which brought many vital databases such as Human Genome into the establishment (Brown, 2002). Human Genome was the biggest landmark establishment, which has revolutionized the genomic era. Fig. 8.3 displays few of these important milestones in the evolution of genomic era. In today's world, as the technology is growing, more researchers are trying to develop more interactive easy-to-use bioinformatics tools. Computational issues rose again with the development of next-generation sequencing (high-throughput sequencing), which enabled millions of DNA sequences to be sequenced in a single run. Roche Diagnostics came up with software, Newbler, that indeed set high standards for its rivals in processing the data from high-throughput sequencers (Silva et al., 2013). This rapid growth of bioinformatics has generated various projects and has required a lot of resources. The expense and expertise required brought several organizations into the picture. In fact, there are several government-sponsored programs running that are taking their respective countries forward in this race.

8.2 Genome

8.2.1 Gene expression

In genetics, DNA expression plays a vital role in various biochemical reactions of eukaryotes, prokaryotes (archaea and bacteria), and viruses. DNA governs the

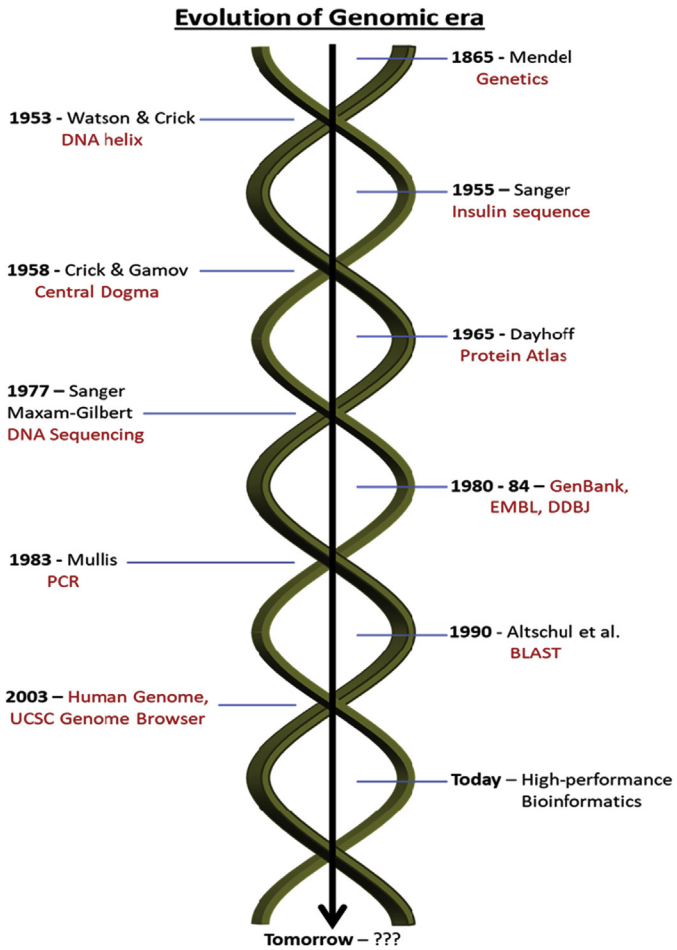


FIGURE 8.3

Evolution of the genomic era and bioinformatics.

synthesis of protein, which is an essential part of our various metabolic activities, consequently regulating gene expression as well as transferring heredity data from one generation to another. A gene is composed of exons (triplets of nucleotides) and introns (the noncoding part). Gene expression is a vital part of the cell cycle; it carries the information for protein synthesis.

In the human genome, only 3%−5% of the entire genome is coding, which results in the production of proteins. Regulation of gene expression affects various processes like transcription, translation, RNA splicing, and posttranslation modifications. It controls cell machinery timing and number of essential products of genes. A gene expresses in different ways in different organisms, which results in variation in

genotypes and phenotypes. Alteration in the gene sequence can cause mutation or various types of genetic disorders like sickle cell anemia, phenylketonuria, etc. For a better understanding of such genetic disorders, mutations, and diseases related to them, large amounts of data need to be analyzed. Therefore the advanced computational approach plays a big role in getting rid of such time-consuming data processing.

8.2.2 Gene prediction

Gene prediction, also known as gene recognition, has been an area of extensive research in bioinformatics. The occurrence of genetic variation within the same species has caught the interest of researchers for biological processes like splice sites, ribosomal binding sites, polyadenylation sites, topoisomerase I cleavage sites, topoisomerase II binding sites, start and stop codons, and various transcription factor binding sites of a genomic DNA (Gelfand, 1995; Sherriff and Ott, 2001). However, in bioinformatics, gene finding is easy in the case of the bacterial gene (prokaryotes) but in the case of eukaryotes, this process is very complex due to the presence of multiple types of intron/exon patterns. Thus it is very difficult to identify the functional sites of a gene in large sequences or in the case of unknown genes. So, machine learning has been developed as a revolutionary step in the prediction of transcription factors and coding regions in the genome (Alipanahi et al., 2015). The selection of bioinformatics tools depends on the nature of analysis required, for example, ATGpr is used to identify the translational sites in cDNA (Li and Leong, 2005), and AUGUSTUS and GLIMMERHMM are used for eukaryotic gene prediction (Keller et al., 2011; Majoros et al., 2004). Another set of tools, BGF, FGENESH, and PRODIGAL (Prokaryotic Dynamic Programming Genefinding Algorithm), are based on the hidden Markov model (HMM) and log-likelihood functions for gene prediction (Hyatt et al., 2010; Salamov and Solovyev, 2000). Some advanced tools like GrailEXP are used for the prediction of exons, CpG islands, promoters, genes, polyAs, and repeating elements in the DNA sequence, and GENOMESCAN predicts the location and exon/intron position in the genes of various organisms (Lambert et al., 2003). Table 8.1 lists some of the commonly used programs.

8.3 Sequence analysis

8.3.1 Nucleotide sequence analysis

Sequence analysis is another important part of the bioinformatics task that plays a vital role in the understanding of various characteristics of biomolecules like DNA, RNA, or proteins. The first-ever DNA-based genomic sequence of insulin proteins was analyzed by Fred Sanger in 1951. The analytical method used in thisstudy is well known as the Sanger method or Sanger sequencing as mentioned earlier. This

Table 8.1 Bioinformatics tools used for gene prediction in various species.

S. no.	Program	Gene prediction available for	References
1.	GeneMark, GenMark.hmm	Human, rat, mouse, chicken, *Drosophila*, *C. elegans*, *Arabidopsis*, rice, yeast, many archaea and bacteria	Besemer et al. (2001), Borodovsky and Mclninch (1993), Lukashin and Borodovsky (1998), Shmatkov et al. (1999)
2.	Glimmer, GlimmerM	Many archaea and bacteria, *Aspergillus*, *Plasmodium*, rice, *Arabidopsis*	Delcher et al. (1999), Salzberg et al. (1998)
3.	Grail, GrailEXP	Human, mouse, *Arabidopsis, Drosophila, Escherichia coli*	Uberbacher et al. (1996)
4.	GenScan	Human, maize, *Arabidopsis*	Burge and Karlin (1997), Burge and Karlin (1998)
5.	GeneBuilder	Human, rat, mouse, *Drosophila*, fugu, *C. elegans*, *Aspergillus, Arabidopsis*	Milanesi et al. (1999)
6.	Genie	Human	Reese et al. (2000)
7.	GeneID	Human, *Drosophila*	Parra et al. (2000)
8.	GeneFinder, Fgenes, Fgenesh	Human, yeast, *C. elegans*, *Arabidopsis*, *Drosophila*	Salamov and Solovyev (2000)
9.	HMMgene	Human, *Arabidopsis*, *C. elegans*	Krogh (2000)
10.	GeneFinder, MZEF	Human, mouse, *Schizosaccharomyces pombe*, *Arabidopsis*	Zhang (1997)

method was one of the biggest breakthroughs for the sequencing of long strand DNA, and later this method was used in the famous Human Genome Project (Sanger et al., 1977). However, Michael Levitt claimed that the first sequence analysis began between 1969 and 1977 (Levitt, 2001). It was believed that Robert Holley's group from Cornell University was the first group to sequence an RNA molecule (Holley et al., 1965). Since that time, many developments have been carried out in this field. In 1977, the complete genomic sequence of bacteriophage was published (Sanger et al., 1977). The bioinformatical analysis of these biomolecules gave an idea of their unique features. In a bioinformatical analysis, first, the sequence of a biomolecule is taken from the databank. After necessary refinement, analytical tools are used to predict the homologous molecules, structure, function, and evolutionary history. So, depending on the nature of analysis, various tools for sequence analysis are available and a few are listed in Table 8.2.

These tools are based on advance statistical and mathematical modeling and are popular for creating databases of genomes and proteomes and their enormous

Table 8.2 Bioinformatics tools used for sequence analysis of protein and nucleic acid.

S. no.	Sequence analyzing tools	Function	References
1.	BLAST	Used for DNA or protein sequence analysis	(Stephen et al., 1990)
2.	HMMER	Used for homologous protein sequences identification	Finn et al. (2011)
3.	Clustal Omega	Used for multiple sequence alignments	Sievers et al. (2011)
4.	Sequerome	Used for sequence profiling	Ganesan et al. (2005)
5.	ProtParam	Used to predict physicochemical properties of proteins	Gasteiger et al. (2005)
6.	JIGSAW	Used to predict the splicing sites in DNA sequences	Allen and Salzberg (2005)
7.	novoSNP	Used to find single nucleotide variation	Weckx et al. (2005)
8.	Virtual Footprint	Used to analyze whole prokaryotic genomes along with promoter regions	Münch et al. (2005)
9.	WebGeSTer	Contains a database of transcription terminator sequences	Unniraman et al. (2002)
10.	Genscan	Used to predict the exon/intron sites	Burge and Karlin (1997)
11.	Softberry Tools	Annotates plant, animal, and bacterial genomes along with function and structure prediction of RNA and proteins	Gangal and Sharma (2005)
12.	ORF Finder	Used to find open reading frame	Rombel et al. (2002)
13.	Prokaryotic promoter prediction	Used to find the promoter sequences in prokaryotes	Kanhere and Bansal (2005)

applications in biological sciences. They help in identifying and analyzing promoter, terminator, and untranslated regions in the genome and also in recognition of exon, intron, open reading frame, and some variable regions that can be used for diagnostic purposes. A bioinformatic study also helps to find similarities and comparisons between two different sequences or variations like point mutation and single nucleotide polymorphism. Until now, millions of sequences of proteins and nucleic acid of different organisms are known. These sequences are divided into protein families and gene families and aligned in two different ways: pairwise sequence alignment and multiple sequence alignment by using the Needleman–Wunsch algorithm and the Smith–Waterman algorithm (Rehm, 2001). Software tools like BLAST (Basic

Local Alignment Search Tool) and ClustalW are used to compare the origin and evolutionary history of species on the basis of genetic and protein databases. There are several tools used by researchers for the graphical view of data like TreeView, Jalview, GeneView, and Genes-Graphs. As mentioned, these tools are mainly based on mathematical modeling and statistical applications like regression analysis, dynamic programming, artificial neural network, HMM, clustering, and sequence mining to study given biological sequences (Mehmood et al., 2014).

8.3.2 Protein sequence analysis

Protein sequences can also be sequenced similar to DNA sequences. Among all the bioinformatics tools available, the more sophisticated tools for protein sequence analysis are motif searching and 3D structural prediction. A motif is a specific sequence in DNA and protein that forms a distinct structure. Motif sequences are patterns of amino acids that have some known function (Bilgen et al., 2004). Several bioinformatics tools contain a collection of motif sequences; one examples is the PROSITE database maintained by the University of Geneva Medical Center (Bairoch et al., 1997). Another example is PFAM, created with the literature data of already available sequence databases for the identification of such sequences (Bateman et al., 2004). Apart from these, many software tools for motif sequences are available, and a few of them are listed in Table 8.3. The databases from various analytical tools help to recognize the specific amino acids with their functional significance even for unknown proteins. The function of a protein is not only decided by its sequence but also depends on its structure. The 3D structure of a protein decides biological activities and thus its functions. One of the major challenges in bioinformatics is to conclude the protein structure by analyzing its sequence.

There have been methods available for the prediction of secondary structure (alpha-helix and beta-sheet) for many years now, for example, Cn3D is an application from NCBI, which was used to predict the 3D structure of the protein. Now, with the advanced version of this application, we can obtain information about protein structure, sequence, and alignment (Wang et al., 2000). Another popular website for protein modeling is EMBL Biocomputing, which provides data on multiple protein modeling (Madeira et al., 2019).

8.3.3 Phylogenetic analyses

The word phylogeny is used for the evolutionary history of any species. Phylogenetic analyses give the evolutionary relationships between the group of organisms. From the molecular sequence database, it was observed that all the species on Earth, including those that are extinct, have arisen from a common ancestor (Khan et al., 2014). Sequence databases provide the data to reveal the role of biomolecules with functions in the evolution of a species. Phylogenetic analysis provides the information to understand the interrelationship between the different species. A better and easier way of understanding the interrelationship between different species is by using a phylogenetic tree or tree of life as shown in Fig. 8.4.

Table 8.3 Bioinformatics tools used for motif finding.

S. no.	Tool name	Description	References
1.	PMS	Motif search	Din and Rajasekaran (2013)
2.	eMOTIF	Extraction of shorter motifs sequence	Huang and Brutlag (2001)
3.	PHI-Blast	Motif alignment tool	Zhang et al. (1998)
4.	PRATT	Used for pattern generation with ScanProsite	Gunduz et al. (2003)
5.	TEIRESIAS	Motif extraction and database	Schwartz and Gygi (2005)
6.	BASALT	Multiple and regular expression motif searches	Redhead and Bailey (2007)
7.	ScanProsite	Motif database tool	Gattiker et al. (2002)
8.	I-sites	Structure motif library	Bystroff et al. (2000)
9.	JCoils	Prediction of leucine zipper and coiled-coil protein structure	Rehman et al. (2017)

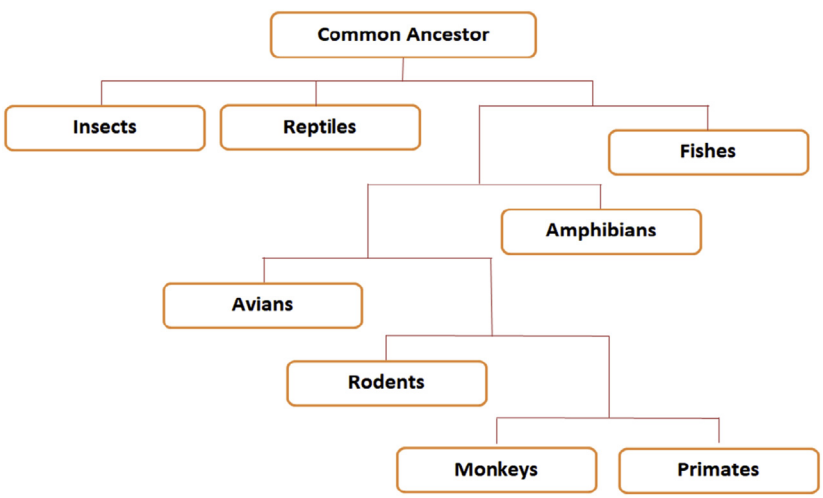


FIGURE 8.4

Interrelation between various species (phylogenetic tree).

Table 8.4 Bioinformatics tools used for phylogenetic analyses.

S. no.	Tools	Function	References
1.	MEGA (Molecular Evolutionary Genetics Analysis)	Used to build a phylogenetic tree to study the evolutionary link between species	Tamura et al. (2007)
2.	MOLPHY	Based on the maximum likelihood method	Adachi and Hasegawa (1992)
3.	PAML	Based on the maximum likelihood method	Yang (2007)
4.	PHYLIP	Used for phylogenetic studies	Retief (2000)
5.	JStree	Database library for viewing and editing the phylogenetic tree	Shank et al. (2018)
6.	TreeView	Used to view the phylogenetic tree	Page (2003)
7.	Jalview	Used to refine alignment	Waterhouse et al. (2009)

The tree is composed of various nodes and branches. The nodes of the tree represent taxonomic units like genes, proteins, species, or populations, whereas branches give the time estimation of the evolutionary relationship. The principle behind the phylogenetic tree is to group the organism on the basis of its genotypic similarity. The organisms that have more genotypic similarities appear much closer in the tree than those having fewer similarities (Freckleton et al., 2002). There are various bioinformatics tools used to draw or analyze various phylogenetic trees, and a few are shown in Table 8.4 (Price et al., 2010).

The following are methods for constructing a tree.

8.3.3.1 Distance-based method

In this method, we calculate the genetic distance between the pairs of sequences to obtain a distance matrix-based tree. The sequences are first aligned using multiple sequence alignment and then the distance of mismatched position is measured to form a distance matrix. This method is very fast and efficient for continuous characters. The two mainly used distance-based methods are the unweighted pair group method with arithmetic mean (UPGMA) and the neighbor-joining (NJ) method. The UPGMA method was originally developed by Sokal and Michener in 1958 for numeric taxonomy. This method uses sequential clustering to form a tree. In this method, the sequences are compared by pairwise alignment to build a distance matrix. The two sequences that have minimum distance are clustered to form a signal and this is repeated with the other sequences to make a new distance. This is the simplest algorithm and thus it is fast, but requires the evolutionary rate to be constant and it works only for ultrametric data (Sokal, 1958).

The other popular distance-based method is NJ. This method was developed by Saitao and Nei in 1987. It works similarly to UPGMA; however, it does not require a

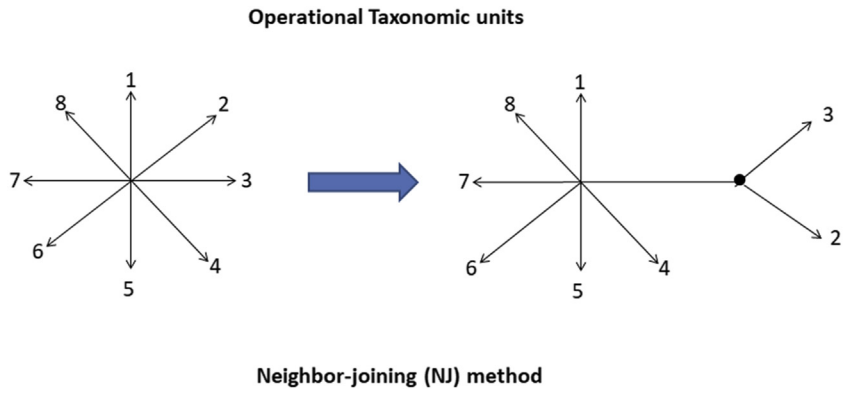


FIGURE 8.5

Phylogenetic tree using Saitou and Nei's neighbor-joining method.

constant evolutionary rate, thus it is used for large sequences as well. In this method, the operational taxonomic units (OTUs) are clustered and formed into a star-like tree. The branches lead to the respective OTUs and radiate from one node and form a star-like pattern. In the next step, another pair of sequences is selected, removed from the star, and joined to a second node, which is connected to the branch, and finally, the distance is calculated as shown in Fig. 8.5. After calculation, the sequences are returned to their respective positions, and then another pair is selected. This minimizes the length of the tree (Gascuel and Steel, 2006).

8.3.3.2 Character-based method

In comparison to the distance-based method, the character-based method is based on utilizing the sequences rather than mismatched distance. Character-based analysis helps in studying the evolutions of the mutations in the sequences. The two commonly known methods are maximum parsimony and maximum likelihood. Maximum parsimony is a simpler and popular criterion. In this method, the phylogenetic tree, which has fewer substitutions to explain evolutionary history, is preferred. Here, all the character changes are considered to be independent of their neighbors (Farris, 1970). On the other hand, maximum likelihood is a statistical-based widely used method. It evaluates data by branch swapping similar to maximum parsimony, but a phylogenetic tree with maximum compound probability is preferred (Chor and Tuller, 2005). This method can be very useful for broad differing datasets. However, this method requires massive computation, which currently limits its usage.

8.4 Sequence database

In the bioinformatics field, a sequence database refers to a biological database comprising a vast collection of information regarding different biological molecule

sequences like nucleic acid sequences, protein sequences, and polymer sequences, which could be classified via a unique key. The information gathered via a sequence database is of great importance for future use, and along with this, it also aids as a primary sequence analyzing tool. With the new developments made in sequence-determining techniques, sequence determination could be acquired even at a single genome level, which has become a huge source of data generation nowadays. Numerous databases are developed all around the world to make collection and submission of sequence data freely accessible to researchers. Every database works as a sovereign depiction of life at a molecular level. Knowledge of these databases will facilitate retrieving information from them a once-only requirement.

Sequencing databases portray a significant role in examining the data of biological organisms. So, they have been categorized into three types, namely primary, secondary, and composite databases based on the information they possess. The primary database consists of data that are attained via experimentation, e.g., through X-ray diffraction and nuclear magnetic resonance (NMR) techniques relevant to a structure or sequence. Examples of primary databases are GenBank (Benson et al., 2008), DDBJ (Miyazaki et al., 2003), Universal Protein (UniProt) sequence database (UniProt, 2014), PIR (Wu et al., 2003), Swiss-Prot (Boeckmann et al., 2003), EMBL (Stoesser et al., 2001), and Protein Data Bank (PDB) (Berman et al., 2000). In contrast, a secondary database consists of information that is derived from the data originated from the examinations and studies done and stored in primary databases, which comprise active sites residues, conserved sequences, and conserved protein secondary motifs (Finn et al., 2014; Gonzalez et al., 2014). Examples of secondary databases are Structural Classification of Proteins (SCOP) database (Fox et al., 2014), PROSITE (Sigrist et al., 2012), Class, Architecture, Topology, and Homology (CATH) database (Pearl et al., 2005), and eMOTIF (Huang and Brutlag, 2001). On a comparative note, primary databases are considered classical databases, whereas secondary databases are considered a more organized form of database. The composite database consists of an array of primary databases, which eradicates the requirement for searching each database separately. Data structures and search algorithms differ from the composite database used. Examples of composite databases are the International Nucleotide Sequence Database (INSD), which is basically a compilation of nucleic acid sequences from GenBank, EMBL, and DDBJ. The nonredundant database is also an example of a composite database, consisting of information from PIR, Swiss-Prot, PRF, GenBank (CDS translations), and PDB. Similarly, UniProt also serves as an example of a composite database representing a collection of sequences obtained from different databases such as Swiss-Prot, PIRPSD (Protein Information Resource Protein Sequence Database), and TrEMBL (Translation from EMBL). A few of the sequence database examples are discussed in Table 8.5.

8.4.1 Genomic database

Bioinformatics, especially genomic informatics, has emerged as a scientific tool of great significance during the postgenomic period with the advancements made in the

Table 8.5 Examples of commonly used sequence databases.

S. no.	Database	About	References
1.	GenBank	Member of International Nucleotide Sequence Database (INSD) and provides an annotated collection of all freely available information regarding nucleotide sequences and their protein translations	Benson et al. (2012a,b)
2.	DNA Data Bank of Japan	Member of INSD and a huge resource for nucleotide sequences	Miyazaki et al. (2003)
3.	Rfam	Contains information on a group of RNA families, depicted by multiple sequence alignments	Burge et al. (2013)
4.	European Nucleotide Archive	Offers free and unrestricted information about annotated DNA and RNA sequences. Also stores the collection of experimental data and sequencing projects metadata.	Amid et al. (2012)

human genomic field. Developments in the field of technology have made it possible to determine a vast number of alterations at the genomic scale in human genes, varying from small mutations to broad rearrangements. With time, it has become clear that an understanding and arrangement of these variations in structure archives would be of huge significance for both diagnosis purposes and the whole scientific community.

The genomic database simply refers to the collection of information regarding genomic mutations or alterations that are available online, represented for a position-specific, general gene, or particular individual or group of people.

8.4.1.1 Advantages of the genomic database are

1.It provides aid in diagnostics at the DNA level to interpret an optimum method for detection of the mutation.
2.It facilitates information regarding particular alterations or mutation-phenotypic motifs.
3.It compares position-specific alteration information with the genomic data already available, such as gene structures, mutation hotspots, recombination frequencies, repeating units, conserved species, and much more (Mehmood et al., 2014).

8.4.1.2 GenBank

GenBank, which comes under the NCBI, is a large compilation of genomic sequences comprising about 250,000 species (Benson et al., 2008). GenBank data can be retrieved via NCBI's integrated system, Entrez, whereas information can be gathered using PubMed (Benson et al., 2012a,b). Enormous amounts of information can be obtained through each genomic sequence regarding the bibliography, organism, literature, and other diverse characteristics. This information contains

promoters, exons, introns, coding regions, translations, untranslated regions, terminators, and repetitive regions. Individual laboratories along with huge genomic sequencing projects contribute to generating sequence data to be collected and stored in GenBank. Similarly, Xenbase is an example of a genomic database that keeps biological and genomic information related to frogs as well as *Xenopus tropicalis* and *Xenopus laevis* (Bowes et al., 2010). In this, *Xenopus* spp. could be considered a model that facilitates new understanding about the advancements made in biology, which can be used for modeling and simulation studies of various human disorders.

8.4.1.3 SGD

Another example of a genomic database is the Saccharomyces Genome Database (SGD), which consists of complete information regarding yeast (*Saccharomyces cerevisiae*) and also facilitates bioinformatics tools to examine its data. This database can be employed to analyze practical relationships between gene sequences and products in the case of fungi and eukaryotes (http://www.yeastgenome.org/). Other than this, genome databases such as FlyBase"facilitate the accessing of information related to genomes and genes of *Drosophila melanogaster* accompanied by search options for alleles, gene sequences, different phenotypes, genetic aberrations, and illustrations of the *Drosophila* class (St. Pierre et al., 2014).

8.4.1.4 Other genomic databases

Other databases such as WormBase and wFleaBase also facilitate accessing data related to genes and genomes. WormBase (http://www. wormbase.org) facilitates accurate, updated, and accessible information regarding *C. elegans* molecular biology, along with that of roundworms. On the other hand, wFleaBase (http://wfleabase.org/) facilitates data accession for the class of genus *Daphnia* (i.e., water flea) where *Daphnia* is treated as a classical system to analyze and gather information on complicated interactions. This information is vital for understanding gene expression, genome structure, individual fitness, and population-level reactions to environmental transformations and chemical pollution. Although wFleaBase consists of a large amount of information on almost all classes of the genus, the fundamental classes are *Daphnia magna* and *Daphnia pulex*. Several other examples of genomic databases are mentioned in Table 8.6.

8.4.2 Protein sequence databases

The examination of proteins on an extensive scale has provided plenty of data because of the information offered via the various genomic projects and the discovery of an advanced range of technologies in science. These advanced technologies have facilitated the straightforward recognition and nature of posttranslational alterations in protein (Sickmann et al., 2003). They have also made it easier to determine a broad range of proteins and outline their interactions to specify their cellular positions (Huh et al., 2003), along with their biological activity interpretation.

Table 8.6 Examples of commonly used genomic databases.

S. no.	Database	About	References
1.	Ensembl	Facilitates integrated genomic information for researchers for genome study	Kersey et al. (2018)
2.	ENCODE	Determines the human genome functional elements	Sloan et al. (2016)
3.	GWAS Central	Free access database providing data for genetic association studies in the case of humans	Beck et al. (2014)
4.	NCBI Genome	Contains information on extensive genomic projects, genome sequences, assemblies, and mapped annotations like alterations, markers, and data obtained through epigenomics analysis	(Sayers et al., 2021)
5.	IGSR (International Genome Sample Resource)	Biggest free access compilation of data for human alterations	1000 Genomes Project Consortium (2015)
6.	DGV (Database of Genomic Variants)	Gives collective information regarding structural alterations in the human genome	MacDonald et al. (2014)
7.	H-InvDB (H-Invitational Database)	Human transcripts and genes unified database	Yamasaki et al. (2010)
8.	GMOD Project	Provides free access for managing, visualizing, storing, and annotating biological data	Stein et al. (2002)

Hence, protein sequence databases portray a significant part of being the hub of data storage and making it accessible to the scientific community. Table 8.7 contains few of the commonly used examples of protein sequence databases.

8.4.2.1 Types of protein sequence databases

Protein databases can be distinguished by having a better understanding and being able to analyze the nature of the data contained in them. Proteins of almost every species existing are available in universal protein databases, while information related to a specified protein family, organism, or certain type of protein is made available by some specialized databases. Universal protein sequence databases can again be divided into sequence archives, which work as storehouses for data consisting of no additional information, and efficiently curated databases, which consist of additional information along with preexisting data (Apweiler et al. 2004).

Table 8.7 Examples of commonly used protein sequence databases.

S. no.	Database	About	References
1.	Swiss-Prot	Curated protein sequence database that provides manually annotated sequences	Boeckmann et al. (2003)
2.	Proteomics Identifications Database	Free access data archive, provides nonredundant data related to functional characterization and posttranslation alterations	(Vizcaíno et al., 2010)
3.	PROSITE	Contains information depicting protein families, domains, and active sites along with their amino acid patterns and profiles	Sigrist et al. (2012)
4.	UniProt	Biggest collection of information regarding protein sequences	UniProt (2008)
5.	Pfam	Contains protein families, their annotations, and multiple sequence arrangements	Finn et al. (2014)
6.	InterPro	Characterizes protein families, active sites, and sustained domains	Quevillon et al. (2005)
7.	Protein Data Bank	A major source of protein information of experimentally determined structures of proteins, nucleic acids, and other composite structures	Berman et al. (2000)

8.4.2.2 Protein sequence archives

A few of the protein sequence databases work as protein sequence archives. Such databases only provide sequence history, details, or reports offering no extra related information and do not even make any attempt to offer necessary information on the sequences. Examples of such databases are GenPept, NCBI's Entrez Protein, and RefSeq. GenPept (GenBank Gene Products Data Bank) is one of the simplest kinds of databases available, which comes under NCBI (Wheeler et al., 2003). The data contained within it are obtained via sequence translations available in the nucleotide databases, managed altogether by the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (Stoesser et al., 2003), the DDBJ, and GenBank, and consist of very few annotations. Another example of such databases is NCBI's Entrez Protein, which consists of a collection of sequence data adapted from already available nucleotide sequences of databases such as EMBL/DDBJ/GenBank and also sequences available in PIR, PDB (Westbrook et al., 2003), Swiss-Prot, and RefSeq (Pruitt et al., 2003). Unlike GenPept, the NCBI's Entrez Protein database also provides supplementary information derived through curated databases like PIR and Swiss-Prot. Other than these examples, the RefSeq database is also a protein sequence archive that is maintained by NCBI. The main objective of this database is to facilitate the necessary compilation of protein- and

nucleotide-linked sequences, provide data confirmation and format regularity, extend distinct series, provide updated information about sequence data and biology, and provide up-to-date curations carried out via NCBI itself and its coworkers (Apweiler et al., 2004).

8.4.2.3 Universal curated database

The other type of protein sequence database is a universal curated database, which contains additional information along with sequences. PIRPSD is one of the most used universal curated databases. This database keeps a collection of extensive, necessary protein sequence data, managed by family and superfamily and annotated through structural, functional, genetic, and bibliographic data. It provides not only sequence data, but also the name and categorization of protein, name of the organism containing the protein naturally, fundamental literature, natural characteristics accompanied with their function's references, and active sites of the sequence. Many times, the database is cross-cited with GenBank nucleic acid/EMBL/DDBJ and protein identifier, MEDLINE IDs and PubMed, and also other database sources.

8.4.2.3.1 Swiss-Prot

Swiss-Prot is an extensively used universal curated database that contains necessary information compiled altogether in one place, and has great unified accordance with other databases (Gasteiger et al., 2001). In this database, the annotation contains information related to protein functions, active sites, domains, posttranslational alterations, resemblance with other proteins, secondary and quaternary protein structures, protein deficiency-related diseases, protein-expressing levels, tissues containing a particular protein, protein corresponding pathways, and competition and modifications in a sequence.

8.4.2.3.2 TrEMBL

Another example of such databases is TrEMBL, which offers sequence searching much faster than Swiss-Prot, because Swiss-Prot requires more manual effort, making it time and labor consuming, which in turn limits the rate of progress of the database. This is because the number of newly acquired sequences is larger than that of the expertly annotated physical sequences entered in the database. It also contains information through computer annotations obtained by translations of all coding sequences present in other databases such as EMBL/DDBJ/GenBank nucleotide sequence database, which were not covered by Swiss-Prot until now.

8.4.2.3.3 UniProt

One of the most significant protein sequence databases is UniProt, which contains an extensive collection of information related to protein sequences, and also grants free and open access to its data. The data contained by PIR, Swiss-Prot, and TrEMBL collectively form UniProt, facilitating all detailed information related to a protein from its sequence to its functions all in one database (UniProt, 2010).

8.5 Structure prediction

Structural biology is the field that concerns how macromolecules, i.e., proteins, RNA, or DNA, acquire their native 3D structure and how their functions change with any change in their native states. Since the number of proteins is vast, the determination of their 3D structures experimentally is a tedious job. The unavailability of information about the structure of proteins can be unsatisfying for researchers. This high demand for knowledge of protein structures was somewhat fulfilled by computational modeling. Protein structure prediction is based on the information provided by already solved protein structures. This gave rise to a new approach, "structural genomics," also referred to as high-throughput structural biology. Structural genomics talks about characterizing the 3D structure of every protein of a given genome rather than focusing on a single protein (Griffiths et al., 2000). With a plethora of information available about a large number of sequenced genomes, structure prediction has become easy with the combination of experimental, bioinformatics, and modeling approaches. Characterization of genomes at the structural level can help in deducing generalizations about the structural organization of genomes (Elsliger and Wilson, 2013). The high-throughput techniques are used to elucidate protein structures of genomes on a large scale. To understand the function of a particular gene or genome, it is important to understand the composition and locus of the gene. This large-scale elucidation can help in cloning and manipulating genes; it also provides insight into potential drug targets for therapeutical purposes. Structural genomics begins with the assignment of genes and markers to specific chromosomes followed by high-resolution chromosome mapping of these genes and markers, and finally physical mapping of genomes and genome sequencing. These genome maps can be used for various genetic analyses, i.e., gene isolation (disease-related genes) and functional genomics.

Predicting the structure and function of a protein using bioinformatics techniques has become a major topic in this field. There are thousands of 3D structures of proteins that are predicted and submitted to protein databases every month. Information on the structure of a protein correlated to the function of that protein as the biological activity is dependent on how a protein folds into a 3D structure. Since the last century, protein just like DNA has been a very complicated macromolecule. It is built from linear sequences of amino acids, specified by nucleotide codons, which ultimately end up in different spatial shapes and structures depending on intramolecular interaction and thus exhibit different biological activities in the biological system. X-ray crystallography and NMR are the two most used experimental techniques to determine the 3D structure of a protein. But their limitations, like cost and time consumption, along with other technical difficulties (sample preparation) make their productivity poor compared to the number of protein sequences submitted in bioinformatics databases. The last two decades have seen exponential growth in the submission of protein sequences. The ongoing advancement in computational methods bridges the gap between the submitted protein sequences and their 3D structures. UniProt and PDB are two major databases that contain millions of

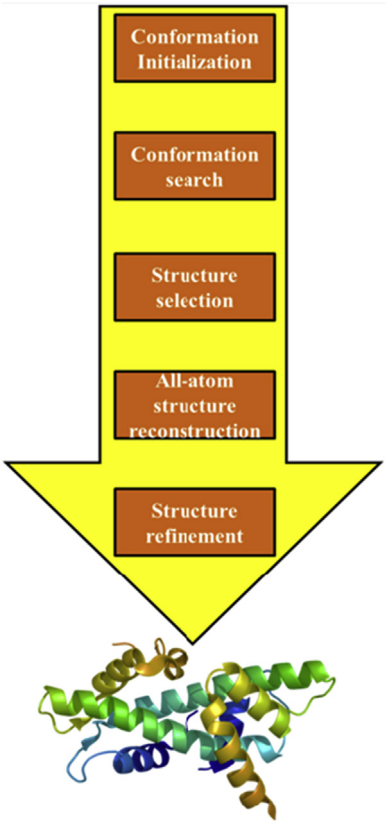


FIGURE 8.6

Fundamental steps of protein modeling.

sequences (Consortium, 2015). Structure prediction is based on "Anfinsen dogma," which says that the native structure of a protein is encoded only by the amino acid sequence. This hypothesis was given by Anfinsen in 1973 and formed the basis of protein folding prediction (Anfinsen, 1973). However, determination of the stable conformation of a protein is still very difficult. The fundamental steps of protein modelling are depicted in Fig. 8.6.

There is a variety of easy-to-use protein sequence databases and web servers for protein structure modeling. A few popular examples are shown in Table 8.8.

Several publicly available online programs are also available that are very convenient and easy to use. Some examples of famous publicly available programs are:

1. BLAST
2. FASTA
3. ClustalW
4. SWISS-MODEL

Table 8.8 Databases for protein structure modeling.

S. no.	Database	About	References
1.	Ensembl	Contains 227 annotated genomes	Yates et al. (2020)
2.	GENBANK	Contains annotated nucleotide sequences with their protein translations for more than 300,000 organisms	Benson et al. (2012a,b)
3.	Protein Information Resource	An integrated database with a variety of protein annotation resources	Huang et al. (2007)
4.	UniProtKB	Contains both UniProtKB/Swiss-Prot, and UniProtKB/TrEMBL sequences	(Consortium, 2015)
5.	Protein Data Bank	A major source of protein information of experimentally determined structures that contain more than 144,000 proteins, nucleic acids, and other composite structures	wwPDB consortium (2019)

There are two types of protein modeling:

1. Template-based modeling
2. Template-free modeling

8.5.1 Template-based modeling

This involves the mimicking and refining of the structural framework (as a template) of a known protein structure to build the structure of the unknown protein. 3D structures for the family of a protein can be built if one protein of the family has a known structure. This assumption is based on the fact that for a small change at the sequence level, there are no significant changes in the 3D structures of the proteins (Chothia and Lesk, 1986).

These methods are very advantageous because of their high-quality and cost-effective 3D structures. However, the requirement of a known structure (template) can limit their application.

8.5.2Comparative protein modeling

Comparative protein modeling is a template-based method that predicts the model of a protein by comparing its alignment with proteins of known structures. It may seem difficult to predict the 3D structure because of the presence of a large number of different proteins. But the limited number of possibilities of tertiary motifs helps in solving the problem (Zhang, 2008).

The general sequencing steps of comparative protein modeling are to:

1. Identify or select the template(s) corresponding to the target sequence.
2. Align the target sequence with the template(s).

3. Construct a model.
4. Evaluate the model for errors in prediction.

Comparative protein modeling is further divided into:

8.5.1.1 Homology modeling

Homology modeling usually involves the target protein sequence, which shares sequence similarities with a related homologous protein whose structure has already been experimentally determined. These types of protein sequences are assumed to share notable similarities in their structures. On the evolution of a protein, it has been demonstrated that the protein structures are more conserved than their amino acid sequences in a homologous family (Illergård et al., 2009). Moreover, difficulties arise in the alignment of the protein sequences. However, this method is ideal for similar protein sequences.

8.5.1.2 Protein threading

On the other hand, protein threading does not limit itself to sequence similarity. Rather, it functions based on fold recognition and can provide better results even in low sequence similarity. It searches the database of known or experimentally determined structures for the unknown or target protein sequence. It uses an algorithm with a scoring function to determine the compatibility of the unknown sequence with a particular solved structure. This 3D structure recognition with 1D protein sequences is unique and aspires to give rise to more advanced methods that can scan structures for a large database (Bowie et al., 1991).

8.5.2 Template-free modeling

This modeling technique is used to model protein sequences that do not contain similarity with existing solved structures. This method is based on fragment assembly of the residues, which uses a database of fragments of already determined proteins to examine the space frame available for the target (unknown) protein. The correlation between the protein sequence and the structure adopted by them is one of the fundamental principles for ab initio modeling. On the other hand, the template-based methods that are most used do not provide any information about the fundamental law of protein folding. The requirement of larger computational resources and sophisticated algorithms by template-free methods has limited their success; therefore they have only been used for small proteins (Zhang, 2008). However, their potential for structural genomics is considered to be very high due to their physics-based atom-by-atom simulation approach, which can help in understanding the principles of protein folding (Shaw et al., 2009).

8.5.2.1 Ab initio protein modeling

Ab initio, also known as de novo, protein modeling methods are based on building 3D models of proteins from their primary sequence. It utilizes physical principles rather than using a known homolog of a protein, thus it requires huge computation.

Ab initio modeling methods function by producing structural conformations (decoys) and as the decoys approach the most stable conformation, the free energy decreases. Conformations with lower free energies are picked by ab initio modeling methods. Examples of these methods include programs like Rosetta by David Baker (Rohl et al., 2004), which can be employed for a complete solution to structures ranging from three to nine residues. Another very useful example is QUARK by Yang Zhang, which is an excellent modeling method for a fragment ranging from 1 to 20 residues. Then, there are other examples that range even more, like FRAGFOLD (Jones, 2001), PROFESY (Lee et al., 2004), SCRATCH (Cheng et al., 2005), etc. However, these methods are not capable of solving large protein problems emerging from side-chain amino acids.

8.6 Bioinformatics and drug discovery

Drug discovery and development is a very tedious, challenging, time-consuming but highly rewarding process. Pharmaceutical companies follow traditional pharmacology and chemistry techniques for drug designing that experience various difficulties. The development of a potent optimum drug requires expertise in machine handling, resources, millions of dollars of investment, and lots of time for its market commercialization, and, in fact, after lots of hard work, it may fail various phase trials (Iskar et al., 2012). So, to eliminate such issues, the involvement of pharmacogenomics and bioinformatics has made the process easier and reduced the cost and time. With the increasing demands of a large number of drugs with reduced potent risk, the interest of the pharmaceutical industry in bioinformatics has increased as it is an easier and faster way to analyze the molecule compared to the experimental approach (Ortega et al., 2012). A new and separate computer-based technique, computer-aided drug design, is used for the discovery of novel drugs (Cordeiro and Speck-Planche, 2012). The entire process of drug discovery can be divided into four types: drug target identification, target validation, lead identification, and lead optimization.

8.6.1 Drug target identification

This is a recent approach used to find biologically active molecules. The drug can be a small molecule that can target the protein, receptor, enzyme, or nucleic acid (Vizovisek et al., 2016). The drug is developed in such a manner so that it can target and inhibit the disease site and deliver therapeutic benefits. The target is the main key to the diagnosis as it allows the drug molecule to act on the metabolism and signaling pathway of the infected cells (Yamanishi et al., 2010). Therefore genetic information is required to understand the nucleotide composition and coding of a target protein and here bioinformatics plays a major role. The genomics and proteomics analysis of a disease is used to locate the target, which shows an abnormal change in gene

Table 8.9 Bioinformatics tools for drug discovery.

S. no.	Tools	Function	References
1.	Potential Drug Target Database	Web-accessible database of proteins for drug target identification	Gao et al. (2008)
2.	Drug Bank	Bioinformatics tools that combine chemical drugs data with a drug target identification database	Wishart et al. (2006)
3.	Therapeutic Target Database	Provides information of the known therapeutic nucleic acids and protein targets	Zhu et al. (2010)
4.	Manually Annotated Targets and Drugs Online Resource	A database for exploring drug–target relationships. It gives direct as well as indirect interactions.	Gunther et al. (2007)
5.	TDR Target Database (Tropical Disease Research)	A tool as well as a genomic database that identifies the gene of interest from the disease. It is part of the World Health Organization special program agenda.	Aguero et al. (2008)
6.	TB Drug Target Database	Contains the database of drugs and targeted protein of tuberculosis	Ekins et al. (2011)
7.	ChEMBL	A large-scale database for drug-like bioactive molecules, this tool gives 2D structure analysis and various calculations like Lipinski parameters, logP, binding constants, molecular mass, pharmacokinetics, etc.	Gaulton et al. (2017)
8.	DrugPort	Gives structural data from the Protein Data Bank related to drugs and their target molecules	Paxman and Heras (2017)

regulation (Katara, 2013). This expression analysis helps in distinguishing between the normal cell and an infected cell (Frantzi et al., 2019). There are various bioinformatics software packages available in the market for fast target identifications and a few of them are listed in Table 8.9.

8.6.2 Drug target validation

Bioinformatics provides algorithms and data to explore new drug targets. After the identification of drug targets, the drug molecule and target must show strong interaction (Yamanishi et al., 2010). The extent of interaction of the drug–target complex decides its success. The establishment of such a relation is known as target validation. Target validation is an area of drug development where bioinformatics helps to avoid the failure of the drugs in clinical trials. Once approved through bioinformatics, the next step is to identify a lead compound.

8.6.3 Lead identification and optimization

After validation of the drug targets, the next step is to find a suitable molecule (drug) that can alter the active site of the target. A large number of bioinformatics databases are available for virtual screening of molecules, which can be used to study the binding and inhibition or activation of the protein. High-throughput screening identifies the promising molecules in the process of drug development and becomes a vital part of it. This screening is a very high-tech approach that exhibits selectivity of the compound for the target site; due to this property it is gaining popularity in pharmaceutical industries (Martis et al., 2011).

the lead data process is followed by lead optimization. To procure unique analogs having enhanced efficacy, lead compound synthesis is the main objective of lead optimization. To deduce the target active sites and improve the optimization of the lead, several properties like metabolic stability, selectivity, etc. are contained therein. Lead optimization is accomplished by chemically modifying its hit structure, followed by alterations by making use of structure−activity relationship, i.e., structure−activity analysis accompanied by design, based upon its structure when the structural data related to the target are accessible (Frye, 1999). Along with this, optimization of the lead mainly involves experimental verification and recognition of compounds depending on the already available animal models and tools offered by absorption, distribution, metabolism, excretion, and toxicity both in vitro and in situ, which can be used for target determination and target confirmation (Wishart, 2005). Furthermore, there should be good correspondence between the lead compound and the drugs, and the lead compound should show no intervention with P-glycoprotein or with enzymes such as cytochrome P450 (Reddy et al., 2007).

8.7 Pharmacogenomics

The development of a precise medicine that can show an equal effect on all patients of the same gene pool is the utmost goal for the medical industry. The study of human genetics from the past 20 years has made the understanding of the relationship between human genetics and diseases somewhat easier (van der Wouden et al., 2020). Genomic information is used in various applications, one of which is the emerging field of pharmacogenomics-informed pharmacotherapy. Pharmacogenomics is a new field of pharmacology and genomics for the development of effective and safe doses of drugs. Pharmacogenomics gives an idea of how the genetic variants respond to a particular dose of a drug (Amstutz and Carleton, 2011; Kaur et al., 2013). The drugs available in the market are "one size fits all," but the effect of the drug is not the same on individuals. Understanding of the role of genetic biomarkers linked with various diseases can help pharmaceutical companies in the development of more effective and precise drugs, therefore identification of the gene−drug regulatory network becomes very important (Eichelbaum et al., 2006). This can only

Table 8.10 Pharmacogenomics tools for drug designing and development.

S. no.	Sources	Information	References
1.	PharmGKB	The knowledge base is used to search and gain knowledge of genotype–phenotype relation, drug dosage, gene–drug interactions, diseases, and pathways	Barbarino et al. (2018)
2.	CYP allele nomenclature	Database on the genetic information of major cytochrome P450s	Gaedigk et al. (2018)
3.	FDA genomic marker table	Database of Food and Drug Administration-approved drugs with their pharmacogenomics information	Schuck and Grillo. (2016)
4.	dbSNP home page	The public domain of single nucleotide variation	Sherry et al. (2001)
5.	HapMap project	An international project to develop a haplotype map to find genes that affect human health, disease, and response to various drugs	Thorisson et al. (2005)

be possible by observing the variations of those genetic elements that take part in drug interactions, which is somewhere linked with the pharmacodynamics and pharmacokinetics of the drugs (Katara, 2013; Whittaker, 2003). Various surveys have reported that the leading cause of patient death every year in hospitals is adverse drug reactions and medical errors (Rawlins, 2004; Rodziewicz and Hipskind, 2020). The pharmacogenomics approach helps to prescribe suitable drugs to patients on the basis of their genetic profiles, which reduces the risk and increases the efficacy of the drugs (Prows and Prows, 2004). Bioinformatics tools provide various information to help researchers to gain a better understanding of gene–drug interactions. Information from various available pharmacogenomics tools is partially summarized in Table 8.10.

8.8 Future aspects

Bioinformatics has become a vital tool in the field of biotechnology and biomedical sciences. This makes it possible to test ideas and hypotheses effectively that can help in decision making prior to any expensive experimental implementation. In the past few years, bioinformatics has shown enormous expansion in various fields and has become a very reliable technique for cost-effective and fast analysis of genomics, proteomics, as well as structure prediction, drug designing, and molecular interaction studies. The large online databases and software tools provide more reliable and accurate results. In addition to this, they also solve various complexities of sophisticated biological pathways and interactions and further help in understanding the relationship between species and the evolution of life. As per the future

perspective, there is still a long way to go, as the possibilities are vast in the computational world. Bioinformatics still requires more advancement to understand the deeper knowledge of principles and functions of biomolecules, which can lead to advancements in drug discovery and other therapies. Thus bioinformatics and other such scientific disciplines need to be explored further for the welfare of the human community.

References

Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526 (7571), 68–74.

Adachi, J., Hasegawa, M., 1992. Protml: Maximum Likelihood Inference of Protein Phylogeny. Computer Science Monographs of the Institute of Statistical Mathematics, Tokyo.

Agüero, F., Al-Lazikani, B., Aslett, M., Berriman, M., Buckner, F.S., Campbell, R.K., et al., 2008. Genomic-scale prioritization of drug targets: the TDR Targets database. Nat. Rev. Drug Discov. 7 (11), 900–907.

Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J., 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. 33 (8), 831–838.

Allen, J.E., Salzberg, S.L., 2005. JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21 (18), 3596–3603.

Amid, C., Birney, E., Bower, L., Cerdeño-Tárraga, A., Cheng, Y., Cleland, I., et al., 2012. Major submissions tool developments at the European Nucleotide Archive. Nucleic Acids Res. 40 (D1), D43–D47.

Amstutz, U., Carleton, B.C., 2011. Pharmacogenetic testing: time for clinical practice guidelines. Clin. Pharmacol. Ther. 89 (6), 924–927.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181 (4096), 223–230.

Apweiler, R., Bairoch, A., Wu, C.H., 2004. Protein sequence databases. Curr. Opin. Chem. Biol. 8 (1), 76–80.

Bairoch, A., Bucher, P., Hofmann, K., 1997. The PROSITE database, its status in 1997. Nucleic Acids Res. 25 (1), 217–221.

Barbarino, J.M., Whirl-Carrillo, M., Altman, R.B., Klein, T.E., 2018. PharmGKB: a worldwide resource for pharmacogenomic information. Wiley Interdiscip. Rev.: Syst. Biol. Med. 10 (4), e1417.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., et al., 2004. The Pfam protein families database. Nucleic Acids Res. 32 (Suppl. l_1), D138–D141.

Beck, T., Hastings, R.K., Gollapudi, S., Free, R.C., Brookes, A.J., 2014. GWAS Central: A Comprehensive Resource for the Comparison and Interrogation of Genome-wide Association Studies.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2008. GenBank. Nucleic acids Res. 36, D25–D30.

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., et al., 2012. GenBank. Nucleic acids Res. 40, D48–D53.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2012. GenBank. Nucleic acids Res. 41 (D1), D36–D42.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235−242.

Besemer, J., Lomsadze, A., Borodovsky, M., 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29 (12), 2607−2618.

Bilgen, M., Karaca, M., Onus, A.N., Ince, A.G., 2004. A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. Bioinformatics 20 (18), 3379−3386.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., et al., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31 (1), 365−370.

Borodovsky, M., McIninch, J., 1993. GENMARK: parallel gene recognition for both DNA strands. Comput. Chem. 17 (2), 123−133.

Bowes, J.B., Snyder, K.A., Segerdell, E., Jarabek, C.J., Azam, K., Zorn, A.M., Vize, P.D., 2010. Xenbase: gene expression and improved integration. Nucleic Acids Res. 38 (Suppl. l_1), D607−D612.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., Gomperts, R., Andres, J.L., Raghavachari, K., 1993. ProMod and Swiss-Model: internet-based tools for automated comparative protein modelling. J. Comput. Chem. 4, 187−217.

Brown, T.A., 2002. The human genome. In: Genomes, second ed. Wiley-Liss.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268 (1), 78−94.

Burge, C.B., Karlin, S., 1998. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8 (3), 346−354.

Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., et al., 2013. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 41 (D1), D226−D232.

Bystroff, C., Thorsson, V., Baker, D., 2000. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J. Mol. Biol. 301 (1), 173−190.

Cheng, J., Randall, A.Z., Sweredoski, M.J., Baldi, P., 2005. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. 33 (Suppl. l_2), W72−W76.

Chor, B., Tuller, T., 2005. Maximum likelihood of evolutionary trees: hardness and approximation. Bioinformatics 21 (Suppl. l_1), i97−i106.

Chothia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5 (4), 823−826.

Consortium, U.P., 2008. The universal protein resource (UniProt). Nucleic Acids Res. 36, D190−D195.

Consortium, U.P., 2010. The universal protein resource (UniProt. In in 2010. Nucleic Acids Res. 38 (Database issue). D142−148.

Consortium, U.P., 2014. Activities at the universal protein resource (UniProt). Nucleic Acids Res. 42 (D1), D191−D198.

Consortium, U.P., 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43 (D1), D204−D212.

Cordeiro, M.N., Speck-Planche, A., 2012. Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. Curr. Top. Med. Chem. 12 (24), 2703.

CRICK, F.H., 1958. On protein synthesis. Symp. Soc. Exp. Biol. 12, 138−163.

Dayhoff, M.O., Ledley, R.S., December 1962. Comprotein: a computer program to aid primary protein structure determination. In: Proceedings of the December 4-6, 1962, Fall Joint Computer Conference, pp. 262−274.

Delcher, A.L., Harmon, D., Kasif, S., White, O., Salzberg, S.L., 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27 (23), 4636–4641.

Dinh, H., Rajasekaran, S., 2013. PMS: a panoptic motif search tool. PloS One 8 (12), e80660.

Edman, P., 1949. A method for the determination of the amino acid sequence in peptides. Arch. Biochem. 22, 475–476.

Eichelbaum, M., Ingelman-Sundberg, M., Evans, W.E., 2006. Pharmacogenomics and individualized drug therapy. Annu. Rev. Med. 57, 119–137.

Ekins, S., Freundlich, J.S., Choi, I., Sarker, M., Talcott, C., 2011. Computational databases, pathway and cheminformatics tools for tuberculosis drug discovery. Trends Microbiol. 19 (2), 65–74.

Elsliger, M.A., Wilson, I.A., 2013. Structural Genomics.

Farris, J.S., 1970. Methods for computing wagner trees. Syst. Biol. 19 (1), 83–92.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39 (Suppl. l_2), W29–W37.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., et al., 2014. Pfam: the protein families database. Nucleic Acids Res. 42 (D1), D222–D230.

Fox, N.K., Brenner, S.E., Chandonia, J.M., 2014. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 42 (D1), D304–D309.

Frantzi, M., Latosinska, A., Mischak, H., 2019. Proteomics in drug development: the dawn of a new era? Proteonomics Clin. Appl. 13 (2), 1800087.

Freckleton, R.P., Harvey, P.H., Pagel, M., 2002. Phylogenetic analysis and comparative data: a test and review of evidence. Am. Nat. 160 (6), 712–726.

Frye, S.V., 1999. Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. Chem. Biol. 6 (1), R3–R7.

Gaedigk, A., Ingelman-Sundberg, M., Miller, N.A., Leeder, J.S., Whirl-Carrillo, M., Klein, T.E., PharmVar Steering Committee., 2018. The pharmacogene variation (PharmVar) consortium: incorporation of the human cytochrome P450 (CYP) allele nomenclature database. Clin. Pharmacol. Ther. 103 (3), 399–401.

Ganesan, N., Bennett, N.F., Velauthapillai, M., Pattabiraman, N., Squier, R., Kalyanasundaram, B., 2005. Web-based interface facilitating sequence-to-structure analysis of BLAST alignment reports. Biotechniques 39 (2), 186–188.

Gangal, R., Sharma, P., 2005. Human pol II promoter prediction: time series descriptors and machine learning. Nucleic Acids Res. 33 (4), 1332–1336.

Gao, Z., Li, H., Zhang, H., Liu, X., Kang, L., Luo, X., et al., 2008. PDTD: a web-accessible protein database for drug target identification. BMC Bioinform. 9 (1), 104.

Gascuel, O., Steel, M., 2006. Neighbor-joining revealed. Mol. Biol. Evol. 23 (11), 1997–2000.

Gasteiger, E., Jung, E., Bairoch, A.M., 2001. SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr. Issues Mol. Biol. 3 (3), 47–55.

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M.R., Appel, R.D., Bairoch, A., 2005. Protein identification and analysis tools on the ExPASy server. In: The Proteomics Protocols Handbook. Humana press, pp. 571–607.

Gattiker, A., Gasteiger, E., Bairoch, A.M., 2002. ScanProsite: a reference implementation of a PROSITE scanning tool. Appl. Bioinf. 1 (2), 107–108.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., et al., 2017. The ChEMBL database in 2017. Nucleic Acids Res. 45 (D1), D945–D954.

Gelfand, M.S., 1995. Prediction of function in DNA sequence analysis. J. Comput. Biol. 2 (1), 87–115.
Gonzalez, S., Binato, R., Guida, L., Mencalha, A.L., Abdelhay, E., 2014. Conserved transcription factor binding sites suggest an activator basal promoter and a distal inhibitor in the galanin gene promoter in mouse ES cells. Gene 538 (2), 228–234.
Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., Gelbart, W.M., 2000. An Introduction to Genetic Analysis. WH Freeman, New York, p. 960.
Gunduz, I., Zhao, S., Dalkilic, M.M., Kim, S., 2003. Motif discovery from large number of sequences: a case study with disease resistance genes in arabidopsos thaliana. In: METMBS, pp. 29–34.
Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., et al., 2007. SuperTarget and matador: resources for exploring drug-target relationships. Nucleic Acids Res. 36 (Suppl. l_1), D919–D922.
Hesper, B., Hogeweg, P., 1970. Bioinformatica: een werkconcept. Kameleon 1 (6), 28–29 (Dutch.) Leiden: Leidse Biologen Club).
Holley, R.W., Everett, G.A., Madison, J.T., Zamir, A., 1965. Nucleotide sequences in the yeast alanine transfer ribonucleic acid. J. Biol. Chem. 240 (5), 2122–2128. https://doi.org/10.1016/S0021-9258(18)97435-1.
Huang, J.Y., Brutlag, D.L., 2001. The EMOTIF database. Nucleic Acids Res. 29 (1), 202–204.
Huang, H., Hu, Z.Z., Arighi, C.N., Wu, C.H., 2007. Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. Front. Biosci. 12, 5071–5088.
Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., O'Shea, E.K., 2003. Global analysis of protein localization in budding yeast. Nature 425 (6959), 686–691.
Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 11 (1), 119.
Illergård, K., Ardell, D.H., Elofsson, A., 2009. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins: Struct., Funct. Bioinform. 77 (3), 499–508.
Iskar, M., Zeller, G., Zhao, X.M., van Noort, V., Bork, P., 2012. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. Curr. Opin. Biotechnol. 23 (4), 609–616.
Jones, D.T., 2001. Predicting novel protein folds by using FRAGFOLD. Proteins: Struct., Funct. Bioinform. 45 (S5), 127–132.
JU, B., Lüthy, R., Eisenberg, D., July, 1991. A method to identify protein sequences that fold into a known three-dimensional structure". Science 253 (5016), 164–170.
Kanhere, A., Bansal, M., 2005. A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinform. 6 (1), 1–10.
Katara, P., 2013. Role of bioinformatics and pharmacogenomics in drug discovery and development process. Network Model. Anal. Health Inform. Bioinform. 2 (4), 225–230.
Kaur, H., Grover, S., Kukreti, R., 2013. Concept of pharmacogenomics and future considerations. CNS Neurosci. Therap. 19 (10), 842.
Keller, O., Kollmar, M., Stanke, M., Waack, S., 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27 (6), 757–763.
Kersey, P.J., Allen, J.E., Allot, A., Barba, M., Boddu, S., Bolt, B.J., et al., 2018. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res. 46 (D1), D802–D808.

Khan, F.A.A., Phillips, C.D., Baker, R.J., 2014. Timeframes of speciation, reticulation, and hybridization in the bulldog bat explained through phylogenetic analyses of all genetic transmission elements. Syst. Biol. 63 (1), 96−110.

Krogh, A., 2000. Using database matches with HMMGene for automated gene detection in Drosophila. Genome Res. 10 (4), 523−528.

Lambert, J.C., Testa, E., Cognat, V., Soula, J., Hot, D., Lemoine, Y., et al., 2003. Relevance and limitations of public databases for microarray design: a critical approach to gene predictions. Pharmacogenomics J. 3 (4), 235−241.

Lee, J., Kim, S.Y., Joo, K., Kim, I., Lee, J., 2004. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins: Struct., Funct. Bioinform. 56 (4), 704−714.

Levitt, M., 2001. The birth of computational structural biology. Nat. Struct. Biol. 8 (5), 392−393.

Li, G.-L., Leong, T.-Y., 2005. Feature selection for the prediction of translation initiation sites. Dev. Reprod. Biol. 3 (2), 73−83.

Lukashin, A.V., Borodovsky, M., 1998. GeneMark. hmm: new solutions for gene finding. Nucleic Acids Res. 26 (4), 1107−1115.

MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., Scherer, S.W., 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 42 (D1), D986−D992.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al., 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47 (W1), W636−W641.

Majoros, W.H., Pertea, M., Salzberg, S.L., 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20 (16), 2878−2879.

Martis, E.A., Radhakrishnan, R., Badve, R.R., 2011. High-throughput screening: the hits and leads of drug discovery-an overview. J. Appl. Pharmaceut. Sci. 1 (1), 2−10.

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U. S. A 74 (2), 560−564. https://doi.org/10.1073/pnas.74.2.560.

Mehmood, M.A., Sehar, U., Ahmad, N., 2014. Use of bioinformatics tools in different spheres of life sciences. J. Data Min. Genom. Proteonom. 5 (2), 1.

Milanesi, L., D'Angelo, D., Rogozin, I.B., 1999. GeneBuilder: interactive in silico prediction of gene structure. Bioinformatics 15 (7), 612−621.

Miyazaki, S., Sugawara, H., Gojobori, T., Tateno, Y., 2003. DNA data bank of Japan (DDBJ) in XML. Nucleic Acids Res. 31 (1), 13−16.

Münch, R., Hiller, K., Grote, A., Scheer, M., Klein, J., Schobert, M., Jahn, D., 2005. Virtual footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. Bioinformatics 21 (22), 4187−4189.

Ortega, S.S., Cara, L.C.L., Salvador, M.K., 2012. In silico pharmacology for a multidisciplinary drug discovery process. Drug Metabol. Personal. Ther. 27 (4), 199−207.

Page, R.D., 2003. Visualizing phylogenetic trees using TreeView. Curr. Prot. Bioinform. (1) 6-2.

Parra, G., Blanco, E., Guigó, R., 2000. Geneid in drosophila. Genome Res. 10 (4), 511−515.

Paxman, J.J., Heras, B., 2017. Bioinformatics tools and resources for analyzing protein structures. In: Proteome Bioinformatics. Humana Press, New York, NY, pp. 209−220.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., et al., 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res. 33 (Suppl. l_1), D247−D251.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PloS One 5 (3), e9490.
Prows, C.A., Prows, D.R., 2004. Medication selection by Genotype: how genetics is changing drug prescribing and efficacy. Am. J. Nurs. 104 (5), 60—70.
Pruitt, K.D., Tatusova, T., Maglott, D.R., 2003. NCBI Reference Sequence project: update and current status. Nucleic Acids Res. 31 (1), 34—37.
Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33 (Suppl. l_2), W116—W120.
Rawlins, M.D., 2004. Cutting the cost of drug development? Nat. Rev. Drug Discov. 3 (4), 360—364.
Reddy, R.N., Mutyala, R., Aparoy, P., Reddanna, P., Reddy, M.R., 2007. Computer aided drug design approaches to develop cyclooxygenase based novel anti-inflammatory and anti-cancer drugs. Curr. Pharmaceut. Des. 13 (34), 3505—3517.
Redhead, E., Bailey, T.L., 2007. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC Bioinform. 8 (1), 385.
Reese, M.G., Kulp, D., Tammana, H., Haussler, D., 2000. Genie—gene finding in *Drosophila melanogaster*. Genome Res. 10 (4), 529—538.
Rehm, B., 2001. Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification. Appl. Microbiol. Biotechnol. 57 (5—6), 579—592.
Rehman, A., Abbas, A., Sarwar, M.A., Ferzund, J., 2017. Need and role of scala implementations in bioinformatics. Int. J. Adv. Comput. Sci. Appl. 8 (02).
Retief, J.D., 2000. Phylogenetic analysis using PHYLIP. In: Bioinformatics Methods and Protocols. Humana Press, Totowa, NJ, pp. 243—258.
Rodziewicz, T.L., Hipskind, J.E., 2020. Medical error prevention. In: StatPearls [Internet]. StatPearls Publishing.
Rohl, C.A., Strauss, C.E., Misura, K.M., Baker, D., 2004. Protein structure prediction using Rosetta. In: Methods in Enzymology, vol. 383. Academic Press, pp. 66—93.
Rombel, I.T., Sykes, K.F., Rayner, S., Johnston, S.A., 2002. ORF-FINDER: a vector for high-throughput gene identification. Gene 282 (1—2), 33—41.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406—425.
Salamov, A.A., Solovyev, V.V., 2000. Ab initio gene finding in Drosophila genomic DNA. Genome Res. 10 (4), 516—522.
Salzberg, S.L., Delcher, A.L., Kasif, S., White, O., 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26 (2), 544—548.
Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94 (3), 441—448. https://doi.org/10.1016/0022-2836(75)90213-2.
Sanger, F., Thompson, E.O.P., 1953. The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochem. J. 53 (3), 353—366.
Sanger, F., Thompson, E.O.P., 1953. The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem. J. 53 (3), 366—374.
Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., et al., 1977. Nucleotide sequence of bacteriophage φX174 DNA. Nature 265 (5596), 687—695.
Sayers, E.W., et al., 2021. Database resources of the national center for biotechnology information. Nucleic Acids Res. 49 (D1) https://doi.org/10.1093/nar/gkaa892.

Schuck, R.N., Grillo, J.A., 2016. Pharmacogenomic biomarkers: an FDA perspective on utilization in biological product labeling. AAPS J. 18 (3), 573−577.

Schwartz, D., Gygi, S.P., 2005. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. 23 (11), 1391−1398.

Shank, S.D., Weaver, S., Pond, S.L.K., 2018. phylotree. js-a JavaScript library for application development and interactive data visualization in phylogenetics. BMC Bioinform. 19 (1), 276.

Shaw, D.E., Dror, R.O., Salmon, J.K., Grossman, J.P., Mackenzie, K.M., Bank, J.A., et al., 2009. November). Millisecond-scale molecular dynamics simulations on Anton. In: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, pp. 1−11.

Sherriff, A., Ott, J., 2001. 20 Applications of Neural Networks for Gene Finding.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29 (1), 308−311.

Shmatkov, M., Melikyan, A.,A., Chernousko, A.,L., F., Borodovsky, M., 1999. Finding prokaryotic genes by the 'frame-by-frame'algorithm: targeting gene starts and overlapping genes. Bioinformatics 15 (11), 874−886.

Sickmann, A., Mreyen, M., Meyer, H.E., 2003. Mass spectrometry—a key technology in proteom research. In: Proteomics of Microorganisms. Springer, Berlin, Heidelberg, pp. 141−176.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7 (1), 539.

Sigrist, C.J., De Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., et al., 2012. New and continuing developments at PROSITE. Nucleic Acids Res. 41 (D1), D344−D347.

Silva, G.G., Dutilh, B.E., Matthews, T.D., Elkins, K., Schmieder, R., Dinsdale, E.A., Edwards, R.A., 2013. Combining de novo and reference-guided assembly with scaffold_builder. Source Code Biol. Med. 8 (1), 1−5.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., et al., 2016. ENCODE data at the ENCODE portal. Nucleic Acids Res. 44 (D1), D726−D732.

Sokal, R.R., 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 38, 1409−1438.

St Pierre, S.E., Ponting, L., Stefancsik, R., McQuilton, P., FlyBase, C., 2014. FlyBase 102—advanced approaches to interrogating FlyBase. Nucleic Acids Res. 42 (D1), D780−D788.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., et al., 2002. The generic genome browser: a building block for a model organism system database. Genome Res. 12 (10), 1599−1610.

Stephen, A., Warren, G., Webb, M., Myers, E., David, L., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403−410.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., et al., 2001. The EMBL nucleotide sequence database. Nucleic Acids Res. 29 (1), 17−21.

Stoesser, G., Baker, W., van den Broek, A., Garcia-Pastor, M., Kanz, C., Kulikova, T., et al., 2003. The EMBL nucleotide sequence database: major new developments. Nucleic Acids Res. 31 (1), 17−22.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24 (8), 1596−1599 (a).

Thorisson, G.A., Smith, A.V., Krishnan, L., Stein, L.D., 2005. The international HapMap project web site. Genome Res. 15 (11), 1592–1593.

Uberbacher, E.C., Xu, Y., Mural, R.J., 1996. [16] Discovering and understanding genes in human DNA sequence using GRAIL. In: Methods in Enzymology, vol. 266. Academic Press, pp. 259–281.

Unniraman, S., Prakash, R., Nagaraja, V., 2002. Conserved economics of transcription termination in eubacteria. Nucleic Acids Res. 30 (3), 675–684.

van der Wouden, C.H., Böhringer, S., Cecchin, E., Cheung, K.C., Dávila-Fajardo, C.L., Deneer, V.H., et al., 2020. Generating evidence for precision medicine: considerations made by the Ubiquitous Pharmacogenomics Consortium when designing and operationalizing the PREPARE study. Pharmacogenetics Genom. 30 (6), 131.

Vizcaíno, J.A., Côte, R., Reisinger, F., Barsnes, H., Foster, J.M., Rameseder, J., Martens, L., et al., 2010. The proteomics identifications database: 2010 update. Nucleic Acids Res. 38 (1), D736–D742. https://doi.org/10.1093/nar/gkp964.

Vizovišek, M., Vidmar, R., Fonović, M., Turk, B., 2016. Current trends and challenges in proteomic identification of protease substrates. Biochimie 122, 77–87.

Wang, Y., Geer, L.Y., Chappey, C., Kans, J.A., Bryant, S.H., 2000. Cn3D: sequence and structure views for Entrez. Trends Biochem. Sci. 25 (6), 300–302.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J., 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25 (9), 1189–1191.

Weckx, S., Del-Favero, J., Rademakers, R., Claes, L., Cruts, M., De Jonghe, P., et al., 2005. novoSNP, a novel computational tool for sequence variation discovery. Genome Research 15 (3), 436–442.

Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H.M., 2003. The protein data bank and structural genomics. Nucleic Acids Res. 31 (1), 489–491.

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., et al., 2003. Database resources of the national center for biotechnology. Nucleic Acids Res. 31 (1), 28–33.

Whittaker, P.A., 2003. What is the relevance of bioinformatics to pharmacology? Trends Pharmacol. Sci. 24 (8), 434–439.

Wishart, D.S., 2005. Bioinformatics in drug development and assessment. Drug Metabolism Reviews 37 (2), 279–310.

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., et al., 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34 (Suppl. l_1), D668–D672.

Wu, C.H., Yeh, L.S.L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., et al., 2003. The protein information resource. Nucleic Acids Res. 31 (1), 345–347.

wwPDB consortium, 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 47 (D1), D520–D528. https://doi.org/10.1093/nar/gky949.

Yamanishi, Y., Kotera, M., Kanehisa, M., Goto, S., 2010. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26 (12), i246–i254.

Yamasaki, C., Murakami, K., Takeda, J.I., Sato, Y., Noda, A., Sakate, R., et al., 2010. H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. Nucleic Acids Res. 38 (Suppl. l_1), D626–D632.

Yang, Z., 2007. Paml 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24 (8), 1586−1591.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al., 2020. Ensembl 2020. Nucleic Acids Res. 48 (D1), D682−D688.

Zhang, M.Q., 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc. Natl. Acad. Sci. U. S. A. 94 (2), 565−568.

Zhang, Y., 2008. Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 18 (3), 342−348.

Zhang, Z., Miller, W., Schäffer, A.A., Madden, T.L., Lipman, D.J., Koonin, E.V., Altschul, S.F., 1998. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res. 26 (17), 3986−3990.

Zhu, F., Han, B., Kumar, P., Liu, X., Ma, X., Wei, X., et al., 2010. Update of TTD: therapeutic target database. Nucleic Acids Res. 38 (Suppl. l_1), D787−D791.

CHAPTER

9 Python, a reliable programming language for chemoinformatics and bioinformatics

Jayadev Joshi

Genomic Medicine, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH, United States

9.1 Introduction

Over the past decade, advancements in scientific fields such as life science and medical science have caused an exponential growth in data generation (Mallappallil et al., 2020; Ezer and Whitaker 2019). Various datasets such as medical imaging data, which is derived from a patient's diagnostics reports (Munn and Jordan, 2011; He et al., 2017), genomics data, which is derived from next-generation sequencing during cancer and other genomics studies (Munn and Jordan, 2011; He et al., 2017), or pharmaceutical datasets that provide biochemical properties of small molecules (Hassan et al., 2006) are not only huge but complex as well. Bioinformatics combines biology and computer science to answer questions derived from life science and biomedical science (Ouzounis and Valencia, 2003). The term bioinformatics, coined by Paulien Hogeweg and Ben Hesper (Hogeweg, 2011), dates back to 1970. However, constant development in this field has dramatically changed the definition of bioinformatics in comparison to its original historic meaning (Bayat, 2002). The rapid development of bioinformatics has given birth to separate research areas such as immunoinformatics (Tomar and De, 2014), computational genomics (Tomar and De, 2014), systems biology (Chuang et al., 2010), computational structural biology (Chuang et al., 2010), and chemoinformatics (Hassan et al., 2006). Chemoinformatics is also known as cheminformatics and deals with the data commonly derived from chemical compounds in various forms (3D structures, chemical fingerprints, activity assays, biomolecular interactions, molecular simulations, etc.). The research involved in chemoinformatics is often focused on data retrieval, database creation, pattern recognition, structure−activity relationship modeling, combinatorial chemistry, molecular docking, and toxicity prediction (Hassan et al., 2006; Joshi et al., 2013, 2015). Drug discovery is a highly laborious, time-consuming, and costly process, therefore pharmaceutical and academic settings now rely increasingly on chemoinformatics approaches (Joshi et al., 2015; Hughes et al., 2011). Chemoinformatics is not limited to biomedical and

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00013-0

pharmacology applications but can also be implemented in chemical and allied industries and environmental science, where chemical processes are evident. Researchers deal with a variety of datasets in different fields but the principle behind data analysis never changes; rather, these computational techniques are highly invariable across different disciplines (Petit et al., 2018; Westra et al., 2017). As might be expected in any area that has access to a massive dataset, researchers are fascinated by what current computational approaches can offer by utilizing these datasets. Exploring worthwhile findings by applying various statistical and machine learning-based approaches with current rapidly developing computational infrastructure is the key strategy to uncover hidden information from these complex datasets (Olson et al., 2018). Python is a widely used and extremely popular programming language that has proved to be a game-changer in recent times (Olson et al., 2018). The primary focus of this chapter is to address two important questions: (1) How can Python programming be adopted in bioinformatics and chemoinformatics research? and (2) What are the available resources (Table 9.1)? In this chapter, we have tried to compile steps to gather resources that are essential in adopting Python programming and popular data science techniques in bioinformatics and chemoinformatics research. This chapter is aimed at an audience that is not very familiar with computer programming and has little idea how and where to start chemoinformatics data analysis using Python. Python (Pilgrim and Willison, 2009) is considered a very popular language in data science and in this chapter we have listed some very popular resources that are required to include Python in research. Here, we have provided a thorough introduction to utilizing open-source resources such as Anaconda and pip to install Python and desired packages, as well as apply data analysis methodologies to biological datasets.

Table 9.1 A curated list of software and Python libraries for bioinformatics and chemoinformatics.

Package name	Description	Installation
Biopython	Molecular biology analysis	conda install -c conda-forge biopython
ChemoPy	Chemoinformatics analysis	https://github.com/salotz/chemopy.git
deepchem	Quantum chemistry with deep learning	conda install -c deepchem deepchem
FragBuilder	Peptide fragment builder	conda install -c bioconda fragbuilder
iFeature	Protein and peptide descriptor calculation	pip install iFeature
Matplotlib	Data plotting library	conda install -c conda-forge matplotlib

Table 9.1 A curated list of software and Python libraries for bioinformatics and chemoinformatics.—*cont'd*

Package name	Description	Installation
Modlamp	Peptide-based analysis	conda install -c bioconda modlamp
Pandas	Data analysis	conda install -c anaconda pandas
PyBioMed	Molecular representation of biomolecules	PyBioMed-1.0.zip
PyDpi	Descriptor calculation	conda install -c biocond pydpi
PyQuante	Quantum chemistry library	conda install -c rpmuller pyquante2
Quantiprot	Quantitative analysis of peptide sequence	conda install -c bioconda quantiprot
RDKit	Chemoinformatics packages	conda install -c rdkit rdkit
Scikit-chem	Chemoinformatics analysis	conda install -c richlewis scikit-chem
Scikit-learn	Machine learning modeling	conda install -c anaconda scikit-learn
Seaborn	Data plotting library	conda install -c anaconda seaborn
Software resources		
Anaconda	Package manager	Downloadable installer is available
Pip	Python package manager	Downloadable installer is available
Jupyter Notebook	Interactive environment to run Python code	Using conda and pip
Jupyter lab	Advanced interactive environment to run Python code	Using conda and pip
Virtualenv	Creates Python virtual environment	Using pip and conda
Miniconda	Lighter version of Anaconda	Downloadable installer is available
PyCharm and spyder	Interactive development environment for Python	Downloadable installer is available

9.2 Desired skill sets

In a constantly developing field like data science, bioinformatics, and chemoinformatics, it is always very challenging to summarize a comprehensive skillset (Wilson Sayres et al., 2018). It may vary across the board, but a basic or minimal skill set can

be defined that drives one's interest in the right direction to adopt these interdisciplinary concepts in academia or industrial research settings. The following are a few important skill sets:

1. Understanding of domain-specific data and data formats.
2. Basic data visualization and presentation.
3. Understanding of application programming interface (API, a set of functions and methods enabling various applications to interact programmatically and provide data access to various website databases and services) methods and their application.
4. Understanding of domain-specific databases and API methods to communicate across various database resources.
5. Generic and domain-specific statistical data analysis methods.
6. Good understanding of a programming language like R or Python.
7. Understanding of supervised and unsupervised methods.
8. Understanding of basic matrix algebra, probability theory, statistics, basic calculus, and familiarity with mathematical terminology associated with the listed mathematical concepts.

The listed core skills are set widely and are useful across various areas, including bioinformatics and chemoinformatics. Additionally, it is worth noting that it is not mandatory to gain expertise in all the fields but a little experience makes a big difference. In addition to technical expertise, a curious and enthusiastic mindset is highly desirable.

9.3 Python

The history of programming languages dates back to the 1950s with very popular classical languages like Regional assembly language, Fortran, COBOL, BASIC, etc. However, since the early 1950s, programming languages have evolved significantly. In comparison to other older languages like C, C++, and Java, Python and R are relatively young. However, as described in the TIOBE programming community index (a measure of the popularity of programming languages), the evolution of Python in data science during the last few decades has been remarkable and has made Python a tough competitor to C++ and Java. Python is a general-purpose language and can be used in any aspect of data science. A few important points that make Python very useful in data science are:

1. Python is an open-source language; hence it is easy to include in any project without being concerned with intellectual property rights issues.
2. Python has a fast learning curve with many open-source resources.
3. Python is a widely accepted general-purpose programming language.
4. Various courses in data science have adopted Python as an introductory programming language.

5. Python is widely used to implement unsupervised and supervised machine learning methods and workflows.
6. More than 235,000 Python packages can be downloaded through PyPI (Python package index, the global repository of open-source Python packages) to extend the capabilities of Python.
7. These packages are implemented on the basis of thousands of peer-reviewed algorithms and tested by a huge user community.
8. Codes are highly redistributable and easy to run if combined with a Jupyter Notebook.
9. Addons like Jupyter Notebook, R studio, and PyCharm make these programming languages very useful and reliable.

Python is an interpreted not a statically typed (a language is statically typed if it is desired to define the type of a variable before compilation instead of at run time) language, hence it is sometimes criticized by the community due to its performance when compared to superfast languages like C/C++; in spite of these imitations the user base of Python is growing significantly day by day.

9.4 Python in bioinformatics and chemoinformatics

As described in an earlier section, the capability of Python can be enhanced by thousands of open-source packages that can be easily incorporated in the research to analyze enormous data types. There are hundreds of packages that can be downloaded using Anaconda and PyPi using simple commands, which are described in detail later in this chapter. In Table 9.1, we have listed very popular and widely utilized Python packages in data science, bioinformatics, and chemoinformatics research.

9.5 Use Python interactively

Whether it is biology or astronomy, data analysis and visualization are two crucial components of data science. Analysis sometimes works as a black box where a user is not very familiar with the underlying algorithms; however, the outcome is sufficient to evaluate a complex hypothesis. Data visualization is the graphical projection of information through various visual counterparts such as charts, graphs, and maps. This visual approach provides an interactive way to explore the data and its features such as patterns, trends, and various statistical parameters. An image is worth a thousand words; hence these visualization techniques are important in data-driven decision making. Commonly, there are three different ways one can write and run a Python code: first, writing a program directly on the interactive shell (Fig. 9.2), second, using a text editor or integrated

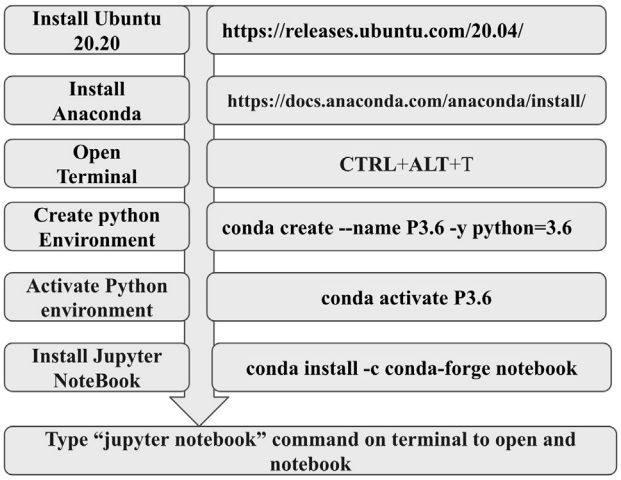


FIGURE 9.1

Flow diagram representing steps to install Python with Jupyter Notebook.

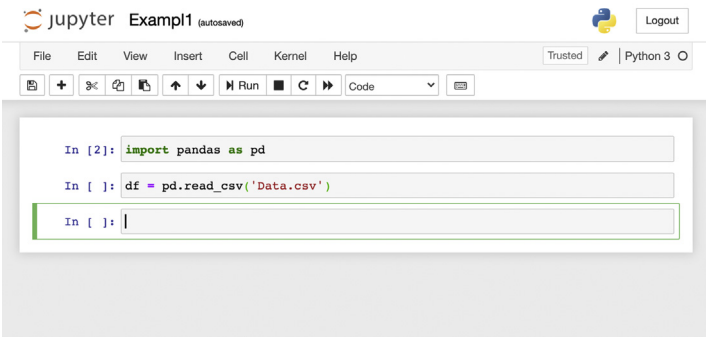


FIGURE 9.2

A standard Python interactive shell.

development environment tools, and third, writing a program via interactive Jupyter Notebook. The least popular approach is writing code directly on the Python interactive shell, which works well with a very basic 5−10 lines of code but is not ideal for a large Python program. The other two methods of writing programming codes are widely adopted approaches in building Python-based software, libraries, or interactive data analysis pipelines. Jupyter Notebook is quite a recent project and was developed by an open-source community to provide a best solution for both analysis and data visualization in one place. All the examples in the upcoming sections are implemented within Jupyter Notebook and example codes are provided with this chapter.

9.6 Prerequisites to working with Python

A flow diagram that explains all the steps to install Python with a Jupyter interactive environment are described in Fig. 9.1.

9.6.1 Linux OS/OSX

Linux is an operating system that is a very popular choice for data science and is highly recommended by the data science community. Several open-source Linux distributions are freely available to download and are easy to install. One of the most popular and widely used Linux distributions is Ubuntu, which has been supported by a large community for almost a decade now. The link https://releases.ubuntu.com/20.04/ can be used to download the latest Ubuntu distribution.

9.6.2 Basic Linux bash commands

Linux command is simply a predefined statement that performs a specific task when entered into a shell (a shell is a user interface tool for access to an operating system's services and is provided with all the operating systems) (Fig. 9.2).

Very frequently used Unix/Linux commands are included in Table 9.2 which may be useful during this chapter.

Table 9.2 Basic Unix/Linux commands.

Type	Command	Function
File and directory	*ls*	Listing directories and files
	ls -al	Formatted listing of hidden files
	cd dir_name	Change directory to "dir_name"
	cd	Return to home directory
	pwd	Show current directory path
	mkdir my_dir	Create a new directory "my_dir"
	rm my_file	Delete a file "my_file"
	rm -rf my_dir	Force remove directory "my_dir"
	cp/path1/file1/path2/file1	Copy file1 to a new path2
	cp -r/path1/dir1/path2/dir1	Copy directory dir1
	mv/path1/file1/path2/file1	Move or rename a file1
	touch my_new_file	Creates a new file
Process management	ps	Display current active process
	top	Display all running processes
	kill pid	Kill a process with a process id
File permission	chmod +x my_script	Change permission of a file
	chmod 777	Permission to read, write, execute for all
Searching	grep pattern files	Search a pattern in files

9.6.3 Anaconda

Anaconda, an open-source package manager for Python and R programming languages for data science, has been widely utilized in machine learning, big-data analysis, predictive modeling, bioinformatics, etc. Anaconda distribution enables easy management of Python and R programming languages and underlying packages.

Installation instructions are as follows:

1. Go to the linkhttps://www.anaconda.com/products/individual#linux.
2. Download the latest version of Anaconda in the download folder "Anaconda3-2020.07-Linux-x86_64.sh".
3. Open the shell terminal (**Ctrl**+**Alt**+**T**) and type this command:

   cd /home/username/Downlaods

4. Type the command to change the permission:

   chmod +x Anaconda3-2020.07-Linux-x86_64.sh

5. Type the command and hit enter:

   ./Anaconda3-2020.07-Linux-x86_64.sh

After hitting enter, just follow the instructions that appear on the terminal screen. It is recommended to accept all the defaults during installation.

9.6.4 Installing Python in the conda environment

The first question that comes to mind is what is a conda environment? In simple language, we can think of a conda environment as a box where one can put desired Python installations and associated packages without affecting the native installation. A conda environment is a directory created on a computer by conda commands that contain a desired collection of Python or R packages for a particular project. Using conda, one can create two separate conda environments and install two different Python versions, for example, 3.6 and 3.8, on one machine without affecting each other's performance.

Type the following command to create a conda environment:

conda create –name P3.6 -y python=3.6

If everything goes as intended we should have a new environment in our conda installation. We can check this by typing a simple command on the terminal "*conda info -e*". If you can see a new environment, "P3.6", on the terminal you are almost ready to follow all the examples provided with this chapter. But do not worry, if you encounter any error, follow the available resources that explain conda in the details.

9.6.5 Jupyter Notebook

In this chapter, all the examples and codes are implemented on Jupyter Notebook, which is a browser-based interactive development environment that facilitates researchers to quickly implement their ideas and analyze the outcomes interactively

at the same time. In addition to this, Jupyter Notebooks can be used to build reproducible data analysis pipelines, and results can be shared easily across the scientific community.

Type the following command to install Jupyter Notebook:

```
conda activate P3.6
conda install -c conda-forge notebook
```

If everything goes as intended you can run Jupyter Notebook by simply typing "*jupyter notebook*" at the command prompt. Jupyter Notebook can be accessed through any browser by typing localhost:8080. Fig. 9.3 describes a standard Jupyter Notebook interface. A few important components are as follows:

File This button can be used to launch a new notebook, save the changes before exit, or download a notebook in various formats.

Cell A cell is a block where you can write and run your code using various options such as run an individual cell or run all cells at once.

Insert Here, you can insert a cell above or below an existing cell.

Kernel This section is important when you need to restart the kernel (a kernel is a program that runs and assesses the code).

The second menu bar contains some important buttons such as add, cut, or copy a new cell to control the behavior of a Jupyter Notebook.

Here, we have provided a compact but simple approach to install all the requirements to start the journey with Python programming language; however, installation of these resources always depends on the operating system and hence several user-defined settings can be made, such as environment variable and installation path. We have selected Ubuntu Linux, which provides several powerful and easy ways of installing resources and does not require complex user-defined settings. For further details, the following are useful weblinks of popular tutorials:

1. Ubuntu installation: https://ubuntu.com/tutorials/install-ubuntu-desktop#1-overview
2. Python installation: https://phoenixnap.com/kb/how-to-install-python-3-ubuntu
3. Advanced conda settings: https://docs.anaconda.com/anaconda-repository/2.23/admin/advanced/

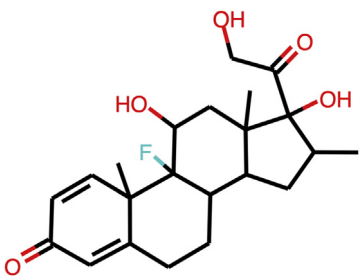


FIGURE 9.3

A standard Jupyter Notebook and the menu bar.

4. Jupyter Notebook installation: https://jupyter.org/install
5. Python virtual environment: https://linuxize.com/post/how-to-create-python-virtual-environments-on-ubuntu-18-04/

9.7 Quick overview of Python components

This section explains the basic concepts and provides a tutorial by summarizing the selected functionalities of the Python programming language. We have not included a detailed tutorial for Python in this chapter; rather, we have focused more on providing a basic understanding of minimal Python utilities that are required to run a simple program and analysis pipeline.

9.7.1 Variable

Let's use an example to understand this. Suppose you have a glass in your home that is empty; this glass has a constant volume that can be filled with water. However, the volume of the water may differ as per your requirement; sometimes the glass is half filled and sometimes fully filled with water, so the glass is a variable that can hold different volumes of water at different times. Similarly, in a programming language, a variable is a reserved memory location to store different types of values. There are several rules that define a variable in Python but the most important is that we will only use "_" special characters to name a variable with the common characters from "a" to "z."

As an example:

```
my_first_variable = 10
my_second_varible = "This is an example"
my_first_variable = 20.5
```

We define two variables: first, we assign an integer value to the first variable, while we assign a string to the second variable. We have overwritten the first variable by assigning a float (a number that has a decimal place) value that replaces the previous integer 10 with 20.5. These variables are very useful and provide different value inputs dynamically during run time.

9.7.2 Operators in Python

Lines in the code are called statements. These statements build upon various expressions: a simple example of an expression is a mathematical expression, for example, 1+1 or a * b. These expressions are defined by operators; in the foregoing example "+" and "*" are operators. Python language supports the following types of operators:

1. Arithmetic operators
2. Relational operators

3. Assignment operators
4. Logical operators
5. Bitwise operators
6. Membership operators
7. Identity operators

Table 9.3 describes some of the selected operators that we will use during this chapter very frequently; however, bitwise and identity operators have been excluded in this chapter.

9.7.3 Control flow and control statements in Python

There are various control flow statements in Python that may appear complex, but are very easy to understand. Control flow statements, such as if, else, assess some condition and control the flow of the code by making a decision. On the other hand, "for" and "while" loops repeat a portion of a code until a certain condition is satisfied. Look at the following example code:

Example 1

```
1 a = 5
2 If a > 4: # statement a > 4 is true
```

Table 9.3 List of commonly used Python operators.

Operator type	Operator	Example
Arithmetic operators	+ Addition − Subtraction * Multiplication / Division % Modulus ** Exponent // Floor division	x + y x − y x * y x / y x % y x ** y = xy 9//2 = 4 and 9.0//2.0 = 4.0, −11//3 = −4, −11.0//3 = −4.0
Comparison operators	== != <> < > <= >=	(x == y) is not true (x != y) is true (x <> y) is true (x > y) is not true (x < y) is true (x >= y) is not true (x <= y) is true
Assignment operators	=	z = x + y assigns the value of x + y into z
Logical operators	and (logical AND) or (logical OR) not (logical NOT)	(x and y) are true (x or y) is true (x and y) are false
Membership operators	in not in	if x is present in the list y = [a,x,n] if x is not present in the list y = [a,b,c]

```
3 print ("Print command executed, 'a' is greater than 4")
Output:
Print command executed, 'a' is greater than 4
```

Example 2

```
1 a = 5
2 If a < 4:
3 print ("Print command executed, 'a' is greater less than 4")
4 else:
5 print("Execute this line, a < 4 is false greater than 5")
Output:
Execute this line, a < 4 is false greater than 5
```

A noticeable point here is the statement after "#", "# statement a > 4 is true", and "#Code never stops but execution passed to the next line". These are the special statements in Python that never contribute any logical operation; rather, they just provide helpful information about the code and are always ignored by the interpreter while executing the program.

In the first example, the second statement "If a > 4" checks if the variable 'a' is greater than 4 or not. We know that this statement is 'true' because "a = 5", therefore the code is executed further and runs the "**print**" command "print (command executed, 'a' is greater than 4)". In the second example, again "a = 5", but the control statement is different, "If a < 4", assessing if variable "a" is less than 4 or not, which we know is incorrect. In this example, a new statement can be observed as "else". When the condition is not satisfied (false) in the second line, technically code should stop, but when the interpreter sees the "else", the code executes further rather than terminating. These examples explain how the "if and else" statement works. The following examples explain the execution of the other two control statements.

Example 3 Multiplying 2 with all the items of a list using for-loop.

```
1 My_list = [2, 4, 6]
2 for x in My_list: # can also be visualized as "for x in [2, 4, 6]:"
3 print(x*2)
Output: 4
  8
  12
```

In the first statement we have created a variable called "My_list"; unlike a normal variable "a" that we defined in examples 1 and 2, lists are used to store multiple items in a single variable and represented by " [ ] ". The second line, "**for x in** My_list", is a for-loop statement that executes a section of code repeatedly until a defined condition is not satisfied. For-loop assigns three values one by one to the variable "x" and the mathematical expression "x*2" performs the multiplication one by one, and prints the results. It is clear that a for-loop makes life easier and minimizes the code by avoiding repetitive lines, but what when we want to stop our for-loop at a certain condition? Example 4 explains similar conditions where the code generated a table of 2 up to 10 values and once the objective is achieved the code stops.

Example 4

```
1 a = 2
2 b = [1,2,3,4,5,6,7,8,9,10,11,12,13,14]
3 for x in b:
4         c = a*x
5         if c > 20:
6            break
7         else:
8            print (c)
```

Output:

```
2
4
6
8
10
12
14
16
18
20
```

Let's dissect this code.

Line 1 Integer 2 assigned to a variable "a", a = 2

Line 2 Integers assigned to a list variable "b";
b = [1,2,3,4,5,6,7,8,9,10,11,12,13,14]

Line 3 Here for-loop "**for** x **in** b:" fetches each value from list "b" one by one and assigns value to a variable "x".

Line 4 Variables "a" and "x" multiply and the result will be saved into a new variable named "c".

Line 5 "**if** c > 20:" is a control statement. Every time "x" receives a new value that overwrites the old value during each cycle, consequently "c" also receives a new value in each cycle.

Line 6 If the "**if** c > 20:" condition is true, the statement **"break",** actually breaks the execution of the code; otherwise, the next statement executes.

Line 7 Every time the condition **"if** c > 20:" is not satisfied, "else" is allowed to execute the last statement "**print** (c)".

Line 8 "**print** (c)" prints the result.

The "while" statement is also somehow similar to the for-loop; it repeatedly executes a block of code as long as a condition is true. The next example explains the execution of the "while" loop statement.

Example 5

```
1 i = 1
2 while i < 5:
3 print(i)
4 i += 1
```

Output:

```
1
2
3
4
```

Let's dissect the code.

Line 1 Value 1 assigned to the variable "i".

Line 2 Unlike for-loop, while loop itself checks the condition and executes a set of statements as long as a condition is true.

Line 3 Prints the updated value of "i", print(i).

Line 4 As the loop progresses with every step, it adds 1 in the variable and at the same time updates the value of "i" by 1, "i += 1". Here, we can see the use of assignment operator "=" in combination with the addition operator "+". We can break this statement as "i = i + 1".

This example provides a fairly good understanding of control flow and control statements. It is recommended to execute these examples using the Jupyter Notebook that is provided with this chapter where you can change values of the variables and operators to get a good understanding of the code.

9.7.4 Python functions

Python functions are simply a chunk of reusable code that can perform similar operations with different inputs. In the previous section, we wrote down a code to produce a table of integer 2. What if we want to produce a table for any given integer without changing the code every time? The solution is a function.

Example 6. Function to write down a table for any given integer.

```
1 def Write_a_table(a):
2 b = [1,2,3,4,5,6,7,8,9,10]
3 for x in b:
4 c = a*x
5 print (c)
```

Calling a function.

```
6 Write_a_table(10) #Call this function with different values
```

Output:

```
10
20
30
40
50
60
70
80
90
100
```

Let's dissect the code.

Line 1 Keyword (predefined words in Python cannot be used as user-defined variables; other examples are print, class, for, while is, etc.) **"def"** is used to define a function named **"Write_a_table"**. Inside the bracket **'**(a)', the variable "a", is a special variable called an argument, which can take inputs during the function call (using a function). A function always ends with ":" and could take no or multiple arguments. At the advanced levels, function arguments can also be defined as a list of dictionaries.

Lines 3, 4, and 5 Similar to previous examples.

Line 6 After defining a function we can use a function by calling "Write_a_table(10)".

9.7.5 Library or a module

Previously, this section described how to reuse code in the form of a function. However, complex and large software does not always depend on function; rather, utilities advance concepts such as "class" and OOPs (object-oriented programming). These advanced programs and software can be distributed in the form of packages, libraries, or modules. We will not discuss this advanced concept here; rather, our focus is to learn how to use libraries and models. More than 100,000 Python packages are available to download and can be integrated into any program script. Additionally, these resources are reusable, redistributable, and built on thousands of advanced available algorithms that enable easy integration inside a Python project. Apart from these external resources, Python already has several prebuilt modules that are provided with every standard Python installation.

Example 1

```
1 cwd = os.getcwd()
```

When this code executes, it returns an expected error.

```
NameError Traceback (most recent call last)
<ipython-input-40-340a96f19465> in <module>
——> 1 cwd = os.getcwd()
NameError: name 'os' is not defined
```

However, "os" is an inbuilt module, but we cannot use it until we "import" it.

```
1 import os
2 cwd = os.getcwd()
3 Print ("cwd")
Output:
  '/home/user'
```

Line 1 Importing the inbuilt module "OS".

Line 2 Call the method "getcwd()" that returns the path to the current working directory and assigns it to a variable cwd.

Line 3 Print the variable value.

Import is an inbuilt function that loads the module "os" to utilize its functionality. Similarly, several inbuilt modules such as sys, time, and global are available in

Python that can be imported directly and utilized. Despite these inbuilt modules, we can download lots of open-source packages provided by the scientific community. In this chapter, we never focus on writing complex functions; rather, we learn how to download, install, import, and utilize these publicly available modules. There are several ways a Python module can be installed. The most preferable method is using conda or through the pip package manager. We prefer conda package managers over pip to install available popular Python libraries.

9.7.6 Indentation

As noticed in the earlier sections, Python code follows a strict structure in terms of margin and indentation. It is important to note that Python is very sensitive to indentation and a single mismatch in indentation level terminates the code execution due to the indentation error. Hence, it is always recommended to verify the indentation before executing the code. A simple example of indentation is when a line in the code after key words like def, if, for, while, etc. always starts with the next indentation level.

```
1 def Write_a_table(a):
2 ….b = [1,2,3,4,5,6,7,8,9,10]
3 ….if a > x:
4…….print(“Something”)
```

Indentation,“….”, in lines 2, 3, and 4 represents the number of spaces that define the indentation level. Execution of the code is always determined by the indentation level. Code execution always starts with the zero indentation level and goes to the second indentation level and so on.

9.7.7 Data structure

The data structure is a huge topic and hence we do not provide details of all the data structures. Usually, a data structure is a way of representing and processing the data in a computer program. In this chapter, we will use a list, dictionary, and data frame, which are the common data structures. A list is defined by square brackets “[ ]” and can hold the string [‘a’, ‘b’], integers [1, 2], and float [5.5, 5.7]. Dictionaries are a bit more advanced than lists, have two components, “keys” and “values” {‘key’: ‘value’}, and can be represented by “{ }”. Keys can be used to access the values of a dictionary. See the following example:

```
1 My_dict = {’fruit’: ‘banana’
2 ‘clothes’: ‘shirt’
3 ‘sports’: ‘cricket’}
4 print (My_dict[’clothes’])
Output:
  shirt
```

Data frame, a 2D data structure that contains columns and rows, is capable of handling very big and complex data files with various file formats such as "csv" and "tsv". "pandas" is a very popular library for handling data frames.

9.8 Bioinformatics and cheminformatics examples

The entire Python functionally and its components is beyond the scope of this chapter. However, the selected topics covered in this chapter concerning the Python programming language are just the tip of the iceberg but sufficient to write simple Python codes to analyze data and generate data analysis plots.

9.8.1 Genomics data handling and analysis

In this section, we utilize a very popular resource, Biopython (Cock et al., 2009), to explore genomics data. The first step is to install the desired modules in the Python environment "P3.6". First, open a terminal (**Ctrl**+**Alt**+**T**) and activate the conda environment by typing this command "*conda activate P3.6.*" The next step will install a Biopython library in the Python environment; the commands are:

```
conda install -c anaconda -y -n P3.6 biopython
```

or

```
pip install biopython
```

However, both commands are capable of installing the desired Python libraries but we recommend conda over pip (pip is a standard package-management system used to install and manage Python-based libraries, modules, and software packages). However, pip can be used if a package is not available in conda. If Biopython installation is successful, it can be imported as follows:

```
1 import Bio
```

This command should import the Biopython module without an error and available classes (classes is an advanced topic that is not covered in this chapter but, in simple words, a class can be imagined as a book and functions are its chapters, while a software package is a library that holds several books) or functions can be assessed using the following command:

```
1 dir(Bio)
```

Output :

['Align', 'AlignIO', 'Alphabet', 'BiopythonDeprecationWarning', 'BiopythonExperimentalWarning', 'BiopythonParserWarning', 'BiopythonWarning', 'Data', 'File', ''GenBank', 'MissingExternalDependencyError', 'MissingPythonDependencyError', 'Nexus', 'Seq', 'SeqFeature', 'SeqIO', 'SeqRecord', 'Sequencing', 'SwissProt', '__builtins__', '__cached__', '__doc__', '__file__', '__loader__', '__name__', '__package__', '__path__', '__spec__', '__version__', '_parent_dir', '_py3k', '_utils', 'Os', 'warnings']

The statement "dir(Bio)" returns a list as described in the output that describes the names of all the available classes and modules in the Biopython library.

Let's explore a bit more.

```
1 from Bio import SeqIO
```

This command imports the underneath functionality; let's try the "dir()" command to explore SeqIO a bit more. This command will return a list of methods associated with "SeqIO".

```
1 dir(SeqIO)
```

Output:

['AbiIO', 'AceIO', 'Alphabet', 'AlphabetEncoder', 'FastaIO', 'GckIO', 'IgIO', 'InsdcIO', 'Interfaces', 'MultipleSeqAlignment', 'NibIO', 'PdbIO', 'PhdIO', 'PirIO', 'QualityIO', 'SeqRecord', 'SeqXmlIO', 'SffIO', 'SnapGeneIO', 'SwissIO', 'TabIO', 'UniprotIO', 'XdnaIO', '_BinaryFormats', '_FormatToIterator', '_FormatToString', '_FormatToWriter', '__builtins__', '__cached__', '__doc__', '__file__', '__loader__', '__name__', '__package__', '__path__', '__spec__', '_dict', '_force_alphabet', '_get_base_alphabet', 'as_handle', 'basestring', 'convert', 'index', 'index_db', 'parse', 'print_function', 'read', 'sys', 'to_dict', 'write']

Furthermore, the build method "help()" is very useful and can be used to explore help related to listed items.

```
1 help(AbiIO)
```

Output:

```
Help on module Bio.SeqIO.AbiIO in Bio.SeqIO:
NAME
    Bio.SeqIO.AbiIO - Bio.SeqIO parser for the ABI format.
DESCRIPTION
    ABI is the format used by Applied Biosystem's sequencing machines to store
    sequencing results.
    For more details on the format specification, visit:
http://www6.appliedbiosystems.com/support/software_community/ABIF_File_
Format.pdf
FUNCTIONS
    AbiIterator(handle, alphabet=None, trim=False)
    Return an iterator for the Abi file format.
DATA
    ambiguous_dna = IUPACAmbiguousDNA()
    tag = {'BufT1': 'Buffer tray heater temperature (degrees C)'}
    unambiguous_dna = IUPACUnambiguousDNA()
FILE
/home/usr/anaconda3/envs/NEWTF/lib/python3.7/site-packages/Bio/SeqIO/
AbiIO.py
```

The foregoing section describes how to install, import, and explore a module and fetch in build help. The following example explains how to write a simple code using Biopython to explore a publicly available database.

Example 1 Function fetch GEO database and extracting records.

```
1 def Extract_records(Entrez_data_ID, Usr_email):
2 from Bio import Entrez
```

```
3 Entrez.email = Usr_email
4 handle = Entrez.esearch(db="gds", term=Entrez_data_ID)
5 record = Entrez.read(handle)
6 handle.close()
7 new_records = record["IdList"]
8 for new_record in new_records:
9 print (new_record)
10 print('total records:', len(new_records))
```

Calling the function with some input

```
11 Extract_recrods("GSE16", 'MyEmail@gmail.com')
```

Output:

```
200000016
100000028
…
300000801
total records: 20
```

Let's dissect this code.

Line 1 Defined a function "**Extract_records"** with two arguments (variable) 'Entrez_data_ID', 'Usr_email'.

Line 2 Imported the 'Entrez' module from "**from** Bio" as "**from** Bio **import** Entrez".

Line 3 Entreze requires user email ID to fetch the data from databases; arguments pass the provided email ID to the Entreze.

Line 4 "Entrez.esearch()" is a function of Biopython that connects to the remote database by providing information such as "db="gds"", "term=Entrez_data_ID", and performs a search to explore the store records.

Line 5 "Entrez.read()" consumes "handle" object and reads the fetched records.

Line 6 Once the reading is completed, a line "handle.close()" closes the connection.

Line 7 The previous line reads the records; records were stored in the form of a dictionary, and dictionary key "IdList" is used to fetch all the associated records.

Line 8 "for-loop" loops over the list that contains records.

Line 9 Printing the records one by one.

Line 10 Inbuilt function "len()" was used to calculate the total number of items in the list 'new_records'.

Line 11 We called the function "Extract_recrods("GSE16", 'MyEmail@gmail.-com')" with the input and it produced some output.

The following examples demonstrate other functions of the Biopython library. Fasta files are the text files that contain biological sequences such as genomic or protein sequences. The example explains how to use a Biopython function to read a fasta file and extract information.

Example 2 Reading a fasta file.

```
1 def ExtractFastaRecords(FastaFile, record):
2 from Bio import SeqIO
```

```
3 Seqs = SeqIO.parse(FastaFile, "fasta")
4 for seq_record in Seqs:
5 if record == 'id':
6 print(seq_record.id)
7 elif record == 'seq':
8 print(repr(seq_record.seq))
9 else:
10 pass
11 print('Enter correct record type')
12 ExtractFastaRecords("Example.fasta", 'seq')
Output:
Seq('CGTAACAAGGTTTCCGTGATCATTGATGAGACCGTGG…CGC',
SingleLetterAlphabet())
Seq('CGTAACAACTGCGGAAGGATCATTGTTGAGACAACAG…GGC',
SingleLetterAlphabet())
…
Seq('CATTGTTGAGATCACATAATAATTGATCGAGTTAATC…GCC',
SingleLetterAlphabet())
13 ExtractFastaRecords("Example.fasta", 'id')
Output:
gi|2765658|emb|Z78533.1|CIZ78533
gi|2765657|emb|Z78532.1|CCZ78532
…
gi|2765564|emb|Z78439.1|PBZ78439
14 ExtractFastaRecords("Example.fasta", 'PID')
Output:
Enter the correct record type
```

This code demonstrates how to read a fasta file, how to create a function, looping over some list items, and how to control the behavior of a program by checking certain conditions.

Let's dissect the code.

Line 1 Constructs a function **'ExtractFastaRecords'** with two arguments.

Line 2 Imports "SeqIO".

Line 3 Parse (extracting desired information from a data file) the input fasta file using the parse method "SeqIO.parse(FastaFile, "fasta")" into a variable called "Seqs".

Line 4 Looping over sequence record list "Seqs".

Line 5 Checking if the record type is "id".

Line 6 If conditions satisfy, print the record id.

Line 7 Checking if the record type is "seq".

Line 8 If conditions satisfy, print the seq.

Line 9 If both conditions do not satisfy "**else**" is executed, and execution is transferred to the next line.

Line 10 "pass" statements execute and continue the execution without interrupting the code.

Line 11 If both conditions do not satisfy execution, exit the loop and reach the first indentation level where it prints the error or a suggestion message "Enter correct record type".

Line 12 Demonstrate the function call.

9.8.2 Chemoinformatics data handling and analysis

Several recent studies have applied modern techniques like machine learning and artificial intelligence to derive new inhibitors and drug molecules to fight disease. Chemoinformatics is becoming a popular approach in drug discovery, toxicology, and environmental chemistry to find answers to complex questions. We will now explore some useful examples of Python in chemoinformatics. Programming languages are useful when a database needs to be explored dynamically by utilizing an API. The next example will explain how to fetch information directly from the ChEMBL database through a Python program. The first step is to install a Python library "chembl_webresource_client", enabling communication with the ChEMBL (Gaulton et al., 2011) database and rdkit to play with chemical structures and small molecules.

```
1 conda install -c chembl -y chembl_webresource_client
2 conda install -c rdkit -y rdkit
```

We can utilize 'dir()' and 'help()' to explore the features of these modules as explained in the previous section.

In the following example we demonstrated how to utilize these popular libraries to explore information and data related to dexamethasone, a well-known steroidal anti-inflammatory drug, from ChEMBL.

Example 1 Exploring dexamethasone, an antiinflammatory drug.

```
1 from rdkit import Chem
2 from chembl_webresource_client.new_client import new_client
3 Chem_1_smiles = "CC1CC2C3CCC4=CC(=O)C=CC4(C3(C(CC2(C1(C(=O)
  CO)O)C)O)F)C"
4 Chem_1_mol = Chem.MolFromSmiles(Chem_1_smiles)
5 Chem_1_mol
```

Output:

Let's dissect the code.

Lines 1 and 2 Libraries being imported.

Line 3 A smile code (a 1D string that represents chemical structure) of dexamethasone is passed in a variable "Chem_1_smiles".

Line 4 Function "MolFromSmiles" converts smiles to "Mol" format that holds 2D and 3D structural information of a chemical structure.

Line 5 Returns the 2D image of the dexamethasone structure (Fig. 9.4).

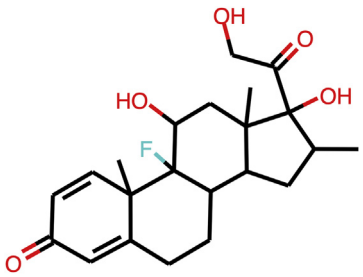


FIGURE 9.4

2D representation of dexamethasone structure.

This example explains how to convert the smile code of dexamethasone into a 2D or 3D structure that can be utilized for further analysis. The following example demonstrates how to fetch the ChEMBL database using "chembl_webresource_client" library.

Example 2 Fetch data related to dexamethasone from the ChEMBL database.

```
1 from chembl_webresource_client.new_client import new_client
2 import pandas as pd
3 molecule = new_client.molecule
4 res = molecule.search('Dexamethasone')
5 df = pd.DataFrame(res)
6 Column_names = df.columns.tolist()
7 print("total Columns:",len(Column_names))
8 for c in Column_names:
9 print(c)
10 df.to_csv('DexaData.csv')
Output:
total Columns: 39
atc_classifications
availability_type
…
Withdrawn_year
Output:
Stores the file in your current directory.
Let's discuss the code.
```

Lines 1 and 2 Importing the Python libraries "chembl_webresource_client.new_client" and "pandas".

Line 3 Creating a molecule instance.

Line 4 Searching the term "Dexamethasone" on the database.

Line 5 Saving the data as a pandas data frame in memory.

Line 6 Accessing individual columns.

Line 7 Counting total column.

Lines 8 and 9 Looping over column names.

Line 10 Saving data on the disc as DexaData.csv file.

The next step is calculating descriptors of a chemical structure (chemical or physiological properties of a chemical compound) that can be utilized in various chemoinformatics methodologies such as prediction of bioactivity, toxicity, and drug efficacy or building a quantitative structure−activity relationship model. First, all the available descriptions can be explored via the given code as follows:

```
1 from rdkit.Chem import Descriptors
2 Des_list = dir(Descriptors)
3 print (Des_list[0:10])
Output:
['BalabanJ', 'BertzCT', 'Chem', 'Chi0', 'Chi0n', 'Chi0v', 'Chi1', 'Chi1n', 'Chi1v', 'Chi2n']
```

Let's dissect the code.

Line 1 Descriptor module of rdkit library being imported.

Line 2 dir(Descriptors) returns all the underneath methods.

Line 3 Prints the top 10 items of the list Des_list[0:10] by slicing the list.

This code explains how to explore the available descriptor to utilize it in the next example.

Example 3 Calculating descriptors for dexamethasone.

```
1 from rdkit import Chem
2 from rdkit.Chem.Descriptors import *
3 compound_1_smiles = "CC1CC2C3CCC4=CC(=O)C=CC4(C3(C(CC2(C1(C(=O)CO)O)C)O)F)C"
4 compound_1_mol = Chem.MolFromSmiles(compound_1_smiles)
5 print ("Log P:",MolLogP(compound_1_mol))
6 print ("Molecular Weight:", MolWt(compound_1_mol))
7 print ("HeavyAtomCount:", HeavyAtomCount(compound_1_mol))
8 print ("HeavyAtomMolWt:", HeavyAtomMolWt(compound_1_mol))
9 print ("Molecular Weight:", round(MolWt(compound_1_mol),3))
10 print ("HeavyAtomMolWt:", round(HeavyAtomMolWt(compound_
   1_mol),3))
Output:
Log P: 1.8957
Molecular Weight: 392.4670000000001
HeavyAtomCount: 28
HeavyAtomMolWt: 363.23500000000007
Molecular Weight: 392.467
HeavyAtomMolWt: 363.235
```

Let's dissect the code.

Lines 1 and 2 Importing libraries "*" work as a wild card and import all the descriptors.

Line 3 Smile code of dexamethasone stored in the variable "compound_1_smiles".

Line 4 Smile converted into a "Mol" format.

Lines 5–10 Descriptor methods "MolLogP", "MolWt", "HeavyAtomCount" and "HeavyAtomMolWt" are called and calculations are performed on "compound_1_mol". We can observe an inbuilt function "round()" that controls the decimal points. "round ()" functions produce a floating-point number that is a rounded version of the original value up to the desired decimal place.

In this example, three different descriptors have been calculated, and how to control the floating-point number has been described.

Example 4 Plotting the descriptor data.

```
1 from rdkit.Chem.Descriptors import *
2 import seaborn as sns
3 Chem_1_smiles = "CC1CC2C3CCC4=CC(=O)C=CC4(C3(C(CC2(C1
  (C(=O)CO)O)C)O)F)C"
4 Comp = Chem.MolFromSmiles(Chem_1_smiles)
5 data = [[HeavyAtomMolWt(Comp), HeavyAtomCount(Comp), MolWt
  (Comp)]]
6 df = pd.DataFrame(data, columns=['HAtmMolW','HAmCount', ' MolWt'])
7 ax = sns.barplot(data=df)
```

Output:

Let's dissect the code.

Lines 1 and 2 Importing the libraries.

Line 3 Smile code of dexamethasone stored in the variable "compound_1_smiles".

Line 4 Smile converted into a "Mol" format.

Line 5 Creating a data array from the descriptor data.

Line 6 Converting data array into a data frame.

Line 7 Plotting the data using the "barplot()" method of the seaborn library (Fig. 9.5).

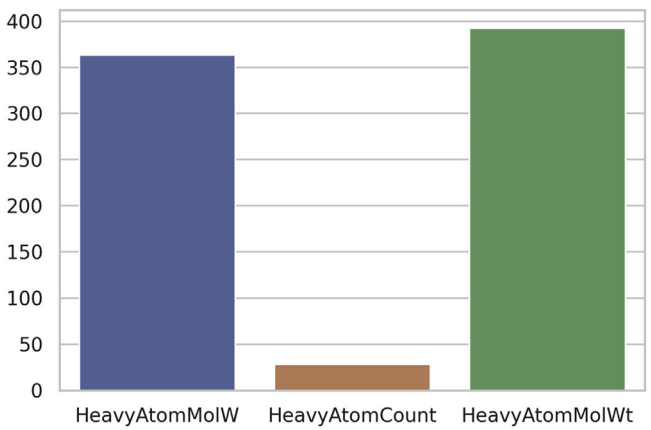


FIGURE 9.5

Bar plot of descriptors.

9.9 Conclusion

Recent advancements in computer infrastructure, data analysis algorithms, and pipelines have been evolving continuously and helping researchers to explore their data rigorously with ease. However, the production of a huge amount of data still needs more and more exploration and demands researchers skilled in advanced programming languages. Bioinformatics and chemoinformatics data are growing exponentially every day and Python, an easy-to-learn but powerful programming language provides the best solution to explore these enormous datasets. Examples included in this chapter prove the simplicity of Python programming, thus providing a start point to adopt Python with ease.

References

Bayat, A., 2002. Science, medicine, and the future: Bioinformatics. Br. Med. J. 324 (7344), 1018−1022.

Chuang, H.-Y., Hofree, M., Ideker, T., 2010. A decade of systems biology. Annu. Rev. Cell Dev. Biol. 26, 721−744.

Cock, P.A., et al., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422−1423.

Ezer, D., Whitaker, K., 2019. Data science for the scientific life cycle. eLife 8.

Gaulton, A., et al., 2011. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr777. D1100−D1107.

Hassan, M., et al., 2006. Cheminformatics analysis and learning in a data pipelining environment. Mol. Divers. 10 (3), 283−299.

He, K.Y., Ge, D., He, M.M., 2017. Big data analytics for genomic medicine. Int. J. Mol. Sci. 18 (2).

Hogeweg, P., 2011. The roots of bioinformatics in theoretical biology. PLoS Comput. Biol. 7 (3), e1002021.

Hughes, J.P., et al., 2011. Principles of early drug discovery. Br. J. Pharmacol. 162 (6), 1239−1249.

Joshi, J., et al., 2013. Cycloxygenase-2 (COX-2)–a potential target for screening of small molecules as radiation countermeasure agents: an in silico study. Curr. Comput. Aided Drug Des. 9 (1), 35−45.

Joshi, J., et al., 2015. Ligand and structure based models for the identification of beta 2 adrenergic receptor antagonists. Curr. Comput. Aided Drug Des. 11 (3), 222−236.

Mallappallil, M., et al., 2020. A Review of Big Data and Medical Research, vol. 8. SAGE open medicine, 2050312120934839.

Munn, Z., Jordan, Z., 2011. The patient experience of high technology medical imaging: a systematic review of the qualitative evidence. JBI Lib. Syst. Rev. 9 (19), 631−678.

Olson, R.S., et al., 2018. Data-driven advice for applying machine learning to bioinformatics problems. Pacific Symp. Biocomput. 23, 192−203.

Ouzounis, C.A., Valencia, A., 2003. Early bioinformatics: the birth of a discipline–a personal view. Bioinformatics 19 (17), 2176−2190.

Petit, C., Bezemer, R., Atallah, L., 2018. A review of recent advances in data analytics for post-operative patient deterioration detection. J. Clin. Monit. Comput. 32 (3), 391−402.
Pilgrim, M., Willison, S., 2009. Dive into Python 3. Springer.
RDKit: Open-source cheminformatics; http://www.rdkit.org.
Tomar, N., De, R.K., 2014. Immunoinformatics: a brief review. Methods Mol. Biol. 1184, 23−55.
Westra, B.L., et al., 2017. Big data science: a literature review of nursing research exemplars. Nurs. Outlook 65 (5), 549−561.
Wilson Sayres, M.A., et al., 2018. Bioinformatics core competencies for undergraduate life sciences education. PloS One 13 (6), e0196878.

CHAPTER

10 Unveiling the molecular basis of DNA–protein structure and function: an in silico view

Anju Singh[1,2], Srishty Gulati[2,a], Md Shoaib[2,a], Shrikant Kukreti[2]

[1]*Department of Chemistry, Ramjas College, University of Delhi, New Delhi, Delhi, India;* [2]*Nucleic Acid Research Lab, Department of Chemistry, University of Delhi, North Campus, New Delhi, Delhi, India*

10.1 Background

DNA–protein complexes have been given utmost importance to uncover the mysteries of various biological processes in vivo. To gain insight into the mechanism of DNA–protein recognition patterns, the prerequisite condition is to understand the structure and function of these biomolecules. DNA cannot exist and function on its own, it requires interaction with proteins to facilitate various functions such as DNA packaging, DNA replication, DNA repair and transcription, etc. The proteins that bind to nucleic acids are composed of nucleic acid-binding domains where interfacing with amino acids takes place in a nonspecific or specific manner. Proteins have entities that may selectively bind to a specific DNA sequence or may recognize any polymorphic structure of DNA. Earlier reports suggested that prokaryotic as well as eukaryotic genomes encode various DNA-binding proteins. There are many DNA-binding motifs that specifically recognize and bind to DNA sequences as well as structures such as leucine zipper, zinc finger, helix-turn-helix, etc. (Luscombe et al., 2000; Walter et al., 2009; Ofran et al., 2007). This selective recognition of DNA structures by proteins is an intriguing process and a challenge for the structural biologist. In vivo, the proteins assemble and adopt 3D complex dynamic structures to facilitate myriad critical functions in cell cycle, as well as other important metabolic functions for cell survival. For a long time, thousands of crystal structures of protein–DNA complexes are deciphered and deposited in the Protein Data Bank (PDB) (Berman et al., 2000). They are made available to the scientific community to understand and crack the codes of various binding modes and mechanisms.

Presently, more than 3868 DNA–protein complex crystal structures are available in the PDB; this is a much smaller number than the number of DNA–protein

[a] Srishty Gulati and Md. Shoaib has contributed equally.

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00008-7

complexes that exist in nature and this number increases day to day. Bioinformatics assists in providing platforms and tools to uncover the complex interaction of macromolecules in a simpler way. With the advent of genome sequencing technology, several genome sequences of various organisms are deciphered and deposited in databanks (Bernstein et al., 1977; Berman et al., 2000). To have an understanding of the clear picture of genomic constitution and functions, the fine details of DNA–protein interactions should be understood. The structural and physical properties of binding sites present on DNA as well as the surface of proteins provide significant information regarding the obstacles limiting their interactions. This helps to discover strategies to overcome the limitations and facilitate the interactions. For a better understanding of DNA–protein interactions, the structural features and sequence context of both the interacting partners should be well explored.

10.2 Structural aspects of DNA

10.2.1 DNA: structural elements

DNA is a dynamic molecule and needs myriad proteins to perform meticulous biological functions such as packaging, replication, transcription, etc. Proteins can interact and bind to DNA structures in a number of ways, i.e., through groove binding or through ionic interaction between the negatively charged sugar phosphate backbone of DNA. It is well known that amino acids are building blocks of proteins comprising many positively charged amino acids such as lysine (K), arginine (R), and histidine (H). These amino acid residues can interact and bind with negatively charged DNA backbone.

DNA exhibits a higher degree of structural polymorphism and exists as canonical (B-DNA or Watson–Crick DNA) as well as many unusual noncanonical (non-B-DNA) structural forms in biological systems such as Z-DNA, hairpin, slipped structures, triplexes, G-quadruplexes, i-motifs, and cruciform structures (Bochman et al., 2012; Kaushik et al., 2016; Spiegel et al., 2020). Fig. 10.1 depicts some of the polymorphic forms of DNA.

10.2.2 DNA: nitrogenous bases of DNA are involved in base pairing

All the DNA unusual structures are stabilized by varied hydrogen-bonding patterns. Apart from the universal antiparallel Watson–Crick base pairing, there exist Hoogsteen, reverse-Hoogsteen, and wobble and parallel Watson–Crick base pairing. The two types of nitrogenous bases, i.e., purines and pyrimidines, present in DNA are adenine/guanine and thymine/cytosine, respectively (Sinden, 1994). Schematic representation of DNA bases facilitating Watson–Crick base pairing is shown in Fig. 10.2.

Nitrogenous bases are bestowed with various potential hydrogen bond donor and acceptor sites, which actively participate in hydrogen bond formation generating various polymorphic structures. Besides the sites on the bases already involved in

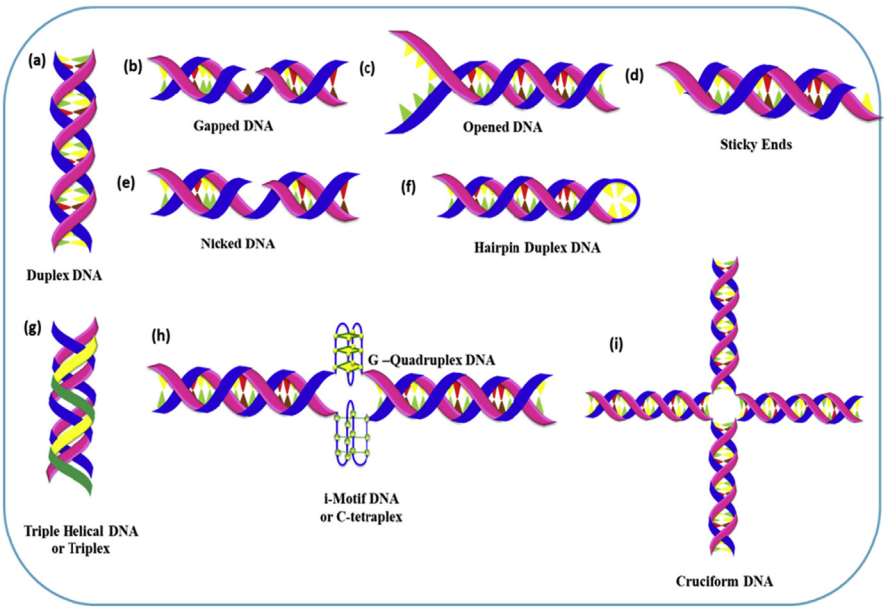


FIGURE 10.1

Canonical (B-DNA): (A) duplex DNA, (B) gapped DNA, (C) opened DNA, (D) sticky ends, (E) nicked DNA, (F) hairpin duplex DNA. Noncanonical DNA (non-B-DNA) structures: (G) triple helical DNA, (H) G-quadruplex and i-motif, and (I) cruciform DNA.

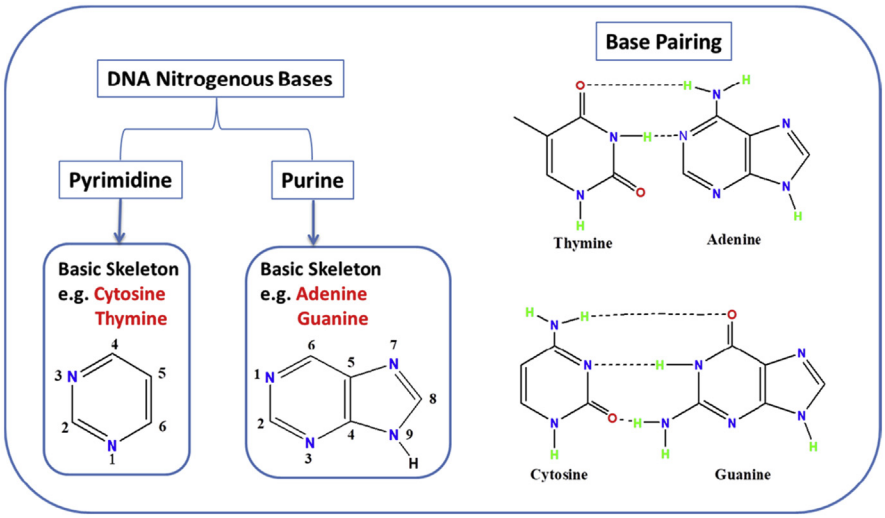


FIGURE 10.2

Nitrogenous bases—pyrimidines and purines—showing the Watson–Crick base pairing scheme involved in a standard DNA duplex.

Watson–Crick base pairing, there also exist other potential hydrogen bond donor and acceptor atoms, which with further interactions with other bases can create higher-order structures. These extra hydrogen bonding sites may also interact specifically with amino acids within the proteins. Tautomeric forms of DNA bases with their hydrogen bonding sites are depicted in Fig. 10.3.

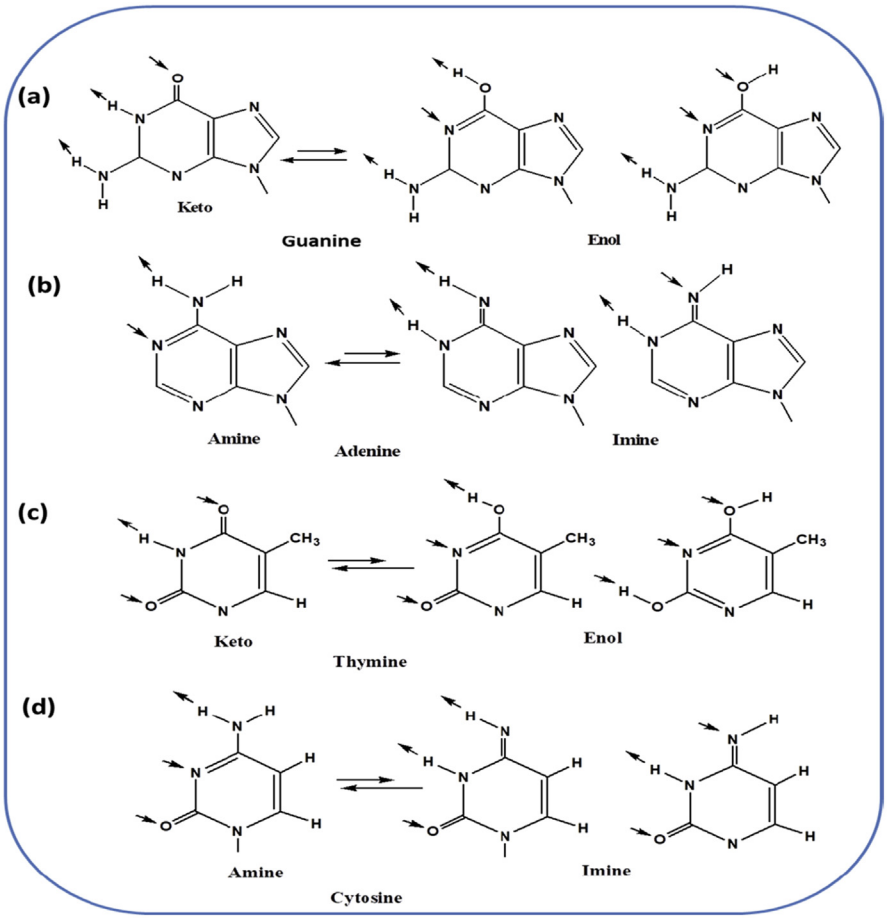


FIGURE 10.3

Tautomerization of bases and various hydrogen-bonding potential sites present on nitrogenous bases. (A) Guanine exists in keto form where a carbonyl group is present at the C6 position. (B) Amine and imine forms of adenine. (C) Keto–enol forms of thymine. (D) Amine and imine forms of cytosine. The *arrows* indicate the hydrogen-bonding properties of the bases. The *arrows* pointing outward from the hydrogen and toward the negative centers represent hydrogen-bonding donors and acceptors, respectively.

The most conspicuous and crucial features of DNA exposed to proteins and other ligands are the two grooves, i.e., major groove and minor groove. The grooves arise due to helical twisting of two strands around each other. The major groove is wide and accessible, while the minor groove is narrow. Since water is an integral part of biological systems, a hydration shell is present around the grooves where secondary structural elements of protein can interact well with DNA bases. The major groove is wider and can harbor an α-helix or two strands of β-ribbon, whereas being narrow, the minor groove can accommodate only a single peptide chain. Hydrogen bond donor and acceptor sites exposed in grooves are accessible by protein directly or indirectly through one or more water molecules (Blackburn et al., 2006). Grooves are well hydrated in DNA, but can be displaced when proteins approach to bind the duplex structures. The structural elements of DNA need to be considered prior to establishing their interaction with proteins.

10.3 Structural aspects of proteins

10.3.1 Characteristic features of amino acids

Protein is a critically important biomolecule, substantial for the myriad biological functions along with building blocks of the body. Protein initiates various biological processes either via interaction with other biomolecules or by directly playing a central role in the processes. Amino acids are the basic units of protein, linked together via peptide bonds ultimately forming a polypeptide chain. Many polypeptides, interlinked with disulfide linkage, may result in forming various domains and ultimately a protein structure. Interaction between DNA and protein are mediated by various hydrogen bond donor and acceptor groups on DNA bases, and amino acids of proteins, through hydrogen bonds. Like the DNA bases, the polar amino acids, at their side chains, contain various hydrogen bond donor and acceptor sites. For example, asparagine and arginine side chains instantly involve hydrogen bonds with guanine and adenine, forming a bidentate complex. The aromatic amino acids tryptophan, phenylalanine, and tyrosine are known to form weaker hydrogen bonds known as "π-hydrogen bonds" with DNA bases.

Water molecules play a substantial role in making bridges between DNA and protein. They can extend the surface of the DNA and link it to an amino acid. The amide backbone of the protein can also hydrogen bonded with nucleic acids (bases). The structural element of a protein found interacting most frequently with the DNA major groove is an α-helix, whereas many proteins such as TATA-binding proteins can interact with DNA via the minor groove (Fig. 10.4).

10.3.2 Characteristic features of proteins

Proteins can be classified on the basis of their structure as well as functions performed by them in biological processes. Based on the information available from

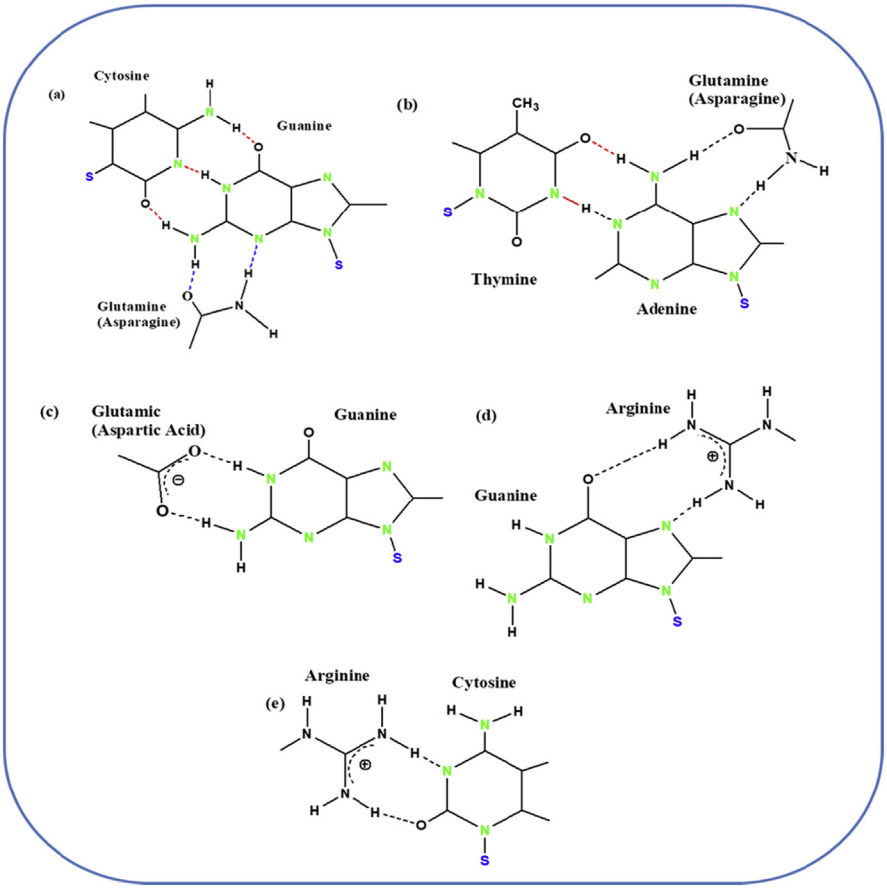


FIGURE 10.4

Basic skeleton of nucleic acid bases, where hydrogen is bonded with amino acids.

previous studies and the literature in PubMed, PDB (Berman et al., 2000), CATH (Orengo et al., 1997), SCOP (Murzin et al., 1995), COPS (Suhrer et al., 2009), NPIDB (Kirsanov et al., 2013), DNAproDB (Sagendorf et al., 2017), etc. databases, three classification categories, namely class, type, and subtypes, are proposed. Protein can be classified in a class on the basis of function and can be further categorized in three subcategories, i.e., enzyme, transcription factor (TF), and structural or DNA-binding proteins, respectively.

10.3.3 Classification of protein-binding motifs

The proteins that involve the modification of DNA are classified as enzymes. Various proteins, which facilitate transcription and regulation of gene expression in biological systems, are classified as TFs. Another class, known as structural or

DNA-binding proteins, comprises proteins involved in DNA packaging, DNA bending, and aggregation. On the basis of reactions catalyzed and the function of enzymes, category type is organized in 15 subcategories: dioxygenase (Yang et al., 2008), endonuclease (Williams, 2003), excisionase (Sam et al., 2004), glucosyltransferase (Lariviere et al., 2005), glycosylase (Fromme et al., 2004), helicase (Lee and Yang, 2006), ligase (Nandakumar et al., 2007), methyltransferase (Brenner and Fuks, 2006), nuclease, photolyase (Mees et al., 2004), polymerase (Brautigam and Steitz, 1998), recombinase (Guo et al., 1997), topoisomerase (Redinbo et al., 1998), translocase (Löwe et al., 2008), and transposase (Davies et al., 2000).

Structural or DNA-binding protein can be categorized in eight different subcategories, i.e., centromeric protein (Verdaasdonk and Bloom, 2011), DNA packaging (Ward and Coffey, 1991), maintenance/protection (Strogantsev and Ferguson-Smith, 2012), DNA bending (Vliet and Verrizer, 1993), repair protein (Ambekar et al., 2017), replication protein (Prakash and Borgstahl, 2012),telomeric protein (Amir et al., 2020), and Zalpha (Yang et al., 2014). TFs include seven categories of proteins that bind to the DNA structures and mediate various functions; they are alpha helix (α-helix) (Doig et al., 2001), α/β protein (Fujiwara et al., 2012), β-sheet (Perczel et al., 2005), helix-turn-helix (Brennan and Matthews, 1989), ribbon/helix/helix (Schreiter and Drennan, 2007), zinc coordinating (Laitaoja et al., 2013), and zipper type (Hakoshima, 2005). The category subtypes include more specific features of proteins such as specific reaction of a particular enzyme, specific DNA binding sites and domains, etc. These classes, types, and characteristic functions are tabulated in Table 10.1.

Apart from these classifications, various protein features are also to be taken into consideration such as number of protein monomers interlinked with double-helix DNA, whether the protein is heteromultimeric or homomultimeric, or both. Numerous proteins present in the vicinity of other DNA-binding proteins may recognize each other and be involved in protein−protein interaction. Another important feature relies on a methodology that gives insight into the location of atoms of DNA and protein in 3D spaces (Ferrada and Melo, 2009). Sequence and structural information of a query protein opens new avenues to explore DNA-binding residues and provide a platform to develop computational strategies or databases. Based on sequence similarity, numerous studies have been considered for the development of databases for DNA-binding domains (Ofran et al., 2007; Hwang et al., 2007; Yan et al., 2006; Ahmad and Sarai, 2005; Wu et al., 2009; Carson et al., 2010). It is observed that the DNA-binding residues are found to be less conserved, which is why if the protein structure is known, some structural biology techniques and other methods can be used to detect DNA-binding residues.

The DNA-binding sites can be predicted by comparing with known putative DNA-binding spots on the query protein. Furthermore, 3D structures of proteins can be used to explore and decipher the binding sites (Alibes et al., 2010; Li et al., 2014; Li et al., 2013; Xiong et al., 2011). Fig. 10.5 displays the schematic diagram of a strategy implemented to explore sequence and structure similarity for binding. Putative sequences (conserved binding motifs) and 3D structural data

Table 10.1 Classes, types, and characteristic function catalyzed by various proteins.

S. no.	Class	Type	Characteristic functions
1.	Enzymes	Dioxygenase	Involved in DNA repair in which lesions are caused, by using a direct oxidative dealkylation mechanism (Yang et al., 2008)
		Endonuclease	DNA cleaves at specific places by restriction enzyme (Williams, 2003)
		Excisionase	Integrase-mediated DNA rearrangement is controlled by this enzyme (Sam et al., 2004)
		Glucosyltransferase	This enzyme interacts and binds on an abasic site of DNA and flips it. Using UDP-glucose, glucosylation happens at 5-methylcytosine in duplex DNA (Lariviere et al., 2005)
		Glycosylase	Involved in base excision repair (a process in which damaged nucleotides in DNA can be removed or replaced) (Fromme et al., 2004)
		Helicase	Helicases are involved in unwinding DNA double helices by using ATP hydrolysis (Lee and Yang, 2006)
		Ligase	This class of enzymes is involved in the recognition of nicks and conditions for strand closure (Nandakumar et al., 2007)
		Methyltransferase	Involved in methylation in the genome and plays a pivotal role in gene silencing (Brenner and Fuks, 2006)
		Nuclease	This is a nucleate or cleave DNA but does not come under the class of endonucleases (Mees et al., 2004)
		Photolyase	In ultraviolet-induced base lesions caused in DNA, this enzyme uses light to repair the lesions (Mees et al., 2004)
		Polymerase	This enzyme is involved in polynucleotide synthesis against a nucleotide template strand using base-pairing interaction (Brautigam and Steitz, 1998)
		Recombinase	This enzyme mediates the recombination process in biological systems (Guo et al., 1997)
		Topoisomerase	Helps to relax DNA superhelical stress or lesions by causing a transient single-stranded break in double helical DNA (Redinbo et al., 1998)
		Translocase	This enzyme is involved in the segregation of circular chromosomes, formed by recombination of monomer sister strands (Löwe et al., 2008)
		Transposase	Mediates the movement of DNA segments known as transposons by a process known as transposition (Davies et al., 2000)

2.	Structural or DNA binding	Centromeric protein	Includes a protein that is part of the centromere (Verdaasdonk and Bloom, 2011)
		DNA packaging	Includes a protein (histone in eukaryotes) that assists DNA in packaging (Ward and Coffey, 1991)
		Maintenance/ protection	These proteins indulge in the protection and maintenance of genomes (Strogantsev and Ferguson-Smith, 2012)
		DNA bending	Protein that helps in the bending of DNA for indirect readout (Vliet and Verrizer, 1993)
		Repair protein	Protein involved in the recognition and repair of damaged DNA (Ambekar et al., 2017)
		Replication	Protein that assists in the replication of DNA (Prakash and Borgstahl, 2012)
		Telomeric protein	These proteins are involved in recognition and binding to the telomere part of the chromosome and impart stability (Amir et al., 2020)
		Zalpha	These proteins specifically recognize and bind to Z-DNA (left-handed DNA) (Yang et al., 2014)
3.	Transcription factor	α-Helix	Protein rich in alpha helices that interacts with DNA via alpha helices (Doig et al., 2001)
		α/β	Includes alpha helices and beta-sheets to interact with DNA (Fujiwara et al., 2012)
		β-Sheet	This protein is rich in beta-sheets and interacts with DNA via the beta-sheets (Perczel et al., 2005)
		Helix-turn-helix	This protein is known as a helix-turn-helix DNA binding protein having a winged helix domain (Brennan and Matthews, 1989)
		Ribbon/helix/helix	Contains ribbons or helices in binding domains and binds to DNA (Schreiter and Drennan, 2007)
		Zinc coordinating	Protein that harbors zinc in structures to interact with DNA (Laitaoja et al., 2013)
		Zipper type	These proteins contain motifs that can bind to DNA as a zipper (e.g., leucine zipper) (Hakoshima, 2005)

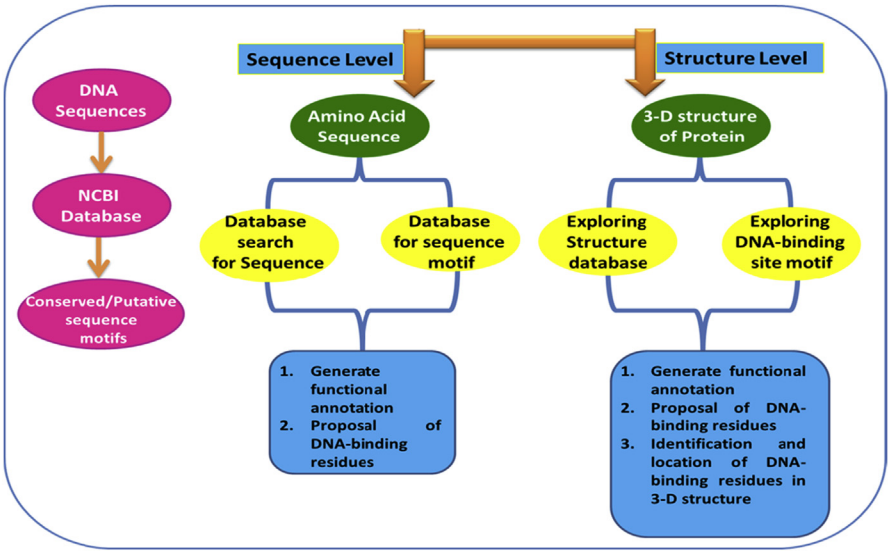


FIGURE 10.5

Strategy to explore the sequence and structure similarity in DNA as well as protein.

are freely available on public databases, which can now be used to construct various datasets to study protein−DNA complexes. Much literature is available showing established and constructed datasets. It is significant to note that for proper binding between protein and DNA, a specific cutoff distance should be present between atoms of amino acid residue and the neighboring atoms of the DNA molecule.

Again, much literature is available on DNA−protein-binding site prediction, which uses different datasets as well as parameters of DNA-binding sites. Datasets used to study/predict DNA-binding proteins and sites are tabulated in Table 10.2.

Using evolutionary and structural information for predicting binding sites, several groups have reported varied cutoff distances for binding and nonbinding residues. While Kuznetsov et al. (2006) established that cutoff distance should be 4.5 Å, Si et al. reported the cutoff distance as 3.5, 4.0, 4.5, 5.0, 5.5, and 6.0 Å, respectively. However, the most suitable cutoff distance was proven to be 3.5 Å for the separation between binding and nonbinding residues (Si et al., 2011). Further development of more software and databases has taken into consideration parameters such as amino acid residues, forces for stabilization, etc.

Each bioinformatics database and tool has its own specific decisive factor and characteristics to set the design datasets. For example, the TRANSFAC database operates by choosing specific DNA sequences depending on the specific class of proteins. It is based on selecting a small stretch of DNA segment (5−25 base pairs) and transferring it to the appropriate relational data model. By using this knowledge, information on protein-binding sites and factors can be predicted. It is necessary to investigate this, as one site can interact with various factors, and many known factors

Table 10.2 Datasets used to predict DNA-binding proteins and binding sites.

S. no.	ID (datasets)	Detailed notes	References
1.	DB179	Contains 179 NA-binding proteins, 40% sequence identity	Gao and Skolnick (2008)
2.	NB3797	Contains 3797 nonbinding proteins, 35% significant sequence identity level (3482 independent clusters)	Gao and Skolnick (2008)
3.	PD138	Consists of 138 DNA-binding proteins, mostly nonredundant at 35% sequence identity, categorized in seven structural classes	Szilagyi and Skolnick (2006)
4.	DISIS	Consists of 78 DNA-binding proteins, redundant at 20% sequence similarity	Ofran et al. (2007)
5.	PDNA62	Contains 62 DNA-binding proteins, 78 chains, approximately 57 nonredundant sequences at 30% identity	Ahmad et al. (2004)
6.	NB110	Consists of 110 nonbinding proteins, nonredundant at 30% sequence similarity. It has entries without DNA derived from RS126 secondary structure dataset.	Ahmad et al. (2004)
7.	BIND54	Contains 54 binding proteins, exactly 58 chains, nonredundant at 30% sequence identity	Stawiski et al. (2003)
8.	NB250	Consists of 250 nonbinding proteins, nonredundant at 35% sequence similarity	Stawiski et al. (2003)
9.	DBP374	Comprises 374 DNA-binding proteins, redundancy at 25% sequence identity level	Wu et al. (2009)
10.	TS75	Consists of 75 DNA-binding proteins, independent from DBP374 and PDNA62 but has redundant entries both with 35% sequence identity level	Wu et al. (2009)
11.	PDNA-316	Comprises 316 target proteins and is used in the metaDBsite web server with 30% sequence similarity	Si et al. (2011)
12.	DNA BindR171	Consists of 171 proteins with sequence identity $\leq$30%, each protein has a minimum of 40 amino acid residues	Yan et al. (2006)

can bind to many sites. The SITES and TABLES, the two separate types of tabular information, can be extracted on the basis of a relational data model in which the SITES table entails the exact position of a regulatory site, the gene where this site is located along with biological relevance of that particular gene (Wingender et al., 1996). Besides this, various other tools are available that use evolutionary aspects and information to design datasets to study DNA–protein interaction.

The web server (DP-Bind) is a kind of user interface comprising three input fields such as query sequence (to be analyzed), selection of encoding method, and email address. Users can investigate and seek detailed information about each field along with output format just by clicking on the help hyperlink. FASTA format can be used for pasting or uploading amino acid sequences in the place of input. Input sequence can be 1000 residues long and the web server accepts 100 sequences for single-sequence-based encoding. The output is seen on DP-Bind, which comprises three parts, i.e., a header predicting the format, input sequence, and results of the query sequence in a tabulated format. The table contains 10 columns where the first column represents a residue index that shows the exact position of the sequence; the second column consists of amino acid residue. DP-Bind also uses different predictors such as support vector machine (SVM) (Vapnik, 1998), kernel logistic regression (KLR) (Zhu and Hastie, 2005), and penalized logistic regression (PLR) (le Cessie and van Houwelingen, 1992) predictors to predict results. Columns 3−8 display output from SVM, KLR, and PLR predictors. Output from each method comprises a predicted binding label and the probability of that label. Different labels, such as label 1 and label 0, represent DNA-binding and nonbinding residues, respectively. Majority and strict consensus are depicted in columns 9 and 10, respectively, and if strict consensus is not obtained, i.e., one method differs from the other two methods, the position is marked with not available (NA). DP-bind is significantly a very informative bioinformatics tool implied for predicting DNA-binding residue, based on sequences present in DNA-binding proteins (Hwang et al., 2007).

10.4 In silico tools for unveiling the mystery of DNA−protein interactions

It is well documented that a number of biophysical techniques are available for gaining insight into the structural details and modes of interaction in DNA−protein complexes. Also, a plethora of computational tools and software is available and being employed presently to predict DNA−protein interactions. These are summarized next.

10.4.1 TRANSFAC

The TRANSFAC database is used to gain information on the TF binding sites from yeast to humans. It is a comprehensive knowledge-based tool employed to enquire about query sequences by comparing with experimentally proven binding sites. TRANSFAC has a broad compilation of binding sites allowing the derivation of matrices, which can be used along with suitable tools to search various DNA sequences. This database includes several entries under different categories. It provides a significant value, which assists in identifying DNA-binding activity and thus leads to decipher a specific factor. TRANSFAC also gives a platform for other tools called Match, which exploits the nucleotide weight matrices of TRANSFAC to

explore potential binding sites in uncharacterized sequences. Other web programs such as AliBaba2 utilize the TRANSFAC database to decipher TF-binding sites in an unknown DNA sequence. This web program exploits the binding sites deposited and collected in TRANSFAC. In addition to TRANSFAC, another tool, P-Match, is also employed to identify TF binding sites in DNA sequences (Matys et al., 2003; Wingender et al., 2001). It uses pattern matching and weight matrices in combination to facilitate a high level of accuracy of recognition. Many other web programs are also available that use the TRANSFAC database to predict and identify unique computational functions.

10.4.2 DISPLAR (DNA site prediction from record of neighboring residues)

The DISPLAR database is based on the neural network that significantly assists in predicting the protein residues involved in recognition and binding to DNA, provided the structure of the protein is known. DISPLAR utilizes position-specific sequence as well as solvent accessibilities along with spatial neighbors to predict the binding residues. It shows prediction accuracy over 80% and interprets the accurate DNA-binding residues (Tjong and Zhou, 2007).

10.4.3 iDBPs (exploration of DNA-binding proteins)

Nimrod et al. (2010), established the iDBPs server for the identification of DNA-binding proteins based on the 3D structure of protein. This server first utilizes PatchFinder to explore the functional region of the protein. Moreover, the PatchFinder algorithm is extensively used to search the clusters of putative conserved residues on the protein surface. By using this algorithm, the maximum-likelihood patches are easy to find and depict the functional regions in protein, as well as DNA-binding regions within the DNA-binding proteins. This information is exploited by users for the investigation of their results by including prediction scores of the proteins with requisite score cutoff (Nimrod et al., 2010).

10.4.4 MAPPER (multigenome analysis of position and patterns of elements of regulation)

MAPPER, utilized for the identification of TF binding sites, is based on the hidden Markov model retrieved from known sites. TF binding sites can be exploited to align with the sites provided by TRANSFAC and other similar databases. It can be used to depict the sites in the genome sequence by scanning various organisms (humans, flies, mice, worms, yeast, etc.) (Marinescu et al., 2005a,b). Compared to other computational models, it is a more specific and sensitive tool. Usually, a query sequence is uploaded and followed by multiple sequence alignments of the TF binding sites.

10.4.5 DP-Bind

In DP-Bind, protein binding sites are predicted on the basis of analysis of amino acid residues. Three support models, namely SVMs, KLR, and PLR, are used to predict the binding sites. The predictions can be made by utilizing single sequence query or cluster of evolutionary conservation of input sequence. These three support models in combination can be exploited to provide consensual, high precision results (Hwang et al., 2007).

10.4.6 PreDs

PreDs is a web-based server where protein molecular surfaces are used to gain information about DNA-binding sites. Atomic coordinates are available in pdb format and are used to generate protein molecular surfaces. This prediction considers electrostatic potential, global curvature, as well as the local curvature of the protein surface (Tsuchiya et al., 2004).

10.4.7 ZIFIBI (zinc finger site database)

As the name indicates, this database includes the zinc finger domains of protein having the potential to recognize DNA sequences. It helps in spotting the C_2H_2 zing finger transcription binding site in the *cis*-regulatory location of the target genes. This tool also exploits the hidden Markov model to perform calculation of the state path for binding sites. Here, specific attention is given to the interaction of amino acid residues of zinc finger and nucleotide sequences (Cho et al., 2008).

10.4.8 Bindn and Bindn+

Bindn is a web-based tool that utilizes SVMs to interpret and depict the nucleic acid (DNA and RNA) binding sites. SVM modes are based on sequence features, such as side chain pKa values, hydrophobicity index, and molecular mass of an amino acid. Basically, the primary sequence data of proteins is exploited extensively to procure binding site information (Wang and Brown, 2006). Similarly, Bindn+ is also a web-based tool that utilizes SVMs but exploits varied protein features to predict DNA-binding sites. In this method, the biochemical property of the amino acids is utilized to prepare a position-specific scoring matrix (Wang et al., 2010).

10.4.9 ProNIT

This database utilizes quantitative parameters in place of DNA−protein structural data. Various thermodynamic parameters (dissociation constant, association constant, Gibbs free energy change, enthalpy change, heat capacity change, etc.), experimental conditions, structural information of protein, as well as nucleic acids are exploited to predict binding sites. It facilitates many output options to provide information to other databases to provide flexibility in searching the binding sites (Prabakaran et al., 2001).

10.4.10 DNA-Prot

This tool utilizes protein sequences for the identification of DNA-binding proteins. DNA-Prot can efficiently differentiate between DNA-binding and non-DNA-binding proteins by specific recognition of nucleotide sequence and chains. This tool employs the random forest method to depict DNA-binding residues and proteins (Kumar et al., 2009).

10.4.11 PDIdb

PDIdb is a tool that acts as a repository containing the relevant structural information of protein–DNA complexes solved by X-ray crystallography. It uses the data deposited in PDBs. This user-friendly database uses a method to classify all the complexes in three hierarchical levels, i.e., classes, types, and subtypes, respectively. Classification is made on the basis of manually curated information gathered from various databases such as PDB, PubMed, CATH, SCOP, and COPS. PDIdb focuses on each atomic interface of both DNA and protein and each entry has a specific protein–DNA interface (Norambuena and Melo, 2010).

10.4.12 PADA1 (protein assisted DNA assembly 1)

PADA1 is a generic algorithm that exploits model structural complexes to depict DNA-binding proteins of already resolved structures. It utilizes a library of protein and duplex DNA pairs harvested from 2103 DNA–protein complexes. To evaluate and filter 3D docking models, a fast-statistical forcefield computed from atom–atom distance is used. The use of PADA1 has established that the quality of docked templates is compatible with the FoldX protein design tool. It also represents DNA–protein conformational changes by predicting DNA-binding regions/nucleotide sequences (Blanco et al., 2018).

10.4.13 DNAproDB

A web-based tool, DNAproDB is specifically designed to gain information on DNA–protein complexes. Structural features of these complexes are extracted by using this tool, which provides an automated structure-processing pipeline. The extracted data are arranged in structured data files that can be utilized by any programming language or easily viewed in a browser. The database can be searched by taking a combination of DNA, protein, or DNA–protein interactions at the interface. DNAproDB facilitates various interactive and customizable tools for generating visualization of the DNA–protein interface (Sagendorf et al., 2017).

10.4.14 WebPDA

WebPDA, where P stands for protein, D stands for DNA, and A indicates analyzer, is a structure analysis program that utilizes PDB files as inputs and carries out structure

analysis, including protein–DNA complex restructuration and double-stranded DNA reconstruction. It provides a new technique to analyze DNA base-pairs, as well as systematic annotation of DNA–protein interactions. This tool also identifies the DNA sequences involved in DNA–protein interactions as well as in binding residues (Kim and Guo, 2009). The current PDB files are easy to process and undergo analysis by using this tool. WebPDA is also used as a web interface for exploiting PDA and for data retrieval.

10.4.15 DOMMINO

The DOMMINO 2.0 database is designed to analyze protein, DNA, RNA, and their interactions. This web interface is utilized to search and depict the subunit interaction network at the atomic level of whole macromolecular assemblies to gain information on binding domains. This software is very efficient as it utilizes the structural classification of the interacting subunits and compares the interactions of several macromolecules (Kuang et al., 2016, Fig. 10.6).

10.4.16 FlyFactorSurvey

This server is specifically designed to identify the TF binding sites in the *Drosophila* genome. It includes 400 recognition motifs and position weight matrices for over

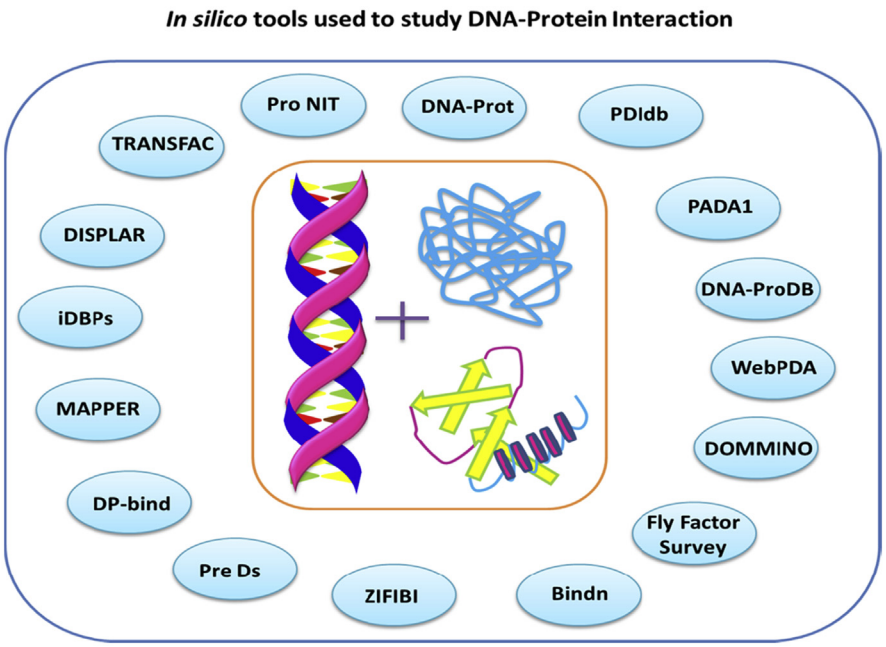


FIGURE 10.6

Summary of databases and tools used for in silico study of DNA–protein interaction.

200 TFs. It utilizes a bacterial one-hybrid system to select TF binding sites in the database. This also facilitates search tools and flat file downloads to retrieve information on binding sites for individual TF binding domains or groups of TF domains with characteristic binding or recognition specificity. Various tools are linked to this database to identify binding motifs similar to the query matrix or for an individual motif throughout the *Drosophila* genome (Zhu et al., 2011).

10.5 Future perspectives

Investigational approaches for defining protein–nucleic acid interactions are expensive and time consuming. Moreover, these approaches will not be able to deal with the ever-increasing numbers of protein sequences requiring annotations for their potential nucleic acid binding ability. Since solving the crystal structure is not possible for each homolog for every DNA-binding protein, even small changes at the sequence level may meaningfully alter the interaction dynamics. Henceforward, computational podiums for the prediction of DNA/RNA binding sites from the sequences can provide an alternative that can be just as reliable to analyze protein–nucleic acid interactions. The computational identification of nucleic acid-binding amino acid residues can greatly contribute to a better understanding of their functions. There are now several prediction methods developed and available for use as web accessible services. One has to make maximum use of these programs for predictions and information on nucleic acid-binding residues. Moreover, their ability to discriminate binding and nonbinding residues can also be accessed.

Macromolecular interactions play a vital role in almost all the biological processes. Quite specifically, DNA and protein biomolecules are an integral and indispensable part of the biological system. DNA–protein interactions are crucial for the proper functioning of every organism. Since DNA alone is a passive molecule, therefore, from packaging to transcription, replication in every arena, DNA needs to interact with proteins to facilitate every biological function. Numerous biophysical techniques are available that explore DNA–protein interaction and mode of binding. Recently, in silico tools have emerged, which give more exquisite and detailed information of DNA–protein interaction and binding residues. Because of the unique structural features of DNA, amino acid residues (protein-binding domains) interact with DNA nucleotides via major and minor grooves. Protein–DNA interaction might be specific or nonspecific. Biophysical and biochemical characterization of DNA–protein complexes is an intriguing as well as laborious process for elucidating the type of interaction taking place between these biomolecules. Bioinformatics tools are more feasible to portray the exact picture of complex formation and mode of binding in vivo.

A plethora of literature is available, indicating the advantages and significance of computational tools to study and depict the DNA-binding residues and their interaction with proteins. Many in silico tools are available that are exclusively exploited to gain insight into the structural architecture and position of atoms of DNA such as the

Nucleic Acid Database (Berman et al., 1992), 3DNA (Lu and Olson, 2003, Lu and Olson, 2008), etc. Based on the UniProtKB/Swiss-Prot database, approximately 21% of proteins are denoted as nucleic acid binding (Boutet et al., 2007). Specific databases are also designed to investigate the recognition between amino acid and nucleotide residues, i.e., the Amino Acid-Nucleotide Interaction Database (Hoffman et al., 2004). The Protein-Nucleic Acid Complex Database is also available to seek information on fine structural details of protein−nucleic acid complexes (An et al., 1998).

In silico tools and databases have further advantages because they overcome the limitations of biophysical techniques by reducing analysis time and displaying hypothetical prediction of binding residue along with depicting binding parameters, energy gap, as well as removing water molecules involved in DNA−protein interaction. To gain a better picture of any macromolecular complex, a 3D structure is required, provided by a computational docking approach based on a fine single detail of an individual entity/macromolecule. In silico tools and databases facilitate a reasonable, precise, and accurate model picture of protein−DNA complexes, which assists users to uncover the mysteries of various biological processes and mechanisms. With the advent of new bioinformatics tools and databases we hope to gain a better understanding of the puzzles of DNA−protein complexes in the future.

10.6 Abbreviations

AANT	Amino Acid-Nucleotide Interaction Database
CATH	Class, Architecture, Topology, Homology
COPS	Co-Occurrence Pattern Search
DISPLAR	DNA site prediction from a list of adjacent residues
DNA-Prot	DNA-Protein
DP-Bind	DNA-Protein Binding
iDBPs	Identification of DNA-binding proteins
MAPPER	Multigenome analysis of position and patterns of elements of regulation
NDB	Nucleic Acid Database
NPIDB	Nucleic Acid−Protein Interaction Database
PADA1	Protein Assisted DNA Assembly 1
PDB	Protein Data Bank
PDIdb	Protein-DNA Interface Database
ProNIT	Protein Nucleic Acid Interactions
ProNuC	Protein-Nucleic Acid Complex Database
SCOP	Structural Classification of Protein
SVM	Support vector machine
TF	Transcription factor
ZIFIBI	Zinc finger site database

References

Ahmad, S., Gromiha, M.M., Sarai, A., 2004. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20, 477−486.

Ahmad, S., Sarai, A., 2005. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinf. 6 https://doi.org/10.1186/1471-2105-6-33.

Alibes, A., Serrano, L., Nadra, A.D., 2010. Structure-based DNA-binding prediction and design. Methods Mol. Biol. 649, 77−88.

Amir, M., Khan, P., Queen, A., Dohare, R., Alajmi, M.F., Hussain, A., Islam, A., Ahmad, F., Hassan, M.I., 2020. Structural features of nucleoprotein CST/Shelterin complex involved in the telomere maintenance and its association with disease. Mutations. Cells 9, 359. https://doi.org/10.3390/cells9020359,1-32.

Ambekar, S.S., Hattur, S.S., Bule, P.B., 2017. DNA: damage and repair mechanisms in humans. Glob. J. Pharmaceu. Sci. 3 (3), 1−8. GJPPS.MS.ID.555613.

An, J., Nakama, T., Kubota, Y., Sarai, A., 1998. 3DinSight: an integrated relational database and search tool for the structure, function and properties of biomolecules. Bioinformatics 14 (2), 188−195.

Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R., Schneider, B., 1992. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J. 63 (3), 751−759.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al., 2000. The protein data bank. Nucleic Acids Res. 28, 235−242. https://doi.org/10.1093/nar/28.1.235PMID:10592235.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., et al., 1977. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur. J. Biochem. 80, 319−324. https://doi.org/10.1111/j.1432-1033.1977.tb11885.xPMID:923582.

Blackburn, G.M., Gait, M.J., Loakes, D., Williams, D.M., 2006. Nucleic Acids in Chemistry and Biology. Published by The Roayl Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge CB4 OWF, UK.

Blanco, J.D., Radusky, L., Climente-González, H., Serrano, L., 2018. FoldX accurate structural protein−DNA binding prediction using PADA1 (Protein Assisted DNA Assembly 1). Nucleic Acids Res. 46 (8), 3852−3863.

Bochman, M.L., Paeschke, K., Zakian, V.A., 2012. DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Genet. 13 (11), 770−780.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A., 2007. UniProtKB/Swiss-Prot methods. Mol. Biol. 406, 89−112.

Brenner, C., Fuks, F., 2006. DNA methyltransferases: facts, clues, mysteries. Curr. Top. Microbiol. Immunol. 301, 45−66.

Brennan, R.G., Matthews, B.W., 1989. The helix-turn-helix DNA binding motif. J. Biol. Chem. 4, 1903−1906.

Brautigam, C.A., Steitz, T.A., 1998. Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes. Curr. Opin. Struct. Biol. 8, 54−63.

Carson, M.B., Langlois, R., Lu, H., 2010. NAPS: a residue-level nucleic acid-binding prediction server. Nucleic Acids Res. 38, W431−W435.

Cho, S.Y., Chung, M., Park, M., Park, S., Lee, Y.S., 2008. ZIFIBI: prediction of DNA binding sites for zinc finger proteins. Biochem. Biophys. Res. Commun. 369, 845−848.

Davies, D.R., Goryshin, I.Y., Reznikoff, W.S., Rayment, I., 2000. Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. Science 289, 77−85. https://doi.org/10.1186/1471-2105-11-262.

Doig, A.J., Andrew, C.D., Cochran, D.A., Hughes, E., Penel, S., Sun, J.K., Stapley, B.J., Clarke, D.T., Jones, G.R., 2001. Structure, stability and folding of the alpha-helix. Biochem. Soc. Symp. (68), 95−110. https://doi.org/10.1042/bss0680095.

Ferrada, E., Melo, F., 2009. Effective knowledge-based potentials. Protein Sci. 18, 1469−1485.

Fromme, J.C., Banerjee, A., Verdine, G.L., 2004. DNA glycosylase recognition and catalysis. Curr. Opin. Struct. Biol. 14, 43−49.

Fujiwara, K., Toda, H., Ikeguchi, M., 2012. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. BMC Struct. Biol. 2, 12. https://doi.org/10.1186/1472-6807-12-18.

Guo, F., Gopaul, D.N., van Duyne, G.D., 1997. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature 389, 40−46.

Gao, M., Skolnick, J., 2008. DBD-Hunter: a knowledge-based method for the prediction of DNA−protein interactions. Nucleic Acids Res. 36, 3978−3992.

Hakoshima, T., 2005. Leucine zippers. Encycl. Life Sci. 1−5. John Wiley & Sons, Ltd. www.els.net.

Hoffman, M.M., Khrapov, M.A., Cox, J.C., Yao, J., Tong, L., Ellington, A.D., 2004. AANT: the amino acid-nucleotide interaction database. Nucleic Acids Res. D174−D181, 32 Database.

Hwang, S., Gou, Z., Kuznetsov, I.B., 2007. DP-bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 23, 634−636.

Kaushik, M., Kaushik, S., Roy, K., Singh, A., Mahendru, S., Kumar, M., Chaudhary, S., Ahmed, S., Kukreti, S., 2016. A bouquet of DNA structures: emerging diversity. Biochem. & Biophys. Rep. 5 (5), 388−395.

Kim, R., Guo, J.T., 2009. PDA: an automatic and comprehensive analysis program for protein-DNA complex structures. BMC Genom. 10 (Suppl. 1), S13.

Kirsanov, D.D., Zanegina, O.N., Aksianov, E.A., Spirin, S.A., Karyagina, A.S., Alexeevski, A.V., 2013. NPIDB: nucleic acid—protein interaction database. Nucleic Acids Res. 41, D517−D523. Database issue.

Kuang, X., Dhroso, A., Han, J.G., Shyu, C.R., Korkin, D., 2016. DOMMINO 2.0: Integrating Structurally Resolved Protein-, RNA-, and DNA-Mediated Macromolecular Interactions. Database (Oxford), 2016, bav114.

Kumar, K.K., Pugalenthi, G., Suganthan, P.N., 2009. DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. J. Biomol. Struct. Dyn. 26, 679−686.

Kuznetsov, I.B., Gou, Z., Li, R., Hwang, S., 2006. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins 64, 19−27.

Laitaoja, M., Valjakka, J., Jänis, J., Laitaoja, M., 2013. Zinc coordination spheres in protein structures. Norg. Chem. 52 (19), 10983−10991.

Larivière, L., Sommer, N., Moréra, S., 2005. Structural evidence of a passive baseflipping mechanism for AGT, an unusual GT-B glycosyltransferase. J. Mol. Biol. 352, 139−150.

Lee, J.Y., Yang, W., 2006. UvrD helicase unwinds DNA one base pair at a time by a two-part power stroke. Cell 127, 1349−1360.
le Cessie, S., van Houwelingen, J.C., 1992. Ridge estimators in logistic regression. Appl. Statist. 41, 191−201.
Li, B.Q., Feng, K.Y., Ding, J., Cai, Y.D., 2014. Predicting DNA-binding sites of proteins based on sequential and 3D structural information. Mol. Genet. Genom. 289, 489−499.
Li, T., Li, Q.Z., Liu, S., Fan, G.L., Zuo, Y.C., Peng, Y., 2013. PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics 29, 678−685.
Löwe, J., Ellonen, A., Allen, M.D., Atkinson, C., Sherratt, D.J., Grainge, I., 2008. Molecular mechanism of sequence-directed DNA loading and translocation by FtsK. Mol. Cell. 31, 498−509.
Luscombe, N.M., Austin, S.E., Berman, H.M., Thornton, J.M., 2000. An overview of the structures of protein−DNA complexes. Genome Biol. 1, 1−37.
Lu, X.J., Olson, W.K., 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 31 (17), 5108−5121.
Lu, X.J., Olson, W.K., 2008. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat. Protoc. 3 (7), 1213−1227.
Marinescu, V.D., Kohane, I.S., Riva, A., 2005a. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinf. 6 (79).
Marinescu, V.D., Kohane, I.S., Riva, A., 2005b. The MAPPER database: a multi-genome catalog of putative transcription factor binding sites. Nucleic Acids Res. 33, D91−D97. Database issue.
Matys, V., Fricke, E., Geffers, R., Gössling, E., et al., 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31, 374−378.
Mees, A., Klar, T., Gnau, P., Hennecke, U., Eker, A.P., Carell, T., Essen, L.O., 2004. Crystal structure of a photolyase bound to a CPD-like DNA lesion after in situ repair. Science 306, 1789−1793.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536−540.
Nandakumar, J., Nair, P.A., Shuman, S., 2007. Last stop on the road to repair: structure of *E. coli* DNA ligase bound to nicked DNA-adenylate. Mol. Cell. 26, 257−271.
Nimrod, G., Schushan, M., Szilagyi, A., Leslie, C., Ben-Tal, N., 2010. iDBPs: a web server for the identification of DNA binding proteins. Bioinformatics 26 (5), 692−693.
Norambuena, T., Melo, F., 2010. The protein-DNA interface database. BMC Bioinf. 11, 262.
Ofran, Y., Mysore, V., Rost, B., 2007. Prediction of DNA-binding residues from sequence. Bioinformatics 23, i347−i353.
Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH–a hierarchic classification of protein domain structures. Structure 5, 1093−1108.
Prabakaran, P., An, J., Gromiha, M.M., Selvaraj, S., Uedaira, H., Kono, H., Sarai, A., 2001. Thermodynamic database for protein−nucleic acid interactions (ProNIT). Bioinformatics 17, 1027−1034.

Prakash, A., Borgstahl, G.E.O., 2012. The structure and function of replication protein A in DNA replication. Subcell. Biochem. 62, 171–196. https://doi.org/10.1007/978-94-007-4572-8_10.

Perczel, A., Gáspári, Z., Csizmadia, I.G., 2005. Structure and stability of beta-pleated sheets. J. Comput. Chem. 26 (11), 1155–1168. https://doi.org/10.1002/jcc.20255.

Redinbo, M.R., Stewart, L., Kuhn, P., Champoux, J.J., Hol, W.G., 1998. Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. Science 279, 1504–1513.

Sam, M.D., Cascio, D., Johnson, R.C., Clubb, R.T., 2004. Crystal structure of the excisionase-DNA complex from bacteriophage lambda. J. Mol. Biol. 338, 229–240.

Sagendorf, J.M., Berman, H.M., Rohs, R., 2017. DNAproDB: an interactive tool for structural analysis of DNA–protein complexes. Nucleic Acids Res. 45, W89–W97. Web Server issue.

Schreiter, E.R., Drennan, C.L., 2007. Ribbon-helix-helix transcription factors: variations on a theme. Nat. Rev. Microbiol. 5 (9), 710–720. https://doi.org/10.1038/nrmicro1717.

Sinden, R.R., 1994. DNA Structure and Function. Academic Press, San Diego.

Si, J., Zhang, Z., Lin, B., Schroeder, M., Huang, B., 2011. MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. BMC Syst. Biol. 5 https://doi.org/10.1186/1752-0509-5-S1-S7.

Spiegel, J., Adhikari, S., Balasubramanian, S., 2020. The structure and function of. DNA G-Quadruplexes Trends Chem. 2, 2. https://doi.org/10.1016/j.trechm.2019.07.002.

Strogantsev, R., Ferguson-Smith, A.C., 2012. Proteins involved in establishment and maintenance of imprinted methylation marks. Brief. Funct. Genom. 11 (3), 227–239.

Stawiski, E.W., Gregoret, L.M., Mandel-Gutfreund, Y., 2003. Annotating nucleic acid-binding function based on protein structure. J. Mol. Biol. 326, 1065–1079.

Suhrer, S.J., Wiederstein, M., Gruber, M., Sippl, M.J., 2009. COPS–a novel workbench for explorations in fold space. Nucleic Acids Res. 37, W539–W544.

Szilagyi, A., Skolnick, J., 2006. Efficient prediction of nucleic acid binding function from low-resolution protein structures. J. Mol. Biol. 358, 922–933.

Tjong, H., Zhou, H.-X., 2007. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res. 35, 1465–1477.

Tsuchiya, Y., Kinoshita, K., Nakamura, H., 2004. PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. Bioinformatics 21, 1721–1723.

Van der Vliet, P.C., Verrizer, C.P., 1993. Bending of DNA by transcription factors. Bioessays 15 (1), 25–33.

Vapnik, V.N., 1998. Statistical Learning Theory. John Wiley and Sons, New York.

Verdaasdonk, J.S., Bloom, K., 2011. Centromeres: unique chromatin structures that drive chromosome segregation. Nat. Rev. Mol. Cell Biol. 12 (5), 320–332. https://doi.org/10.1038/nrm3107.

Walter, M.C., Rattei, T., Arnold, R., Guldener, U., Munsterkotter, M., Nenova, K., Kastenmuller, G., Tischler, P., Wolling, A., Volz, A., et al., 2009. PEDANT covers all complete RefSeq genomes. Nucleic Acids Res. 37, D408–D411.

Wang, L., Brown, S.J., 2006. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. 34, W243–W248.

Wang, L., Huang, C., Yang, M.Q., Yang, J.Y., 2010. BindN? for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst. Biol. 4, S3.

Ward, W.S., Coffey, D.S., 1991. DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells. Biol. Reprod. 44 (4), 569–574. https://doi.org/10.1095/biolreprod44.4.569.

Williams, R.J., 2003. Restriction endonucleases: classification, properties, and applications. Mol. Biotechnol. 23, 225–243.

Wingender, E., Dietze, P., Karas, H., Knüppel, R., 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 24 (1), 238–241.

Wingender, E., Chen, X., Fricke, E., Geffers, R., et al., 2001. The TRANSFAC system on gene expression regulation. Nucleic Acids Res. 29, 281–283.

Wu, J., Liu, H., Duan, X., Ding, Y., Wu, H., Bai, Y., Sun, X., 2009. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics 25, 30–35.

Xiong, Y., Xia, J., Zhang, W., Liu, J., 2011. Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures. PLoS One 6, e28440.

Yang, C.G., Yi, C., Duguid, E.M., Sullivan, C.T., Jian, X., Rice, P.A., He, C., 2008. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. Nature 452, 961–965.

Yang, Y., Ramelot, T.A., Lee, H.-W., Xiao, R., Everett, J.K., Montelione, G.T., Prestegard, J.H., Kennedy, M.A., 2014. Solution structure of the free Zα domain of human DLM-1 (ZBP1/DAI), a Z-DNA binding domain. J. Biomol. NMR 60, 189–195.

Yan, C., Terribilini, M., Wu, F., Jernigan, R.L., Dobbs, D., Honavar, V., 2006. Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinf. 7 https://doi.org/10.1186/1471-2105-7-262.

Zhu, J., Hastie, T., 2005. Kernel logistic regression and the import vector machine. J. Comput. Graph Stat. 14, 185–205.

Zhu, L.J., Christensen, R.G., Kazemian, M., et al., 2011. FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Res. 39, D111–D117.

CHAPTER

Computational cancer genomics

11

Aman Chandra Kaushik[1], **Shakti Sahi**[2]

[1]*Wuxi School of Medicine, Jiangnan University, Wuxi, Jiangsu, China;* [2]*School of Biotechnology, Gautam Buddha University, Greater Noida, Uttar Pradesh, India*

11.1 Introduction

Cancer is characterized by abnormal and uncontrolled cell growth infiltrating healthy tissues. These cancerous cells can metastasize and destroy other tissues and organs. Cancer genomics aims to understand the proliferation differences between normal and malignant cells, the underlying mutations, and the role of the immune system. It focuses on deciphering the genetic basis of tumor cells. There are different types of cancers, and an understanding of cancer's genetic basis is essential for effective therapeutic interventions. In the 20th century, the unregulated growth in sea urchin eggs was attributed to chromosomal aberrations. This observation highlighted the hypotheses of the role of chromosomes in cancer (Harris, 2008). The discovery of Philadelphia chromosomes confirmed the role of genetic alterations (Rowley, 1973; Nowell and Hungerford, 2004). Various studies have established that each cancer is unique and characterized by different cells having varied mutational spectra. Selected subclonal mutations are the driver mutations (Knudson, 1971; Nowell, 1976; ER, 1990).

Somatic evolution is dependent on the rate of mutation and clonal expansion. The different genomic regions have substantially different mutational rates. One of the distinguishing features of most cancer types is chromosomal instability caused by amplifications, deletions, translocations, or other structural changes. The relevance of genetic aberrations in cancer physiology is established by the discovery of various nucleotides responsible for the oncogenic phenotype (Reddy et al., 1982; Tabin et al., 1982; Taparowsky et al., 1982). For example, the v-*src* gene is an oncogene. Reports have shown that the v-*src* gene, responsible for cancer induction, is present in viruses that are oncogenic transforming like Rous sarcoma virus (RSV). Furthermore, studies on different mutants of RSV have shown that of all the genes present, only the v-*src* gene is involved in cancer. Similarly, hereditary colorectal cancer occurs due to mutations in the *POLD1* and *POLE* genes. These genes are known to encode DNA polymerases δ and ϵ.

The identification of genes involved in cancer using techniques like positioning cloning, candidate gene studies, and biological screening assays (Futreal et al., 2004) coupled with an understanding of the human genome (Consortium, 2001;

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00007-5

Venter et al., 2001) has facilitated deeper understanding at chromosomal structural levels. Cancer genetics studies have led to the identification of many genes (Bardelli et al., 2003; Davies et al., 2005; Stephens et al., 2005; Bignell et al., 2006; Wang et al., 2002) involved in different types of cancers. Large-scale genomics studies have been carried out to understand the genetic and epigenetic changes in cancer. Such studies have revealed the aberrations in genes leading to the development and growth of cancer and paved the way for better diagnostics and therapeutic interventions. One prominent example is the discovery of a common mutation in the *BRAF* gene in several types of cancer (Edwards et al., 2004). This resulted in the development of targeted Food and Drug Administration (FDA)-approved drugs vemurafenib and dabrafenib for the treatment of cancer patients having specific *BRAF*-V600E mutation in the *BRAF* gene (Hauschild et al., 2012). A comparison of genomic changes observed in different tumors has shown certain similarities. For example, the *HER2* mutations gene is known to occur across bladder, breast, pancreatic, and ovarian cancers. The capillary-based sequencing and targeting polymerase chain reaction techniques have led to the identification of cancer genes in colorectal cancers (Sjöblom et al., 2006), breast cancers, pancreatic cancers (Jones et al., 2008), and glioblastoma multiforme tumors (Parsons et al., 2008). The advances in sequencing technologies, including next-generation sequencing (NGS) coupled with computational data analysis, are revolutionizing our understanding of cancer biology. These have led to several novel targeted therapies for cancer treatment whose efficacy is dependent mainly on the mutation profile of tumors in patients (Jing et al., 2019). Fig. 11.1 depicts a strategy for the identification of

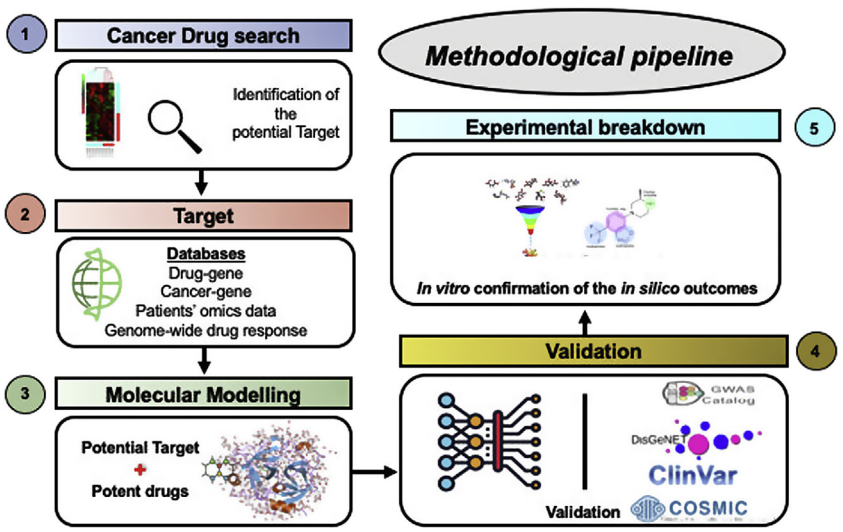


FIGURE 11.1

Strategies for identifications of anti-cancer drugs.

specific drugs for cancer therapy. Larotrectinib is an FDA-approved inhibitor of tropomyosin receptor kinases for any cancer with specific *TRK* gene fusion (Drilon et al., 2018).

11.2 Cancer genomics technologies

The investigation of cancer biology is growing faster due to the human genome's availability. Completion of the Human Genome Project (Collins et al., 2003) has revolutionized biomedical research and practices. The groundbreaking newer sequencing technologies are changing the way cancer genomics is being approached. The advantage is the improvement of DNA sequencing. The cost to generate a whole-genome sequence high-quality draft is reduced to below $1000. Massive parallel sequencing or NGS technologies, including pyrosequencing (Margulies et al., 2005), ligation (Shendure et al., 2005), sequencing by synthesis (Bentley et al., 2008), single polymerase (Eid et al., 2009), patterned nanoarrays (Drmanac et al., 2010), and semiconductor pH-based pyrosequencing (Rothberg et al., 2011) have led to sequencing multiple samples across genes. Newer methods have been developed for parallel target selection to target coding regions of multiple genes, for example, molecular inversion probes (Porreca et al., 2007), microarray-based genomic selection (Albert et al., 2007), and solution hybrid selection (Okou et al., 2007; Teer et al., 2010; Clark et al., 2011). NGS platforms can sequence millions of whole-genome fragments, producing a large number of short reads in a shorter time frame and at a reduced cost. The nucleotides are added and deleted in a sequential approach. However, samples from cancerous tissues pose a challenge to the analysis. First, samples of solid tumor cells are a mixture of healthy cells and tumor cells. Ideally, for sequencing, the majority of the sample should consist of tumor cells. Techniques like laser capture microdissection and flow sorting are used to extract genomic DNA/RNA from tumor cells. Second, formalin fixation and paraffin embedding preservation techniques are used for tissue examination. Third, tissues are available in a low amount. To counter this, NGS libraries are constructed from samples having low DNA content (Mardis, 2019).

The decoding of cancer exomes and genomes from a whole-genome library is possible with "hybrid capture" methods. Hybrid capture methods use synthetic DNA or RNA probes that are complementary to known coding sequences (Gnirke et al., 2009; Hodges et al., 2009; Bainbridge et al., 2010). The newer library construction techniques in NGS facilitate RNA sequencing (*RNA-seq*). These can identify the mutations in cancer cells from DNA, chromatin packaging, and other epigenomic mechanisms, and evaluate their expression.

11.3 Computational cancer genomics analysis

There is considerable diversity in the genetic abnormalities found within cancers of a single type. The process of identifying specific and rare mutations inducing cancer development and progression is a major challenge. The development of cell lines and animal models that can mimic human cancer diversity is also needed. NGS sequencing of genes in a high-throughput manner with low cost is done. The management and analysis of the enormous amounts of genomics data generated require efficient and robust computational algorithms and tools. The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have DNA sequences from many different types of tumors. Many computational methods have been developed to detect sequence reads in the genetic variation(s). These tools employ an algorithm with three steps: processing the read, mapping and alignment, and variant calling to identify variants/mutations.

11.3.1 Mapping and alignment

NGS technologies generate billions of overlapping short reads. These reads are assembled into contigs. Sequence alignment is one of the most commonly employed techniques for assembling these reads. Alignment helps in assembling the sequence, identifying its location in the genome, and understanding the differences compared to the normal genome.

A high degree of structural variation is observed in the cancer genome compared to the germline. Structural variation refers to changes in the genome sequence due to duplication, copy number variation, inversion, or translocation. The small structural variants (SVs) range from single base pair to 1 kb, whereas large-scale variants can include 1 million base pairs. These SVs can result in chromosomal aberrations and are significant markers for cancer. Therefore the alignment techniques employed should be able to detect these SVs. Identification of large-scale variants is a complicated task. The two commonly used alignment techniques are (1) reference assembly, and (2) de novo assembly.

Generally, reference-based alignment techniques are used where the analysis of variant is dependent on initial alignment of reads with a reference genome and then clustered using various methods. Reference-based alignment algorithms like Burrows–Wheeler Aligner (Burrows and Wheeler, 1994), Bowtie (Langmead et al., 2009), MOSAIK (Lee et al., 2014), and SOAP2 (Li et al., 2009a) work well for reads with fewer alignments. The Burrows–Wheeler transform is the conventional method for the data lossless compression technique. They have limited use in the case of reads having multiple alignments when there are high rates of variation.

De novo assembly is useful in identifying SVs and complex rearrangements. These assemblers are based on graph theory. They can be classified into three classes: (1) the Overlap-Layout-Consensus (OLC), (2) the de Bruijn graph (DBG) or Eulerian, and (3) the greedy graph algorithms. The greedy graph algorithms use

either OLC or DBG. In the de novo assembly, the equal regions are first identified, and then these regions are overlapped by fragmented sequenced ends. This method works very well for long reads; however, incorrect alignments may occur for short reads. It is computationally more intensive compared to the reference assembly.

11.3.2 *RNA-seq* data for pan-cancer

RNA-seq reads can be analyzed for single nucleotide polymorphisms (SNPs), fusion genes, transcript abundance, splice variants, and quantification of differential expression of isoforms. The *RNA-seqs* lack introns, making *RNA-seq* alignment challenging to classify all sequence read outputs. The intronless regions appear as large gaps in alignment with the reference genome. Many alignment methods have been developed, taking into consideration the absence of introns like GSNAP, MapSplice, SOAPsplice (Huang et al., 2011), STAR, CRAC, FineSplice (Gatto et al., 2014), and Tophat2 (Kim et al., 2013). Some of these methods have higher sensitivity and specificity. However, they all have splice−junction misalignment. The tools STAR and HISAT2 have improved accuracy. Both these tools differ in the methods to align the reads against genome assembly. A dataset of known splice sites is used for the identification of probable spliced sequencing reads. STAR (Dobin et al., 2013) does the alignment in two steps. First, it aligns the first region, known as "seed," for a specific read sequence. This "seed" is aligned to the maximum mappable length of the read against the reference genome. In the second step, the rest of the region, known as "second seed," is aligned to the maximum mappable length. In the next step, the two or more "seeds" are connected. The scoring is done based on mismatches, insertions, and deletions. The tool HISAT2 (Pertea et al., 2016) uses two indices for alignment-whole-genome Ferragina−Manzini (FM) index for "seed" and multiple overlapping local FM indices for extending the alignment.

RNA-seq is used to identify and quantify differentially expressed genes (Rapaport et al., 2013; Seyednasrollah et al., 2013; Soneson and Delorenzi, 2013). The programs *edgeR* and *DESeq2*, both R packages, have good accuracy for quantifying data and differential expression. They minimize the differences between the array and the sequence data. *DESeq2* (Anders and Huber, 2010) uses a generalized linear model for normalizing the count of each gene. Next, for the correction of dispersion and $\log_2$-fold change, it uses an empirical Bayes shrinkage method. The program *edgeR* (Robinson et al., 2010) estimates the ratio of RNA production using a trimmed mean of the log expression ratio.

The quantification of transcriptomic features is done by *RNA-seq* data analysis. Many programs have been developed for a comparative, relative, or differential abundance of *RNA-seq* isoform data. These methods examine either the read counts on each exon (DEXseq) or the exon−exon junction like ALEXA-seq, MISO, MATS, and SpliceSeq (Ryan et al., 2012; Shen et al., 2012; Aschoff et al., 2013; Ge et al., 2011; McPherson et al., 2011; Chen et al., 2012). Exon quantification pipelines are also available for both the alignment and the quantification steps of an *RNA-seq* workflow (Schuierer and Roma, 2016).

11.3.3 Databases

Large-scale genomic datasets are sequenced from tumor samples and experimental models. These datasets can be analyzed and reanalyzed to understand the mechanisms underlying the development and progression of different types of cancers.

Analysis of the genomic data involves the processing of raw data generated by experiments, normalization of data, analyses, and meaningful biological interpretation.

The Gene Expression Omnibus (GEO) and the NCI Genomic Data Commons are highly essential repositories for genomics data. NCI is a unified data repository for several cancer genome programs, including TCGA and Tumor Alterations Relevant for Genomics-Driven Therapy. The other significant cancer datasets are the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), the Genomics of Drug Sensitivity in Cancer (Garnett et al., 2012), the BROAD LINCS dataset (http://www.lincscloud.org/), the Fantom Consortium datasets (Forrest et al., 2014), the ENCyclopedia Of DNA Elements (ENCODE) project (ENCODE_Project_Consortium, 2012), and the Epigenome Roadmap (Kundaje et al., 2015). Oncomine is a database focusing on collecting, standardizing, analyzing cancer transcriptome data (Rhodes et al., 2007). UALCAN uses data from TCGA and is an interactive web portal facilitating the study of variations in gene expression. It also assesses survival associations (Chandrashekar et al., 2017). cBioPortal utilizes cancer gene sequencing data and provides visualization and analysis of data (Gao et al., 2013).

There are several databases containing information on cancer genes and their function. The Catalogue Of Somatic Mutations In Cancer (COSMIC) is a comprehensive resource for somatic mutations in human cancer (Tate et al., 2019). It currently has approximately 6 million coding mutations across 1.4 million tumor samples. The cancer gene census of COSMIC includes 719 genes, the types of mutations causing dysfunction of the gene, and details of the cancers in which increased frequency of mutations is observed. Out of the 719 genes, 554 are oncogenic or have tumor-promoting or -suppressing activity, depending on the tissue of origin, tumor stage, and various other factors. The cell lines project of COSMIC has mutation profiles of more than 1000 cell lines used in cancer research. COSMIC also comprises COSMIC-3D, a tool for understanding cancer mutations by mapping protein missense, in-frame deletion, and nonsense mutations to protein sequence and structure. The cancer mutation data can be correlated with known small-molecule binding sites, and druggable binding sites (Le Guilloux et al., 2009) that can help in the mutation-guided design of lead molecules to specific cancer mutants (Jubb et al., 2018).

The Atlas of Genetics and Cytogenetics is a database having information on cancer genes, genetic abnormalities, histopathology, and clinical diagnosis (Huret et al., 2013). Such information is helpful to clinicians and the pharmaceutical industry for developing a therapy. The Network of Cancer Genes is a catalog of 2372 known or predicted cancer driver genes (Repana et al., 2019) identified from cancer sequencing screens. It has in-depth information on the various distinguishing

features of somatic mutations in these driver genes, including duplicability, origin, RNA and protein expression, miRNA and protein interactions, and protein function. It has wide application in the identification and validation of cancer genes and miRNA biomarkers (Andres-Leon et al., 2017; Jäkel et al., 2017; Xion et al., 2017; Woollard et al., 2016). The Cancer 3D database includes about 1,457,702 mutations from 9079 samples from 32 different cancer types (Sedova et al., 2019). These mutations have been mapped to 18,425 unique proteins. This database is useful for analyzing patterns of cancer missense mutations occurring in the proteins and their correlation to patients' clinical data. The functional modules of a protein coded by gene and mutation positions in the protein can be studied. The interactive domain and significant protein−protein interactions can be assessed.

Databases like CCLE focus primarily on therapeutic agents for cancer. The CCLE is a database of gene expression, genotype, and drug sensitivity data for human cancer cell lines. It has wide applications in the systematic interpretation of single nucleotide variants (SNVs), copy number aberrations (CNAs), and mRNA expression and utilizing them to develop therapeutics. It helps determine genetic variables leading to drug resistance, drug awareness, and genomics of drug susceptibility in cancer. Recently, about 1072 cancer cell lines have been studied to extract information regarding RNA splicing, DNA methylation, histone H3 modification, microRNA expression, and reverse-phase protein array. The data obtained was integrated with drug sensitivity data, short hairpin RNA knockdown, and CRISPR−Cas9 knockout data. The correlation helped identify potential targets and associated biomarkers (Ghandi et al., 2019). The Drug-Gene Interaction Database (DGIdb) is one of the largest databases for drug−gene interactions and potential gene druggability (Cotto et al., 2018).

The crossing of information for analysis from one database to another is not an easy task. Each database or resource uses different identifiers. The datasets are signified in a tabular format in these databases for genome analysis. The de facto standard and shared resources of the database are termed tab-separated value files. These files have advantages in computation as they are easier to read and write, and are straightforward. The contents from the file typically grip the columns. The proteins and the drugs targeting germline differences and the diseases regulate their transcription. The experimental databases, such as the Gene Expression Omnibus and Array Express, contain statistics from the microarray experiments.

Different identifiers have been used in different databases for very similar entries. For example, genes in Entrez and Ensembl have different identifiers for similar entities. These resources borrow identifiers from *HGNC* gene symbols. *KEGG* and its library have their identifiers for genes. These offer maps for gene symbols. Translating identifiers and their incompatibilities exist between resources, i.e., MutationAssessor predicts protein mutations, pathogenicity, and uses identifiers of UniProt. Analysis systems use Ensembl data for mutations using Ensembl Protein IDs. Coordinate mappings, and in some cases translating identifiers, are used for defining mutations to the incorrect isoforms. Incorrect predictions help to reduce the number of predictions. One of the crucial tasks in computational biology is to

analyze and correlate data from various resources. A sound translation coding system or program is required for meaningful data integration.

Pharmacological evidence combined with the genomic data in these databases could be a significant instrument for clinicians to convert new findings into medicines. A comparative study of seven databases of variants in cancer has been reported (Pallarz et al., 2019). They are compared in terms of genes, drugs, and gene−drug associations. The study revealed that although many of the databases had largely overlapping information, each also had its specific features. Therefore a comprehensive analysis of data from multiple databases should be used for developing precision medicine.

11.3.4 Genomics landscape for oncogenic mutations

The economical cost and shorter time needed to sequence a sample are advantageous as multiple samples from different types of tumors can be analyzed for common and rare variants. Consortiums like *TCGA* and *ICGC* catalog and characterize the common somatic genetic alterations in different types of tumors (Uhlén et al., 2015; Kotelnikova et al., 2016; Nagarajan et al., 2019). A variety of mutational patterns such as kataegis, chromothripsis, and chromoplexy result in complex germline and somatic SVs and are key signatures of cancer (Pellestor, 2019). The genetic variations observed in cancer are categorized as germline and somatic mutations. Many programs have been developed for germline and somatic variant calling.

11.3.4.1 Germline mutations

The germline or inherited mutations are analyzed by programs using the Bayesian model such as SAMtools (Li et al., 2009b), GATK (DePristo et al., 2011), MPG (Garrison and Marth, 2012), and FreeBayes (Walsh et al., 2010). Healthy tissues are generally used to identify inherited variants. Many studies Walsh et al. (2010) (Johnston et al., 2012; Chang et al., 2013; Kanchi et al., 2014) examining germline variants have been reported. The germline variant algorithms have a 50% or 100% allele frequency rate. These tools can also be used for somatic variants; however, analysis of somatic variants poses various challenges. It is challenging to accurately separate somatic variants from inherited variants without a matched normal sample.

11.3.4.2 Somatic mutations

The various kinds of somatic mutation range from SNVs, to bigger CNAs (>50 bp), to tiny insertions and deletions (indels). These genomic changes are examined by low-performance methods, including targeted gene sequencing, cytogenetic methods, systemic mutagenesis, and DNA linkage assessment. One of the significant challenges associated with somatic variant calling methods is distinguishing between the variants with low frequency precisely. Usually, to distinguish between the germline variant and somatic variant, both the tumor sample and normal sample are sequenced, taken from the same individual. The variants that are detected only in the tumor samples are referred to as somatic mutations, whereas the variants

observed both in tumor sample as well as in normal sample are referred to as inherited mutations. Such an approach enhances the sensitivity and specificity. Some of the software based on this approach are VarScan/VarScan 2 (Koboldt et al., 2009, 2012), Strelka (Saunders et al., 2012), Somatic Sniper (Larson et al., 2011; Cibulskis et al., 2013), MuTect, and Shimmer (Hansen et al., 2013). There are many algorithms based on machine learning frameworks: e.g., MutationSeq (Ding et al., 2012), SomaticSeq (Fang et al., 2015), SNooPer (Spinella et al., 2016), and BAYSIC (Cantarel et al., 2014). The genetic variants are grouped into (1) insertion and deletion, (2) SVs like duplication, translocation, copy number variation, and (3) SNV. The analysis of each of these types of variants requires specific algorithms. However, the algorithms using a minimum number of variant callers are used to analyze all such genetic variants. In the case of SNVs and short indels, non-reference nucleotide bases are checked from the sequences that cover each position. The reads are too short in case of SVs, and long indels are present; therefore the algorithms use the patterns of misalignment with paired-end reads to detect the breakpoints. Split reads assembly and de novo methods are used for the analysis of SVs.

11.3.4.2.1 Somatic mutations in pan-cancer

Numerous studies utilize data from TCGA and report a pan-cancer analysis of somatic mutations (Narayan et al., 2016; Kandoth et al., 2013). Significantly, mutated genes with high confidence levels have been reported (Tomczak et al., 2015). Such studies have shown that tumors have distinct mutational profiles and identified genes within and across all specific types of tumors. Co-occurrence and mutual exclusivity tests were performed for mutated gene pairs. Mutation status with clinical outcomes was correlated across tumors. The copy number, mutation, and DNA methylation data have been used to classify different subclasses of tumors (Ciriello et al., 2013). Their study resulted in the identification of significant functional mutations and subclasses based on alteration signatures. Such a hypothesis supports that based on the alteration profile, and some combinations of drugs may show good activity across different types of tumors. Using data from 21 types of tumor, 224 cancer driver genes have been reported for one or more tumor type (Lawrence et al., 2014). They were able to identify additional genes that could not be detected through individual analysis.

Patterns of nucleotide changes within tumors have been studied (Alexandrov et al., 2013). In this study, 21 single-nucleotide mutational signatures and their flanking bases were identified. Characterization of mutational signatures is significant for the understanding of cancer genomics. Studies linking mutation profile to a potential cause have been done. The different signatures have been co-related in cancer samples. The "localized substitution hypermutations" were also observed (Nik-Zainal et al., 2012; Burns et al., 2013; Roberts et al., 2013).

11.3.5 Noncoding mutations

Cancer genomics studies are also focusing on noncoding mutations causing cancer. The noncoding region comprises more than 98% of the genome. The majority of the somatic mutations lie in the noncoding region. The population genomics model using 1000 genomes and ENCODE project data identified functionally important noncoding regions and rare polymorphism enrichment. These suggest an approach that effectively identifies cancer driver genes. Transcription factors (TFs) were found to contain many disease variants and binding motifs for specific TF families (Supek et al., 2014).

The cancer drivers have been identified in regulatory regions like promoters and enhancers. The mutations in the *TERT* gene promoter (Huang et al., 2013; Vinagre et al., 2013) indicated the role of mutations in the noncoding region in cancer progression. Telomerase reverse transcriptase is repressed in healthy somatic cells and catalyzes the lengthening of telomeres. In acute lymphoblastic leukemia within the enhancer of *TAL1*, recurrent noncoding mutations have been reported (Mansour et al., 2014). Furthermore, ChIP-seq data indicated that *TAL1* enhancer mutations create new binding sites that are important for the binding of MYB. So noncoding mutations affect gene expression and may create novel pathways by altering the transcriptional networks. In chronic lymphocytic leukemia, mutations in a potential enhancer region close to the *PAX5* gene involved in B-cell differentiation have been reported (Puente et al., 2015).

Noncoding mutations in untranslated regions (UTRs) are likely to contain cancer driver genes. The 3′ UTR mutations of CD274 are reported to disrupt miRNA-mediated degradation of the mRNA transcript, leading to overexpression of CD274 in gastric cancer (Wang et al., 2012). In melanoma, 5′ UTR mutations in the gene *RPS27* are known to occur (Dutton-Regester et al., 2014).

Noncoding mutations in functional RNA molecules, such as miRNAs and some long noncoding RNAs (lncRNAs), are significant. The miRNA mutations, both somatic and germline, are known to drive cancer (Wojcicka et al., 2014). MALAT1, a lncRNA, has shown mutations in estrogen receptor-positive breast cancer (Ellis et al., 2012).

The noncoding cancer mutations can be detected by (1) annotation-based methods, (2) rate-based methods, and (3) correlation-based methods. They are used for identifying somatic mutations by mapping them to regions of interest. Tools like LARVA and FunSeq2 are used for detecting noncoding mutations.

Coding and noncoding mutation information needs to be correlated to decode how somatic mutations cause cancer progression.

11.3.6 Variant annotation

Many tools have been developed for variant annotation. They catalog various mutations and calculate the frequencies of these mutations across different samples and different types of tumors. These help in identification of those positions or genes that

are frequently more mutated than expected. Studies have revealed that some somatic mutations are functional, while some are incidental mutations. It is the functional mutations that drive the oncogenic process. There are methods for detecting potential functional mutations by finding positions with a higher frequency of mutation. Some of the tools based on this include MuSiC, MutSig, and Gistic (Mermel et al., 2011; Dees et al., 2012; Lawrence et al., 2013). Other methods employ nonrecurrence approaches such as Multi-Dendrix to identify pathways based on mutated genes in patients. Oncodrive-FM (Gonzalez-Perez and Lopez-Bigas, 2012; Leiserson et al., 2013) identifies the potential driver genes based on mutations having a considerable effect on function. OncodriveCLUST uses localized mutation clusters to find specific driver genes (Reimand and Bader, 2013).

COSMIC (Forbes et al., 2008), TCGA, and ICGC (Lathrop et al., 2010) are used for the determination of variants and their frequencies observed in different types of tumor. A better understanding of variants' potential impact can be determined with datasets of genotype−phenotype relation. Online Mendelian Inheritance in Man, ClinVar (Landrum et al., 2013), Human Gene Mutation Database (Stenson et al., 2014), and My Cancer Genome are some of the databases. The DGIdb (Griffith et al., 2013) database of drug−gene interactions leads to a better understanding of function and therapeutic relevance. The various tools for the prediction of function based on mutations include SIFT (Ng and Henikoff, 2003; Bromberg and Rost, 2007), SNAP, PolyPhen2 (Adzhubei et al., 2010), CHASM (Carter et al., 2009), CHASMPlus (Tokheim and Karchin, 2019), mCluster, and transFIC (Gnad et al., 2013).

11.3.7 Structural variants

SVs, inversions, deletions, duplications, and translocations are distinguishing markers of cancer. These critical mutational rearrangements delete, amplify, or reorder genomic fragments. SVs can disrupt gene function and regulate gene expression. There are several methods based on the type of information they utilize. Copy number variants (CNVs) are a subtype of SVs and include deletions and duplications. The read depth differences are used in identifying CNVs. In a genomic sequence, the read depth data are homogeneous. The shift in values from mean depth enables the identification of CNVs. Tools like RDXplorer (Yoon et al., 2009) and CNVnator (Abyzov et al., 2011) are based on read depth differences for CNV detection. In targeted sequencing, heterogeneity is observed in read depth over different regions. In such cases, CNVs are detected by comparing the read depth of tumor sample against normal samples using tools like ExomeCNV (Sathirapongsasuti et al., 2011) and VarScan2 (Koboldt et al., 2012). CoNIFER and XHMM are programs for singular value decomposition used for normalizing target regions (Fromer et al., 2012; Krumm et al., 2012).

De novo assembly is the preferred method for the identification of SVs. They have the advantage that they can detect larger insertions. The limitation of standard de novo assembly is that it represents only one haplotype, thereby missing heterozygous SVs.

Cortex is a method that uses combinations of SNVs, indels, and rearrangements to detect SVs by using DBG (Iqbal et al., 2012). The SGVar (Tian et al., 2018) method uses a string graph-based de novo assembly pipeline and also uses short reads. It makes use of the read length and read quality and has better results for identifying insertion and deletion. Tools like BlasR (Chaisson and Tesler, 2012), MUMmmer (Delcher et al., 1999), or Minimap2 (Li, 2018) utilize previously assembled contigs and scaffolds. The tool DELLY (Rausch et al., 2012) analyzes the split reads for detecting abnormal distances and orientations among pairs of reads. The efficiency is enhanced for the detection of smaller deletions. LUMPY (Layer et al., 2014) and Manta (Chen et al., 2016) analyze the read depth, paired-end read, and split reads. They can be used on single samples and also for comparison with a tumor sample. They build graphs across regions and identify specific variations. The tool TARDIS (Soylev et al., 2019) can detect tandem duplication.

All of these methods are suitable for detecting specific variants. However, the accuracy rate varies for different types and sizes of SVs. Meta-methods combine features from different tools using varied methods. Thereby, using multiple methods, variants can be detected. Tools like MetaSV (Mohiyuddin et al., 2015), Parliament2 (Zarate et al., 2018), and SURVIVOR (Jeffares et al., 2017) give multiple types of the variant compared to single variant calling methods.

11.4 Pathway analysis

Various callers and instruments to predict the functional effect of mutations only focus on individual genes, mutations, and functional impacts of their DNA. However, genes do not operate in isolation but communicate by complicated cellular responses that change their regular patterns in cancer. They are structured into organizations, often called pathways, based on these relationships. Pathway analysis of high-throughput data is an essential tool for understanding the pathways being regulated. The association between genes or proteins needs to be measured. The confidence scores retrieved from protein−protein interaction databases have recently been reported for association studies (García-Campos et al., 2015). Various methods have been reported to identify significant pathways from high-throughput biological data. Somatic transformations are interpreted by comparing pathways involving variants with recognized pathway databases. The overlap of mutated genes and genes with known functional notations can be calculated, and the probability of their occurrence is assessed by statistical measures such as the exact Fischer or hypergeometric tests (Lai et al., 2017).

Another widely used strategy, Gene Set Enrichment Analysis (GSEA), determines whether a defined set of genes shows statistically significant, concordant differences between two biological states such as a healthy and diseased state. GSEA works on groups of genes rather than a single gene. GSEA has been used to screen the common pathways and differentially expressed genes in lung cancer at the

transcriptional level (He et al., 2019). Four datasets of lung adenocarcinoma from GEO were analyzed using the GSEA approach. The analysis led to the identification of crucial genes and pathways involved in cell cycle and DNA replication. Reactome is another effective method for high-performance pathway analysis. It uses in-memory data structures and algorithms for genome-wide dataset analysis in a short time span (Fabregat et al., 2017).

One of the most commonly used databases for enrichment assessment is Gene Ontology (GO). It comprises three ontologies that are hierarchically organized to define protein in terms of related biological procedures and cellular and molecular activities (GO terms). By considering them independently, the simple approach to measuring the enrichment of GO terms cannot account for GO's hierarchical framework. Pathway enrichment testing has been used to characterize transcriptional subpopulations from single-cell *RNA-seq* data (Fan, 2019). The Goeman and Mansmann test technique (Goeman and Mansmann, 2008) was suggested to preserve the GO chart's composition. It needs an individual to select a focus level in the GO chart representing the specific number of terms selected by the user. The collective sets of mutated genes in a biological pathway contribute to the development of a tumor. In such cases, the algorithm must be able to differentiate between highly mutated genes and those that show very few mutations. To resolve these, evaluation techniques are used for screening patient-related genes (for example, mutated genes) versus recognized pathways to detect mutated pathways in all patients. PathScan is one such statistical method that considers the variations in gene lengths within a pathway and also distribution probabilities of mutations among samples for significance test (Wendell et al., 2011). Thus the results are biologically more relevant. It uses mathematical concepts of convolution and Fisher–Lancaster theory. Likewise, Boca et al. calculated enhancement results for perpetual mutation and combined them with the general classification. Various methods have been developed to integrate known biological interactions. These methods have led to the improved performance of network inference and better differential dependency network approaches. A computational network-based method known as Evaluation of Differential DependencY (EDDY) combines GSEA's gene-set-assisted advantages and differential network dependency for identifying biological associations in pathways. The method was applied to gene expression data of 202 glioblastoma samples to identify pathways enriched with differential dependency (Speyer et al., 2016). One of the major problems in identifying subtype-specific drug vulnerabilities is to assess how the key signaling networks are affected by genetic alterations and gene expression. With the application of EDDY, the subtype-specific network and gene dependencies were identified in glioblastoma samples. The results showed 57 pathways with a statistically significant divergence between mesenchymal and nonmesenchymal samples. These results have applications in the identification of subtype-specific drug vulnerabilities.

With the development of newer methods, it is becoming possible to integrate various pathways without analyzing individual pathways. Moreover, not every gene is similarly essential for a pathway, and the topology of the relationships

that can catch the dependency between the genes on a particular trajectory is not considered. Cross-talking of various paths has to be taken into consideration while developing new methods.

11.5 Network analysis

In contrast to processes that evaluate pathways with well-established features, interaction networks utilize network-based methods to infer new cancer genes and pathways. Gene functional similarity networks have found wide application in predicting protein−protein interactions, cellular localization, and identifying genes involved in diseases. A refined gene functional similarity network method has been proposed. Gene−gene association networks were created from the Protein-Protein Interaction data (Tian et al., 2017). Protein networks can be either undirected (physical protein−protein relationships) or guided (high-level functional relationships). While most present methods use undirected networks, the use of guided networks is significant since the various relationships leading to cancer development can be demonstrated. The identification of biomarkers from microarray gene expression data of breast cancer was done using network analysis (Khunlertgit and Yoon, 2016). The progression of cancer involves the dysregulation of multiple genetic processes. Therefore genes present in common pathways or the protein coded by these genes known to be functionally related in protein−protein interaction networks should be treated as a single feature. Various association coefficients may be applied to estimate the topological similarity. These topological attributes are known to enhance the prediction of potential subnetwork markers, which can be helpful in the prediction of cancer prognosis. Reactome includes a human protein network of functional interacting proteins, gene coexpression, protein−field relationships, and other sources. Either experimentally examined and deemed more confident, or computationally based, the changes in the protein communication networks are predictable. Examples of databases containing relationships include HPRD and BioGRID. Other databases, such as STRING, comprise predicted and experimentally determined protein−protein interactions, which are either direct (physical) or indirect (functional) associations. KEGG and Reactome are reliable resources for pathway information. iRefWeb covers the largest full network of interaction protein because it integrates protein interaction data from 10 different interaction databases: BioGRID, BIND, CORUM, DIP, HPRD, INTACT, MINT, MPPI, MPACT, and OPHID. iRefWeb is an interface to a relational database.

Various techniques have been suggested incorporating somatic mutations with communication networks to identify interaction populations of mutated DNA. The concept behind these techniques is that mutations in the DNA vary from person to person with the same disease form, but the cells impacted engage in the same biological procedures. Any disease is generally not due to an abnormality of a gene, protein, or cell but as a result of the interactions of genes, proteins, or cells in a complex network. The network biomarkers and dynamic network biomarkers with

protein−protein or gene−gene interactions are useful in studying the progression of cancer. Identification of cancer subtypes predictor of clinical results such as patient survival and treatment reaction using the network-based stratification (NBS) methodology has been reported. NBS is less precise in showing the significance of biological networks in operation when clustering without network data (Chen et al., 2015; Papanikolaou et al., 2015). Similarly, with three distinct interaction networks, Leiserson et al. and Vandin et al. conducted a TCGA pan-canner assessment of 12 cancer types. Its technique, HotNet2, utilizes network propagation to detect heavily linked parts of aberrant DNA and a statistical experiment for assessing the importance of subnet numbers and sizes.

11.5.1 Data integration and methodological combination

By applying several information sources via information inclusion, cancer genome information can be better predicted and interpreted. The pipelines, like Mutex, MEMo, and MEMCover, include multiple information sources and combine distinct approaches. In particular, both network and shared exclusivity analyzes are conducted. MEMo detects genetically aberrant cliques in a protein−protein network.

MEMCover, on the other hand, in its first stage detects mutually exclusive trends of mutations in many tissue kinds and then utilizes interaction information to evaluate the possibility of fresh pan-cancer-dysregulated subnetworks for identified, mutually exclusive communities. The combination of various methodologies is especially helpful in producing interpretable listings of cancer genes or clusters. For instance, the memo-derived modules identify genes that are mutually exclusive and communicating, thus enabling their biological interpretation and validation tests to be designed. The inclusion of multiple kinds of carcinogenic information like contact nets, expression of mRNA, abundance of phosphoproteome, genetic aberrations, and microRNAs may, on the whole, provide an insight into the fundamental molecular processes of cancer.

In addition to network assessment, identification of combinatorial models is another successful strategy to detect cancer genes. The methods for cancer driver mutations and pathways use known pathways, network information, or de novo techniques for identifying pathways (Dimitrakopoulos and Beerenwinkel, 2017). The mutually exclusive gene alterations can be identified in a given set of genomic profiles (Babur et al., 2015). Co-occurring mutations, by comparison, show favorable mutations in the gene. There are simultaneous mutations in two or more genes to gain a competitive benefit for the cell. The RME algorithm detects gene modules whose components are mutated repeatedly and shows mutually exclusive models. Besides, mutually exclusive occurrences of unusual genes tend to occur by chance, which is why they are harder to identify.

MEMo utilizes a statistical permutation examination that permits mutated genes between the specimens for mutual exclusion between genetic aberrations. The permutation experiment is conducted on gene communities identified as cliques from a network of protein interactions. The MEMo, therefore, incorporates various sources of information and incorporates various methodologies.

Dendrix (De novo driver exclusivity) and Multi-Dendrix are the methods that solve the restriction caused by unusual errors. By its elevated exclusivity and high visibility, Dendrix recognizes driver paths. In contrast to RME, instead of each gene individually, Dendrix requires a substantial level of coverage of the gene components found. Multi-Dendrix simulcast is a linear integer programming strategy to identify various mutually exclusive gene sets. Genome-scale information is much faster than Dendrix, and mutually exclusive mutations have been recognized in well-studied cancer processes, like p53 and PI3K/AKT. Mutex and CoMEt are two of the latest techniques for the identification of reciprocal genomic occurrences. By establishing a shared exclusivity rule that prevents significant imbalances in the contribution of each gene to the general shared exclusivity model, Mutex detects clusters of chromosomes with a downstream impact on a signaling network. To accomplish this, each gene is screened against the association of the other community changes for shared exclusivity (Zhang and Zhang, 2018).

CoMEt conducts a precise statistic experiment on the frequency of each modification and therefore can identify both unusual mutations more efficiently and various sets of mutually exclusive changes in conjunction, which can overlap, differ in magnitude, and relate to various types of disease. Compared with Dendrix, Multi-Dendrix, MEMo, and RME, Mutex and CoMEt have shown enhanced efficiency in forecasting mutually excluding occurrences. Progress of cancer may be considered to accumulate mutations in various genes. Nevertheless, more solid models can be obtained primarily due to the large-scale gene-mutational heterogeneity by considering dependencies between modified processes. The concept is that developmental limitations among pathways are explicitly taken into consideration, which would otherwise confuse the identification of the mutually exclusive gene communities. Candidate cancer genes and pathways can be suggested by combinatorial mutational models unbiased without any previous understanding. The combined approach can also provide an understanding of the genes' functional relationships and play an essential role in developing drugs for targeted therapy. Fig. 11.2 shows an integrated approach for targeted therapy.

11.5.2 Software resources (workflow and visualization interfaces)

The analysis of cancer genome data involves several tasks that necessitate the use of secondary software to support the analysis in a cancer pipeline. This software could be used for data mining searches of germline mutations, SNPs, and identifications of protein−protein interaction subnetworks. Many cancer data analysis pipelines have in-built tools for such analysis or may be using third-party software for these analyses. The workflow of cancer data analysis uses web services, local applications, browser-based applications, command-line tools, or application programming interfaces (APIs). Many of the resources accept data in multiple ways. For example, in Ensembl, the data can be accessed using the web interface, FTP server, or through the PERL API.

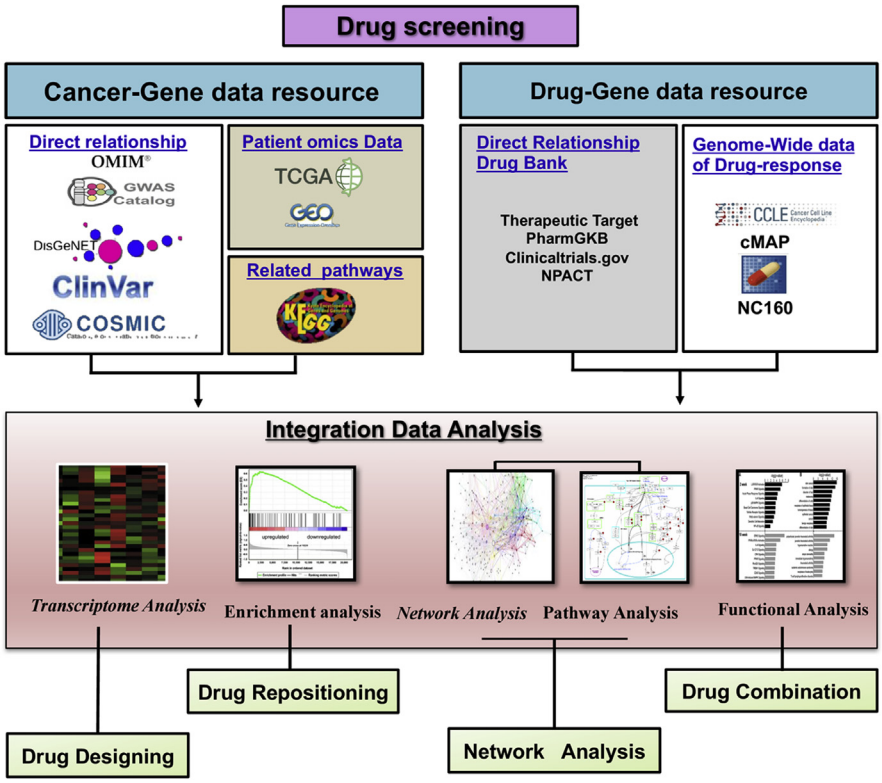


FIGURE 11.2

Integrated approach for targeted cancer therapy.

Any cancer genome analysis involves either a single dataset or multiple datasets; therefore the pipeline requires many interconnected analysis steps. Consequently, in the pipeline, analytical steps should be scripted so that all are interlinked, and reproducible results are produced. A comprehensive analysis should be carried out, and there should be features in the workflow to adapt to specific data requirement analyses from different experiments. Systems like Taverna and Galaxy have been designed to build pipelines with visual interfaces, including a range of functionalities.

11.6 Conclusion

A large amount of data is generated by the newer sequencing technologies of the second and third generations. The analysis involves the application of sophisticated methods. These data, from multiple samples and different types of tumors, are heterogeneous. Heterogeneity of the data, coupled with the disparity of the software

implementations, adds to the complexity in analysis for producing meaningful biological results. Computational cancer genomics applies algorithms and statistical models to the datasets. Considerable progress is being made in cancer genome analysis systems with newer and better algorithms to manage the complexity, taking into consideration-specific characteristics of each analysis. Computational cancer genomics through the development of computational methods and tools, and utilizing platforms, datasets, and resources, aim to help in the deep understanding of cancer biology. These methods and tools are used to analyze cancer genomics data across populations to identify genes, regions, and pathways that are altered and subtypes of the disease. Numerous tools are available for detecting somatic mutations and structural variants. As specific pathways are capable of complex rewiring between conditions, methods involving pathway analysis and network-based analyses are highly useful. While the findings of computational techniques are essential for understanding cancer, it is vital to integrate them with experimental approaches to generate meaningful and interpretable results.

References

Abyzov, A., Urban, A.E., Snyder, M., Gerstein, M., 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21, 974–984.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., 2007. Direct selection of human genomic loci by microarray hybridization. Nat. Methods 4, 903.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., Boyault, S., Burkhardt, B., Butler, A.P., Caldas, C., Davies, H.R., Desmedt, C., Eils, R., Eyfjord, J.E., Foekens, J.A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Jager, N., Jones, D.T., Jones, D., Knappskog, S., Kool, M., Lakhani, S.R., Lopez-Otin, C., Martin, S., Munshi, N.C., Nakamura, H., Northcott, P.A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J.V., Puente, X.S., Raine, K., Ramakrishna, M., Richardson, A.L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T.N., Span, P.N., Teague, J.W., Totoki, Y., Tutt, A.N., Valdes-Mas, R., Van Buuren, M.M., Van 'T Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L.R., Zucman-Rossi, J., Futreal, P.A., Mcdermott, U., Lichter, P., Meyerson, M., Grimmond, S.M., Siebert, R., Campo, E., Shibata, T., Pfister, S.M., Campbell, P.J., Stratton, M.R., 2013. Signatures of mutational processes in human cancer. Nature 500, 415–421.

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol. 11, R106, 110.

Andres-Leon, E., Cases, I., Alonso, S., Rojas, A.M., 2017. Novel miRNA-mRNA interactions conserved in essential cancer pathways. Sci. Rep. 7, 46101.

Aschoff, M., Hotz-Wagenblatt, A., Glatting, K.-H., Fischer, M., Eils, R., König, R., 2013. SplicingCompass: differential splicing detection using RNA-seq data. Bioinformatics 29, 1141−1148.

Babur, Ö., Gönen, M., Aksoy, B.A., Schultz, N., Ciriello, G., Sander, C., Demir, E., 2015. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. Genome Biol. 16 (1), 45.

Bainbridge, M.N., Wang, M., Burgess, D.L., Kovar, C., Rodesch, M.J., D'Ascenzo, M., Kitzman, J., Wu, Y.Q., Newsham, I., Richmond, T.A., Jeddeloh, J.A., Muzny, D., Albert, T.J., Gibbs, R.A., 2010. Whole exome capture in solution with 3 Gbp of data. Genome Biol. 11 (6). R62.

Bardelli, A., Parsons, D.W., Silliman, N., Ptak, J., Szabo, S., Saha, S., Markowitz, S., Willson, J.K., Parmigiani, G., Kinzler, K.W., 2003. Mutational analysis of the tyrosine kinome in colorectal cancers. Science 300, 949−949.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M.F., Monahan, J.E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F.A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I.H., Cheng, J., Yu, G.K., Yu, J., Aspesi Jr., P., de Silva, M., Jagtap, K., Jones, M.D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R.C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J.P., Gabriel, S.B., Getz, G., Ardlie, K., Chan, V., Myer, V.E., Weber, B.L., Porter, J., Warmuth, M., Finan, P., Harris, J.L., Meyerson, M., Golub, T.R., Morrissey, M.P., Sellers, W.R., Schlegel, R., Garraway, L.A., 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483 (7391), 603−607.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53.

Bignell, G., Smith, R., Hunter, C., Stephens, P., Davies, H., Greenman, C., Teague, J., Butler, A., Edkins, S., Stevens, C., 2006. Sequence analysis of the protein kinase gene family in human testicular germ-cell tumors of adolescents and adults. Gene Chromosom. Cancer 45, 42−46.

Bromberg, Y., Rost, B., 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 35, 3823−3835.

Burns, M.B., Lackey, L., Carpenter, M.A., Rathore, A., Land, A.M., Leonard, B., Refsland, E.W., Kotandeniya, D., Tretyakova, N., Nikas, J.B., Yee, D., Temiz, N.A., Donohue, D.E., Mcdougle, R.M., Brown, W.L., Law, E.K., Harris, R.S., 2013. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature 494, 366−370.

Burrows, M., Wheeler, D.J., 1994. A block-sorting lossless data compression algorithm. Tech. Rep. 124. Digital Equipment Corporation, 1994.

Cantarel, B.L., Weaver, D., McNeill, N., Zhang, J., Mackey, A.J., Reese, J., 2014. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinf. 12 (15), 104.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Karchin, R., 2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 69, 6660−6667.

Chaisson, M.J., Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinf. 13, 238.

Chandrashekar, D., Bashel, B., Balasubramanya, S., Creighton, C., Ponce-Rodriguez, I., Chakravarthi, B., Varambally, S., 2017. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia 19, 649−658.

Chang, V.Y., Basso, G., Sakamoto, K.M., Nelson, S.F., 2013. Identification of somatic and germline mutations using whole exome sequencing of congenital acute lymphoblastic leukemia. BMC Cancer 13, 55.

Chen, K., Wallis, J.W., Kandoth, C., Kalicki−Veizer, J.M., Mungall, K.L., Mungall, A.J., Jones, S.J., Marra, M.A., Ley, T.J., Mardis, E.R., 2012. BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. Bioinformatics 28, 1923−1924.

Chen, H., Zhu, Z., Zhu, Y., Wang, J., Mei, Y., Cheng, Y., 2015. Pathway mapping and development of disease-specific biomarkers: protein-based network biomarkers. J. Cell Mol. Med. 19 (2), 297−314.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A.J., Kruglyak, S., Saunders, C.T., 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220−1222.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G., 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., Sande, t, 2013. Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. 45 (10), 1127−33.

Clark, M.J., Chen, R., Lam, H.Y., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J., Snyder, M., 2011. Performance comparison of exome DNA sequencing technologies. Nat. Biotechnol. 29, 908.

Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S., 2003. A vision for the future of genomics research. Nature 422, 835.

Consortium, I.H.G.S., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860.

Cotto, K.C., Wagner, A.H., Feng, Y.Y., Kiwala, S., Coffman, A.C., Spies, G., Wollam, A., Spies, N.C., Griffith, O.L., Griffith, M., 2018. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. Nucleic Acids Res. 46 (D1), D1068−D1073.

Davies, R.J., Miller, R., Coleman, N., 2005. Colorectal cancer screening: prospects for molecular stool analysis. Nat. Rev. Cancer 5 (3), 199−209.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., 2012. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 22, 1589−1598.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., Salzberg, S.L., 1999. Alignment of whole genomes. Nucleic Acids Res. 27, 2369−2376.

Depristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491.

Dimitrakopoulos, C.M., Beerenwinkel, N., 2017. Computational approaches for the identification of cancer genes and pathways. Wiley Interdiscip. Rev. Syst. Biol. Med. 9 (1), e1364.

Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M.A., Condon, A., Aparicio, S., Shah, S.P., 2012. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. Bioinformatics 28 (2), 167−175.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29 (1), 15−21.

Drilon, A., Laetsch, T.W., Kummar, S., DuBois, S.G., Lassen, U.N., Demetri, G.D., Nathenson, M., Doebele, R.C., Farago, A.F., Pappo, A.S., Turpin, B., Dowlati, A., Brose, M.S., Mascarenhas, L., Federman, N., Berlin, J., El-Deiry, W.S., Baik, C., Deeken, J., Boni, V., Nagasubramanian, R., Taylor, M., Rudzinski, E.R., Meric-Bernstam, F., Sohal, D.P.S., Ma, P.C., Raez, L.E., Hechtman, J.F., Benayed, R., Ladanyi, M., Tuch, B.B., Ebata, K., Cruickshank, S., Ku, N.C., Cox, M.C., Hawkins, D.S., Hong, D.S., Hyman, D.M., 2018. Efficacy of larotrectinib in TRK fusion-positive cancers in adults and children. N. Engl. J. Med. 378 (8), 731−739.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327, 78−81.

Dutton-Regester, K., Gartner, J.J., Emmanuel, R., Qutob, N., Davies, M.A., Gershenwald, J.E., Robinson, W., Robinson, S., Rosenberg, S.A., Scolyer, R.A., Mann, G.J., Thompson, J.F., Hayward, N.K., Samuels, Y., 2014. A highly recurrent RPS27 5′UTR mutation in melanoma. Oncotarget 5 (10), 2912−2917.

Edwards, R., Ward, M., Wu, H., Medina, C., Brose, M., Volpe, P., Nussen-Lee, S., Haupt, H., Martin, A., Herlyn, M., 2004. Absence of BRAF mutations in UV-protected mucosal melanomas. J. Med. Genet. 41, 270−272.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133−138.

Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J., Goiffon, R.J., Goldstein, T.C., Ng, S., Lin, L., Crowder, R., Snider, J., Ballman, K., Weber, J., Chen, K., Koboldt, D.C., Kandoth, C., Schierding, W.S., McMichael, J.F., Miller, C.A., Lu, C., Harris, C.C., McLellan, M.D., Wendl, M.C., DeSchryver, K., Allred, D.C., Esserman, L., Unzeitig, G., Margenthaler, J., Babiera, G.V., Marcom, P.K., Guenther, J.M., Leitch, M., Hunt, K., Olson, J., Tao, Y., Maher, C.A., Fulton, L.L., Fulton, R.S., Harrison, M., Oberkfell, B., Du, F., Demeter, R., Vickery, T.L., Elhammali, A., Piwnica-Worms, H., McDonald, S., Watson, M., Dooling, D.J., Ota, D., Chang, L.W., Bose, R., Ley, T.J., Piwnica-Worms, D., Stuart, J.M., Wilson, R.K., Mardis, E.R., 2012. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature 486 (7403), 353−360.

ENCODE_Project_Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57−74.

Er, F., 1990. Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 61, 759−767.

Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., D'Eustachio, P., Stein, L., Hermjakob, H., 2017. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinf. 18 (1), 142.

Fan, J., 2019. Differential pathway analysis. Methods Mol. Biol. 97−114, 1935.

Fang, L.T., Afshar, P.T., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J.C., Gibeling, G., Barr, S., Asadi, N.B., Gerstein, M.B., Koboldt, D.C., Wang, W., Wong, W.H., Lam, H.Y., 2015. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol. 16 (1), 197.

Forbes, S., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J., Futreal, P., Stratton, M., 2008. The catalogue of somatic mutations in cancer (COSMIC). Curr. Protoc. Human Genet. 57, 10.11. 11−10.11. 26.

Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., Andersson, R., Mungall, C.J., Meehan, T.F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y.A., Plessy, C., Vitezic, M., Severin, J., Semple, C., Ishizu, Y., Young, R.S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G.M., Arakawa, T., Archer, J.A., Arner, P., Babina, M., Rennie, S., Balwierz, P.J., Beckhouse, A.G., Pradhan-Bhatt, S., Blake, J.A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Burroughs, A.M., Califano, A., Cannistraci, C.V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H.C., Dalla, E., Davis, C.A., Detmar, M., Diehl, A.D., Dohi, T., Drabløs, F., Edge, A.S., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M.C., Faulkner, G.J., Favorov, A.V., Fisher, M.E., Frith, M.C., Fujita, R., Fukuda, S., Furlanello, C., Furino, M., Furusawa, J., Geijtenbeek, T.B., Gibson, A.P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T.J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K.J., Ho Sui, S.J., Hofmann, O.M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B.R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A.S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y.I., Kawashima, T., Kempfle, J.S., Kenna, T.J., Kere, J., Khachigian, L.M., Kitamura, T., Klinken, S.P., Knox, A.J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A.T., Laros, J.F., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-Sim, A., Manabe, R., Mar, J.C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D.A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C.L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Noma, S., Noazaki, T., Ogishima, S., Ohkura, N., Ohimiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D.A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J.G., Rackham, O.J., Ramilowski, J.A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M.B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E.A., Schulze-Tanzil, G.G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J.W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R.K., 't Hoen, P.A., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyodo, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L.M., Verado, R., Vijayan, D., Vorontsov, I.E., Wasserman, W.W., Watanabe, S., Wells, C.A., Winteringham, L.N., Wolvetang, E., Wood, E.J., Yamaguchi, Y., Yamamoto, M.,

Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S.E., Zhang, P.G., Zhao, X., Zucchelli, S., Summers, K.M., Suzuki, H., Daub, C.O., Kawai, J., Heutink, P., Hide, W., Freeman, T.C., Lenhard, B., Bajic, V.B., Taylor, M.S., Makeev, V.J., Sandelin, A., Hume, D.A., Carninci, P., Hayashizaki, Y., 2014. FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. Nature 507 (7493), 462−470.

Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., Mccarroll, S.A., O'donovan, M.C., Owen, M.J., 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet. 91, 597−607.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R., 2004. A census of human cancer genes. Nat. Rev. Cancer 4, 177.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., Schultz, N., 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6 (269), pl1.

García-Campos, M.A., Espinal-Enríquez, J., Hernández-Lemus, E., 2015. Pathway analysis: state of the art. Front. Physiol. 17 (6), 383.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R.J., Bignell, G.R., Tam, A.T., Davies, H., Stevenson, J.A., Barthorpe, S., Lutz, S.R., Kogera, F., Lawrence, K., McLaren-Douglas, A., Mitropoulos, X., Mironenko, T., Thi, H., Richardson, L., Zhou, W., Jewitt, F., Zhang, T., O'Brien, P., Boisvert, J.L., Price, S., Hur, W., Yang, W., Deng, X., Butler, A., Choi, H.G., Chang, J.W., Baselga, J., Stamenkovic, I., Engelman, J.A., Sharma, S.V., Delattre, O., Saez-Rodriguez, J., Gray, N.S., Settleman, J., Futreal, P.A., Haber, D.A., Stratton, M.R., Ramaswamy, S., McDermott, U., Benes, C.H., 2012. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483 (7391), 570−575.

Garrison, E., Marth, G., 2012. Haplotype-based Variant Detection from Short-Read Sequencing arXiv:1207.3907 [q-bio.GN].

Gatto, A., Torroja-Fungairino, C., Mazzarotto, F., Cook, S.A., Barton, P.J., Sanchez-Cabo, F., Lara-Pezzi, E., 2014. FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. Nucleic Acids Res. 42 e71−e71.

Ge, H., Liu, K., Juan, T., Fang, F., Newman, M., Hoeck, W., 2011. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics 27, 1922−1928.

Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald 3rd, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., Hu, K., Andreev-Drakhlin, A.Y., Kim, J., Hess, J.M., Haas, B.J., Aguet, F., Weir, B.A., Rothberg, M.V., Paolella, B.R., Lawrence, M.S., Akbani, R., Lu, Y., Tiv, H.L., Gokhale, P.C., de Weck, A., Mansour, A.A., Oh, C., Shih, J., Hadi, K., Rosen, Y., Bistline, J., Venkatesan, K., Reddy, A., Sonkin, D., Liu, M., Lehar, J., Korn, J.M., Porter, D.A., Jones, M.D., Golji, J., Caponigro, G., Taylor, J.E., Dunning, C.M., Creech, A.L., Warren, A.C., McFarland, J.M., Zamanighomi, M., Kauffmann, A., Stransky, N., Imielinski, M., Maruvka, Y.E., Cherniack, A.D., Tsherniak, A., Vazquez, F., Jaffe, J.D., Lane, A.A., Weinstock, D.M., Johannessen, C.M., Morrissey, M.P., Stegmeier, F., Schlegel, R.,

Hahn, W.C., Getz, G., Mills, G.B., Boehm, J.S., Golub, T.R., Garraway, L.A., Sellers, W.R., 2019. Next-generation characterization of the cancer cell line encyclopedia. Nature 569 (7757), 503−508.

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., Zhang, Z., 2013. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genom. 14, S7.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S., Nusbaum, C., 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27 (2), 182−189.

Goeman, J.J., Mansmann, U., 2008. Multiple testing on the directed acyclic graph of gene ontology. Bioinformatics 24 (4), 537−544.

Gonzalez-Perez, A., Lopez-Bigas, N., 2012. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 40 e169−e169.

Griffith, M., Griffith, O.L., Coffman, A.C., Weible, J.V., Mcmichael, J.F., Spies, N.C., Koval, J., Das, I., Callaway, M.B., Eldred, J.M., 2013. DGIdb: mining the druggable genome. Nat. Methods 10, 1209.

Hansen, N.F., Gartner, J.J., Mei, L., Samuels, Y., Mullikin, J.C., 2013. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics 29, 1498−1503.

Harris, H., 2008. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. Preface. J. Cell Sci. 121, v−vi.

Hauschild, A., Grob, J.J., Demidov, L.V., Jouary, T., Gutzmer, R., Millward, M., Rutkowski, P., Blank, C.U., Miller Jr., W.H., Kaempgen, E., Martín-Algarra, S., Karaszewska, B., Mauch, C., Chiarion-Sileni, V., Martin, A.M., Swann, S., Haney, P., Mirakhur, B., Guckert, M.E., Goodman, V., Chapman, P.B., 2012. Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial. Lancet 380 (9839), 358−365.

He, W., Fu, L., Yan, Q., Zhou, Q., Yuan, K., Chen, L., Han, Y., 2019. Gene set enrichment analysis and meta-analysis identified 12 key genes regulating and controlling the prognosis of lung adenocarcinoma. Oncol. Lett. 17 (6), 5608−5618.

Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Benjamin Gordon, D., Brizuela, L., Richard McCombie, W., Hannon, G.J., 2009. Hybrid Selection of Discrete Genomic Intervals on Custom-Designed Microarrays for Massively Parallel Sequencing, vol. 4, pp. 960−974.

Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T.-W., Peng, Z., Yiu, S.-M., 2011. SOAPsplice: genome-wide ab initio detection of splice junctions from RNA-Seq data. Front. Genet. 2, 46.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., Garraway, L.A., 2013. Highly recurrent TERT promoter mutations in human melanoma. Science 339 (6122), 957−959.

Huret, J.L., Ahmad, M., Arsaban, M., Bernheim, A., Cigna, J., Desangles, F., Guignard, J.C., Jacquemot-Perbal, M.C., Labarussias, M., Leberre, V., Malo, A., Morel-Pair, C., Mossafa, H., Potier, J.C., Texier, G., Viguié, F., Yau Chun Wan-Senon, S., Zasadzinski, A., Dessen, P., 2013. Atlas of genetics and cytogenetics in oncology and Haematology in 2013. Nucleic Acids Res. 41 (Database issue), D920−D924.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., McVean, G., 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat. Genet. 44, 226−232.

Jäkel, C., Bergmann, F., Toth, R., Assenov, Y., van der Duin, D., Strobel, O., Hank, T., Klöppel, G., Dorrell, C., Grompe, M., Moss, J., Dor, Y., Schirmacher, P., Plass, C., Popanda, O., Schmezer, P., 2017. Genome-wide genetic and epigenetic analyses of pancreatic acinar cell carcinomas reveal aberrations in genome stability. Nat. Commun. 8 (1), 1323.

Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bahler, J., Sedlazeck, F.J., 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. 8, 14061.

Jin, J., Wu, X., Yin, J., Li, M., Shen, J., Jing, L., Zhao, Y., Zhao, Q., Wu, J., Wen, Q., Cho, C.H., Yi, T., Xiao, Z., Qu, L., 2019. Identification of genetic mutations in cancer: challenge and opportunity in the new era of targeted therapy. Front Oncol 9, 263.

Johnston, J.J., Rubinstein, W.S., Facio, F.M., Ng, D., Singh, L.N., Teer, J.K., Mullikin, J.C., Biesecker, L.G., 2012. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. Am. J. Hum. Genet. 91, 97−108.

Jones, S., Zhang, X., Parsons, D.W., Lin, J.C.-H., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., 2008. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321, 1801−1806.

Jubb, H.C., Saini, H.K., Verdonk, M.L., Forbes, S.A., 2018. COSMIC-3D provides structural perspectives on cancer genetics for drug discovery. Nat. Genet. 50 (9), 1200−1202. https://doi.org/10.1038/s41588-018-0214-9.

Kanchi, K.L., Johnson, K.J., Lu, C., Mclellan, M.D., Leiserson, M.D., Wendl, M.C., Zhang, Q., Koboldt, D.C., Xie, M., Kandoth, C., 2014. Integrated analysis of germline and somatic variants in ovarian cancer. Nat. Commun. 5, 3156.

Kandoth, C., Mclellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., Mcmichael, J.F., Wyczalkowski, M.A., Leiserson, M.D.M., Miller, C.A., Welch, J.S., Walter, M.J., Wendl, M.C., Ley, T.J., Wilson, R.K., Raphael, B.J., Ding, L., 2013. Mutational landscape and significance across 12 major cancer types. Nature 502, 333−339.

Khunlertgit, N., Yoon, B.J., 2016. Incorporating topological information for predicting robust cancer subnetwork markers in human protein-protein interaction network. BMC Bioinf. 17 (Suppl. 13), 351.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36.

Knudson, A.G., 1971. Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U S A 68, 820−823.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., Mclellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 2283−2285.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22 (3), 568−576.

Kotelnikova, E.A., Pyatnitskiy, M., Paleeva, A., Kremenetskaya, O., Vinogradov, D., 2016. Practical aspects of NGS-based pathways analysis for personalized cancer science and medicine. Oncotarget 7, 52493.

Krumm, N., Sudmant, P.H., Ko, A., O'roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A., Eichler, E.E., Project, N.E.S., 2012. Copy number variation detection and genotyping from exome sequence data. Genome Res. 22, 1525−1532.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.C., Pfenning, A.R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R.A., Shoresh, N., Epstein, C.B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R.D., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Hansen, R.S., Kaul, R., Sabo, P.J., Bansal, M.S., Carles, A., Dixon, J.R., Farh, K.H., Feizi, S., Karlic, R., Kim, A.R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthall, K.T., Sinnott-Armstrong, N.A., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Beaudet, A.E., Boyer, L.A., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J., Li, W., Marra, M.A., McManus, M.T., Sunyaev, S., Thomson, J.A., Tlsty, T.D., Tsai, L.H., Wang, W., Waterland, R.A., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J.A., Wang, T., Kellis, M., 2015. Integrative analysis of 111 reference human epigenomes. Nature 518 (7539), 317−330.

Lai, E.Y., Chen, Y.H., Wu, K.P., 2017. A knowledge-based T2-statistic to perform pathway analysis for quantitative proteomic data. PLoS Comput. Biol. 13 (6), e1005601.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., Maglott, D.R., 2013. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42, D980−D985.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., Ding, L., 2011. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28, 311−317.

Lathrop, M., Gut, I., Heath, S., Tost, J., Gress, T., Hudson, T., 2010. International Network of Cancer Genome Projects (The International Cancer Genome Consortium).

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., Getz, G., 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495−501.

Layer, R.M., Chiang, C., Quinlan, A.R., Hall, I.M., 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15, R84.

Le Guilloux, V., Schmidtke, P., Tuffery, P., 2009. Fpocket: an open source platform for ligand pocket detection. BMC Bioinf. 10, 168.

Lee, W.-P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., Marth, G.T., 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9, e90581.

Leiserson, M.D., Blokh, D., Sharan, R., Raphael, B.J., 2013. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput. Biol. 9, e1003054.

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094−3100.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J., 2009a. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966−1967.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009b. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078−2079.

Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B., Loh, M.L., Hunger, S.P., Sanda, T., Young, R.A., Look, A.T., 2014. Oncogenic Regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science 346 (6215), 1373−1377.

Mardis, E.R., 2019. The impact of next-generation sequencing on cancer genomics: from discovery to clinic. Cold Spring Harb. Perspect. Med. 9 (9), a036269.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376.

Mcpherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M.G., Griffith, M., Moussavi, A.H., Senz, J., Melnyk, N., 2011. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput. Biol. 7, e1001138.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., Getz, G., 2011. GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, R41.

Mohiyuddin, M., Mu, J.C., Li, J., Bani Asadi, N., Gerstein, M.B., Abyzov, A., Wong, W.H., Lam, H.Y., 2015. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics 31, 2741−2744.

Nagarajan, N., Yapp, E.K.Y., Le, N.Q.K., Kamaraj, B., Al-Subaie, A.M., Yeh, H.Y., 2019. Application of computational biology and artificial intelligence technologies in cancer precision drug discovery. Biomed. Res. Int. 11, 8427042.

Narayan, S., Bader, G.D., Reimand, J., 2016. Frequent mutations in acetylation and ubiquitination sites suggest novel driver mechanisms of cancer. Genome Med. 8, 55.

Ng, P.C., Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812−3814.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K.W., Mudie, L.J., Varela, I., Mcbride, D.J., Bignell, G.R., Cooke, S.L., Shlien, A., Gamble, J., Whitmore, I., Maddison, M., Tarpey, P.S., Davies, H.R., Papaemmanuil, E., Stephens, P.J., Mclaren, S., Butler, A.P., Teague, J.W., Jonsson, G., Garber, J.E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerod, A., Tutt, A., Martens, J.W., Aparicio, S.A., Borg, A., Salomon, A.V., Thomas, G., Borresen-Dale, A.L., Richardson, A.L., Neuberger, M.S., Futreal, P.A., Campbell, P.J., Stratton, M.R., 2012. Mutational processes molding the genomes of 21 breast cancers. Cell 149, 979−993.

Nowell, P.C., 1976. The clonal evolution of tumor cell populations. Science 194, 23−28.

Nowell, P., Hungerford, D., 2004. A minute chromosome in human chronic granulocytic leukemia. Landmarks Med Genet. Class. Pap. Comment. 132, 103.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., Zwick, M.E., 2007. Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4, 907.

Pallarz, S., Benary, M., Lamping, M., Rieke, D., Starlinger, J., Sers, C., Wiegandt, D.L., Seibert, M., Seva, J., Schäfer, R., Keilholz, U., Leser, U., 2019. Comparative analysis of public knowledge bases for precision oncology. JCO Precis. Oncol. https://doi.org/10.1200/PO.18.00371, 23:PO.18.00371.

Papanikolaou, N., Pavlopoulos, G.A., Theodosiou, T., Iliopoulos, I., 2015. Protein-protein interaction predictions using text mining methods. Methods 74, 47−53.

Parsons, D.W., Jones, S., Zhang, X., Lin, J.C.-H., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G.L., 2008. An integrated genomic analysis of human glioblastoma multiforme. Science 321, 1807−1812.

Pellestor, F., 2019. Chromoanagenesis: cataclysms behind complex chromosomal rearrangements. Mol. Cytogenet. 11 (12), 6.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., Salzberg, S.L., 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, Stringtie and Ballgown. Nat. Protoc. 11, 1650−1667.

Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., Leproust, E.M., Peck, B.J., Emig, C.J., Dahl, F., 2007. Multiplex amplification of large sets of human exons. Nat. Methods 4, 931.

Puente, X.S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J.I., Munar, M., Rubio-Pérez, C., Jares, P., Aymerich, M., Baumann, T., Beekman, R., Belver, L., Carrio, A., Castellano, G., Clot, G., Colado, E., Colomer, D., Costa, D., Delgado, J., Enjuanes, A., Estivill, X., Ferrando, A.A., Gelpí, J.L., González, B., González, S., González, M., Gut, M., Hernández-Rivas, J.M., López-Guerra, M., Martín-García, D., Navarro, A., Nicolás, P., Orozco, M., Payer, Á.R., Pinyol, M., Pisano, D.G., Puente, D.A., Queirós, A.C., Quesada, V., Romeo-Casabona, C.M., Royo, C., Royo, R., Rozman, M., Russiñol, N., Salaverría, I., Stamatopoulos, K., Stunnenberg, H.G., Tamborero, D., Terol, M.J., Valencia, A., López-Bigas, N., Torrents, D., Gut, I., López-Guillermo, A., López-Otín, C., Campo, E., 2015. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 526 (7574), 519−524.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., Betel, D., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 14, 3158.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., Korbel, J.O., 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333−i339.

Reddy, E.P., Reynolds, R.K., Santos, E., Barbacid, M., 1982. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature 300, 149.

Reimand, J., Bader, G.D., 2013. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol. Syst. Biol. 9, 637.

Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S.K., Tourna, A., Yakovleva, A., Palmieri, T., Ciccarelli, F.D., 2019. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 220 (1), 1.

Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B.B., Barrette, T.R., Anstet, M.J., Kincead-Beal, C., Kulkarni, P., Varambally, S., Ghosh, D., Chinnaiyan, A.M., 2007. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 9 (2), 166−180.

Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G.V., Carter, S.L., Saksena, G., Harris, S., Shah, R.R., Resnick, M.A., Getz, G., Gordenin, D.A., 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat. Genet. 45, 970−976.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139−140.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T., Bustillo, J., 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475 (7356), 348−352.

Rowley, J.D., 1973. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 243, 290.

Ryan, M.C., Cleland, J., Kim, R., Wong, W.C., Weinstein, J.N., 2012. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. Bioinformatics 28, 2385−2387.

Sathirapongsasuti, J.F., Lee, H., Horst, B.A., Brunner, G., Cochran, A.J., Binder, S., Quackenbush, J., Nelson, S.F., 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 27 (19), 2648−2654.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., Cheetham, R.K., 2012. Strelka: accurate somatic small-variant calling from sequenced tumor−normal sample pairs. Bioinformatics 28, 1811−1817.

Schuierer, S., Roma, G., 2016. The exon quantification pipeline (EQP): a comprehensive approach to the quantification of gene, exon and junction expression from RNA-seq data. Nucleic Acids Res. 44 (16), e132.

Sedova, M., Iyer, M., Li, Z., Jaroszewski, L., Post, K.W., Hrabe, T., Porta-Pardo, E., Godzik, A., 2019. Cancer3D 2.0: interactive analysis of 3D patterns of cancer mutations in cancer subsets. Nucleic Acids Res. 47 (D1), D895−D899.

Seyednasrollah, F., Laiho, A., Elo, L.L., 2013. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings Bioinf. 16, 59−70.

Shen, S., Park, J.W., Huang, J., Dittmar, K.A., Lu, Z.-X., Zhou, Q., Carstens, R.P., Xing, Y., 2012. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res. 40 e61−e61.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., Mccutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., Church, G.M., 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728−1732.

Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314, 268−274.

Soneson, C., Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinf. 14, 91.

Soylev, A., Le, T., Amini, H., Alkan, C., Hormozdiari, F., 2019. Discovery of tandem and interspersed segmental duplications using high throughput sequencing. Bioinformatics 35 (20), 3923−3930.

Speyer, G., Kiefer, J., Dhruv, H., Berens, M., Kim, S., 2016. Knowledge assisted approach to identify pathways with differential dependencies. Pac. Symp. Biocomput. 21, 33−44.

Spinella, J.F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., Ouimet, M., Healy, J., Sinnett, D., 2016. SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC Genom. 17 (1), 912.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A.D., Cooper, D.N., 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. 133, 1−9.

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C., 2005. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nat. Genet. 37, 590.

Supek, F., Minana, B., Valcarcel, J., Gabaldon, T., Lehner, B., 2014. Synonymous mutations frequently act as driver mutations in human cancers. Cell 156, 1324−1335.

Tabin, C.J., Bradley, S.M., Bargmann, C.I., Weinberg, R.A., Papageorge, A.G., Scolnick, E.M., Dhar, R., Lowy, D.R., Chang, E.H., 1982. Mechanism of activation of a human oncogene. Nature 300, 143.

Taparowsky, E., Suard, Y., Fasano, O., Shimizu, K., Goldfarb, M., Wigler, M., 1982. Activation of the T24 bladder carcinoma transforming gene is linked to a single amino acid change. Nature 300, 762.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S.C., Kok, C.Y., Noble, K., Ponting, L., Ramshaw, C.C., Rye, C.E., Speedy, H.E., Stefancsik, R., Thompson, S.L., Wang, S., Ward, S., Campbell, P.J., Forbes, S.A., 2019. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 47 (D1), D941−D947.

Teer, J.K., Bonnycastle, L.L., Chines, P.S., Hansen, N.F., Aoyama, N., Swift, A.J., Abaan, H.O., Albert, T.J., Margulies, E.H., Green, E.D., 2010. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome Res. 20, 1420−1431.

Tian, Z., Guo, M., Wang, C., Liu, X., Wang, S., 2017. Refine gene functional similarity network based on interaction networks. BMC Bioinf. 18 (Suppl. 16), 550.

Tian, S., Yan, H., Klee, E.W., Kalmbach, M., Slager, S.L., 2018. Comparative analysis of de novo assemblers for variation discovery in personal genomes. Briefings Bioinf. 19, 893−904.

Tokheim, C., Karchin, R., 2019. CHASMplus reveals the scope of somatic missense mutations driving human cancers. Cell Syst 9 (1), 9−23 e8.

Tomczak, K., Czerwińska, P., Wiznerowicz, M., 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. 19, A68.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., 2015. Tissue-based map of the human proteome. Science 347, 1260419.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., 2001. The sequence of the human genome. Science 291, 1304−1351.

Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., Melo, M., da Rocha, A.G., Preto, A., Castro, P., Castro, L., Pardal, F., Lopes, J.M., Santos, L.L., Reis, R.M., Cameselle-Teijeiro, J., Sobrinho-Simões, M., Lima, J., Máximo, V., Soares, P., 2013. Frequency of TERT promoter mutations in human cancers. Nat. Commun. 4, 2185.

Walsh, T., Lee, M.K., Casadei, S., Thornton, A.M., Stray, S.M., Pennil, C., Nord, A.S., Mandell, J.B., Swisher, E.M., King, M.-C., 2010. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc. Natl. Acad. Sci. U S A 107, 12629−12633.

Wang, T.-L., Rago, C., Silliman, N., Ptak, J., Markowitz, S., Willson, J.K., Parmigiani, G., Kinzler, K.W., Vogelstein, B., Velculescu, V.E., 2002. Prevalence of somatic alterations in the colorectal cancer cell genome. Proc. Natl. Acad. Sci. U S A 99, 3076.

Wang, W., Sun, J., Li, F., Li, R., Gu, Y., Liu, C., Yang, P., Zhu, M., Chen, L., Tian, W., Zhou, H., Mao, Y., Zhang, L., Jiang, J., Wu, C., Hua, D., Chen, W., Lu, B., Ju, J., Zhang, X., 2012. A frequent somatic mutation in CD274 3′-UTR leads to protein over-expression in gastric cancer by disrupting miR-570 binding. Hum. Mutat. 33 (3), 480−484.

Wendl, M.C., Wallis, J.W., Lin, L., Kandoth, C., Mardis, E.R., Wilson, R.K., Ding, L., 2011. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics 27 (12), 1595−1602.

Wojcicka, A., de la Chapelle, A., Jazdzewski, K., 2014. MicroRNA-related sequence variations in human cancers. Hum. Genet. 133 (4), 463−469.

Woollard, W.J., Pullabhatla, V., Lorenc, A., Patel, V.M., Butler, R.M., Bayega, A., Begum, N., Bakr, F., Dedhia, K., Fisher, J., Aguilar-Duran, S., Flanagan, C., Ghasemi, A.A., Hoffmann, R.M., Castillo-Mosquera, N., Nuttall, E.A., Paul, A., Roberts, C.A., Solomonidis, E.G., Tarrant, R., Yoxall, A., Beyers, C.Z., Ferreira, S., Tosi, I., Simpson, M.A., de Rinaldis, E., Mitchell, T.J., Whittaker, S.J., 2016. Candidate driver genes involved in genome maintenance and DNA repair in Sezary syndrome. Blood 127 (26), 3387−3397.

Xiong, D., Pan, J., Zhang, Q., Szabo, E., Miller, M.S., Lubet, R.A., You, M., Wang, Y., 2017. Bronchial airway gene expression signatures in mouse lung squamous cell carcinoma and their modulation by cancer chemopreventive agents. Oncotarget 8, 18885−18900.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J., 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19 (9), 1586−1592.

Zarate, S., Carroll, A., Krasheninina, O., Sedlazeck, F.J., Jun, G., Salerno, W., Boerwinkle, E., Gibbs, R., 2018. Parliament2: fast structural variant calling using optimized combinations of callers. bioRxiv 424267.

Zhang, F., Wang, M., Xi, J., et al., 2018. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci. Rep. 8, 3355.

CHAPTER

12 Computational and functional annotation at genomic scale: gene expression and analysis

Srishty Gulati[1], **Anju Singh**[1,2], **Md Shoaib**[1], **Shrikant Kukreti**[1]

[1]*Nucleic Acid Research Lab, Department of Chemistry, University of Delhi, North Campus, New Delhi, Delhi, India;* [2]*Department of Chemistry, Ramjas College, University of Delhi, New Delhi, Delhi, India*

12.1 Introduction: background (history)

Genetics was considered a discipline of biology that studied hereditary traits, and dealt with the discovery and research of units involved in central dogma, discovery of reverse transcriptase, Mendel's laws of inheritance, genes, and genetic variations (Plomin et al., 2008; McClean, 2011). Later, a new discipline of modern biology was evolved called *genomics*, which involved sequencing, mapping, and analysis of the genome, and was considered to be a part of genetics (McKusick and Ruddle, 1987; Weissenbach, 2016). With the progression of technology, sequencing became practicable especially with the discovery of polymerase chain reaction (PCR), availability of enzymes to modify nucleic acids, and fluorescent techniques (Saiki et al, 1985, 1988).

In 1977, Sanger et al. at Cambridge University developed a new method of sequencing for nonviral DNA (Sanger et al.,1977b). The first gene (encodes protein) was sequenced from bacteriophages by using RNA-sequencing (RNA-seq) (Jou et al., 1972; Fiers et al., 1976) and a few months later DNA-sequencing of viral DNA was successfully done (Sanger et al., 1977a). Sanger's team initially discovered the complete nucleotide DNA sequence of unicellular microorganisms and then successfully worked on eukaryotes. *Haemophilus influenzae* (bacteria) was the first prokaryote to be sequenced in 1995 (Fleischmann et al., 1995) followed by *Saccharomyces cerevisiae* (budding yeast), which was the first eukaryote sequenced in 1996 (Galibert et al., 1996). The development of technologies for genome sequencing provided massive genomic data, which was an exceptional challenge for researchers. As a genome is sequenced, it is required to store, organize, and analyze huge genomic datasets, which seems impossible without computers.

In 1980, the global digital revolution gave rise to the adoption of computers for digital storage of biological data. Development in computer science led to advances in the field of genomics. Therefore genomics is now considered a

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00014-2

multidisciplinary science that includes biology, genetics, biochemistry, computer sciences, and statistics (McClean, 2011). Bioinformatics is a crucial part of genomics, which helps to collect, store, organize, and analyze gigantic biological datasets by using computational, statistical, and mathematical tools (Baxevanis et al., 2020). Rapid development of advanced software gave researchers new programs to manage these enormous datasets (De Filippo et al., 2012). This resulted in the genesis of a new era of molecular biology (Weissenbach, 2016).

The growing interest in existing genetic technologies led the US Department of Energy to officially release the *Human Genome Project* (HGP) to map the complete human genome in 1987. At the beginning of the 21st century, the first draft containing sequencing of around three billion base-pairs that make up the whole human genome was published (Lander et al., 2001). Consequently, the HGP bestowed researchers with unprecedented information for the study of genetic diseases and evolution. The emergence of computers resulted in more advanced techniques compared to Sanger's sequencing such as HiSeq, Ion Torrent, PacBio, etc., which are cost effective, high throughput, and have parallel sequencing capacity (Liu et al., 2012; Pareek et al., 2011; Reuter et al., 2015). This led to the challenge of evaluating this large influx of data.

In the past 20 years, several innovative techniques allowed the conversion of the data obtained by sequencing techniques for annotation of the genome (Koonin and Galperin, 2003; de Sá et al., 2018). Various platforms, methods, and bioinformatics tools have been used to manage data obtained by high-throughput DNA hybridization microarray and sequencing techniques (Edgar et al., 2002; Clough and Barrett, 2016). Techniques involved in the functional analysis of the genome were able to recognize over- or under-expressed genes, gene characterization, and the development of biological profiles (ENCODE Project Consortium, 2012). This helped biologists to effortlessly understand the relation of cellular events with variable genomic conditions. In eukaryotes, annotation is more challenging due to the presence of several repeats, variable lengths of intergenic regions (IGRs), and inconsistent behavior of protein coding genes (Yandell and Ence, 2012). Computational methods of annotation are highly efficient quantitatively but sometimes they become unreliable qualitatively due to a large number of errors (Salzberg, 2019). The demands of data production and analysis necessitate bioinformatics to provide continuous upgraded computing, storage, and data analysis tools. In addition, the data should be easily accessible to the public.

Gene expression is a critical process in which information from genes is expressed as a final functional gene product (protein) (Crick, 1970). Traditionally, the expression of only one gene could be measured at a time, but today several computational methods have been used based on mRNA expression. RNA and DNA gene expression microarray techniques empowered biologists to quickly understand the aspects of life, genetic abnormalities, and evolution. Computational methods have been used in genomics to measure and identify gene expression data obtained from high-throughput technologies like DNA microarray, RNA-seq, etc. (Fyad et al., 2016). Clinically, the blooming technologies in the world of biology have been fruitful in recognizing genetic disorders (Guttmacher and Collins, 2005; Steward et al., 2017).

12.2 Genome sequencing

After the discovery of DNA, several methods were discovered for the detection of nucleotide sequences of DNA in the genome. Prediction of a gene is related to sequencing and assembly approaches. The discovery of DNA sequencing techniques deciphered all the codes of life and inspired many researchers to understand genetic diseases and evolution. Advancement in technology rapidly improves the potentiality and volume in the field of genome sequencing (Levy and Myers, 2016). Fifteen years ago sequencing was based only on Sanger's "chain termination method," but from 2005 new technologies with high throughput, reduced cost, and increased efficiency have evolved and are considered as the second generation of sequencing, often called "next-generation sequencing" (NGS) (McPherson, 2014; Liu et al., 2012). Further development in technology in the last few decades has emerged at an incredible pace. More advanced sequencing techniques have evolved with higher speed and volume of genomic sequencing. Therefore based on time and improvement in the sequencing techniques, three generations of sequencing technologies exist so far as shown in Table 12.1 (Levy and Myers, 2016; Kchouk et al., 2017).

12.2.1 First generation (Sanger's generation): an old but reliable approach

Sanger's and Maxam–Gilbert's discoveries of sequencing techniques were the breakthrough in the field of genomics to decode all the codes of biological systems. Maxam–Gilbert's (degradation method) and Sanger's techniques (synthesis method) for sequencing were considered as the "first generation of sequencing." Sanger was awarded a Nobel Prize in 1980 for his first and common sequencing technique with low radioactivity and high efficiency. The HGP utilized Sanger's "chain termination method" to sequence the entire human genome comprising nearly three billion bps due to its better quality. Sanger's method utilized one strand of DNA to serve as a template, dideoxy nucleotides (dNTPs), radioactive primer, and DNA polymerase. Different size fragments of dNTPs were obtained with DNA polymerase, which were separated by gel electrophoresis to achieve a final sequence (Sanger et al., 1977b). To make it easier, in 1995 Applied Biosystems Inc. built an improved and updated version of automated Sanger sequencing technology. Various projects of sequencing, including plants (Goff et al., 2002) and humans (1000 Genomes Project Consortium, 2010), used Sanger's sequencing technology. Even now, Sanger sequencing has been used to verify variants in sequence because of its better quality and accuracy.

The Maxam–Gilbert method of sequencing was based on degradation of sequences by using chemicals (Maxam and Gilbert, 1977). Fragments obtained after degradation are separated by gel electrophoresis. This method was less popular because of its slow speed, complexity, and toxicity (Kchouk et al., 2017). Thus Sanger's sequencing technique was the most efficient genome sequencing technique until the new era of technologies.

Table 12.1 Evolution of sequencing techniques with the development of technologies.

Sequencing technique	Manufacturer	Instruments	Chemical method	Detection method	Read type	Error rate	URL
First generation							
ABI Sanger	Applied Biosystems Inc. (ABI)	3730xl	Sequencing by synthesis	Electrophoresis	Single end	0.3%	–
Maxam –Gilbert	–		Sequencing by degradation	Electrophoresis	Single end	–	–
Second generation							
454 (Roche)	Roche 454	GS20, GS FLX, GS FLX Titanium, Titanium^{+}, GS Junior	Sequencing by synthesis	Optical	Single end, paired end	1%	http://www.454.com/
Illumina	Illumina Inc.	MiniSeq, MiSeq, NextSeq, HiSeq, HiSeq X	Sequencing by synthesis	Optical	Single end, paired end	0.1%–1%	http://www.illumina.com/
Ion Torrent	Thermo Fisher Scientific	PGM 314 chip v2, PGM 316 chip v2, PGM 318 chip v2, Ion Proton, Ion S5/S5XL 520, Ion S5/S5XL 530, Ion S5/S5XL 540	Sequencing by synthesis	Solid state	Single end	1%	http://www.thermofisher.com/us/en/home/brands/ion-torrent.html
SOLiD	Life Technologies, ABI Thermo Fisher Applied Biosystems	5500 W, 5500xlW	Sequencing by ligation	Optical	Single end	~0.1%	http://www.lifetechnologies.com; http://www.thermofisher.com/us/en/home/brands/applied-biosystems.html
Third generation							
PacBio	Pacific Biosciences	RS CI, RS C2, RS C2 XL, RS II C2 XL, RS II P5 C3, RS II P6 C4, Sequel	Sequencing by synthesis	Optical	Single end	12%–15%	http://www.pacificbiosciences.com/
Oxford Nanopore	Oxford Nanopore Technologies	MinION Mk, PromethION	Nanopore	Nanopore	ID, 2D	12%	http://www.nanoporetech.com

12.2.2 Second-generation/next-generation sequencing

Sanger's sequencing technique was utilized for three decades in the world of genomics until "second- or next-generation sequencing" introduced a new perspective to genome analysis in 2005 (de Sá et al., 2018). NGS has several advantages over first-generation sequencing methods in terms of speed, cost, high throughput, and effortlessness (Liu et al., 2012). NGS can generate parallel analysis of millions of reads in less time and at a lower cost (Shendure, 2008). Additionally, there is no requirement for electrophoresis as NGS directly detects the output. Among the large number of NGS technologies, we briefly discuss some of the commonly used sequencing technologies.

12.2.2.1 454 (Roche) sequencing

The new era of sequencing starts with 454 Roche sequencing technology, which is based on the sequencing-by-synthesis approach, which utilizes the pyrosequencing technique. This technique involves the release of pyrophosphate, which emits light when nucleotides are incorporated in DNA and detects the fragments of DNA (Margulies et al, 2005). This technique easily generates millions of long reads in parallel fashion (Kchouk et al., 2017; Liu et al., 2012). The only drawback to this method was that it could not detect insertions and deletions (indels) in the sequence accurately (Huse et al., 2007).

12.2.2.2 Illumina sequencing

This is a widely used sequencing technique of NGS. It also uses a sequencing-by-synthesis approach. This technique decodes the given DNA fragment from each end. The overall procedure of sequencing by Illumina is illustrated in Fig. 12.1 (Kchouk et al., 2017). The length of base-pair reads is 150−250 bp with an output of more than 600 Gbp for the latest Illumina sequencer. Errors in Illumina sequencing techniques are due to an excessive requirement of samples, which causes overlapping and results in substitution nucleotide errors (Liu et al., 2012; Kulski, 2016).

12.2.2.3 Ion Torrent sequencing

This sequencing technology was initially commercialized by Life Technologies, but in 2014 Thermo Fisher Scientific acquired Ion Torrent (IT) sequencing. IT sequencing involves detection of nucleotides based on change in H^+ ion concentration (Rothberg et al, 2011). During synthesis, when a nucleotide is incorporated in DNA by DNA polymerase, there is a release of H^+ ions that changes the pH of the overall solution, and that change is detected by the sensor and finally changes the voltage that detects the nucleotide incorporated. The IT sequencer uses a chip that can produce a throughput of 10 Gb with base-pair reads of length 200, 400, and 600 bp. Unlike other NGS techniques, it does not use fluorescence for labeling nucleotides (de Sá et al, 2018). Like the 454 Roche sequencing technique, it also has indel error (Reuter et al., 2015).

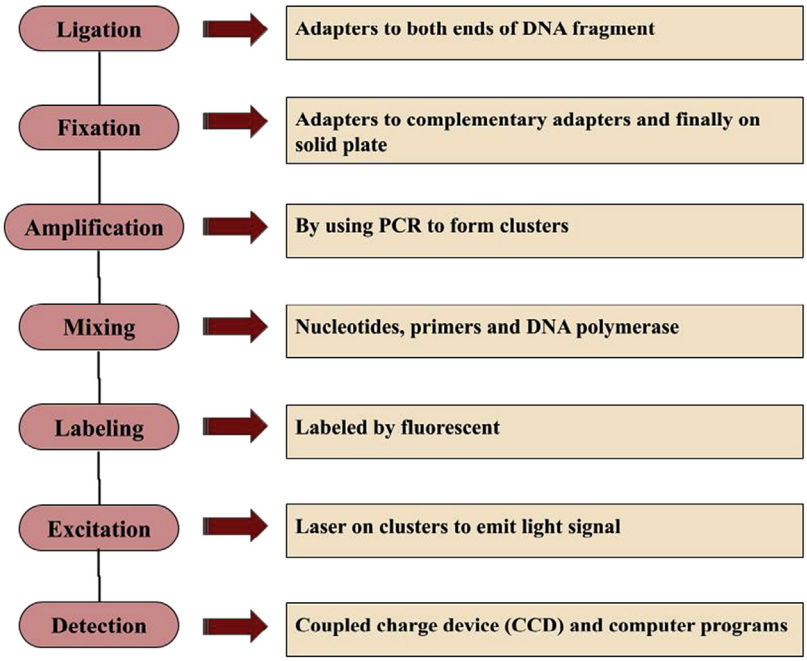


FIGURE 12.1

Illumina sequencing technique.

12.2.2.4 SOLiD sequencing

Life Technologies commercialized the Supported Oligonucleotide Ligation and Detection (SOLiD) sequencing technique, but in 2007 this technique was acquired by Applied Biosystems Inc. (ABI) (Shendure, 2008). It includes five steps: (1) attachment of adapter to DNA, (2) fixation on beads, (3) cloning by PCR, (4) sequential ligation to DNA fragment after fluorescent labeling, and (5) detection by color. Error type is substitution and sometimes error in recognition of bases is due to noise in the ligation cycle. Moreover, it is a slow process and only applicable to short reads (Kchouk et al., 2017).

12.2.3 Third generation (current generation)

The stumbling blocks in NGS were short reads and difficult assembly of the genome. Third-generation sequencing is a considerable improvement over first and second generation as it can quickly produce long reads at low cost and easily prepare samples without the requirement of PCR (Kchouk et al., 2017). The current generation uses two approaches, i.e., single-molecule real time (SMRT) and synthetic approaches (Goodwin et al., 2016). The following are the latest methods of sequencing that utilize the SMRT approach.

12.2.3.1 PacBio

PacBio. is a commonly used third-generation sequencing technique that uses the SMRT approach launched by Pacific Biosciences. This sequencing technique recognizes the sequence of DNA molecules during replication. It involves many SMRT cells, which are composed of zero-mode waveguide (ZMW). DNA polymerase and DNA fragments are attached to the bottom of each ZMW. Every time a nucleotide is incorporated into the DNA fragment by using polymerase a luminance signal is liberated, which is detected by sensors (Rhoads and Au, 2015). PacBio sequencing can produce longer reads (∼10,000 bp/read) compared to the second generation in a shorter period but with exceptionally large error rates (Kchouk et al., 2017). Indel errors in PacBio are randomly distributed in long reads (Kulski, 2016; Koren et al., 2012).

12.2.3.2 Oxford Nanopore

Oxford Nanopore Technologies promises to produce longer reads with high-resolution repeats and variants. They generate a device called MinION, which is a portable single-molecule nanopore sequencer device that can connect to a laptop by USB 3.0 (Mikheyev and Tin, 2014). The sample is simply loaded in the device, and without delay data from longer reads (>150 kbp) are generated on the screen. This technique involves passing a fragment of DNA through a nanopore (protein nanopore) after attaching to the complementary strand of a hairpin. An ionic current is generated when a DNA fragment enters the pore, and the variation in ionic current is measured and recorded in graphical model (Jain et al, 2016). It has various advantages like portability and low cost over any other sequencing techniques but at the cost of high error (∼12%) (Kchouk et al., 2017; Ip et al., 2015).

12.3 Genome assembly

Grouping together data obtained from the foregoing sequencing techniques is the next challenge for researchers in the process of genome analysis. Fragments of sequences are subjected to pretreatment because it is necessary to put all the fragments together for genome annotation and minimize the possibility of error (Fig. 12.2). Reads obtained from the sequencing technique are trimmed and bases of low-quality are filtered by using quality filters. Then all the fragments are assembled after estimation of size and correcting reads (Wojcieszek et al, 2014.; Ekblom and Wolf, 2014).

12.3.1 De novo assembly

This process involves the assembling of short reads and long reads to make a complete genome without a reference sequence (de Sá et al., 2018). Several software packages were developed based on the following computational approaches for de novo assembly of the genome:

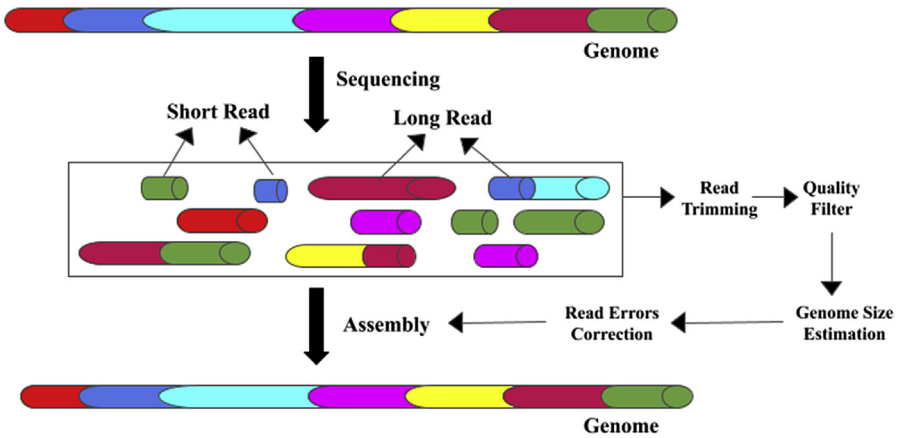


FIGURE 12.2

Schematic representation of genome assembly.

1. *Greedy algorithm:* This approach is based on alignment of reads for the analysis of genome (Wojcieszek et al, 2014; Ekblom and Wolf, 2014). Software for genome assembly based on greedy algorithms are VCAKE (Jeck et al, 2007), SSAKE (Warren et al, 2007), and SHARCGS (Dohm et al, 2007).
2. *Overlap-layout-consensus (OLC):* The OLC approach involves construction of graphs after recognition of overlaps in the reads and ultimately creating a consensus sequence (Wojcieszek et al., 2014; Ekblom and Wolf, 2014). Software programs for genome assembly based on the OLC approach are Mira (Chevreux et al., 2004), Edena (Hernandez et al., 2008), and Newbler (Reinhardt et al., 2009).
3. *De Bruijn graphs:* This approach identifies the overlaps by creating k-mer and (k-1)-mer read lengths (k represents original read lengths). Afterward, they determine (k-1)-mers out of k-mers and represent them graphically (Wojcieszek et al., 2014; Ekblom and Wolf, 2014). Software programs for genome assembly based on De Bruijn are Velvet (Zerbino and Birney, 2008), SPAdes (Bankevich et al., 2012), SOAPdenovo (Luo et al., 2012), and ALL-PATHS-LG (Gnerre et al., 2011).

These consensus sequences are actually different parts of a genome and need to be arranged by the process of scaffolding as they are composed of gap regions. After genome assembly, software programs like GAPFILLER were applied for gap closure and produced scaffold (Boetzer and Pirovano, 2012). The assembled contiguous sequences are then evaluated by mapping.

12.3.2 Reference assembly

This involves allocation of sequences at a specific position on a genome. Sequence mapping is a challenge for computational methods as most of the data obtained from

computational sequencing techniques are short reads. Software packages for alignments are classified based on the portion aligned in the reads, i.e., local alignment (only small parts of reads) and global alignment (full length of reads) (de Sá et al, 2018). Challenges for reference assembly are:

1. Complexity is due to huge amounts of data from sequencing techniques.
2. Quality of mapping for reference is important, therefore errors are minimized.
3. Original genetic variation and errors produced during sequencing should be distinguishable.

Based on the analysis of DNA, RNA, and miRNA, various software packages are available for mapping such as Bowtie (Langmead et al., 2009), BWA (Li and Durbin, 2009), SHRiMP (Rumble et al., 2009), SOAP2 (Li et al., 2009), TopHat2 (Kim et al., 2013), and mrsFAST (Hach et al., 2010).

12.4 Genome annotation

Genome sequencing platforms paved the way for researchers to understand the features, functions, and structures of genes. With the advancement in technology and improvement in computational methods, it is now possible to provide unprecedented information about the genome. Annotation is a very crucial part of genome analysis and there are many ways to define genome annotation like "genome annotation is identification and interpretation of features and functions of the genome by using biological facts and computational methods" (Fyad et al., 2016) or "genome annotation is a subfield of genome analysis which can be done by using computational tools" (Koonin and Galperin, 2003) or "genome annotation is the ability to interpret the effects of variations on the function of gene by gathering information from structure of the gene" (Steward et al, 2017).

Genome annotation is preceded by a gene prediction algorithm that determines gene structures associated with transcription, coding protein, splicing, etc. (Mudge and Harrow, 2010). When a genome is sequenced, it is important to annotate it, as annotation describes the functions associated with the product of the gene. But many researchers still believe that it is an unreliable process in genome analysis due to its inaccuracy. Annotation of the genome is diversified into interpreting assorted features of genes. Genome annotation can extract the name of the gene, its characteristic function, physical behavior, and altogether metabolic activities performed by genes in the organism (Koonin and Galperin, 2003).

12.4.1 Levels of genome annotation

Overall, the process of genome annotation is organized in three levels for a better understanding of the outcomes of the genome (Stein, 2001). Annotation of the genome is precisely associated with the following three levels, which are inextricably linked (Fig. 12.3).

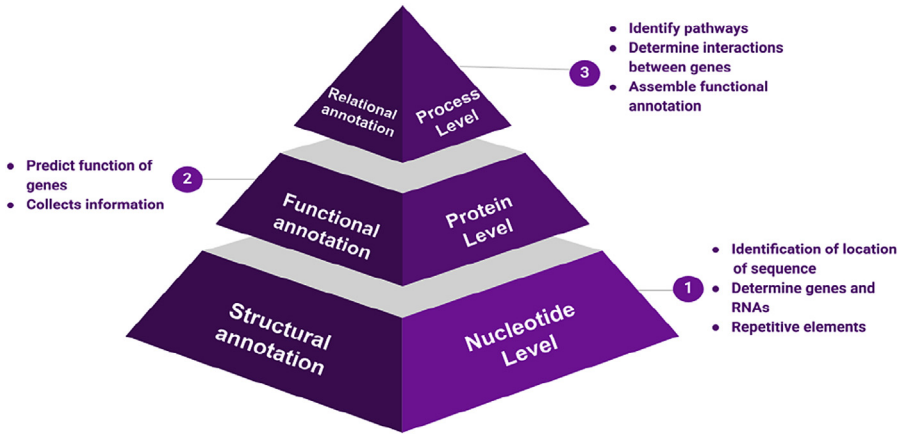


FIGURE 12.3

Levels of genome annotations.

12.4.1.1 Nucleotide level

This is also considered a syntactic or structural level of genome annotation as it determines the structure and component of a gene. This level involves the identification of location of relevant DNA sequences. It also seeks to locate repetitive elements in genes and RNAs. This level is associated with error during the sequencing processes, generally indel errors (Fyad et al., 2016; de Sá et al., 2018).

12.4.1.2 Protein level

This is the functional level of annotation that determines the function of genes identified in the nucleotide level. It predicts the functions by identifying the similarities in structures and patterns of sequence with the experimental data obtained from the literature with computational protein/gene datasets (Fyad et al., 2016; de Sá et al., 2018).

12.4.1.3 Process level

Process level annotation is relational and contextual as it identifies the process and pathways that interact with various genes and other biological elements. This results in several regulatory networks, various gene families, and other metabolic networks (Fyad et al., 2016; de Sá et al., 2018).

The genome annotation-utilized in silico approach explains all the foregoing requirements (Médigue et al., 2002). Bioinformatics software has been developed for ab initio genome annotations based on the following three specific features:

1. *Signal sensor:* This recognizes the functional site in the gene and is generally a small sequence motif, e.g., start codons, stop codons, branch points, TATA box, polypyrimidine tract, splicing sites, etc.

2. *Content sensor:* This senses the different regions of DNA and classifies them based on their content and codon structure, e.g., codon usage, dicodon (AAG-AAG) frequency, G + C content, etc.
3. *Similarity detection:* This determines the degree of similarity between different DNA, RNA, and protein sequences, e.g., similarities in mRNAs of the same organism, similarity in proteins between related individuals, etc. It is the ratio of similar residue over the length of the aligned sequence and similarity is measured in percentage (Stein, 2001; de Sá et al., 2018).

The schematic representation of genome annotation is shown in Fig. 12.4. Genome annotation started with genome sequencing. Predicting functions of genes is associated with statistical gene methods such as GeneMark and GLIMMER, which are successfully used (based on hidden Markov models) for prokaryotes, and less successfully GENSCAN for eukaryotes. This results in a set of data for genes, proteins, or RNAs, which is a combined result of general database searching in the National Center for Biotechnology Information (NCBI), statistical gene prediction, and structural feature prediction. Structural features are predicted by

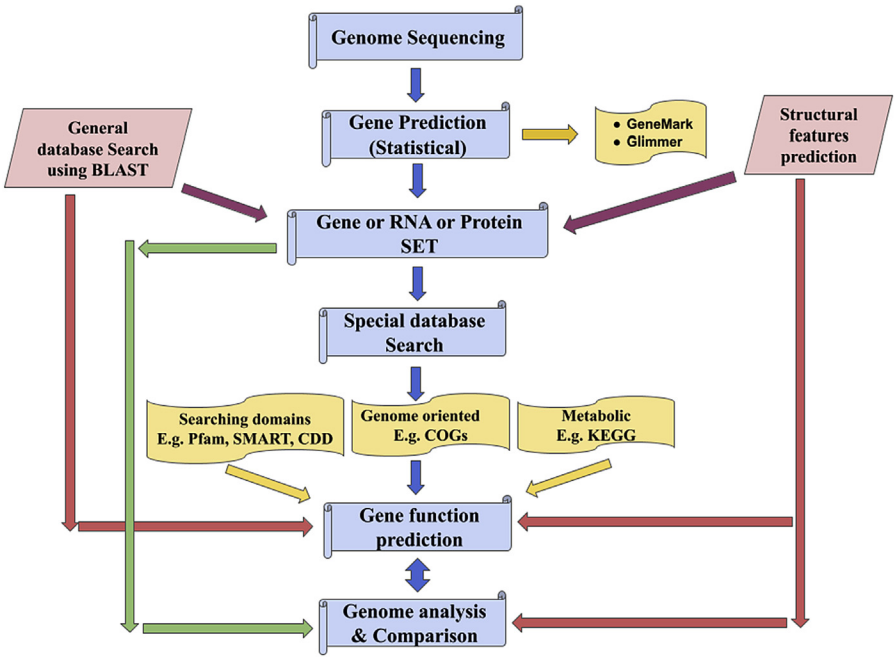


FIGURE 12.4

Genome annotation flow chart. *BLAST*, Basic Local Alignment Search Tool; *CDD*, Conserved Domains Database; *COGs*, clusters of orthologous groups of proteins; *KEGG*, Kyoto Encyclopedia of Genes and Genomes; Pfam, protein families; *SMART*, Simple Modular Architecture Research Tool.

peptide signal, coiled domain transmembrane, and other features in protein. These sets help to determine the errors in the sequences, which might arise during genome sequencing and thus provide feedback to genome sequencing techniques. Function of a gene is predicted by using special database searches, which are categorized into three parts: (1) searching domain databases for conserved domains like SMART, CDD, and Pfam; (2) genome-oriented databases for determining the relationship between homologous genes and predicting functions like COGs; and (3) metabolic databases for reconstructing metabolic pathways. Final analysis and comparison of genome is interlinked with gene function prediction and other database searches (Koonin and Galperin, 2003).

12.4.2 Tools for genome annotation

Since genome annotation requires different levels of predictions and identifications (Stein, 2001), manual pasting of sequences in the computer is practically infeasible. Gaasterland and Sensen predicted that genome annotation of a sequence requires a minimum of 1 year per person for one megabase by hand (Gaasterland and Sensen, 1996). Due to this, genome annotation is the limiting step in several genome projects. Therefore requirement of automation at several levels of genome annotation is essential. Hence, to make it practicable, every genome project uses software to achieve automatic run of routine tasks and then organizes the outcomes in a convenient fashion (Reeves et al., 2009).

The first automated tool for genome analysis was GeneQuiz (Scharf et al., 1994). This project was fully automatic and ran databases, analyzed sequences, evaluated results, and generated functional annotations automatically. Several other automated tools were created after GeneQuiz such as PEDANT (Frishman et al., 2001), MAGPIE (Gaasterland and Sensen, 1996), ERGO (Overbeek et al., 2003), and Imagene (Médigue et al., 1999) but only GeneQuiz was open to the public (Table 12.2). Unfortunately, annotation by GeneQuiz generates a significant number of errors in the sequence similarity analysis. The latest automatic genome annotation tool for eukaryotes is Ensembl; it obtains data from mRNA, protein sequence, and RNA-seq (Zerbino et al., 2018). However, manual annotation is still considered standard because of its high accuracy. Thus a project named ENCODE was developed to determine the accuracy of computational annotation methods by comparing with manual gene annotation assembled by the Human and Vertebrate Analysis and Annotation (HAVANA) group. Surprisingly, only 3.2% transcripts estimated by computational means were found to be valid (ENCODE Project Consortium, 2007; Harrow et al., 2006). There are two main groups that generate manual annotations: (1) HAVANA at Wellcome Trust Sanger Institute in the United Kingdom, and (2) RefSeq at NCBI in the United States (Pruitt et al., 2014). HAVANA is well known for its excellent quality of manual transcript and gene annotation. HAVANA is associated with other computational groups and identifies pitfalls in annotation by experimentally annotating transcripts and providing feedback

Table 12.2 Genome annotation tools.

Genome annotation tools	Automated/manual	URL
GeneQuiz	Automated	https://www.osti.gov/biblio/377162-genequiz- workbench-sequence-analysis
PEDANT	Automated	https://academic.oup.com/nar/article/31/l/207/24 01158
MAGPIE	Automated	https://bioinformatics.tugraz.at/sensencw/magpie.htm
ERGO	Automated	http://ergo.integratedgenomics.com/ERGO
Imagene	Automated	http://www.imagene.eu/
Ensembl	Automated	http://www.ensembl.org/
PFAM	Manual + automated	http://pfam.xfam.org/
HAVANA	Manual	https://www.sanger.ac.uk/project/manual-annotation/
RefSeq	Manual	https://www.ncbi.nlm.nih.gov/RefSeq

to computational groups so that they can improve their analysis (Harrow et al., 2012). However, RefSeq is not completely manual; 45% of transcripts of RefSeq are computationally annotated (Steward et al., 2017).

12.4.3 Reliability of genome annotation

For many scientists, genome annotation is considered an unreliable process as it produces errors and incorrect annotation of the genome. Despite the low error rate, flaws are highly visible due to the massive size of the genome. Production of enormous data leads to the accumulation of errors. These errors can be algorithm errors due to bugs in the script or program, and clerical errors due to humans (Koonin and Galperin, 2003). Therefore several challenges and problems are faced by researchers when converting quantity into quality. These automated and manual annotation tools are easy and accurate at low levels, e.g., for 10 sequences, but when their number changes to 100,000 their accuracy decreases and labor increases dramatically (Salzberg, 2019). Furthermore, error rate increases at a much higher rate in computational methods compared to manual methods. As discussed earlier, the ENCODE project conceived a huge error in computational annotation methods when compared with the same data with manual annotators (ENCODE Project Consortium, 2007).

Besides this, computational methods have several advantages over manual methods because they are economical and determine unknown information quickly, but on the other hand manual annotation methods are reliable and have better quality. So, why not take benefits from both? For improved results, scientists made a hybrid/mixed annotation process, e.g., in the GENCODE project, and took

advantage from both automated and manual tools. Annotation by computational means provides information to the manual annotation groups and gives hints for unannotated features in the gene, whereas manual annotators identify the errors made by automated annotators and help them to make further improvements. GENCODE predicts the genome annotation by making use of manual (HAVANA) and automated (Ensembl) annotators (Harrow et al., 2012).

The GENCODE model is now clinically used as it can describe IncRNA, sRNA, protein coding gene, and pseudogene with good-quality annotation (Coffey et al., 2011) Moreover, GENCODE is collaborated with RefSeq to recognize CoDing sequences agreed by both in the protein coding gene and hence further improves the consensus CoDing sequence (Farrell et al., 2014). UCSC browser (Casper et al., 2018) and Ensembl (Zerbino et al., 2018) both display GENCODE models that are accessible to the public and update after every 6 months.

12.5 Techniques for gene expression analysis

Gene expression in simple terms is a process by which the information for proteins is procured by coding genes in the form of three bases (Crick, 1970). Gene expression regulates the appearance of phenotype and mechanisms that are responsible for functions in living organisms. Gene expression techniques have numerous advantages in biomedical research of cancer diagnosis and treatment, and subdivision of various other diseases. Several techniques have been developed to date to determine the expression possessed by various genes. Northern blot (Pall and Hamilton, 2008), Western blot (Morash et al, 1999), reverse transcription polymerase chain reaction (RT-PCR) (Muller et al., 2002), microarray analysis (Schena, 1996), fluorescent in situ hybridization (Zenklusen and Singer, 2010), serial analysis of gene expression (SAGE) (Yamamoto et al., 2001), and RNA-seq (Ji and Sadreyev, 2018) are some techniques used to determine the expression of genes. However, these techniques work in two ways, i.e., either they evaluate protein level (e.g., Western blot) or evaluate mRNA level (e.g., Northern blot, RT-PCR, microarray analysis, etc.). However, techniques involved in measuring mRNA showed improved results compared to techniques involved in measuring protein. Some of the automated techniques are discussed next.

12.5.1 SAGE

SAGE stands for serial analysis of gene expression. This technique allows digital analysis of genome-wide expression. The gene expression profiles produced by SAGE are sensitive and comprehensive. SAGE does not require prior knowledge of sequence and it works well even in low quantities of mRNA transcripts. This technology generates a collection of short sequence tags that can uniquely recognize transcripts. The expression level of each transcript in SAGE is dependent on the number of detection cycles of a tag (Yamamoto et al., 2001) (https://www.thermofisher.com/).

12.5.2 DNA microarray

DNA microarray or DNA chip or biochip is an extremely useful technique to discover the gene expression profile of several genes from different sites of genome at one time. This technique involves formation of an array of DNA spots on a solid surface, which can be silicon chip or glass slide. Microarray is based on hybridization of mRNA to template DNA. Gene discovery and gene expression are determined simultaneously by binding of complementary sequences to the DNA spots. A DNA chip contains thousands of DNA spots. Thus a DNA chip furnishes information of all the samples at one time. Printing of probes/samples on solid surfaces is now automatic (Schena, 1996). This technique is reliable and cost effective and therefore it has been used in gene discovery, disease diagnosis like cancers, pharmacogenomics for relating drugs with genomic profiles, and toxicological research (Russo et al., 2003).

12.5.3 RNA-seq

This is a highly comprehensive and sensitive technique for identification, comparison, and measurement of gene expression (Ji and Sadreyev, 2018). It is based on data obtained from NGS methods. Reads created in NGS methods first undergo mapping for characterization and then identification of differential gene expression (de Sá et al., 2018). In the presence of a reference, Bioscope (Pinto et al., 2014) is used first to map the reads, then DEGseq (Wang et al., 2010) or TopHat-Cufflinks pipelines (Trapnell et al., 2012) are used to analyze differential gene expressions. However, if a reference is unavailable, SOAPdenovo-Trans (Xie et al., 2014), Trinity (Grabherr et al., 2011), and Trans-Abyss (Robertson et al., 2010) are used to represent transcripts. These transcripts are then mapped for quantification of expression and finally expression is identified in different conditions. There are two advantages of RNA-seq: (1) errors in the genome annotation can be corrected by RNA-seq, and (2) gene prediction, e.g., GeneMark-ET, is possible (Lomsadze et al., 2014; de Sá et al., 2018).

12.6 Gene expression data analysis

12.6.1 Data analysis by data mining

High-throughput technologies like RNA-seq, microarray, etc. erupt enormous data that are rich in information. As discussed in the previous section, analyzing these massive data is a challenge for researchers. Datasets generated from sequencing and gene expression require large-scale data mining of the genome. Data mining is considered a process of analyzing biological relevance information from large datasets obtained from sequencing, gene expressions, and interaction studies (Fyad et al., 2016). Methodologies involved in data mining are divided into two categories, i.e., clustering and classification techniques (Fig. 12.5). These methods have been extensively used to determine the characteristics of genes (Lee et al., 2008).

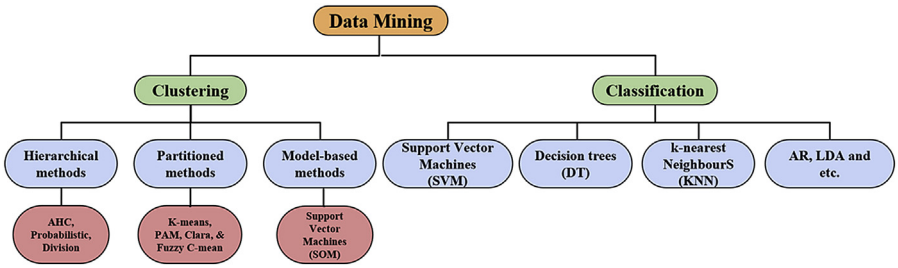


FIGURE 12.5

Categorization of data mining methods. *AR*, Association rule; *LDA*, Latent Dirichlet Allocation.

12.6.1.1 Clustering method

This technique has various applications in biomedical sciences. It helps to reduce data by grouping and clustering. Clustering analysis determines clusters and converts high-dimensional data produced in gene expression techniques into 2D data (Lee et al., 2008). Clustering techniques have several applications in cancer diagnostics (Smolkin and Ghosh, 2003). This technique is only effective when some prior knowledge of data is known. Hierarchical, partitioned, and model based are three methods involved in the clustering of data. Their principles, advantages, and disadvantages are discussed next.

12.6.1.1.1 Hierarchical

This method is used to recognize differential gene expression in sarcopenia and to visualize and categorize data obtained from proteomics of pathogenic species of bacteria (Meunier et al., 2007). Hierarchical methods include agglomerative hierarchical methods, probabilistic methods, and division methods.

Principle: This method assigns elements to other closely related elements by merging small groups into large clusters or splitting large clusters into smaller groups.

Advantages: No requirement of initial input of parameters. Prior knowledge of the number of clusters is not compulsory. It makes a complete hierarchy with intuitive visual distribution of data.

Disadvantages: Clusters are not clear, and they cannot automatically discover required clusters (Fyad et al., 2016).

12.6.1.1.2 Partitioned

This method is used to identify gene expression in yeast during cell cycles. It gives a better engrossed distribution of clusters during the diagnosis of cancers like leukemia and melanoma. It includes K-means, partitioning around medoids, Clara, and fuzzy C-mean methods (Kim et al., 2005).

Principle: It is based on degrading the whole datasets into separate clusters (subsets) and then determining the similarities in behavior of genes in those subsets.

Advantages: This method has several advantages over other methods of clustering. For example, K-means is best suited for large-scale data analysis and is easy to execute and highly efficient.

Disadvantages: Prior knowledge of the number of clusters is compulsory. It has high sensitivity toward outliers, start point, and noisy data (Fyed et al., 2016).

12.6.1.1.3 Model based

This is used to determine gene groups in the differentiation mechanism of cells involved in intestinal absorption (enterocytes), e.g., self-organizing map (Bédrine-Ferran et al., 2004).

Principle: It is based on the division of genes in a partitioning experiment into geometrically present structures of subgroups.

Advantages: Similarity in the data is a function of position of clusters/groups, hence close-lying data have similar expression profiles.

Disadvantages: First, outcome is dependent on the distance and second, an expected number of groups need to be specified (Fyad et al., 2016).

12.6.1.2 Classification methods

Classification methods make use of preclassified genomic data to generate predicted genomic models for various categories. Such analyses have been employed in cancer research as an alternative technique for the diagnosis of cancer. However, prediction of gene expressions faces problems during classification in biomedical sciences. Various methods have been developed for classification modeling such as support vector machines (SVMs), gene voting, Bayesian regression models, decision trees (DTs), partial least squares, association rule (AR), kernel estimation (KNN), linear discriminant analysis, etc. (Lee et al., 2008; Fyad et al., 2016). Some of the classification methods are discussed next.

12.6.1.2.1 KNN

Kernel estimation or k-nearest neighbors based on searching the k-nearest neighbor of a given sample depends on distance measure. Due to its easy accessibility and interpretability, it is commonly used in classification of diseases and other clinical purposes. It has been reported that KNN modeling is implicated for systematic analysis of data obtained from the gene expression microarray of various cancers (Parry et al., 2010).

12.6.1.2.2 SVM

Support vector machine method is used for the classification of tumors. The SVM classification technique involves the separation of two classes by identifying the hyperplane and is based on mapping data in kernel space. A combination of SVM and mutual information have been reported for classifying lymphoma and colon cancer (Ca and Mc, 2015). A major drawback in this classification technique is that the SVM is only applicable to a maximum of two classes.

12.6.1.2.3 DT

Decision tree is one of the well-known algorithms that have been used in bioinformatics due to its simplicity, easy interpretability, and ability to handle huge amounts of data. The DT genomic model is developed in the form of a tree where the large dataset is split into smaller and smaller subsets, and a side-by-side decision tree is progressively developed. Results obtained from DT are better than SVM. Moreover, its performance for identification of cancers further improves when it is used in combination with particle swarm optimization (Chen et al., 2014).

12.6.2 Data analysis by ontology

Ontology is a representation of information and knowledge in a domain. Gene ontology (GO) is a representation of biological knowledge that shows the properties of genes and their product and relates the molecular components and cellular components with the biological processes (Fyad et al., 2016). GO helps the research community to enhance and enable the sharing of data (Gene Ontology Consortium, 2012).

GO is accessible on the AmiGO portal, which consists of lots of information and references (Carbon et al., 2009). For biomedical research, the Open Biomedical Ontology project (Ghazvinian et al., 2011) has been created for reference ontologies and the National Center for Biomedical Ontology made a bioportal (Whetzel et al., 2011) for biomedical researchers. Furthermore, the Sequence Ontology project (Eilbeck et al., 2005) was created for the characterization of genomic sequencing for genome annotation. The Gene Expression Omnibus (GEO) project at the NCBI is designed as a public repository for storage, submission, and retrieval of hybridization array and gene expression high-throughput data (Barrett et al., 2007). Similarly, ArrayExpress at the European Bioinformatics Institute is another source for gene ontologies (Parkinson et al., 2007). The Microarray Gene Expression Data Society furnishes all aspects of gene expression data by DNA chips for annotation, management, and sharing.

Besides this, GO provides static representation of biological components like DNA or RNA sequence, genes, and gene products but it cannot permit a proper visualization. Hence, the required functional analysis of gene expression is established by a combination of data mining and GO. Analysis of gene expression data is performed as a first determination of the cluster or group of coexpressed genes using data mining and then its functional analysis by GO (Chabalier et al., 2007).

12.7 Software for gene expression analysis

Several software programs have been developed based on clustering and classification techniques to analyze gene expression data (Fyad et al., 2016). These software programs are the combination of clustering and graphical methods that provide comprehensible images of outcomes generated by gene expression. Some of these programs that have been developed for grouping data from gene expression are listed in Table 12.3 with their applications, language, and URL.

Table 12.3 Software for gene expression analysis.

Software	Language	Applications	URL
MAGIC Tools	Java	Microarray data analysis	http://www.mybiosoftware.com/magic-tool-2-1-microarray-genome-imaging-clustering-tool.html
Cluster and Treeview	C	Organization and analysis of datasets from microarray and other experiments	http://bonsai.hgc.jp/~mdehoon/software/cluster/manual/Introduction.html#Introduction
Weka	Java	Data classification, preprocessing, clustering, association rule, and visualization	http://www.cs.waikato.ac.nz/ml/weka/
SAS	C	Data visualization and analysis	https://www.sas.com/en lu/home.html
IBM SPSS	Java	Data mining	http://www.spss.com/software/modeling/modeler-pro/
MeV	Java	Data stratification, clustering, classification, visualization, and analysis especially for microarray and RNA-seq	http://mev.tm4.org/#/welcome
LIBSVM	C++ and Java	Data classification	https://www.csie.uni.edu.tw/~cjlin/1ibsvm/
SVMlight	C	Data classification and analysis	http://svmlight.joachims.org

12.8 Computational methods for clinical genomics

DNA sequencing, genome annotation, and gene expression analysis are now basic requirements for biomedical research involved in the classification, diagnosis, and treatment of diseases. For example, once the genome of a patient is sequenced, it is compared with the reference for the analysis of functional variants. Genome analysis is particularly important for clinical applications. Researchers are seeking to perceive knowledge to understand the accurate annotation and function of genomic data. But in some cases, current technologies disappoint clinicians when identifying correct pathogenic variants responsible for diseases in patients. Errors in genome annotation strongly influence the identification of variants in the gene of a patient (Steward et al., 2017). Therefore it is necessary to resolve these abnormalities by further improving and reanalyzing the technologies involved in genome sequencing, annotation, and gene expression analysis.

12.9 Conclusion

Bioinformatics tools provide computational screens for identification, storage, and analysis of massive genomic data. Their low cost, easy accessibility, and speed have attracted biological researchers to better understand cellular processes. Progress and advancement in technology with computers yield better sequencing, annotation, data mining, and gene expression techniques. The launch of a new technique in the field of genome analysis promises high coverage, low cost, and better quality. The in silico approach has benefited the insightful description of raw as well as analyzed data. Moreover, all these requirements in addition to human resources must be upgraded according to the demands due to increased data production.

This chapter provided an overview of all the steps and techniques involved in genome data analysis and how improvement in technologies led to the complete and exact assembly of genomic data. Clinically, these current technologies in genomics play a crucial role in disease diagnostics and treatment. They help in the treatment of patients having genetic abnormalities by determining pathogenic variants. Especially in the case of cancer, sequencing and differential gene expression techniques decipher the behavior of the gene involved. However, inaccuracy in the computational methods of genome analysis is a major pitfall. Enhancement in genome annotation methods is necessary to resolve problems in the present techniques of genome analysis.

Abbreviations

ABI	Applied Biosystems Inc.
BLAST	Basic Local Alignment Search Tool
BWA	Burrows−Wheeler alignment
CCD	coupled charge device
COGs	clusters of orthologous groups of proteins
DEGseq	differentially expressed genes or isoforms for RNA-seq
dNTPs	dideoxy nucleotides
DT	decision tree
EBI	European Bioinformatics Institute
ENCODE	ENCyclopedia Of DNA Elements
GLIMMER	Gene Locator and Interpolated Markov Model ER
GO	gene ontology
HAVANA	Human and Vertebrate Analysis and Annotation
HGP	Human Genome Project
IGR	intergenic regions
indel	insertion and deletion
IT	Ion Torrent
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	k-nearest neighbor

MAGIC	Mining Algorithm for GenetIc Controllers
MAGPIE	Multipurpose Automated Genome Project Investigation
MeV	MultiExperiment Viewer
mRNA	messenger RNA
mrsFAST	microread (substitutions only) fast alignment and search tool
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
OLC	overlap-layout-consensus
PCR	polymerase chain reaction
Pfam	protein families
RefSeq	reference sequence
RNA-seq	RNA sequencing
RT-PCR	reverse transcription polymerase chain reaction
SAGE	Serial Analysis of Gene Expression
SAS	Statistical Analysis System
SEALS	System for Easy Analysis of Lots of Sequences
SMART	Simple Modular Architecture Research Tool
SMRT	Single-Molecule Real Time
SO	Sequence Ontology
SOAP	Short Oligonucleotide Analysis Package
SOLiD	Supported Oligonucleotide Ligation and Detection
SPSS	Statistical Package For The Social Sciences
SVM	support vector machines
UCSC	University of California Santa Cruz
VCAKE	Verified Consensus Assembly by K-mer Extension
Weka	Waikato Environment for Knowledge Analysis
ZMW	zero-mode waveguide C

References

1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature 467 (7319), 1061.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., Pyshkin, A.V., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19 (5), 455–477.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Edgar, R., 2007. NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res. 35 (Suppl. l_1), D760–D765.

Baxevanis, A.D., Bader, G.D., Wishart, D.S. (Eds.), 2020. Bioinformatics. John Wiley & Sons.

Bédrine-Ferran, H., Le Meur, N., Gicquel, I., Le Cunff, M., Soriano, N., Guisle, I., Le Gall, J.Y., et al., 2004. Transcriptome variations in human CaCo-2 cells: a model for enterocyte differentiation and its link to iron absorption. Genomics 83 (5), 772–789.

Boetzer, M., Pirovano, W., 2012. Toward almost closed genomes with GapFiller. Genome Biol. 13 (6), R56.

Ca, D.A.V., Mc, V., 2015. Gene expression data classification using support vector machine and mutual information-based gene selection. Proc. Comp. Sci. 47, 13–21.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., 2009. The AmiGO hub & web presence working group. AmiGO 25 (2), 288–289.

Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Hinrichs, A.S., et al., 2018. The UCSC genome browser database: 2018 update. Nucleic Acids Res. 46 (D1), D762–D769.

Chabalier, J., Mosser, J., Burgun, A., 2007. A transversal approach to predict gene product networks from ontology-based similarity. BMC Bioinform. 8 (1), 1–12.

Chen, K.H., Wang, K.J., Tsai, M.L., Wang, K.M., Adrian, A.M., Cheng, W.C., Yang, T.S., Tan, K.P., Chang, K.S., 2014. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinform. 15 (1), 49.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E., Wetter, T., Suhai, S., 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14 (6), 1147–1159.

Clough, E., Barrett, T., 2016. The gene expression omnibus database. In: Statistical Genomics. Humana Press, New York, NY, pp. 93–110.

Coffey, A.J., Kokocinski, F., Calafato, M.S., Scott, C.E., Palta, P., Drury, E., Joyce, C.J., LeProust, E.M., Harrow, J., Hunt, S., Turner, D.J., Hubbard, T.J., Palotie, A., Lehesjoki, A.E., 2011. The GENCODE exome: sequencing the complete human exome. Eur. J. Hum. Genet. 19 (7), 827–831.

Crick, F., 1970. Central dogma of molecular biology. Nature 227 (5258), 561–563.

De Filippo, C., Ramazzotti, M., Fontana, P., Cavalieri, D., 2012. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Briefings Bioinf. 13 (6), 696–710.

de Sá, P.H., Guimarães, L.C., das Graças, D.A., de Oliveira Veras, A.A., Barh, D., Azevedo, V., da Costa da Silva, A.L., Ramos, R.T., 2018. Next-generation sequencing and data analysis: strategies, tools, pipelines and protocols. In: Omics Technologies and Bio-Engineering. Academic Press, pp. 191–207.

Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H., 2007. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res. 17 (11), 1697–1706.

Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30 (1), 207–210.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., Ashburner, M., 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 6 (5), R44.

Ekblom, R., Wolf, J.B., 2014. A field guide to whole-genome sequencing, assembly and annotation. Evolut. Appl. 7 (9), 1026–1042.

ENCODE Project Consortium, 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447 (7146), 799.

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489 (7414), 57–74.

Farrell, C.M., O'Leary, N.A., Harte, R.A., Loveland, J.E., Wilming, L.G., Wallin, C., Hiatt, S.M., et al., 2014. Current status and new features of the Consensus Coding Sequence database. Nucleic Acids Res. 42 (D1), D865–D872.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Jou, W.M., Molemans, F., Raeymaekers, A., Van den Berghe, A., Ysebaert, M., Volckaert, G., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260 (5551), 500−507.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Merrick, J.M., et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269 (5223), 496−512.

Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., Mewes, H.W., 2001. Functional and structural genomics using PEDANT. Bioinformatics 17 (1), 44−57.

Fyad, H., Barigou, F., Bouamrane, K., 2016. Computational methods for functional analysis of gene expression.

Gaasterland, T., Sensen, C.W., 1996. MAGPIE: automated genome interpretation. Trends Genet. 12 (2), 76−78.

Galibert, F., Alexandraki, D., Baur, A., Boles, E., Chalwatzis, N., Chuat, J.C., Coster, F., Cziepluch, C., De Haan, M., Domdey, H., Entian, K.D., Gatius, M., Goffeau, A., Grivell, L.A., Hennemann, A., Herbert, C.J., Heumann, K., Hilger, F., Hollenberg, C.P., Huang, M.-E., Jacq, C., Jauniaux, J.-C., Katsoulou, C., Kirchrath, L., Kleine, K., Kordes, E., Kotter, P., Liebl, S., Louis, E.J., Manus, V., Mewes, H.W., Miosga, T., Obermaier, B., Perea, J., Pohl, T., Portetelle, D., PujoI, A., Purnelle, B., Rad, M.R., Rasmussen, S.W., Rose, M., Rossau, R., Schaaff-Gerstenschlager, I., Smits, P.H.M., Scarcez, T., Soriano, N., Tovan, D., Tzermia, M., Broekhoven, A.V., Vandenbol, M., Wedler, H., Wettstein, D.V., Wambutt, R., Zagulski, M., ZolIner, A., Karpfinger-Hartl, L., Durand, P., 1996. Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X. EMBO J. 15 (9), 2031−2049.

Gene Ontology Consortium, 2012. The gene ontology: enhancements for 2011. Nucleic Acids Res. 40 (D1), D559−D564.

Ghazvinian, A., Noy, N.F., Musen, M.A., December 2011. How orthogonal are the OBO Foundry ontologies?. In: Journal of Biomedical Semantics, vol. 2. BioMed Central, p. S2. No. S2.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U.S.A 108 (4), 1513−1518.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Hadley, D., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296 (5565), 92−100.

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17 (6), 333.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Chen, Z., et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29 (7), 644−652.

Guttmacher, A.E., Collins, F.S., 2005. Realizing the promise of genomics in biomedical research. Jama 294 (11), 1399−1402.

Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E.E., Sahinalp, S.C., 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat. Methods 7 (8), 576−577.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G.R., Storey, R., Swarbreck, D., Ucla, C., Hubbard, T., Antonarakis, S.E., Guigo, R., Rossier, C., 2006. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 7 (1), 1–9.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Barnes, I., et al., 2012. GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res. 22 (9), 1760–1774.

Hernandez, D., François, P., Farinelli, L., Østerås, M., Schrenzel, J., 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18 (5), 802–809.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., Welch, D.M., 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8 (7), R143.

Ip, C.L., Loose, M., Tyson, J.R., de Cesare, M., Brown, B.L., Jain, M., Piazza, P., et al., 2015. MinION analysis and reference Consortium: phase 1 data release and analysis. F1000Research 4.

Jain, M., Olsen, H.E., Paten, B., Akeson, M., 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17 (1), 239.

Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L., Jones, C.D., 2007. Extending assembly of short DNA sequences to handle error. Bioinformatics 23 (21), 2942–2944.

Ji, F., Sadreyev, R.I., 2018. RNA-seq: basic bioinformatics analysis. Curr. Protoc. Mol. Biol. 124 (1), e68.

Jou, W.M., Haegeman, G., Ysebaert, M., Fiers, W., 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 237 (5350), 82–88.

Kchouk, M., Gibrat, J.F., Elloumi, M., 2017. Generations of sequencing technologies: from first to next generation. Biol. Med. 9 (3).

Kim, S.Y., Choi, T.M., Bae, J.S., February 2005. Fuzzy types clustering for microarray data. WEC (2), 12–15.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14 (4), R36.

Koonin, E.V., Galperin, M.Y., 2003. Genome annotation and analysis. In: Sequence—Evolution—Function. Springer, Boston, MA, pp. 193–226.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., Phillippy, A.M., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. 30 (7), 693–700.

Kulski, J.K., 2016. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. Next Gen. Sequen. Adv. Appl. Chall. 3–60.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J.,., Funke, R., 2001. Initial Sequencing and Analysis of the Human Genome.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (3), R25.

Lee, J.K., Williams, P.D., Cheon, S., 2008. Data mining in genomics. Clin. Lab. Med. 28 (1), 145–166.

Levy, S.E., Myers, R.M., 2016. Advancements in next-generation sequencing. Annu. Rev. Genom. Hum. Genet. 17, 95–115.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows−Wheeler transform. Bioinformatics 25 (14), 1754−1760.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25 (15), 1966−1967.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. BioMed Res. Int. 2012.

Lomsadze, A., Burns, P.D., Borodovsky, M., 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 42 (15) e119-e119.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., Tang, J., et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1 (1), 2047-217X.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Dewell, S.B., et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437 (7057), 376−380.

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. 74, 560−564.

McClean, P., 2011. A History of Genetics and Genomics.

McKusick, V.A., Ruddle, F.H., 1987. A new discipline, a new name. New J.

McPherson, J.D., 2014. A defining decade in DNA sequencing. Nat. Methods 110, 1003−1005.

Médigue, C., Bocs, S., Labarre, L., Mathé, C., Vallenet, D., 2002. In silico annotation of genomic-Bioinformatics sequences (1). Med. Sci. 18 (2), 237−250.

Meunier, B., Dumas, E., Piec, I., Bechet, D., Hebraud, M., Hocquette, J.F., 2007. Assessment of hierarchical clustering methodologies for proteomic data mining. J. Proteome Res. 6 (1), 358−366.

Mikheyev, A.S., Tin, M.M., 2014. A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Res. 14 (6), 1097−1102.

Morash, B., Li, A., Murphy, P.R., Wilkinson, M., Ur, E., 1999. Leptin gene expression in the brain and pituitary gland. Endocrinology 140 (12), 5995−5998.

Mudge, J., Harrow, J., 2010. Methods for Improving Genome Annotation. Knowledge Based Bioinformatics: From Analysis to Interpretation. John Wiley & Sons, Chichester, West Sussex, pp. 209−214.

Muller, P.Y., Janovjak, H., Miserez, A.R., Dobbie, Z., 2002. Short technical report processing of gene expression data generated by quantitative real-time RT-PCR. Biotechniques 32 (6), 1372−1379.

Medigue, C., Rechenmann, F., Danchin, A., Viari, A., 1999. Imagene: an integrated computer environment for sequence annotation and analysis. Bioinformatics 15 (1), 2−15.

Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov Jr., E., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., Burd, H., Gardner, W., Hanke, P., Kapatral, V., Mikhailova, N., Vasieva, O., Osterman, A., Vonstein, V., Fonstein, M., Ivanova, N., Kyrpides, N., Bhattacharyya, A., 2003. The ERGO TM genome analysis and discovery system. Nucleic Acids Res. 31 (1), 1*64*−171.

Pall, G.S., Hamilton, A.J., 2008. Improved northern blot method for enhanced detection of small RNA. Nat. Protoc. 3 (6), 1077.

Pareek, C.S., Smoczynski, R., Tretyn, A., 2011. Sequencing technologies and genome sequencing. J. Appl. Genet. 52 (4), 413−435.

Parkinson, H.K., Shojatalab, M., Abeygunawardena, M., Coulson, N., Farne, R., Holloway, A.E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., Brazma, A., 2007. ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 35 (Suppl. 1), D747–D750.

Parry, R.M., Jones, W., Stokes, T.H., Phan, J.H., Moffitt, R.A., Fang, H., Shi, L., Oberthuer, A., Fischer, M., Tong, W., Wang, M.D., 2010. K-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. Pharmacogenomics J. 10 (4), 292–309.

Pinto, A.C., de Sá, P.H.C.G., Ramos, R.T., Barbosa, S., Barbosa, H.P.M., Ribeiro, A.C., Silva, W.M., Rocha, F.S., Santana, M.P., de Paula Castro, T.L., Schneider, M.P.C., Silva, A., Azevedo, V., Miyoshi, A., 2014. Differential transcriptional profile of Corynebacterium pseudotuberculosis in response to abiotic stresses. BMC Genom. 15 (1), 1–14.

Plomin, R., DeFries, J.C., McClearn, G.E., 2008. Behavioral Genetics. Macmillan.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Murphy, M.R., et al., 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 42 (D1), D756–D763.

Reeves, G.A., Talavera, D., Thornton, J.M., 2009. Genome and proteome annotation: organization, interpretation and integration. J. R. Soc. Interface 6 (31), 129–147.

Reinhardt, J.A., Baltrus, D.A., Nishimura, M.T., Jeck, W.R., Jones, C.D., Dangl, J.L., 2009. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. oryzae. Genome Res. 19 (2), 294–305.

Reuter, J.A., Spacek, D.V., Snyder, M.P., 2015. High-throughput sequencing technologies. Mol. Cell. 58, 586597. https://doi.org/10.1016/j.molcel.2015.05.004. Available from:

Rhoads, A., Au, K.F., 2015. PacBio sequencing and its applications. Dev. Reprod. Biol. 13, 178–289.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Griffith, M., et al., 2010. De novo assembly and analysis of RNA-seq data. Nat. Methods 7 (11), 909–912.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Hoon, J., et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475 (7356), 348–352.

Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A., Brudno, M., 2009. SHRiMP: accurate mapping of short color-space reads. PLoS Comput. Biol. 5 (5), e1000386.

Russo, G., Zegar, C., Giordano, A., 2003. Advantages and limitations of microarray technology in human cancer. Oncogene 22 (42), 6497–6507.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., et al., 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230, 1350–1354.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., et al., 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239, 487–491.

Salzberg, S.L., 2019. Next-generation Genome Annotation: We Still Struggle to Get it Right.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., Smith, M., 1977a. Nucleotide sequence of bacteriophage φX174 DNA. Nature 265 (5596), 687–695.

Sanger, F., Nicklen, S., Coulson, A.R., 1977b. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A 74 (12), 5463–5467.

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C., Sander, C., 1994. GeneQuiz: a workbench for sequence analysis. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, ISMB, vol. 94. AAAI Press, pp. 348–353.

Schena, M., 1996. Genome analysis with gene expression microarrays. Bioessays 18 (5), 427–431.

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26 (10), 1135–1145.

Smolkin, M., Ghosh, D., 2003. Cluster stability scores for microarray data in cancer studies. BMC Bioinform. 4 (1), 36.

Stein, L., 2001. Genome annotation: from sequence to biology. Nat. Rev. Genet. 2 (7), 493–503.

Steward, C.A., Parker, A.P., Minassian, B.A., Sisodiya, S.M., Frankish, A., Harrow, J., 2017. Genome annotation for clinical genomic diagnostics: strengths and weaknesses. Genome Med. 9 (1), 49.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7 (3), 562–578.

Wang, L., Feng, Z., Wang, X., Wang, X., Zhang, X., 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26 (1), 136–138.

Warren, R.L., Sutton, G.G., Jones, S.J., Holt, R.A., 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23 (4), 500–501.

Weissenbach, J., 2016. The rise of genomics. Comp. Rendus Biol. 339 (7–8), 231–239.

Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A., 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 39 (Suppl. l_2), W541–W545.

Wojcieszek, M., Pawełkowicz, M., Nowak, R., Przybecki, Z., 2014. Genomes correction and assembling present methods and tools. SPIE Proc 9290, 92901X. https://doi.org/10.1117/12.2075624. Available from:

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Zhou, X., et al., 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30 (12), 1660–1666.

Yamamoto, M., Wakatsuki, T., Hada, A., Ryo, A., 2001. Use of serial analysis of gene expression (SAGE) technology. J. Immunol. Methods 250 (1–2), 45–66.

Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13 (5), 329–342.

Zenklusen, D., Singer, R.H., 2010. Analyzing mRNA expression using single mRNA resolution fluorescent in situ hybridization. In: Methods in Enzymology, vol. 470. Academic Press, pp. 641–659.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18 (5), 821–829.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Gil, L., et al., 2018. Ensembl 2018. Nucleic Acids Res. 46 (D1), D754–D761.

CHAPTER

13 Computational methods (in silico) and stem cells as alternatives to animals in research

Nishant Tyagi, Subodh Kumar, Gurudutta Gangenahalli, Yogesh Kumar Verma

Stem Cell and Gene Therapy Research Group, Institute of Nuclear Medicine & Allied Sciences (INMAS), Defence Research and Development Organisation (DRDO), New Delhi, Delhi, India

13.1 Introduction

Annually, 100 million animals are used for testing, scientific research, and educational purposes throughout the world. These animals are bred, wounded, cut open, injected, infected, genetically modified, and ultimately sacrificed (killed) (Badyal and Desai, 2014). Countries like the United States, Japan, China, the United Kingdom, Brazil, and Germany are the topmost nations using animals. According to reports emerging from the British Union for the Abolition of Vivisection and Dr. Hadwen Trust, about 115 million vertebrates are used a year worldwide (Dua and Dua, 2013). In the United Kingdom alone, over 3.6 million animals were used for experiments in 2010. This number is 37% higher than the figures in 2000. At the beginning of the last decade (2011) more than 2,600,000 mice, 2,700,000 rats, 162,618 birds, 563,903 fishes, 37,714 sheep, 15,900 amphibians, 15,000 rabbits, 11,500 guinea pigs, 84,000 horses, 4550 dogs, 2700 primates, 4340 pigs, 383 reptiles, and 235 cats were used in experiments; these figures have significantly increased in the last 9 years (Fig. 13.1, Statistics of Scientific Procedures on Living Animals Great Britain 2011, 2011; Dua and Dua, 2013). In the Indian scenario, more than 50,000 animals are used in different laboratories and institutes per year. Hyderabad-based National Centre for Laboratory Animal Science is one of the leading animal suppliers to 175 institutes, including educational institutes and pharmaceutical companies. Such a vast quantity of animals used in scientific research, education, and testing has always raised concerns in the minds of environmentalists and animal lovers to limit or eliminate animal use in experiments. This has led to many movements and legislative initiatives. In the 18th century, a group of people in the United Kingdom initiated an animal protection movement against the use of animals in experiments. In 1975, Societies for Protection and Care of Animals opposed the utilization of all kinds of animals in research worldwide (Arora et al., 2011). Also, consideration of nonanimal procedures (reasonable and practicably available) or justification of minimum animal use had

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00003-8

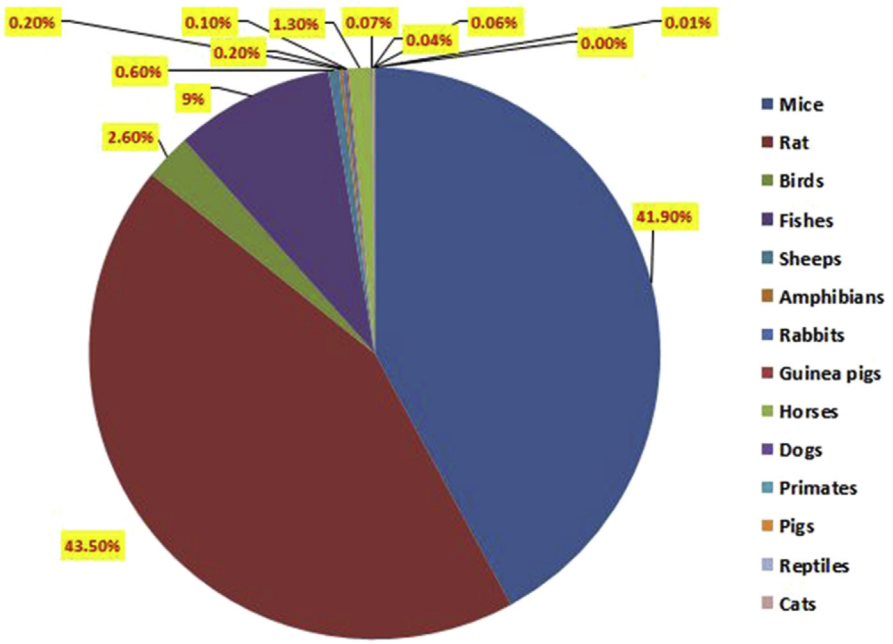


FIGURE 13.1

Percentage of animals used in research. Worldwide, mice and rats are the most experimented creatures.

been suggested by the Council of Europe in 1986 for minimization of suffering and pain (Gruber and Thomas, 2004; Knight, 2008). To monitor animal pain, suffering, and poor treatment—before, during, or after an experiment—the Committee for the Purpose of Control and Supervision of Experiments on Animals (CPCSEA), India, was formed. For animal welfare, the Government of India promulgated the Breeding of and Experiments on Animals (Control and Supervision) Rules, 1998, which were further amended in 2001 and 2006 for the proper regulation of animal procedures (CPCSEA Guideline, 2010). With the recommendation of the CPCSEA, constituted under the provision of Section 15 of the Prevention of Cruelty to Animals Act of 1960, the Union Ministry of Environment and Forests banned the use of living animals in research and educational institutes for dissection purposes. However, using live animals in new scientific discoveries was exempted from this ban (Dua and Dua, 2013; CPCSEA Guideline, 2010). The CPCSEA has also designed rules, procedures, and guidelines for the care and use of animals in experiments. It monitors animal research through ethical committees, like the Institutional Animal Care and Use Committees, formed at the institutional level.

Concern regarding scientific research in India has always been raised but overthrown many times due to inadequate experimental procedures, lack of proper lab

practice, and reduced animal care. There are many incidences where regulatory bodies like the People for Ethical Treatment to Animals and the CPCSEA have intervened because of substandard animal care, e.g., more than 30 monkeys were rescued from the National Institute of Virology, Pune, India, due to inadequate use of animal records, appalling conditions, and lack of proper care. Similarly, in 2002, the CPCSEA inspected Ranbaxy Laboratories animal facilities in Delhi, India, and observed animals suffering from infectious disease and inbreeding defects (Dua and Dua, 2013).

In 1959 Russell and Burch first proposed the fundamentals for good animal practice in laboratories and their idea about alternatives in the form of "the 3Rs"—reduction (in the number of animals), refinement (decrease in animal suffering and pain), and replacement (to replace with nonanimal model) (Fig. 13.2). As they stated, "Refinement is never enough, and we should always seek further reduction and if possible replacement … . Replacement is always a satisfactory answer" (Arora et al., 2011; Knight, 2008; Herrmann, 2019). The 3Rs are defined as follows:

- Reduction: Accounts for limits in animal numbers like sharing animals, phylogenetic reduction, and improved statistical design.
- Refinement: Implemented by improved animal handling, control of pain, proper instrumentation, and limited invasiveness.
- Replacement: Associated with the use of nonanimal models to limit animal use.

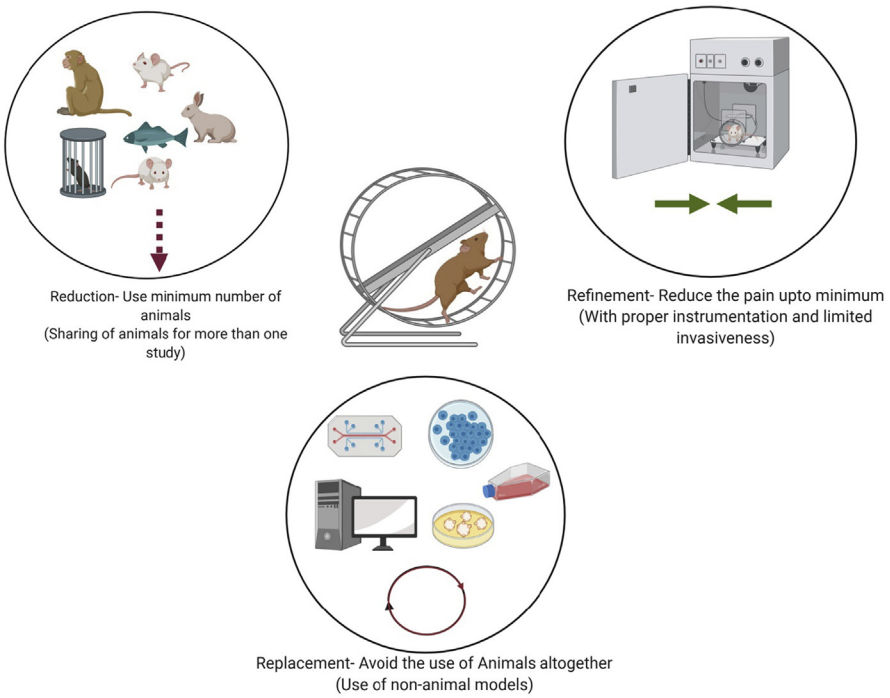


FIGURE 13.2

3Rs principle of animal use in research.

13.2 Need for alternatives

Scientific limitation linked with animal models for testing toxicopharmacokinetics or pharmacodynamics (PD) is very prominent. This may include differences in species, restriction to single gender, unrealistic dose, drug exposure time, inappropriate group size, loss of variation due to inbred strain, stress-related physiological or immunological distortion, and lack of comorbidities (Van Norman, 2019). Such limitations cause variation in drug response in the human system and produce ambiguous results. Substitutes for animal use in research has opened a new field of medical science where various therapeutic drugs or chemicals have been checked for their therapeutic effect to negate the use of animals. Replacement of animals in research does not expose humans to health risks. This helps to improve the quality of research as well as the time taken to report experimental findings. In recent times, many scientists have seen it ethical to support alternatives to animals in research (Doke and Dhawale, 2015). Nowadays, substitutes for animal testing have been exploited in various fields of science like toxicity testing, neuroscience, and drug development (Exner et al., 2007).

In the early stages of candidate drug analysis, animal experimentation was a prior requirement of the Food and Drug Administration (FDA) (Doke and Dhawale, 2015). Pharmacokinetics (PK) and absorption/distribution/metabolism/excretion (ADME) studies are generally done using more than two species to understand both the effects of a drug in the living system and how the human body will process the drug (Li et al., 2019). Due to variation in the genome sequence of animals concerning humans, animal experiments may not always give 100% accuracy. This nonhomology may further lead to changes in their biochemistry, genetics, and physiological properties, e.g., P450-dependent monooxygenase enzyme is one of the best examples to understand interspecies variation. This enzyme catalyzes the oxidation of drugs or toxins and is a significant enzyme in xenobiotic metabolism. Oxidation generates nontoxic metabolites that are soluble in blood and suitable for renal elimination (Baillie et al., 2016; Knight, 2008). However, metabolic pathways and the rate of metabolite generation may differ due to interspecies difference. This is one of the main reasons for drug failure during human trials (Knight, 2008).

Stand-ins for animals like computational models, human cell lines, and tissue culture can provide more accurate results. Early prediction of ADME properties aiding computer-based tools prevents wastage of resources and time in the field of drug research. Traditionally, animals are the main sources to conduct ADME study of the drugs but the development of new ADME-based software programs (listed in Section 13.3.1) do not just eliminate the use of animals (including in vivo bioavailability, absorption reactive metabolites, and metabolic identification), they also reduce the cumbersome process of drug development (Andrade et al., 2016a).

Recently, researchers have discovered that tests using human skin cells grown in vitro are more accurate than the traditional animal test when it comes to

identifying chemicals that are irritant to the skin, e.g., a model of collagen-containing human dermal fibroblasts as a substrate to primary human keratinocytes is readily being employed as 3D skin (Carlson et al., 2008).

The merits of the alternatives to animal testing are the production of fast and more accurate results. Rapid testing means researchers can evaluate five to six products with proxy testing simultaneously compared to studying a single product by animal testing. The other advantage of the substitute for animal testing is cost reduction, which reduces the need for animal purchasing, housing, feed, and care (Digges, 1986; Doke and Dhawale, 2015). The most relevant answer to animal replacement is that it does not create any ethical problem because research and discovery of any new drug is always a painful experience for animals (Akhtar, 2015).

Nonanimal models as alternatives to animals are used whenever required. These models could be physicochemical techniques, microbiological systems, tissue/organ culture preparations, in silico techniques, epidemiological surveys, metabolism assays in plants, stem cells, microdosing, DNA chips, microfluidics chips, imaging technologies, etc. Here, we discuss in detail the use of in silico techniques and stem cells as alternatives to animal use.

13.3 What are the alternative methods to animal research

Scientists have suggested various methods to avoid the use of animals in experimentation. These scientific practices provide alternative means for drug and chemical testing to obtain better results. There are multiple advantages associated with these techniques, such as time efficiency, less workforce requirement, cost effectiveness, better results, and high-throughput screening (Table 13.1).

Table 13.1 Alternative methods of animal research.

S. no.	Methods	Examples
1.	Computational methods	CADD, expert system
2.	Cell and tissue culture	Human dopaminergic neurons are used as a model of Parkinson's disease and for transgenic models with modified expression of PARK genes
3.	Epidemiology	Smoking linked to cancer; high cholesterol linked to heart disease
4.	Plants	Drug-induced defense response and activation of detoxification mechanisms as a result of oxidative stress in *Brassica juncea*
5.	Microorganisms	*Cunninghamella elegans* and *Vibrio vulnificus*
6.	DNA chip	Microarray analysis
7.	Microfluidics	Organ-on-a-chip
8.	Noninvasive imaging	MRI, AMS, MEG, DTI, ultrasound

AMS, *Accelerator mass spectroscopy;* CADD, *computer-aided drug designing;* DTI, *diffusion tensor imaging;* MEG, *magnetoencephalography;* MRI, *magnetic resonance imaging.*

13.3.1 Physicochemical techniques

These techniques are used to assess human response to a chemical or biological substance in a cost-effective manner. Tests like absorption, organ concentration, toxicity, PK, and PD are crucial for evaluating a drug in the preclinical stage, e.g., chitosan-based films are used as a substitute for human epidermal sheets to assess polar and nonpolar drugs for in vitro permeation studies (Arora et al., 2011; Knight, 2008).

During drug development, ADME PK studies are critical for the selection and efficacy benchmarking for multiple drug candidates. These assays are necessary for the movement of drug candidates into clinical programs. Keeping the selectivity and potency of lead compounds intact, continuous improvement in ADME properties is attained through lead optimization. However, in vivo potency of such drug candidates could vary even after acceptable ADME properties (Andrade et al., 2016a).

- Absorption: Extent to which the drug administered has absorbed and reached the site of action. It can be affected by multiple factors like solubility of drug, intestinal transit, chemical stability, etc. Route of administration is crucial for drug efficacy.
- Distribution: Drug distribution takes place through the blood stream in multiple organs. It is essential to know the extent to which a drug has reached the target site.
- Metabolism: The parent drug molecule breaks down into smaller molecules through the action of enzymes. It leads to active and inactive metabolites. Inactive metabolites are inert and reduce the effect of the parent drug, while active metabolites enhance the effect at the target site.
- Excretion: Excretion prevents the accumulation of foreign substances in the body. The risk of adverse effects on multiple metabolic processes is high if the excretion of foreign entities has not happened.
- Toxicity: ADME studies estimate the harmful effect on organisms with the extent of administered drugs. A toxicity study is the part of the last phase of ADME studies.

PK behavior is significant for drug development as it is linked to the efficacy of that compound (drug). In silico tools, such as VolSurf+ and GRID, are ideal for optimizing compounds simultaneously on multiple criteria.

VolSurf+ creates 128 molecular descriptors from 3D molecular interaction fields produced by the software GRID, which are specific to ADME prediction and are easy to interpret, e.g., for membrane permeability, an interaction energy moment descriptor between hydrophilic and hydrophobic regions is crucial and can be created by VolSurf+, which is further used to build statistical models. VolSurf+includes models such as blood−brain barrier permeation, solubility, protein binding, volume of distribution, metabolic stability, and passive intestinal absorption, and has key functions like calculation of relevant ADME descriptors. It also

performs statistical modeling with experimental data, selection of compounds based on similar ADME properties, and predicts behavior of new compounds on the basis of existing or new ADME models.

Computer-based tools are currently more prevalent for PK/PD analysis. Furthermore, PK/PD analysis determines optimal dosing regimens in clinical trials or describes the kinetic and dynamic relation for new drugs (de Velde et al., 2018). Initially, it obtains population PK information of the selected drug (for instance, antimicrobial drug) in the target population (Tuntland et al., 2014). Using in silico tools like Boomer and the PKPD software server, the PK/PD indices are calculated (Nielsen et al., 2011). PK/PD indices determine the dose regimen using simulation. If the index is more than 90%, the dose regimen of the drug is passed and selected (Rizk et al., 2019). Physicochemical properties of a compound, such as water solubility, log P (octanol−water partition coefficient), rotatable bonds, nonpolar surface area, etc., are the primary considerations under Lipinski's rule of five (Lipinski, 2004). According to this rule, a drug should not violate more than one property to be considered as a lead for further development. The compounds that fail to comply with Lipinski's rule of five possess poor pharmacokinetic properties. Such drugs may show poor absorption, faster metabolism and excretion, unfavorable distribution, and might be toxic (Rautio et al., 2008; Asín-Prieto et al., 2015). Common in silico software programs and servers are:

- BiokmodWeb: Contains some features of the mathematical tool BIOKMOD to be applied in PK.
- Boomer: A simulation and modeling program for PK and PD data analysis.
- ADME-AP: Stands for ADME-Associated Protein software, and is used to find the relation between ADME of its associated proteins.
- PKPD software server: Evaluates PK/PD of a particular drug.
- ChemTree-: Helps in the prediction of absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) properties of a drug molecule and is freely available for researchers and the scientific community.
- MDL (metabolite database): This software program provides metabolic information on drugs.
- MDL (toxicity database): This database gives an insight into the structure-searchable bioactivity database of toxic chemical substances.
- MetaSite: A computational procedure specially designed to predict the site of metabolism for xenobiotics starting from the 3D structure of a compound.
- GRID: Determines energetically favorable binding sites on molecules of known structure.
- Shop: Useful in guiding the scaffold-hopping procedure during the drug discovery process.
- ADME/Toxicity Property Calculator: In silico screening-based program from the known ADME/Tox knowledge base.

13.3.2 Cell and tissue culture

Cells and tissues that are isolated from animal bodies and cultured in laboratory conditions can be used without the direct use of animals for research. The cells and tissues directly obtained from various sources, such as skin, kidney, liver, etc., are cultured in a suitable growth medium for a few days to several months or even a few years (Doke and Dhawale, 2015). The prominent nature of cells in the body is their limited division to make new daughter cells. This limitation is known as Hayflick's limit (Bartlett, 2014; Varela and Blasco, 2010). Most of the cells that cross this limit become immortal. Such immortal cells can divide infinite times to generate a cell line, a powerful tool for experimental studies.

Benefits associated with cell and tissue culture techniques are easy to follow, less time consuming, and economical. These cultures are routinely in use for preliminary screening of potential drug molecules/chemicals for their toxicity and efficacy studies (Doke and Dhawale, 2015), e.g., the Organization for Economic Cooperation and Development (OECD) has approved viability tests for 3T3 cells to test the toxicity of drugs such as skin irritants and phototoxicity (Kim et al., 2015).

13.3.3 Tissue engineering

This approach is very similar to the cell and tissue culture technique described earlier. However, it has a major limitation. Due to the 2D culture system, cell/tissue culture fails to mimic the internal environment of the experimental animal. Tissue-engineering techniques can simulate the exact 3D model for drug testing. One of the best examples is the formation of spheroids. These spheroids are made of normal cells in 3D, and are currently being used as a cancer model to test antitumor drugs (Sant and Johnston, 2017). A simple and miniature version of organs, formed by the self-organization of cells like stem cells in 3D, is called an organoid, and introduces new horizons to drug evaluation. Organoids are models of organs used to better understand the effect of drugs on a particular organ. To gain insight into SARS-CoV-2 (causing the COVID-19 pandemic) versatile invading behavior, from lung to liver, kidney, and guts, these miniorgans provide significant findings (de Souza, 2018; Mallapaty, 2020).

Advancements, such as 3D bioprinting in tissue engineering, provide a boost in various aspects, like drug testing and tissue/organ transplantation. Bioprinting is a process in which tissue-like structures are created using biological agents, such as cells, growth factors, and a biocompatible matrix (biomaterial) to mimic natural tissue. As the name suggests, it works like a printing process. Bioink (a composite of biological factors and biomaterial) is used as material, through the layer-by-layer deposition of bioink, to produce a 3D structure of a tissue. Currently, different types of bioprinting methods are available like inkjet (Foyt et al., 2018), acoustic (Sriphutkiat et al., 2019), extrusion (Pati et al., 2015), and laser technique (Foyt et al., 2018). Despite different methods, the process of bioprinting is conserved:

- Blueprint or model: 3D imagining scans, like computed tomography or magnetic resonance imaging scans, are used to obtain the precise dimensions of tissue (Bishop et al., 2017). The process eliminates further fine adjustment to fit in the

tissue. Software programs, like AutoCAD, provide the blueprint for the printing process (Bishop et al., 2017). This blueprint offers essential details related to layer-by-layer instruction.

- Bioink: This is the combination of biological factors, i.e., living cells and factors with compatible matrix, i.e., biomaterial like alginate, gelatin, collagen, silk, etc. The role of the matrix is to provide the former with scaffolding or a structure on which to grow. Bioink is viscous.
- Printing and solidification: The deposition of bioink layer by layer leads to the formation of the physical structure of the blueprint. The thickness and size of each layer can be controlled by the incorporation of different nozzles and are dependent on the type of tissue being printed. As viscous bioink layers start to solidify, the structure starts to hold the shape. Crosslinking is a widely accepted method of solidification where specific chemicals, known as crosslinkers, are added. The aid of ultraviolet light or heat is also well known for crosslinking (Knowlton et al., 2017).

This innovative approach can develop vascularized 3D tissues for tissue regeneration, tissue/organ transplantations (as implants), disease testing models, and drug-testing models (organ-on-a-chip). These structures provide a good insight into the specific effect of a drug on a particular tissue/organ. Hence, 3D bioprinting is a promising nonanimal model.

13.3.4 Microbiological analysis

Microbiological analysis has enormous potential to limit animal usage, if not substituted for scientific testing. The microbes can be easily handled, and are cost-effective, nonhuman, and predictive systems for drug screening, e.g., fungi are used for drug metabolism studies. Bacterium, like *Vibrio vulnificus*, has been used to evaluate the cytotoxicity of RtxA1 (Guo et al., 2018). Microbiology systems help to check the carcinogenicity and toxicity of an experimental molecule. Ames test is a famous test to check the carcinogenicity of any compound. Ames test, developed by B.N. Ames in 1970, is an assay to assess the mutagenic potential of any substance with the assumption that if a compound is a mutagen (induces mutation) in bacteria, then it may also be carcinogenic (causes cancer). This test employs bacteria, like *Salmonella typhimurium* and *Escherichia coli*. A point mutation is introduced in histidine in the case of *S. typhimurium* (or in tryptophan in *E. coli*) operon to obtain a histidine (or tryptophan) strain of respective bacteria. Due to the point mutation, these strains are unable to synthesize histidine, hence they limit growth in a medium that is deprived of histidine. If the test sample reverts this mutation in histidine operon genes, then the investigational substance is a mutagen and may cause cancer. The mutagenic potential of a test sample is assessed by culturing amino acid-deficient organisms in medium lacking that particular amino acid with different concentrations of sample substance for the reversion mutation event. Selection is based on the survival of bacteria in media lacking particular amino acids (histidine or tryptophan) (Ames et al., 1973).

Also, microbial cultures play a significant role in testing the antimicrobial property of any drug. However, testing of antimicrobial activity of compounds derived from various sources based on the agar disk diffusion technique (Kirby–Bauer test) is time consuming (Hudzicki, 2009), and screening thousands of compounds is very difficult. Hence, in silico methods are convenient. The use of in silico techniques is based on the development of 3D structures of the compounds (inhibitors) using the graphical user interface and implemented in the molecular operating environment software program (Vilar et al., 2008). The enzyme–inhibitor complexes are developed with ligands–proteins energy minimization using the Merck molecular forcefield. Furthermore, a quantitative structure –activity relationship (QSAR) module calculates the molecular descriptors. Finally, docking is performed to select the best inhibitory molecule. It helps to select the best inhibitory compounds (drug) and decreases microbial lab testing.

13.3.5 Mathematical models and computer simulations

In this alternative, a biological effect is depicted in the form of codes or equations, e.g., BIOKMOD is one such tool that can be used to analyze PK (Sanchez, 2005). These methods can accurately predict the effect of drugs in humans and several times have bypassed animal testing due to the urgent requirement of treatment (protease inhibitors in the case of human immunodeficiency virus patients).

The continuum model is one of the mathematic models that are based on the principle of fluid and continuum mechanics. It describes cancer-related variables, such as cell population, nutrient concentration, oxygen distribution, and growth factor concentration through continuum differential equations. This mathematical modeling provides data and helps in the development of the hypothesis of cancer progression, and mathematical simulation helps in the development of new drug candidates based on mathematical modeling. With the advancement in technology, computer-aided drug designing (CADD) gives an opportunity to allow targeting of specific receptors or molecules. In light of this, new drugs like captopril and dorzolamide have already been approved based on robust data received from the virtual hearts study as animal data were inconclusive (Talele et al., 2010). By using this method, sensitive anatomical functions, like heart rate, can be simulated in computer models to determine predisposition to certain illnesses.

13.3.6 Epidemiological surveys

These surveys are helpful in limiting a large number of investigational drugs on animals. These estimates correlate previous data of chemical exposure with lifestyle factors in a population, e.g., the risk of glioblastoma is linked to alcohol consumption in a dose–response relationship (Baglietto et al., 2011). In an epidemiological analysis, raw data are collected from the patient. In the study design, the data are categorized into age distribution, status of patients, and sex distribution. Central tendencies for the age of patients of different categories are also calculated. After this,

mathematical analysis is performed by multiple study models. SIR (more prevalent in infectious disease; S stands for susceptible, I stands for infected, and R stands for recovered) is one of them. In this model, a fixed population of N individuals is divided into various "compartments" that vary as a function of time. The SIR model describes the change in the population of each of these compartments in terms of two parameters, β and γ. β describes the effective contact rate of the disease (a susceptible individual comes into contact with an infectious individual and acquires the disease). γ is the mean removal rate (it is calculated using the removed cases as against the new claims daily) (Mazumder et al., 2020).

13.3.7 Plant analysis

Certain compounds show relatively similar effects after exposure to plants compared to the mammalian system. According to the "green liver" concept, proposed by Sandermann in 1994, some plants (e.g., *Brassica juncea*) show a tantamount detoxification scheme of mammalian liver. Xenobiotics metabolism in plants occurs in multiple phases. Phase 1 activates the compounds that occur through a specific set of enzymatic reactions, followed by phase 2 where conjugation reactions take place, and later sequestration of substances from specific organelles completes phase 3 of detoxification. However, this model has limited success as an alternative to animals in research due to significant differences between animals and plants (Arora et al., 2011; Sandermann, 1999).

13.3.8 Microdosing

A new drug can be administered in humans at such an ultralow dose that it creates enough impact to be measured on individual cells instead of a huge physiological concussion. This approach has been shown as cost effective and safe for Investigational New Drug submission and hence is accepted by the FDA subject to their guidelines. It is based on ultrasensitive accelerator mass spectroscopy. Human metabolism data, obtained by using this method, can be helpful to screen out drugs in the early phase of a trial. Microdosing cuts the cost and time required for drug testing apart from providing excellent accuracy (Tewari and Mukherjee, 2010). Psychedelic compounds are very prominent for testing under microdosing, e.g., LSD and psilocybin have been tested against alcohol and tobacco dependence, depression, and end-of-life anxiety, while relative research for disorders related to posttraumatic stress, 3,4-methylenedioxymethamphetamine, has shown great promise (Anderson et al., 2019).

13.3.9 Microfluidics chips

These are small chips having a series of tiny chambers connected to fine channels (diameter in micrometers, Fig. 13.3). Each chamber contains specific tissue from different body parts. To mimic the human body at the microscale level, blood

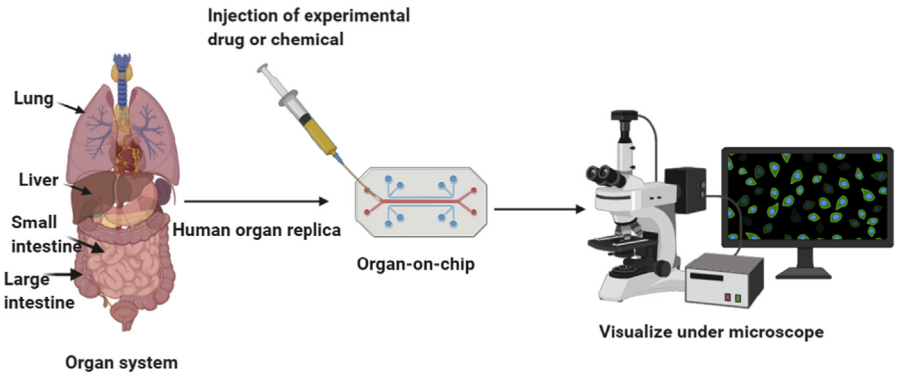


FIGURE 13.3

Microfluidics chip and its role in research.

substitute medium is allowed to flow in these microchannels. The investigational drug is added to the blood substitute medium and circulated through the chip. Sensors in the chip provide information to the computer for further analysis (Han and Kang, 2016).

13.3.10 Tissue chips in space

Currently, a unique strategy is employed to observe the effect of microgravity on the human body, called tissue chip in space. Tissue chip in space is a joint venture of the National Centre for Advancing Translational Sciences from the National Institute of Health and the International Space Station (ISS) national laboratory to obtain better insight into human disease models and potential drug testing in the low or microgravity of the ISS. Tissue chips are designed to mimic living human tissue and organs, e.g., the immune system chip takes account of specific immune cells, progenitor immune cells from bone marrow, and infection encountering immune cells (cells from the lining of blood vessels). These chips travel to the space station and stay there for 2 or more weeks in an incubator. Later, the chips are preserved and sent back to Earth for further analysis. This initiative is likely to unfold physiological changes encountered by astronauts in Earth's low orbital gravity, like aging, bone loss, muscle deterioration, and immune system alteration. Such a low-gravity environment causes an alteration in cell behavior, cell signaling, proliferation, aggregation, and differentiation due to rehabilitated movement of fluids and specific stress stimulation of a space environment (Yeung et al., 2020).

13.3.11 Noninvasive imaging techniques

Techniques, like computed tomography, ultrasound, nuclear imaging, magnetic resonance imaging, magnetoencephalography, diffusion tensor imaging, and accelerator mass spectroscopy provide real-time and very sophisticated measurements of

relations between structure and function in humans compared to unreliable animal models. These noninvasive techniques are very sensitive and may be used up to single-cell resolution (Arora et al., 2011).

13.4 Potential of in silico and stem cell methods to sustain 3Rs

13.4.1 In silico

There are various computer-based methods available, which are focused on the basic principle of biology (Table 13.2). These specialized computer models and tools/software programs help to design new medicines and chemical compounds. New computational approaches rely on various biological and toxic effects of a chemical

Table 13.2 List of software.

S. no.	Biological tools	Uses	Software	References
1.	Basic local alignment search tool	Sequence similarity for significant matches	Online available tool at NCBI	https://blast.ncbi.nlm.nih.gov/
2.	Multiple sequence alignment tool	Sequence similarity of multiple sequences simultaneously	Cluster, omega, MEGA-X, and MEGA 7	https://www.ebi.ac.uk/Tools/msa/clustalo/
3.	Molecular modeling	Mimics the behaviors of the molecules	Chime version 2.0, Macro Model, etc.	Daugelaite et al. (2013)
4.	Molecular docking	Searches for the best interacting molecules	AutoDock, SwissDock, etc.	Deckha and Xie (2008)
5.	QSAR	Predicts biological activity of new molecules before their synthesis.	VEGA platform, CAESAR, DEMETRA, etc.	CPCSEA Guideline (2010)
6.	Microarray	Gene expression analysis	MetaCore, Cytoscape, etc.	Exner et al. (2007)
7.	Artificial intelligence/ machine learning	Future prediction of the nature of compounds from publicly available data of previous studies	RASAR	https://analyticsindiamag.com/machine-learning-may-soon-be-an-alternative-to-animal-testing/

QSAR, *Quantitative structure–activity relationship.*

or potential drug candidate without an animal's dissection. In silico models help to screen out multiple compounds and provide the best molecule for in vivo experimentation. Interaction of drug molecules for particular receptors has become of high interest to the scientific community to save time and money involved in the drug development process. To understand the receptor binding site of a drug, extensive in vivo experiments are generally performed. However, a technique like CADD can predict receptor binding sites for a potential drug molecule. CADD works on the principle of the best-fitting model in terms of binding energy minimization and hence avoids testing of unwanted chemicals without biological activity (Baig et al., 2017). Energy optimization, also known as energy minimization, is an operation to obtain an arrangement of atoms in space in such a way that interatomic force on each atom is close to or zero. Software programs that are commonly used for binding energy estimation are Hyde, X-score, NNScore, etc. With the help of such software, one can tailor-make a new drug for specific receptor binding, and later, animal testing can be done to obtain confirmatory results. Hence, the requirement of total number of experimental animals decreases significantly.

Reduction in the use of animals utilized in research is due to the use of highly sophisticated in silico tools. This not only reduces the use of animal experiments for drug testing and drug safety, but also lowers the risk for patients during clinical trials and minimizes delays in the research for novel drugs (Swaminathan et al., 2019; Arora et al., 2011).

13.4.1.1 BLAST (basic local alignment search tool)

Initial information on new nucleotides (DNA/RNA) and proteins is obtained from their sequence. This information helps researchers to infer function of nucleotide/protein by comparing sequence with homologous molecules. BLAST is based on the use of sequence information to search for similarity using the heuristics technique to generate quick results. BLAST programs have been designed to compare nucleotide or protein databases with the unknown sequence. There are algorithms, which have been incorporated within the BLAST program, for searching similarity in protein and nucleotide sequence (Altschul et al., 1997). There are different variants of BLAST search (Altschul et al., 1997):

- Blastn: Helps in comparing a nucleotide query sequence with nucleotide database. It has high speed but less sensitivity.
- BlastP: Used for comparing a protein query with a database.
- BlastX: Used for comparing a nucleotide query with a protein database by translating the query sequence into six possible frames, and comparing each against the database.
- tblastn: Compares a protein query to a nucleotide database in six possible frames.
- tblastX: Used for comparing protein encoded by a query nucleotide to the protein encoded in a nucleotide database.
- blast2: An advance version of BLAST. It can also perform gapped alignments.

- PSI-Blast (Position Specific Iterated BLAST): Performs iterative database searching.
- RPSBLAST (Reverse-Position-Specific BLAST): Quickly searches a protein query against a database of position-specific scoring matrices (PSSMs) that were usually produced by PSI-BLAST.
- DELTA-BLAST: Produces a PSSM with a fast RPSBLAST search of the query, followed by searching this PSSM against a database of protein sequences.

13.4.1.2 Multiple sequence alignment tools

Multiple sequence alignment (MSA) is generally used to align more than three biological sequences of protein or nucleotides of similar length. The result helps to discover homology and the evolutionary relationship between the sequences. Similarity of the sequences can also help to identify functional, structural, and/or evolutionary relationships among the biological sequences (Daugelaite et al., 2013).

- Clustal Omega: A new MSA tool that uses a seeded guide tree to generate alignments. It is particularly suited for medium to large alignments of sequences.
- EMBOSS Cons: Utilizes a consensus sequence for protein or nucleotide multiple alignments.
- Kalign: A high-speed MSA tool that concentrates on local regions of the sequence. It is suitable for large-scale alignments.
- MAFFT: Stands for Multiple Alignments using Fast Fourier Transform. It is based on fast Fourier transformation for medium–large alignments.
- MUSCLE: Stands for Multiple Sequences Comparison by Log-Expectation. It is a very accurate MSA tool, especially good with proteins. It is suitable for medium-range alignments.
- MViewTransform: A sequence similarity search that results in an MSA using the MView program.

13.4.1.3 Structure–activity relationship

Structure–activity relationship is a popular computer-aided tool. A quantitative SAR (QSAR)-based model has continuously been employed in medicinal chemistry for drug discovery and lead optimization (Fig. 13.4). QSAR helps in the development of the mathematical relationship of physiochemical properties of a drug and its biological activity. This technique is based on the chemical moieties present on the parent compound and how they interact with other compounds. The QSAR project is based on four principles: selection of suitable molecules or drugs (adequate number of compounds, a wide range of activities, and consistent biological activity), construction of a model of the selected compound, validation of the models, and their application model. There is another related technique of QSAR, known as multitarget QSAR, which is used to predict the activity of an investigational drug on various targets simultaneously (Cherkasov et al., 2014).

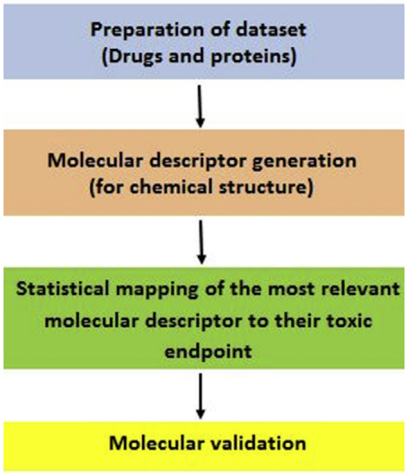


FIGURE 13.4

Quantitative structure–activity relationship model.

These computer-based methods have an advantage over traditional techniques due to their efficiency and time-saving capability. Their speed limit is also high, so thousands of compounds can be validated in a short span of time.

QSAR helps in the reduction of animals used in research because it enables researchers or clinicians to predict/select the best drug before resorting to animal study. It also lowers the risk for patients during clinical trials. Strike is a software program of Schrödinger, which is used for structure–activity relationship study (Andrade et al., 2016b). Software programs are mentioned in Table 13.2.

13.4.1.4 Molecular modeling

Molecular modeling (MM) is a computer-based technique for drawing, manipulating structures, reaction of molecules, and other properties of compounds that are dependent on 3D structures. MM incorporates various fields, such as computational chemistry, drug design, computational biology, nanostructures, and material science (Pimentel et al., 2013). It helps in understanding the fundamentals of physical and chemical interactions, which are difficult to calculate using experimental procedures. It also helps in the development of new theories, models, processes, and products. Molecular dynamics, Monte Carlo, and geometry optimization are the most commonly used simulation techniques in MM. Monte Carlo simulation differs from traditional simulation techniques of MM because it treats random variables for model parameters, while others use fixed variables. RiskAMP is the Monte Carlo simulation engine for Microsoft Excel. Various industrial applications of MM are being exploited. One such example is prediction of hydrocarbon composition in crude oil assay. MM uses crude oil raw data, like distillation curve, American Petroleum Institute gravity, and peptide nucleic acids content, etc., to build a model of

hydrocarbon molecules that mimics the measurable physiochemical properties of crude oil. The chemical compositions of model hydrocarbon molecules as derived from profile data of the crude oil are further used to interpolate, extrapolate, and predict crude oil assays and properties based on molecular thermodynamic models (Pimentel et al., 2013).

13.4.1.5 Computer simulation in organ modeling

The application of computer simulation has the potential to improve drug development and reduce the need for animal testing. Recently, computer simulations have become prevalent in drug development. With the help of computer simulation, many virtual organs have been modeled for drug studies. Virtual heart is one such organ (Trayanova and Chang, 2016; Hurmusiadis, 2007). Small variations in the metabolism of drugs in animal and human cells can amplify the risk to the patient and may lead to drug withdrawals from the market because of safety issues. Researchers in the Department of Computer Science from the University of Oxford have demonstrated that computational models of human heart have higher accuracy (89%−96%) than animal models and can serve as an alternative tool for predicting side effects of drugs (Passini et al., 2017). Ultimately, such tools have the advantages of reducing the use of animals in the preliminary testing of drugs.

13.4.1.6 Molecular docking

The main goal of molecular docking is to predict molecular recognition, binding modes, and binding affinity. One of the software programs to predict binding affinity is Liaison. Molecular docking is performed between small molecules and target macromolecules (Fig. 13.5), such as protein−protein docking and protein−drug docking. Glide and induced fit are two main software programs of Schrödinger used for docking studies. Apart from Schrödinger, GOLD is a protein−ligand docking software program. Molecular docking utilizes a molecular descriptor tool to analyze physiochemical properties. This tool reduces and enriches the library of ligands available in a database for molecular docking. The molecule to be docked is used at the final stage for virtual screening to provide a 3D hypothesis of how a ligand interacts with its target. Molecular docking has a wide array of applications in drug discovery, like structure−activity studies, lead optimization, and finding potential leads by virtual screening (Sethi et al., 2020).

13.4.1.7 Structure-based virtual screening

The rapid increase in 3D structures of proteins and drugs requires the use of highly advanced computational programs for the screening of thousands of compounds simultaneously for selection of suitable lead molecules (Fig. 13.6). Screening of compounds against a target protein is carried out in less time by virtual screening. It can run using parallel computing because protein−ligand docking events are completely independent of each other. Virtual screening utilizes a database for hit identification and lead optimization. High-throughput docking is used as a hit identification method when the structure of a target and its active or binding site is

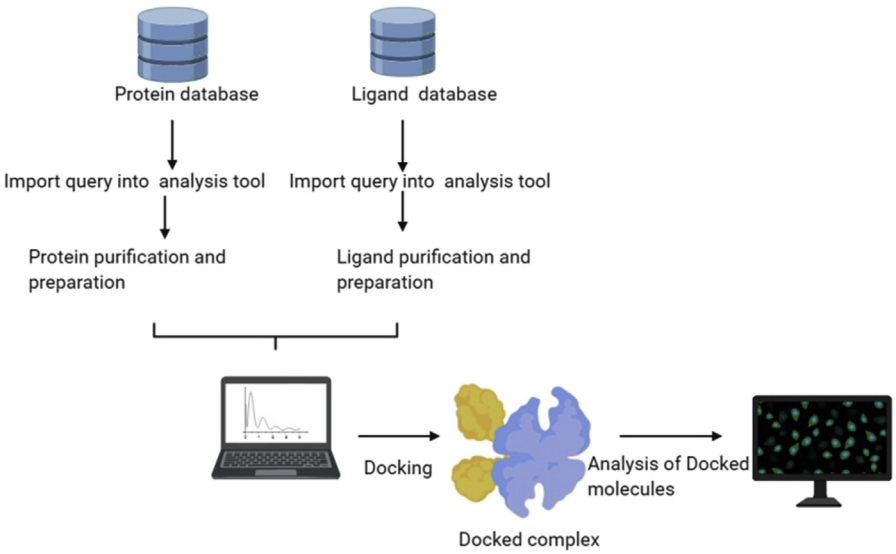


FIGURE 13.5

Docking of molecules.

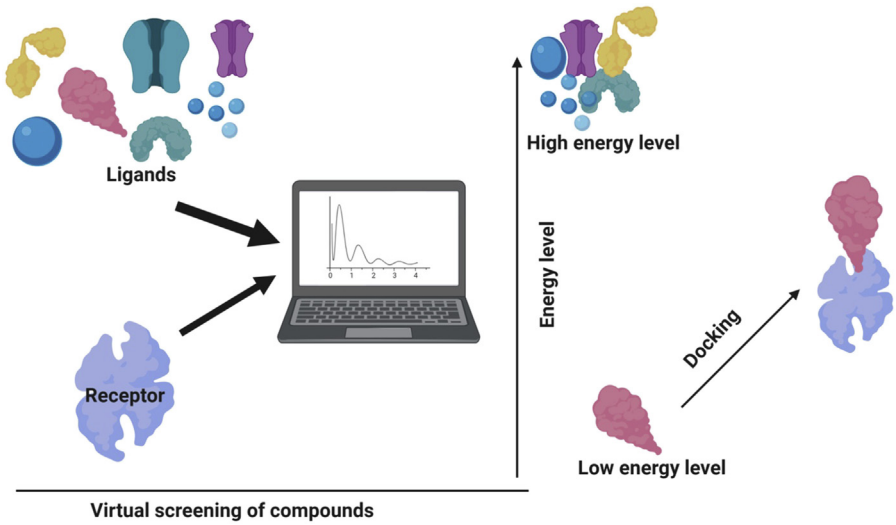


FIGURE 13.6

Virtual screening of lead molecules (ligands) binding to receptors by docking.

available. However, similar calculations are often used during lead optimization when a modification to a known active structure can quickly be tested in computer models before compound synthesis (Lionta et al., 2014).

The necessary steps for docking/virtual screening include protein structure preparation, ligand database preparation, docking calculation, and postprocessing (Guedes et al., 2014). The preparation of protein for virtual screening experiments necessitates different conformations of proteins to be considered. The receptor site of the protein needs to be determined and specific charges have to be assigned. The modeled protein should be accurate. Surface atoms and interaction data such as marking hydrogen-bond donors/acceptors and so forth are incorporated separately (Pantsar and Poso, 2018). Because many ligand molecules are involved in docking, manual steps in preparation for such a database have to be avoided. Starting typically from a 2D structure, bond types have to be checked, protonation states must be determined, charges must be assigned, and solvent molecules should be removed before docking. 3D configuration can be generated using a program such as CORD. Scoring steps involve docking search for one ligand interacting with a given protein, search for ligands binding to one protein, and search for one or different ligands with respect to their binding affinity (Lionta et al., 2014; Pantsar and Poso, 2018; Sliwoski et al., 2014).

13.4.1.8 Microarray or DNA-based chip

Microarray technology has become one of the indispensable tools used to analyze and monitor expression levels of genes in a given organism. A microarray platform is typically a glass slide made by the process of photolithography. A sequence of genes or DNA is coated on the slide in an orderly manner at a specific location known as a spot. Each microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules uniquely corresponding to a gene or DNA sequence. Gene expression profile can be linked to external information to gain insight into biological processes and assist in the discovery of pathways and functions of genes. Microarray helps to understand variation in gene level due to treatment and transcriptional control (Kaliyappan et al., 2012; Bumgarner, 2013). Besides, there are various applications of microarray data analysis. It can:

- Help in predicting binding sites.
- Identify statistically overrepresented sequence patterns.
- Assess the quality of the discovered pattern using statistical significance criteria.
- Find out coexpressed genes in two studied organisms.
- Identify conserved proteins.
- Find instances where conserved proteins are coexpressed in both organisms.
- Map information on protein interaction pathways or metabolic pathways available for one organism to predict interacting proteins or function of identified proteins in other organisms.

- Find out evolutionary conservation of proteins in more than two organisms; this provides knowledge of functional modules that have been conserved in evolution.

13.4.1.9 Microarray data analysis

The process of microarray data analysis involves various steps, like feature extraction, quality control, normalization, differential expression analysis, biological interpretation of results, and submission of data to a public database (Zahurak et al., 2007; Rhodius and Grossutz, 2011):

- Feature extraction: Feature extraction is the process of converting the scanned image of the microarray data into quantifiable values and annotating them with gene IDs, sample names, etc. The output of this process is raw data files that can be in binary or text format. After the feature extraction process, the data can be analyzed. The downstream process of data analysis is done using software like GenePattern and statistics software R.
- Quality control: Quality control is the process to inspect the visual scanned microarray images and make sure that there should not be any blank areas available in the data. After feature extraction, the data analysis software packages can be used to make diagnostic plots to help identify problematic arrays, reporters, or samples.
- Normalization: Normalization of microarray data is carried out to eliminate technical variation present in the raw data while processing. It is also used to secure the biological variation of the assay. Microarray data are normalized using multiple methods, such as robust multiarray average, quantile normalization, and Loess normalization.
- Differential expression analysis: Differential expression analysis aims to identify genes whose expressions vary in different conditions. An essential consideration for differential expression analysis is correction for multiple testing. It increases false positive results. For identification of differentially expressed genes, multiple testing methods can be employed, e.g., Log2 fold change ratio between the test and control condition and an adjusted *P*-value that rates the significance of the difference.
- Biological interpretation of data: Once significant genes have been identified, the relevant genes and their core pathways can be analyzed using several publicly available databases and tools, like DAVID, GO, pathfinR, etc., and the most significant pathways are further used for experimental study.
- Submission of data: Once the microarray data are entirely analyzed, they can be published publicly to the ArrayExpress database.

13.4.1.10 Artificial intelligence and machine learning

Artificial intelligence (AI) and machine learning (ML) are very advanced technologies of the 21st century and have several applications across a wide range of industries and R&D projects. AI and ML have revolutionized biological research leading

to the development of innovations across biotechnology. Sometimes newly synthesized compounds interact with living cells in an unexpected way to harm the system. There are multiple applications of AI and ML in biology such as gene prediction, functional annotation of genes, systems biology, microarray data analysis, pathway analysis, genomic data analysis, future prediction, etc. that give AI and ML an edge over time-consuming animal testing. Applications mainly consist of two broad categories:

1. Identification of coding gene: Continuous advancement in sequencing technique of a genome in a short time, like next-generation sequencing. In this area, ML and AI are used to identify coding regions within the genome. It is highly sensitive compared to typical homolog-based searches.
2. Prediction of structure: The use of ML- and AI-based tools has gained 70%−80% accuracy in results. ML and AI help in the identification of new or novel drug targets using predicted structure of proteins.

Al makes it possible to automate some tests using previous knowledge of chemical interactions available in databases. AI uses various algorithms to predict a more reliable compound based on previous animal tests. It can also correlate the biological effect of any compound based on its structure and other relevant details present in the database. Hartung's software is one of the AI-based software programs that determine the toxicity of a new compound by comparing it with similar compounds available in the database and making predictions based on their properties (Luechtefeld et al., 2018).

The process of ML is quite similar to predictive modeling and data mining. ML searches data to identify patterns and alter the action of the program accordingly (Kourou et al., 2015; Kavakiotis et al., 2017; Riordon et al., 2019). ML-based tools are:

- Cell Profiler: This software is used for biological image analysis. It only measures single parameters from a group of images, like quantitatively individual features similar to fluorescent cell number in the microscopy field.
- Deep Variant: Used for genomic data analysis and helps in the prediction of common genetic variations.
- Atomwise: This tool helps researchers to convert 2D molecules into 3D pixels. It works with highly atomic precision.
- Deep neural network: Used for validation of biomarkers that reveal the disease state. It also helps in the identification of potential biomarkers from genome and proteome data.

13.4.1.11 In silico has the edge over animal testing

Commencement of the in silico technique and advancement of computational knowledge create a new opportunity for scientists to overcome the shortage of animals and ethical concerns associated with animals used in research. Development of

a novel drug is a time-consuming process. Traditionally, whenever a new drug is developed, it requires extensive animal study before human trials. Selection of suitable animal models in terms of genomic similarity to humans is a crucial step to achieve better results of newly developed drugs. BLAST and MSA are used to discover the level of homology, evolutionary relationship, and similarity among groups of species as well as intergroup species. They help in the selection of suitable animal models that lead to the reduction of unnecessary preliminary animal studies or unspecific animal pilot studies (de Aguilar-Nascimento, 2005). In silico techniques, like QSAR, virtual screening, molecular docking, molecular modeling, ADME, PK/PD, and ML/AI, have proven very effective in predicting the effect of drugs. Using virtual screening, scientists are able to screen thousands of compounds simultaneously and select the best candidate for further study. Techniques like AutoDock can be used to check the binding potential of developed drugs against particular proteins, which helps researchers to modify all the unnecessary moieties present in the drugs. Hence, they provide the pace to drug development. AI and ML are, however, in the initial phases of their development but are still very effective in reducing the use of animals in research. Their working principles are based on algorithms and advanced computing tools that are generally used for the identification of coding regions of genomic data obtained from next-generation sequencing, which is highly sensitive. AI and ML are also used for future prediction of the effect of drugs based on previously available data. The accuracy of AI and ML is also very high, about 70%−80%. Using these two tools, predictions of the effects of drugs are now evident, which ultimately reduces the use of animals in research (Webb, 2018; Murphy, 2014). Undoubtedly, in silico methods have the edge in terms of efficiency, ethics, and economic aspects. However, the success rate is limited to predictions only. The best-predicted candidates are further required to undergo various in vitro and in vivo tests for validation. Apart from this limitation, in silico methods surely shorten the lengthy process and also help in achieving refinement and reduction, if not replacement, of animals.

13.4.2 Stem cells: an emerging alternative to animal research

The use of stem cells has shown immense potential as in vitro models for testing the disease and toxicity of drugs to minimize animal testing.

13.4.2.1 Stem cells and their types

Stem cells are self-renewable and can differentiate into a particular lineage under specific stimulus. So, in this way, stem cells not only maintain their pool but also replenish the lost cells to regrow tissue or organs. Here, three types of stem cells are mentioned, which are generally used as animal alternatives: embryonic stem (ES) cells, adult stem (AS) cells, and induced pluripotent stem cells (iPSC) (Nugud et al., 2018).

ES cells are pluripotent and considered as the initiator cells from which all other types of cells are derived. These are present in the inner cell mass of an embryo, at the blastocyst stage of development. These cells form the three germ layers: ectoderm, mesoderm, and endoderm, whereas AS cells are present in the adult tissue and support tissue homeostasis by replenishing the lost cells. These cells also display a self-renewable property but with limited proliferation. AS cells can differentiate into most body cells and are considered multipotent. Bone marrow, peripheral blood, dental pulp, and gastrointestinal tract are some popular niches of AS cells. Bone marrow is the residence of two distinct types of AS cell subpopulations: hematopoietic stem cells (HSCs) and mesenchymal stem cells (MSCs). HSCs give rise to various types of blood cells, including both myeloid (monocytes, macrophages, neutrophils, basophils, eosinophils, erythrocytes, and megakaryocytes) lineage and lymphoid (T-cells, B-cells, natural killer cells, and some dendritic cells) lineage. MSCs, on the other hand, have the potential to differentiate into mesodermal lineages like chondrocytes, osteocytes, and adipocytes. With specific stimuli, MSCs display the potential to differentiate into some ectodermal and endodermal lineages as well. In contrast to other stem cells, iPSCs/reprogrammed cells are the adult cells that gain pluripotency by transfecting a set of genes known as 4F (4 factors: OCT4, SOX2, KLF4, and c-MYC). iPSCs combine the advantage of both ES and AS cells, and this makes them an excellent model for drug testing with no concern for ethical issues (Kondo et al., 2009; Ullah et al., 2015; Shi et al., 2017).

13.4.2.2 Stem cells as a promising alternative

According to the OECD test guidelines, assessment of teratogenicity and embryotoxicity requires multigeneration studies that are not only expensive and time consuming but may also use about 3000 animals per substance (Knight, 2008). In vitro methods are an attractive replacement to test such developmental toxicity studies, especially in ES cells with the establishment of endpoints for screening compounds toxic to embryos (Luz and Tokar, 2018). Toxicogenomics, "loss of function" assays for cells having a homozygous mutation of specific genes, "gain of function" assays for overexpressing foreign genes, pharmaceutical assays, and models to test the function of pathological cells can also be performed in ES cells (Fig. 13.7). Embryotoxicity is assessed by three endpoints, i.e.: (1) inhibition of 3T3 cell growth (cytotoxicity) in MTT assay, (2) undifferentiated ES cells after 10 days of treatment with test compounds, and (3) inhibition of ES cell differentiation into myoblast (cardiac muscle cells; α-actinin as a marker) precursors after 10 days of treatment. Also, gene expression profiling at different stages of ES cell differentiation could be checked for chemical vulnerability.

ES cells also find application in toxicogenomics (application of genomics to toxicology) (McHale et al., 2014), including transcriptomics, proteomics, and metabolomics. In transcriptomics, cDNA microarrays are used to detect carcinogens and hepatotoxicants (Joseph, 2017). The term proteomics is defined as the systematic

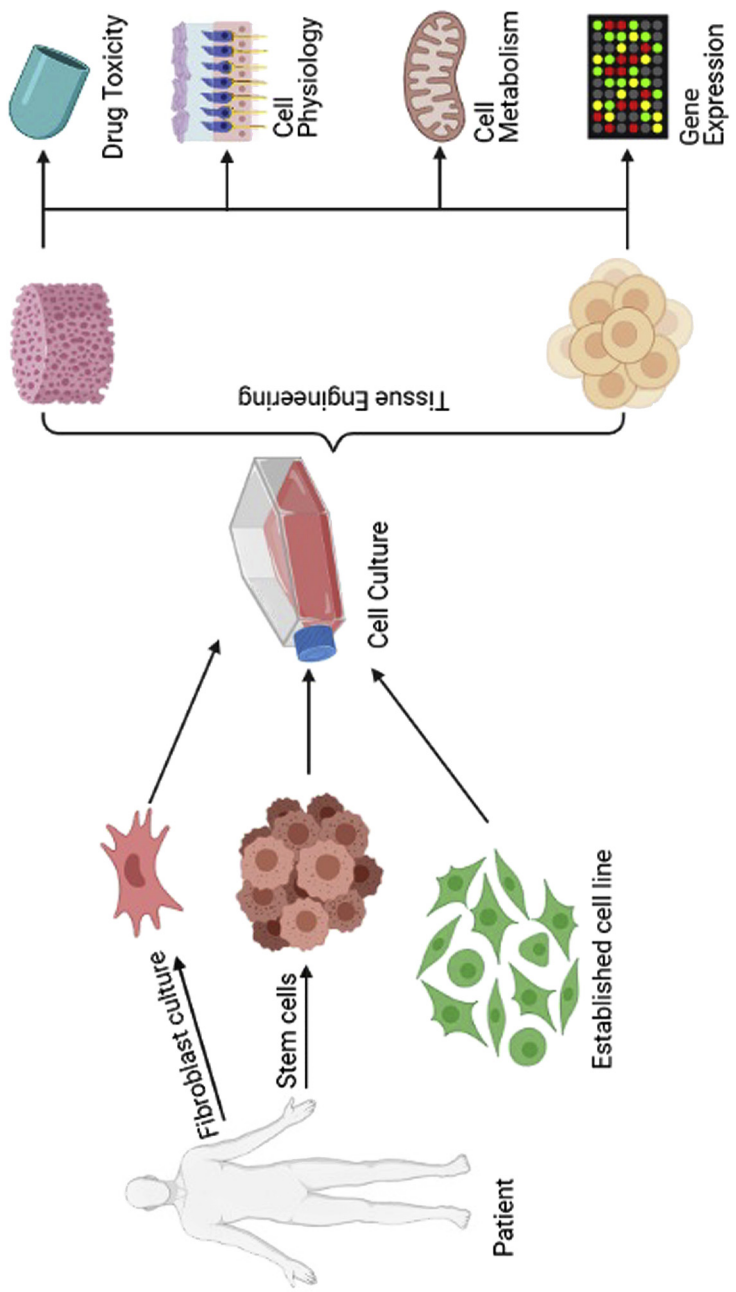


FIGURE 13.7

Stem cell research as a substitute for animal use.

analysis of the protein profile of tissues, whereas metabolomics is a large-scale study of small molecules (metabolites) and their interaction within cells, biofluids, tissues, and organisms (Martins-de-Souza, 2014). ES cells could also serve as a specific tissue graft by exploiting their growth factor-induced differentiation potential (Metallo et al., 2008). Differentiation of ES cells into a number of potential target tissues makes them one of the best candidates for the screening of teratogenicity, growth retardation assay, and embryotoxicity (Schumann, 2010). The resultant assays of ES cells are comparatively straightforward and reproducible. In vitro culture of ES cells may provide a good model for human development and circumvent interspecies difference (Vazin and Freed, 2010). However, the use of human ES (hES) cells is still ethically arguable (Knight, 2008).

To assess toxicity, embryoid body outgrowth derived from murine ES cells is widely used, e.g., morphological analysis of cardiomyocyte contraction under toxic inhibition is carried out by this method (Denning et al., 2016). Another assay includes the detection of changes in sarcomeric myosin heavy chain and α-actinin using intracellular staining and flow cytometry during cardiac differentiation of ES cells (Mummery et al., 2012). Detection of teratogenicity is possible with in vitro culture of transgenic ES cells expressing green fluorescence protein. In vitro culture of differentiated mouse ES cells, BLC-6, into synaptically coupled neurons has shown that these cells carry complex electric properties of postmitotic neurons and are more efficient than tumor cells or primary embryonic cells extracted from animals (Gruber and Thomas, 2004). Molecular endpoints to screen out embryotoxic compounds have been successfully established using stem cells and are efficiently validated by the European Centre for the Validation of Alternative Methods (Brown, 2002). By introducing the genes of Parkinson's patients into ES cells, models have been established that resemble the degenerative potential of this disease (Li et al., 2018). Other disease models have also been developed using ES cells, including two spinal cord diseases, i.e., spinal muscular atrophy and Lou Gehrig's disease for drug screening.

13.4.2.3 Shortcomings of stem cells

Besides many advantages, stem cells fall short of predicting the effect of drug metabolites inside the human body. These metabolites might have a modest impact on different organs, and because of this, stem cells fail as a model of systemic toxicity. Moreover, the growth and maintenance of hES cells depend on calf serum and mouse "feeder" cells. These two components secrete molecules that enable hES cells to maintain their stemness. Nevertheless, this does not eliminate animal use. In the current scenario, ES cell research is relatively expensive and faces ethical issues that limit their use as a complete alternative to animal research. Hence, stem cells would not replace animal use completely. However, these cells can definitely refine and reduce the number of experimental animals (Deckha and Xie, 2008).

13.5 Challenges with alternatives

With all the pros, alternative techniques also have some drawbacks. Alternative methods, as of now, can only fulfill 2Rs, i.e., reduction and refinement, out of 3Rs. Complete replacement is not yet achievable. Some of the limitations are:

- One of the most common challenges is unnecessary experimental duplication. Repetition of studies is one of the leading causes of neglecting animal welfare (Knight, 2008).
- To extract complete knowledge of drug metabolism inside an organism, a metabolic response is required. This cannot be entirely fulfilled by any of the alternatives (Arora et al., 2011).
- Due to a lack of immune response, acceptance or rejection of implants cannot be established using alternative techniques. Similarly, replacement methods fail to determine the idiosyncratic response of a substance that generates an unpredicted response (Mak and Uetrecht, 2018).
- Companies fail to take advantage of previous studies or records from other companies due to the unavailability of data in the public domain and other legal issues (Adibuzzaman et al., 2017).
- ES cell culture is not wholly devoid of animal use. Serum and animal cells (mouse "feeder" cells) are still the significant requirements of maintaining ES cell culture.
- The inability to find proxy due to lack of awareness and improper protocols also contributes toward ignoring alternative techniques.

13.6 Conclusion

Animal experiments are performed to understand disease/disorder and find possible therapies for humans. However, the difference between species and genders with subsequent effects on toxicity, PK, and PD has led to a high rate of clinical trial failures. Furthermore, regulatory limitations have prompted the scientific community to use alternatives to animal experimentation to decrease the number of animals used and minimize pain and suffering to animals. The 3Rs, replacement (with nonanimal models), reduction (of animal numbers), and refinement (to decrease animal suffering), were proposed in this regard by Russel and Burch in 1959. Nowadays, a broad range of tools exists that may replace animal use within biomedical research. Apart from others, in silico methods and stem cell use are the focus of this chapter. In silico ways are used for physicochemical evaluation and computerized modeling of drugs, proteins, and evaluating ADME/Tox properties. Using these tools, unnecessary use of animals could be minimized, which may require thousands of animals and wastage of money and labor (Fig. 13.8). Similarly, stem cells are being used for embryotoxicity, teratogenicity, and growth retardation studies.

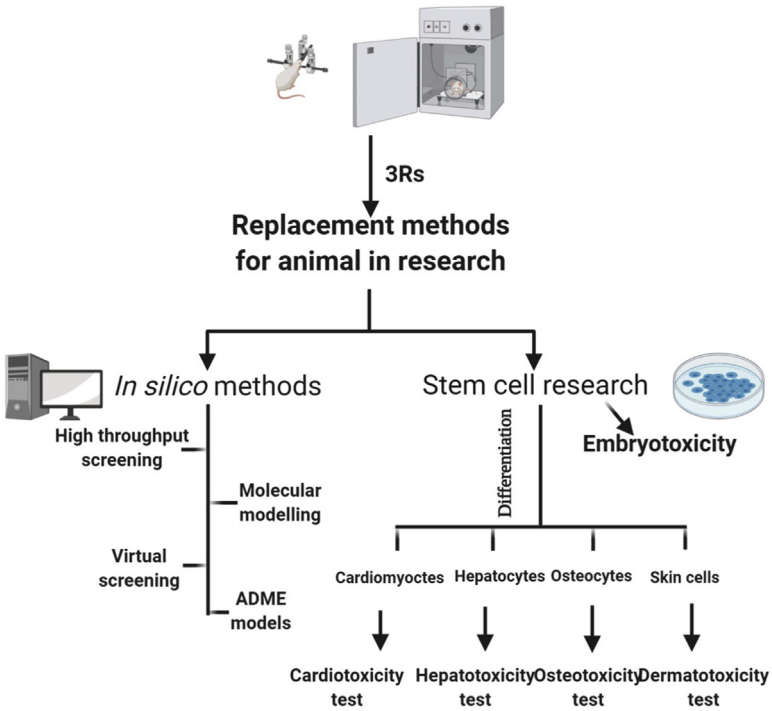


FIGURE 13.8

In silico and stem cell-based alternatives to animals in research. *ADME*, Absorption/distribution/metabolism/excretion.

However, due to limitations with both of these alternatives, complete replacement is not yet possible. But we should always endeavor to reduce the pain and suffering of experimental animals.

Acknowledgments

All the images are created by BioRender.com, invoiced #5C8F22DD-0001, and receipt #2835–999.

References

Adibuzzaman, M., DeLaurentis, P., Hill, J., Benneyworth, B.D., 2017. "Big data in healthcare - the promises, challenges and opportunities from a research perspective: a case study with a model database. In: Annual symposium proceedings. AMIA Symposium 2017, pp. 384–392.

de Aguilar-Nascimento, J.E., 2005. Fundamental steps in experimental design for animal studies. Acta Cir. Bras. 20 (1), 2−8. https://doi.org/10.1590/s0102-86502005000100002.

Akhtar, A., 2015. The flaws and human harms of animal experimentation. Camb. Q. Healthc. Ethics 24 (4), 407−419. https://doi.org/10.1017/S0963180115000079.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped blast and PSI-blast: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389−3402. https://doi.org/10.1111/j.1365-3059.1995.tb02715.x.

Ames, B.N., Durston, W.E., Yamasaki, E., Lee, F.D., 1973. Carcinogens are mutagens: a simple test combining liver homogenates for activation and bacteria for detection. Proc. Natl. Acad. Sci. U. S. A 70 (8), 2281−2285. https://doi.org/10.1073/pnas.70.8.2281.

Anderson, T., Petranker, R., Adam, C., Rosenbaum, D., Weissman, C., Dinh-Williams, L.A., Hui, K., Hapke, E., 2019. Psychedelic microdosing benefits and challenges: an empirical codebook. Harm Reduct. J. 16 (1), 1−10. https://doi.org/10.1186/s12954-019-0308-4.

Andrade, E.L., Bento, A.F., Cavalli, J., Oliveira, S.K., Freitas, C.S., Marcon, R., Schwanke, R.C., Siqueira, J.M., Calixto, J.B., 2016a. Non-Clinical studies required for new drug development − part I: early in silico and in vitro studies, new target discovery and validation, proof of principles and robustness of animal studies. Braz. J. Med. Biol. Res. 49 (11), 1−9. https://doi.org/10.1590/1414-431X20165644.

Andrade, E.L., Bento, A.F., Cavalli, J., Oliveira, S.K., Schwanke, R.C., Siqueira, J.M., Freitas, C.S., Marcon, R., Calixto, J.B., 2016b. Non-clinical studies in the process of new drug development - part II: good laboratory practice, metabolism, pharmacokinetics, safety and dose translation to clinical studies. Revista Brasileira de Pesquisas Medicas e Biologicas Braz. J. Med. & Biol. Res. 49 (12), e5646. https://doi.org/10.1590/1414-431X20165646.

Arora, T., Mehta, A., Joshi, V., Mehta, K., Rathor, N., Mediratta, P., Sharma, K., 2011. Substitute of animals in drug research: an approach towards fulfillment of 4R′s. Indian J. Pharmaceut. Sci. 73 (1), 1−6.

Asín-Prieto, E., Rodríguez-Gascón, A., Isla, A., 2015. Applications of the pharmacokinetic/pharmacodynamic (PK/PD) analysis of antimicrobial agents. J. Infect. Chemother. 21 (5), 319−329. https://doi.org/10.1016/j.jiac.2015.02.001.

Badyal, D.K., Desai, C., 2014. Animal use in pharmacology education and research: the changing scenario. Indian J. Pharmacol. 46 (3), 257−265. https://doi.org/10.4103/0253-7613.132153.

Baglietto, L., Giles, G.G., English, D.R., Karahalios, A., Hopper, J.L., Severi, G., 2011. Alcohol consumption and risk of glioblastoma; evidence from the melbourne collaborative cohort study. Int. J. Canc. 128 (8), 1929−1934. https://doi.org/10.1002/ijc.25770.

Baig, M.H., Ahmad, K., Rabbani, G., Danishuddin, M., Choi, I., 2017. Computer aided drug design and its application to the development of potential drugs for neurodegenerative disorders. Curr. Neuropharmacol. 16 (6), 740−748. https://doi.org/10.2174/1570159x15666171016163510.

Baillie, T.A., Dalvie, D., Rietjens, I.M.C.M., Cyrus Khojasteh, S., 2016. Biotransformation and bioactivation reactions − 2015 literature highlights. Drug Metabol. Rev. 48 (2), 113−138. https://doi.org/10.1080/03602532.2016.1195404.

Bartlett, Z., 2014. The Hayflick limit. In: Embryo Project Encyclopedia, pp. 1−5. http://embryo.asu.edu/handle/10776/8237.

Bishop, E.S., Mostafa, S., Pakvasa, M., Luu, H.H., Lee, M.J., Moriatis Wolf, J., Ameer, G.A., Chuan He, T., Reid, R.R., 2017. 3-D bioprinting technologies in tissue engineering and regenerative medicine: current and future trends. Genes & Dis. 4 (4), 185−195. https://doi.org/10.1016/j.gendis.2017.10.002.

Brown, N.A., 2002. Selection of test chemicals for the ECVAM international validation study on in vitro embryotoxicity tests. ATLA 30, 177−198. https://doi.org/10.1177/026119290703500608.

Bumgarner, R., 2013. Overview of DNA microarrays: types, applications, and their future. Curr. Protoc. Mol. Biol. 6137 (Suppl. 101), 1−17. https://doi.org/10.1002/0471142727.mb2201s101.

Carlson, M.W., Alt-Holland, A., Egles, C., Garlick, J.A., 2008. Three-dimensional tissue models of normal and diseased skin. Curr. Protoc. Cell Biol. 23 (1), 1−7. https://doi.org/10.1038/jid.2014.371.

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., et al., 2014. QSAR modeling: where have you been? Where are you going to? J. Med. Chem. 57 (12), 4977−5010. https://doi.org/10.1021/jm4004285.

CPCSEA Guideline, 2010. CPCSEA Guideline. http://cpcsea.nic.in/WriteReadData/userfiles/file/SOP_CPCSEA_inner_page.pdf.

Daugelaite, J., O' Driscoll, A., Sleator, R.D., 2013. An overview of multiple sequence alignments and cloud computing in bioinformatics. ISRN Biomath. 2013, 1−14. https://doi.org/10.1155/2013/615630.

Deckha, M., Xie, Y., 2008. The stem cell debate: why should it matter to animal advocates? Stanf. J. Anim. Law & Pol. 1, 69−100.

Denning, C., Borgdorff, V., James, C., Karl, S., Firth, A., George, V., Kalra, S., Alexander, K., et al., 2016. Cardiomyocytes from human pluripotent stem cells: from laboratory curiosity to industrial biomedical platform. Biochim. Biophys. Acta Mol. Cell Res. 1863 (7), 1728−1748. https://doi.org/10.1016/j.bbamcr.2015.10.014.

Digges, K.H., 1986. Economic considerations. In: Alternatives to Animal Use in Research, Testing, and Education, vol. 5, pp. 531−542. https://doi.org/10.1002/9783527620869.ch24.

Doke, S.K., Dhawale, S.C., 2015. Alternatives to animal testing: a review. Saudi Pharmaceut. J. 23 (3), 223−229. https://doi.org/10.1016/j.jsps.2013.11.002.

Dua, P., Dua, P., 2013. Use of animals in research: do we have any alternatives? Am. J. Phytomed. Clin. Ther. 1 (9), 740−750. www.ajpct.org.

Exner, C., Bode, H.-J., Blumer, K., Giese, C., 2007. Animal Experiments in Research. Lemmens Medien GmbH, Bonn. https://doi.org/10.1038/187977b0.

Foyt, D.A., Norman, M.D.A., Yu, T.T.L., Gentleman, E., 2018. Exploiting advanced hydrogel technologies to address key challenges in regenerative medicine. Adv. Healthc. Mater. 7 (8) https://doi.org/10.1002/adhm.201700939.

Gruber, F.P., Thomas, H., 2004. Alternatives to animal experimentation in basic research. ALTEX 21 (Suppl. 1), 3−31.

Guedes, I.A., de Magalhães, C.S., Dardenne, L.E., 2014. Receptor-ligand molecular docking. Biophys. Rev. 6 (1), 75−87. https://doi.org/10.1007/s12551-013-0130-2.

Guo, R.H., Lim, J.Y., Nu Tra My, D., Jin Jo, S., Up Park, J., Haeng Rhee, J., Ran Kim, Y., 2018. Vibrio vulnificus RtxA1 toxin expression upon contact with host cells is RpoS-dependent. Front. Cell. & Infect. Microbiol. 8 (MAR), 1−11. https://doi.org/10.3389/fcimb.2018.00070.

Han, S., Kang, S., 2016. Next stage of alternative approaches to animals testing. Int. J. Pharm. Rev. Res. 5 (May), 54−56.

Herrmann, K., 2019. Refinement on the Way towards Replacement: Are We Doing what We Can? Animal Experimentation: Working towards a Paradigm Change. https://doi.org/10.1163/9789004391192.

Hudzicki, J., 2009. Kirby-bauer disk diffusion susceptibility test protocol author information. Am. Soc. Microbiol. 1−13. https://www.asm.org/Protocols/Kirby-Bauer-Disk-Diffusion-Susceptibility-Test-Pro.

Hurmusiadis, V., 2007. Virtual ert: simulation-based cardiac physioloy for education. Comput. Cardiol. 34, 65−68.

Joseph, P., 2017. Transcriptomics in toxicology pius. Food Chem. Toxicol. 109 (1), 650−662. https://doi.org/10.1016/j.gde.2016.03.011.

Kaliyappan, K., Palanisamy, M., Govindarajan, R., Duraiyan, J., 2012. Microarray and its applications. J. Pharm. BioAllied Sci. 4 (6), 310. https://doi.org/10.4103/0975-7406.100283.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15, 104−116. https://doi.org/10.1016/j.csbj.2016.12.005.

Kim, K., Park, H., Lim, K.M., 2015. Phototoxicity: its mechanism and animal alternative test methods. Toxicol. Res. 31 (2), 97−104. https://doi.org/10.5487/TR.2015.31.2.097.

Knight, A., 2008. Non-animal methodologies within biomedical research and toxicity testing. ALTEX 256 (3).

Knowlton, S., Yenilmez, B., Anand, S., Tasoglu, S., March 2017. Photocrosslinking-based bioprinting: examining crosslinking schemes. Bioprinting 5, 10−18. https://doi.org/10.1016/j.bprint.2017.03.001.

Kondo, M., O'Brien, T.F., Lai, A.Y., 2009. Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. In: Cell Determination during Hematopoiesis, 238, pp. 25−51.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8−17. https://doi.org/10.1016/j.csbj.2014.11.005.

Li, H., Jiang, H., Zhang, B., Feng, J., 2018. Modeling Parkinson's disease using patient-specific induced pluripotent stem cells. J. Parkinsons Dis. 8 (4), 479−493. https://doi.org/10.3233/JPD-181353.

Li, Y., Meng, Q., Yang, M., Liu, D., Hou, X., Tang, L., Wang, X., et al., 2019. Current trends in drug metabolism and pharmacokinetics. Acta Pharm. Sin. B 9 (6), 1113−1144. https://doi.org/10.1016/j.apsb.2019.10.001.

Lionta, E., George, S., Vassilatis, D., Cournia, Z., 2014. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Curr. Top. Med. Chem. 14 (16), 1923−1938. https://doi.org/10.2174/1568026614666140929124445.

Lipinski, C.A., 2004. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov. Today Technol. 1 (4), 337−341. https://doi.org/10.1016/j.ddtec.2004.11.007.

Luechtefeld, T., Craig, R., Hartung, T., 2018. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. Toxicol. Res. 7 (5), 732−744. https://doi.org/10.1039/c8tx00051d.

Luz, A.L., Tokar, E.J., 2018. Pluripotent stem cells in developmental toxicity testing: a review of methodological advances. Toxicol. Sci. 165 (1), 31−39. https://doi.org/10.1093/toxsci/kfy174.

Mak, A., Uetrecht, J., April 11−15, 2018. Idiosyncratic adverse drug reactions. Compr. Toxicol. 681−716. https://doi.org/10.1016/B978-0-12-801238-3.64242-3 third ed.

Mallapaty, S., 2020. Mini organs reveal how the coronavirus ravages the body. Nature 2.

Martins-de-Souza, D., 2014. Proteomics, metabolomics, and protein interactomics in the characterization of the molecular features of major depressive disorder. Dialogues Clin. Neurosci. 16 (1), 63−73.

Mazumder, A., Arora, M., Bharadiya, V., Berry, P., Agarwal, M., Gupta, M., Behera, P., 2020. SARS-CoV-2 epidemic in India: epidemiological features and in silico analysis of the effect of interventions. MedRxiv 9, 315. https://doi.org/10.1101/2020.04.05.20053884.

McHale, C.M., Smith, M.T., Zhang, L., 2014. Application of toxicogenomic profiling to evaluate effects of benzene and formaldehyde: from yeast to human. Ann. N. Y. Acad. Sci. 1310 (1), 74−83. https://doi.org/10.1038/jid.2014.371.

Metallo, C.M., Azarin, S.M., Ji, L., De Pablo, J.J., Palecek, S.P., 2008. Engineering tissue from human embryonic stem cells: tissue engineering review series. J. Cell Mol. Med. 12 (3), 709−729. https://doi.org/10.1111/j.1582-4934.2008.00228.x.

Mummery, C.L., Zhang, J., Ng, E., Elliott, D.A., Elefanty, A.G., Kamp, T.J., July 2012. Differentiation of human ES and IPS cells to cardiomyocytes: a methods overview. Circ. Res. 111, 344−358. https://doi.org/10.1161/CIRCRESAHA.110.227512.Differentiation.

Murphy, R.F., 2014. An active role for machine learning in drug development. Nat. Chem. Biol. 7 (6), 327−330. https://doi.org/10.1038/nchembio.576.

Nielsen, E.I., Otto, C., Friberg, L.E., 2011. Pharmacokinetic/pharmacodynamic (PK/PD) indices of antibiotics predicted by a semimechanistic PKPD model: a step toward model-based dose optimization. Antimicrob. Agents Chemother. 55 (10), 4619−4630. https://doi.org/10.1128/AAC.00182-11.

Nugud, A., Sandeep, D., El-Serafi, A.T., 2018. Two faces of the coin: minireview for dissecting the role of reactive oxygen species in stem cell potency and lineage commitment. J. Adv. Res. 14 https://doi.org/10.1016/j.jare.2018.05.012. Cairo University.

Pantsar, T., Poso, A., 2018. Binding affinity via docking: fact and fiction. Molecules 23 (8). https://doi.org/10.3390/molecules23081899, 1DUMMY.

Passini, E., Britton, O.J., Lu, H.R., Rohrbacher, J., Hermans, A.N., Gallacher, D.J., Greig, R.J.H., Bueno-Orovio, A., Rodriguez, B., 2017. Human in silico drug trials demonstrate higher accuracy than animal models in predicting clinical pro-arrhythmic cardiotoxicity. Front. Physiol. 8 (SEP), 1−15. https://doi.org/10.3389/fphys.2017.00668.

Pati, F., Jang, J., Woo Lee, J., Woo Cho, D., 2015. Extrusion Bioprinting. Essentials of 3D Biofabrication and Translation. Elsevier Inc. https://doi.org/10.1016/B978-0-12-800972-7.00007-4.

Pimentel, A.S., Guimarães, C.R.W., Miller, Y., 2013. Molecular modeling: advancements and applications. J. Chem. 2013 (001), 2−4. https://doi.org/10.1155/2013/875478.

Rautio, J., Hanna, K., Heimbach, T., Oliyai, R., Oh, D., Järvinen, T., Savolainen, J., 2008. Prodrugs: design and clinical applications. Nat. Rev. Drug Discov. 7 (3), 255−270. https://doi.org/10.1038/nrd2468.

Rhodius, B.V.A., Grossutz, C.A., 2011. Using DNA microarrays to assay part function. Methods Enzymol. 497 https://doi.org/10.1016/j.physbeh.2017.03.040.

Riordon, J., Sovilj, D., Scott, S., Sinton, D., Young, E.W.K., 2019. Deep learning with microfluidics for biotechnology. Trends Biotechnol. 37 (3), 310−324. https://doi.org/10.1016/j.tibtech.2018.08.005.

Rizk, M.L., Bhavnani, S.M., Drusano, G., Dane, A., Eakin, A.E., Guina, T., Jang, S.H., et al., 2019. Considerations for dose selection and clinical pharmacokinetics/pharmacodynamics for the development of antibacterial agents. Antimicrob. Agents Chemother. 63 (5), 1–13. https://doi.org/10.1128/AAC.02309-18.

Sanchez, G., 2005. BIOKMOD: a mathematica toolbox for modeling biokinetic systems. Math. Educ. Res. 10 (2). http://web.usal.es/~guillermo/publications/Articles/MathematicaEandR(10)-2.pdf.

Sandermann, H., 1999. Plant metabolism of organic xenobiotics. Status and prospects of the 'green liver' concept. In: Plant Biotechnology and In Vitro Bioiogy in the 21 St Century, pp. 321–328. https://doi.org/10.1007/978-94-011-4661-6_74.

Sant, S., Johnston, P.A., 2017. The production of 3D tumor spheroids for cancer drug discovery. Drug Discov. Today Technol. 23 (xx), 27–36. https://doi.org/10.1016/j.ddtec.2017.03.002.

Schumann, J., 2010. Teratogen screening: state of the art. Avicenna J. Med. Biotechnol. (AJMB) 2 (3), 115–121.

Sethi, A., Khusbhoo Joshi, K.S., Alvala, M., 2020. Molecular Docking in Modern Drug Discovery: Principles and Recent Applications. Drug Discovery and Development - New Advances. https://doi.org/10.5772/intechopen.85991.

Shi, Y., Inoue, H., Wu, J.C., Yamanaka, S., 2017. Induced pluripotent stem cell technology: a decade of progress. Nat. Rev. Drug Discov. 16 (2), 115–130. https://doi.org/10.1038/nrd.2016.245.

Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E.W., 2014. Computational methods in drug discovery. Pharmacol. Rev. 66 (1), 334–395. https://doi.org/10.1124/pr.112.007336.

Swaminathan, S., Kumar, V., Kaul, R., 2019. Need for alternatives to animals in experimentation: an Indian perspective. Indian J. Med. Res. 149, 149. https://doi.org/10.4103/ijmr.IJMR.

de Souza, N., 2018. Organoids. Nat. Methods 15 (1), 23. https://doi.org/10.1038/nmeth.4576.

Sriphutkiat, Y., Kasetsirikul, S., Ketpun, D., Zhou, Y., 2019. Cell alignment and accumulation using acoustic nozzle for bioprinting. Sci. Rep. 9 (1), 1–12. https://doi.org/10.1038/s41598-019-54330-8.

Statistics of Scientific Procedures on Living Animals Great Britain 2011, 2011.

Talele, T., Khedkar, S., Rigby, A., 2010. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. Curr. Top. Med. Chem. 10 (1), 127–141. https://doi.org/10.2174/156802610790232251.

Tewari, T., Mukherjee, S., 2010. Microdosing: concept, application and relevance. Perspect. Clin. Res. 1 (2), 61–63. http://www.ncbi.nlm.nih.gov/pubmed/21829784%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3148612.

Trayanova, N.A., Chang, K.C., 2016. How computer simulations of the human heart can improve anti-arrhythmia therapy. J. Physiol. 594 (9), 2483–2502. https://doi.org/10.1113/JP270532.

Tuntland, T., Ethell, B., Kosaka, T., Blasco, F., Zang, R., Jain, M., Gould, T., Keith, H., July 5, 2014. Implementation of pharmacokinetic and pharmacodynamic strategies in early research phases of drug discovery and development at novartis institute of biomedical research. Front. Pharmacol. 1–16. https://doi.org/10.3389/fphar.2014.00174.

Ullah, I., Baregundi Subbarao, R., Rho, G.J., 2015. Human mesenchymal stem cells - current trends and future prospective. Biosci. Rep. 35 https://doi.org/10.1042/BSR20150025.

Van Norman, G.A., 2019. Limitations of animal studies for predicting toxicity in clinical trials: is it time to rethink our current approach? JACC (J. Am. Coll. Cardiol.): Basic Transl. Sci. 4 (7), 845–854. https://doi.org/10.1016/j.jacbts.2019.10.008.

Varela, E., Blasco, M.A., 2010. 2009 Nobel prize in physiology or medicine: telomeres and telomerase. Oncogene 29 (11), 1561–1565. https://doi.org/10.1038/onc.2010.15.

Vazin, T., Freed, W.J., 2010. Human embryonic stem cells: derivation, culture, and differentiation: a review tandis. Restor. Neurol. Neurosci. 28 (4), 589–603. https://doi.org/10.3233/RNN-2010-0543.Human.

de Velde, F., Mouton, J.W., de Winter, B.C.M., van Gelder, T., Koch, B.C.P., July 2018. Clinical applications of population pharmacokinetic models of antibiotics: challenges and perspectives. Pharmacol. Res. 134, 280–288. https://doi.org/10.1016/j.phrs.2018.07.005.

Vilar, S., Cozza, G., Moro, S., 2008. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr. Top. Med. Chem. 8 (18), 1555–1572. https://doi.org/10.2174/156802608786786624.

Webb, S., 2018. Deep learning for biology. Nature 554 (7693), 555–557. https://doi.org/10.1038/d41586-018-02174-z.

Yeung, C.K., Paul, K., Countryman, S., Thummel, K.E., Himmelfarb, J., Kelly, E.J., 2020. Tissue chips in space—challenges and opportunities. Clin. & Trans. Sci. 13 (1), 8–10. https://doi.org/10.1111/cts.12689.

Zahurak, M., Parmigiani, G., Yu, W., Scharpf, R.B., Berman, D., Schaeffer, E., Shabbeer, S., Cope, L., 2007. Pre-processing agilent microarray data. BMC Bioinf. 8, 1–13. https://doi.org/10.1186/1471-2105-8-142.

CHAPTER

14 An introduction to BLAST: applications for computer-aided drug design and development

Tanmay Arora, Asrar A. Malik
School of Chemical and Life Sciences (SCLS), Jamia Hamdard, New Delhi, Delhi, India

14.1 Basic local alignment search tool

Biology is more like history than it is like physics. You have to know the past to understand the present. And you have to know it in exquisite detail. There is as yet no predictive theory of biology, just as there is not yet a predictive theory of history. The reasons are the same: both subjects are still too complicated for us. But we can know ourselves better by understanding other cases.

— Carl Sagan, *COSMOS*

What is Area 51? What is quid pro quo? What is a Brexit? These were among the top 10 most-searched queries on Google throughout 2019.

So what happens when we search "what is a Brexit" on Google? The Google search engine, like any other search engine, is a tool that facilitates search across databases on the World Wide Web and delivers specific websites/webpages as its output, depending on the query or search value entered in the search bar. So, when we ask what a Brexit is, a list of internet links related to the brexit news is delivered to us from a server. Similarly, asking Google the difference between alligator and crocodile would get us a list of sites suggesting the differences and commonalties between the two reptiles. This framework is extremely simple and acts as a beautiful analogy for understanding what BLAST is. In layman's terms, BLAST is the program/method of comparing a sequence of DNA or protein with other sequences of DNA and protein present in different databases around the world to further understand the molecular biology of that sequence. That very particular sequence might lead to the breakthrough discovery of a drug target to a menacing disease. It is for this reason that BLAST remains crucial and is at the very core of drug designing. However, to completely comprehend the functions of BLAST and how it works, it is imperative to have an understanding of certain biological concepts.

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00015-4

14.2 Building blocks

To understand what and how BLAST works, we ought to understand what the building blocks are, what the importance of comparing these building blocks is, and how they are compared. Gregor Mendel and Charles Darwin laid down the framework for us to follow, a framework of heredity and inheritance and evolution over time by adapting to one's natural environment. These adaptations have now been studied extensively and characterized as genetic changes taking place in DNA and/or RNA that translate functionally and structurally through proteins along generations. Structurally, nucleic acids (DNA and RNA) are made up of chemical building blocks called nucleotides. Each nucleotide consists of one of the four nitrogenous bases (cytosine [C], guanine [G], adenine [A], and thymine [T]) plus a phosphate group and a pentose sugar molecule. These nucleotides are linked into chains, with alternating phosphate and sugar groups that form a strand of DNA or RNA. Conversely, proteins are made up of another set of building blocks, strings of amino acids (e.g., proline, alanine, arginine glycine, etc.) joined together by specialized chemical bonds (peptide bonds) to form linear sequences that determine the primary structure of the protein.

These biomolecules, DNA, RNA, and proteins have a specific arrangement of sequences in which they occur in nature. The difference in the arrangement of sequences of nucleotides in DNA differentiates one gene from another; similarly, a difference in the arrangement of sequences of amino acids in proteins distinguishes between the structure and functionality of one protein from another (Fig. 14.1A and B). In the field of bioinformatics, a single, continuous molecule of nucleic acid or amino acid (protein) is considered a biological sequence. These biological sequences are molecular products of evolution, since they undergo random changes during the evolutionary process, leading to the structural and functional changes in organs, pathways, and/or mechanisms. Scientific research, however, has suggested that the traces of evolution remain in certain sequences despite the accumulation of mutations and divergence over the course of time. Evolutionary traces remain because certain amino acids that are crucial for the survival of organisms do not mutate, thus they are preserved by natural selection. On the other hand, amino acids whose function and/or structure is not crucial for the survival of organisms have a greater tendency to be mutated (possibly by change in codon by deletion, insertion, or substitution). These traces can lead researchers to the identification of common ancestry between several different sequences, which can be understood and analyzed by aligning two or more biological sequences together.

14.2.1 Sequence alignment

Sequence alignment is the procedure of arranging and comparing two or more sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. To carry out the structural and functional analysis of newly identified biological sequences or the sequence of interest,

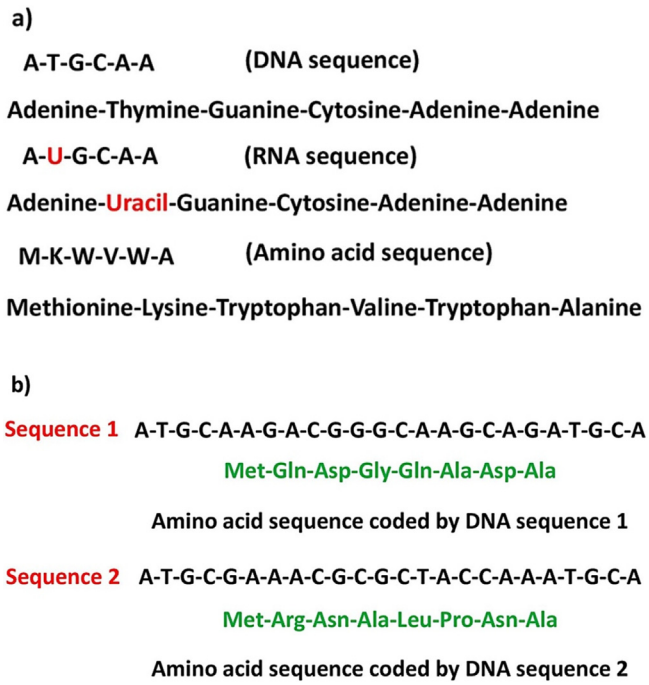


FIGURE 14.1

(A) The three biological sequences of DNA, RNA, and protein (amino acid sequence); RNA has uracil instead of thymine. (B) How two different DNA sequences code for different amino acids and how amino acids with different orders of sequence become different proteins (three-nucleotide codon coding for a single amino acid).

sequence alignment is the first step. The evolutionary relatedness between the two sequences can be identified by the degree of evolutionary traces left in between them (conserved residues in sequences), while identification of mutations (deletions, insertions, and substitution of amino acids) through sequence alignment helps in studying the degree of variance between the sequences. Sequence alignment is often done by lining up two sequences with the intention of achieving the maximal levels of identity and conservation in the case of amino acid alignments. This is often called pairwise alignment. In this process, a nucleotide or protein sequence is placed over another biological sequence and arranged by sliding one sequence over the other to find the best pairing of the two sequences, such that there is maximum correspondence or matching among residues. When two sequences are aligned with each other, alignments can reflect evolutionary conservation, structural similarity, functional similarity, or a random event, which can be measured through the following three parameters (Fig. 14.2):

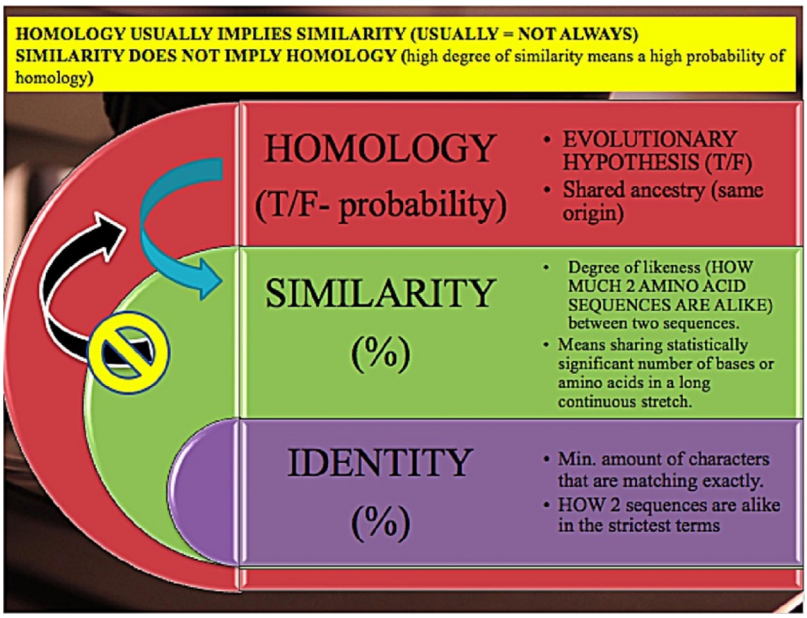


FIGURE 14.2

Correlation between sequence similarity and sequence identity and the connection they share with sequence homology.

1 **Homology:** Homology is an evolutionary hypothesis that indicates shared ancestry or same origin. Since we can rarely be certain about homology, it is therefore usually a hypothesis that may be more or less probable and can be understood as a true/false probability parameter. Either sequences are homologous (i.e., they have a common ancestor; the hypothesis is true) or the sequences are not homologous (i.e., they do not have a common ancestor; the hypothesis is false). Homology is a more generic term that encompasses similarity and identity. A common mistake that is often made is quantifying the level of homology in percentage value. Homology is a qualitative and not a quantitative measure; two sequences can never be 80% homologous or 20% homologous (for a detailed explanation, check the important notes later).

2 **Similarity:** Similarity refers to the degree of likeness between two sequences. It is an indication of the extent to which two biological sequences are alike and is expressed in terms of percentage. It means sharing a statistically significant number of bases or amino acids in long continuous stretches of nucleotide or protein sequences. Similarity between protein sequences is often described as the percentage of aligned residues that are similar in physiochemical properties such as size, charge, and hydrophobicity. Similarity is more specific than homology but less specific than identity.

   Sequence similarity (percentage) is calculated as:

$$S = [(Ls \times 2) / (La + Lb)] \times 100$$

where S is percentage sequence similarity, Ls is the number of aligned residues with similar characteristics, and La and Lb are the total lengths of each individual sequence.

3 **Identity:** Identity refers to the minimum number of characters that match exactly in the two sequences and indicates how two sequences are alike in the strictest terms. It is also expressed in percentage. Identity is the most specific of the three parameters of evolutionary conservation. It is therefore more useful to consider sequence identity shared by two sequences rather than similarity.
Sequence identity (percentage) is calculated as:

$$I = [(Li \times 2) / (La + Lb)] \times 100$$

where L is percentage sequence identity, Li is the number of aligned identical residues, and La and Lb are the total lengths of each individual sequence.

Sequences of evolutionary significance are often characterized by different names, based on the different specific characteristics they share. These common terms are as follows (Fig. 14.3):

1. **Homologs:** Sequences that have a common origin or shared ancestry are usually termed homologs or homologous sequences.

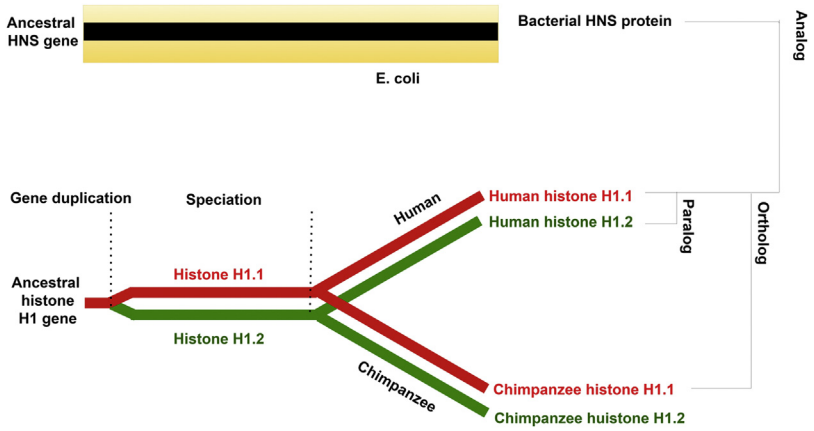


FIGURE 14.3

The concept of homology. Ancestral histone H1 gene upon gene duplication has two copies of the same gene histone H1.1 (red) and histone H1.2 (blue), which upon speciation (an event in evolution where a species diverges into two different species) passes on histone H1.1 (red) or histone H1.2 (blue) to humans and chimpanzees. When this happens, the two genes are said to share common ancestry and are called orthologs. With the passing of both copies of genes (histone H1.1 [red] or histone H1.2 [blue]) to the same species, the two genes are said to be paralogs. Contrary to homology, analogs are those two genes that may perform a similar function but do not share common ancestry: bacterial histone-like nucleoid-structuring protein and human/chimp histone protein.

2. **Paralogs:** Homologs in the same common organism arise through gene duplication, for example, hemoglobin A and hemoglobin F are paralogs.
3. **Orthologs:** Homologs in different organisms arise through speciation (divergent copies of a solitary gene), e.g., histone H1 of humans and histone H1 of chimpanzees.
4. **Xenologs:** Orthologs arise through horizontal gene transfer.
5. **Analogs:** Sequences that have a different origin or no shared ancestry are usually termed analogs or analogous sequences. Analogs and homologs are thus antonymous, and are the very opposite of each other, e.g., wings of a bat and wings of a sparrow.

Important notes

Homology is an evolutionary concept and not a quantifiable entity, and thus cannot be expressed in terms of percentage. It is incorrect to express homology by using phrases such as "significant homology," "less homologous," "more homologous," or "x% homologous."

1. Homologous sequences are not necessarily similar; similar organs are not necessarily homologous. Two sequences (Alignment A) with 94% similarity and two sequences (Alignment B) with 98% similarity are both homologous. The extent of homology cannot be determined and it cannot be said that the two sequences in Alignment A are more homologous than the two sequences in Alignment B. Thus a high degree of similarity implies a high probability of homology.
2. If two sequences are not similar, we cannot say with certainty if they are homologous.
3. If two sequences are not homologous, their sequences are usually not similar (but they may be similar by chance).
4. If two sequences are homologous, their sequences may or may not be similar.
5. Detection of similarity between sequences allows us to transfer information about one sequence to other similar sequences with reasonable, though not always total, confidence.
6. Our ability to perform rapid automated comparisons of sequences facilitates everything from assignment of function to a new sequence, to prediction and construction of model protein structures, to design and analysis of gene expression experiments.

Whenever statistically significant sequence or structural similarity between proteins or protein domains is observed, this is an indication of their divergent evolution from a common ancestor or, in other words, evidence of homology.

14.2.2 Note

It is a common inference that protein sequences are more sensitive than DNA sequences in homology. This is primarily because of the following reasons:

1. DNA is composed of four characters: A, G, T, and C. Hence, two unrelated DNA sequences are expected to have a default of 25% similarity.
2. In contrast, a protein sequence is composed of 20 amino acids, thus the sensitivity of comparison is improved.
3. It is also accepted that convergence of proteins is rare, meaning that high similarity between the two proteins *almost always* means homology.
5. DNA databases are much larger and grow faster than protein databases. Bigger databases imply more entries and therefore more random hits.

So far, understanding alignment of biological sequences allows a deeper understanding of evolutionary background of given sequences, which sheds light on structural and functional similarity between them. However, the questions arise: How is the alignment done? What are the parameters? What are the rules that decide which alignment is more successful and significant than other possible alignments among the residues of two sequences? The answer to all these questions is "scoring system". Each sequence is aligned with other sequences and then scores are given based on the fixed matrixes (explained later); alignment with the best score is considered to be the most significant alignment. The sequence alignments of different biological sequences like DNA and proteins have their own virtues and drawbacks. Sequence alignment is carried out by giving scores depending on match, deletion, insertion, and/or substitution, and based on the best scores, the final alignment is predicted. Alignment of DNA sequences is straightforward as there are only four bases (A, T, G, and C), so scoring the building blocks of two sequences is comparatively easy since the possibility of mismatch or substitution is one in four (Fig. 14.4A). On the other hand, scoring the alignment of amino acids is more complicated as there are 20 amino acids (meaning a one in 20 possibility of mismatch/substitution), which means substitution mutation must have a properly defined scoring system (this is where PAM [point accepted mutation] and BLOSUM [BLOcks amino acid SUbstitution Matrices] come to rescue and will be explained later) (Fig. 14.4B). It is important to understand that scoring the residues of two sequences is not as simple as aligning the numerator of one sequence directly on the denominator residue of the sequence below. However, this is possible in an ideal sequence alignment, where both sequences are of the same length and sequence identity is 100%. Nevertheless, if both sequences are of the same length but not 100% identical, then the evolutionary mutations of deletion, insertion, and substitution complicate the alignment process, for which the concept of gap (−) was introduced in sequence alignment algorithms as explained in Figs. 14.4A, B and 14.5 (to understand the deletion/insertion gap concept better). Gap is introduced during sequence alignment where an algorithm predicts the putative mutation and gives a score as per the rules of the algorithm, which by having the final score in different possible alignments predicts the best possible alignment with the best score. The addition of gaps in an alignment may be biologically relevant, since gaps would reflect the evolutionary changes that may have occurred to the sequence. So, when a sequence alignment shows significantly high similarity among the group of sequences, it can be considered that the sequences belong to the same family. This is where sequence alignment shines because having significant similarity among sequences would also suggest having higher similarity in structure as well, thus if structure and function of one of the family members is known, the structure and function of the query sequence can also be elucidated. Such similarity suggests evolutionary relatedness.

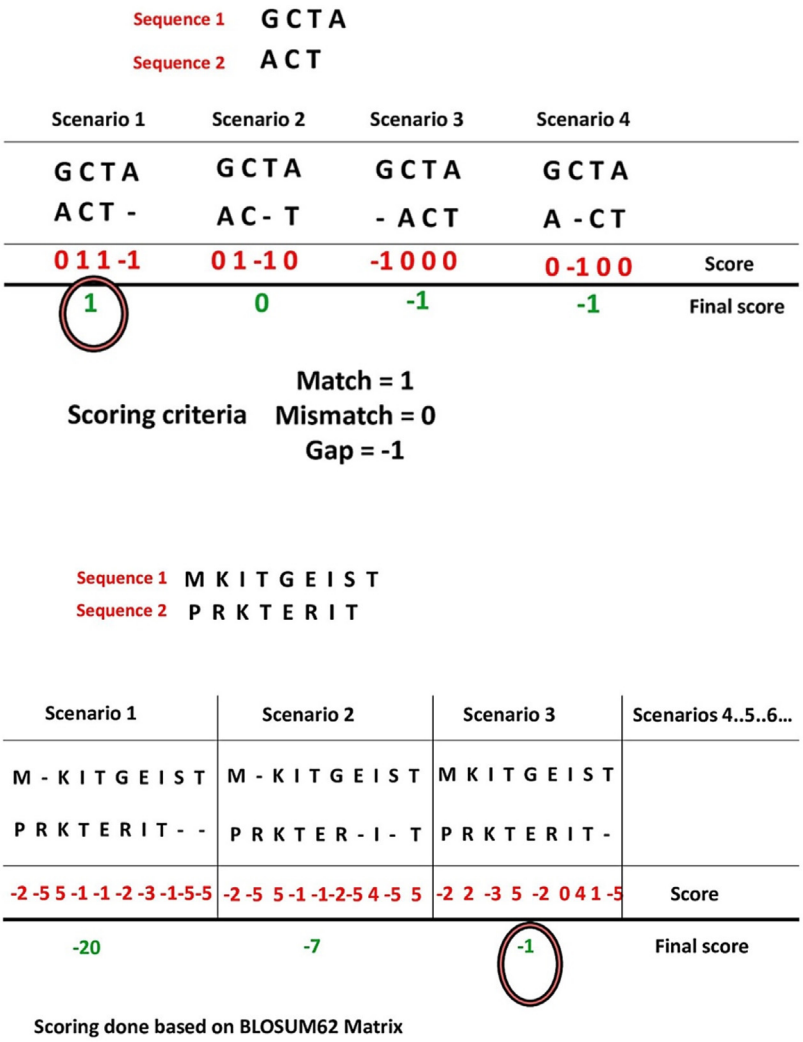


FIGURE 14.4

(A) The sequence alignment of two DNA sequences. Assuming sequence 1 and 2 are the two sequences that need to be aligned, there are several ways two sequences can be glided over one another and aligned (four scenarios as shown in the figure, given that this sequence is small; for larger sequences, the number of scenarios also increases). Each nucleotide from one sequence is aligned with the nucleotide from the other sequence and the scores are given as per Match (1), Mismatch (0), and Gap (−1). Gap receives a negative score as it implicates a strong mutation (insertion or deletion), which normally cells are reluctant to acquire. Moreover, the higher the number of gaps would mean a more negative score, which would suggest the least final score. Since it is impossible to know the position of mutation of an indel (as it might have happened billions of years ago),

14.2.3 Types of mutations

The level of similarity, to any extent, existing between the sequences of different organisms can be attributed to the theory of evolution that all genetic material has eventually come from one common ancestral DNA. So, over the course of billions of years, with every evolving step, mutations kept on occurring and diverging the species at each point, creating differences and diversity between once closely related individuals of the same species. Most mutations are considered to be local mutations, modifying the DNA sequence in a specific manner. These modifications between nucleotide sequences or strings can be of the following types:

1. **Insertion:** An insertion of a base (A, T, G, or C) or several bases between the already existing bases in the sequence, or an amino acid residue in the case of a protein sequence. Insertion is denoted by the amino acid/nucleotide that is inserted.

different scenarios are created by inserting the gap in different positions. Then, each scenario is scored as per the scoring matrix and the scenario that receives the best score is considered to be the best possible alignment. Scenario 1 is the best possible alignment for the given two sequences. It is important to remember that gaps can only be included in between the nucleotides, and care must be taken not to interfere with the sequential presence of extant sequences, e.g., in the given sequence 2 a single gap or multiple gaps can be introduced before or after A or C or T; however, a gap cannot remove the already existing nucleotide sequence present in the sequence, i.e., A—T— (removing C) or —CT— (removing A) would be wrong. (B) Sequence alignment of two amino acid sequences. Assuming sequence 1 and 2 are the two sequences that need to be aligned, there are several ways two sequences can be glided over one another and aligned (three scenarios as shown in the figure, with more possible scenarios, given that this sequence is small; for larger sequences, the number of scenarios also increases). Each amino acid residue from one sequence is aligned with the amino acid residue from the other sequence and scores are given as per scoring matrix (BLOSUM), while gaps are given a heavy negative score (−5). Gap receives a negative score as it implicates a strong mutation (insertion or deletion), which normally cells are reluctant to acquire. Moreover, the higher the number of gaps would mean a more negative score, which would suggest the least final score, thus there is less possibility of there being a significantly good alignment. Since, it is impossible to know the position of mutation of an indel (as it might have happened billions of years ago), different scenarios are created by inserting the gap in different positions. Then, each scenario is scored as per the scoring matrix and the scenario that receives the best score is considered to be the best possible alignment. Scenario 3 is the best possible alignment for the given two sequences. It is important to remember that gaps can only be included in between the residues, and care must be taken not to interfere with the sequential presence of extant sequences, e.g., in the given sequence 2 a single gap or multiple gaps can be introduced before or after any of the residue; however, a gap cannot remove the already existing amino acid from the sequence, i.e., PR-TERIT— (removing the third amino acid K) or PRKT__IT__ (removing E and R) would be wrong.

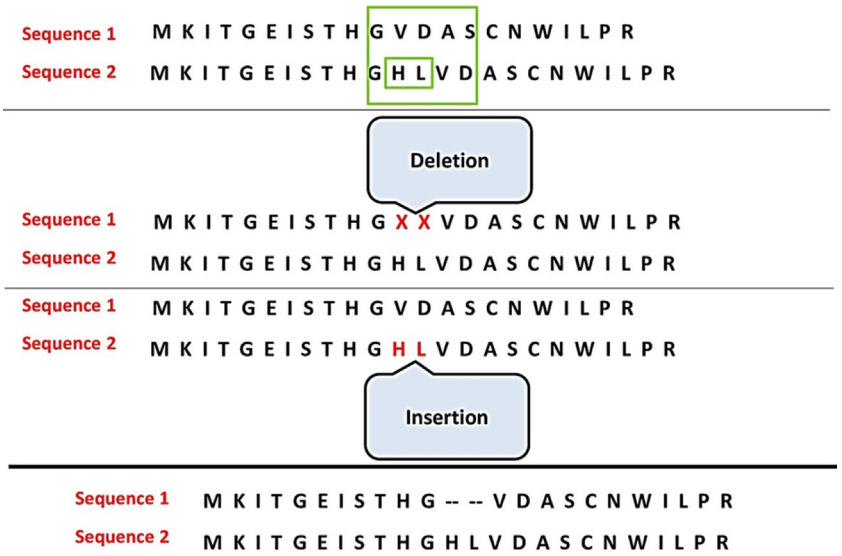


FIGURE 14.5

How deletions and insertions are to be interpreted when two sequences with certain differences are aligned. Practically, there are more complicated sequences in nature; however, for the sake of simplicity, in the figure sequence is exemplified to have two mutations highlighted with a green box. Sequence 2 seems to have an extra histidine and leucine at positions 11 and 12, respectively. Since two sequences are extant and might have gone through mutations over the course of evolution, it is difficult to decipher which one is original and which one is mutated. This scenario can be presented in two different ways: either sequence 2 was always the ancestral sequence and sequence 1 had deletion mutation over the course of millions of years at positions 11 and 12, or sequence 1 was always the original ancestral sequence and sequence 2 acquired insertion mutations at positions 11 and 12. Both scenarios are correct in their own right, thus in scenarios like this, the region in question with missing residues is denoted as a gap, which could suggest either deletion in sequence 1 or insertion in the other.

2. **Deletion:** A deletion of a base (A, T, G, or C) or several bases from the already existing sequence, or an amino acid residue in the case of a protein sequence. Deletion is denoted by (−) and (.) for DNA and protein, respectively.
3. **Substitution:** Substitution of a base (A, T, G, or C) or several bases between the already existing bases in the sequence, or an amino acid residue in the case of a protein sequence. Substitution is denoted by the new nucleotide/amino acid that substitutes the one that was present before.

When two sequences are aligned, it is impossible to know what mutations either of them has gone through over the course of billion years of evolution. Hence, the deletion and insertion concepts are interchangeable when two sequences are aligned. That is, from sequence 1, what is assumed to be deletion compared to sequence 2

could actually be insertion mutation in sequence 2 while sequence 1 was not mutated; similarly, what is considered as an insertion mutation in sequence 2 might actually be the residues that were deleted from sequence 1 during evolution while sequence 2 was the original all along (Fig. 14.5). This is why the missing residue, be it a deletion or insertion in either of the sequences, is referred as "GAP (indel)."

14.2.4 Scoring matrices

Scoring matrices also known as substitution matrixes, which are used to give scores for sequence alignment. Scoring matrices deserve a separate chapter, so we will not be addressing it here in its entirety. However, for the sake of understanding BLAST, we will make a brief introduction to the scoring system. As previously mentioned, the scoring matrices for nucleotide sequences are relatively simple given that only four nucleotides are available for substitution. Usually, a high/positive score is given for a match between the two nucleotides, while a low/negative score is given for a mismatch. However, slight complicacy may originate due to transitions (substitutions between purines and purines or between pyrimidines and pyrimidines) and transversions (substitutions between purines and pyrimidines), but the former is observed to occur more frequently than the latter.

Conversely, it is more complicated to create scoring matrices for amino acids because scoring has to reflect the physicochemical properties of amino acid residues. Margaret Dayhoff (considered as the Mother of Bioinformatics) studied thousands of closely related proteins from several protein families. Upon sequence alignment of homologous proteins, she observed that specific amino acid substitution had taken place between the two sequences. She identified the amino acids that were replaced by other amino acids and were accepted by natural selection, which she called point accepted mutation, PAM. (In reality it was called accepted point mutation, but PAM rolled well with the tongue.) To understand which of the amino acid substitutions were accepted by nature, she studied around 71 groups of families and observed 1572 changes. She created a matrix based on these observations, which is used even today given that massive high-throughput sequencing data have made it easier to study substitution changes in sequences and Dayhoff's observation was correct. She observed that amino acids with similar physiochemical properties (like alanine and glycine) had increased tendency to mutate (because the function they carried out could be easily performed by other amino acids with similar physicochemical properties), while certain unique amino acids like cysteine and tryptophan rarely mutated (possibly due to their uniqueness in structure: sulfur bridges in the case of cysteine and single codon coding for tryptophan, which upon mutation would render protein functionless). Based on these mutational probabilities, she created matrices commonly known as PAM matrices, which are used to give scores to each residue matchup in an alignment. Similarly, studying more than 2000 conserved amino acid patterns representing 500 groups of protein sequences, Henikoff and Henikoff constructed a BLOSUM matrix where ungapped alignments of less than 60 amino acid residues in length were used. BLOSUM and PAM were derived from

alignments of higher extant similar sequences. Using these matrices, a scoring system was developed that gave a high score for a more likely substitution, whereas a low score was given for a rare substitution.

With the advent of technology surging through most spheres of life, revolutionary advancements have been made in the field of computational biology and bioinformatics. Over time, different alignment algorithms have been developed to achieve the aforementioned objective of pairwise alignment. These algorithm approaches are mainly of two types, namely exhaustive search approaches and heuristic search approaches.

However, before we dive deeper into these two approaches, it is important to note that although the nexus between biology and computer science has been strengthened with modern technologies, the two still use different nomenclature (Table 14.1).

Though subsequence has different connotations in biology and computer sciences, bioinformatics researchers usually use the biological terminology associated with subsequence. Thus "subsequence" mainly implies a contiguous sequence of letters. Now that we have an understanding of the commonly used terminology in biology and computational sciences, we can discuss the different algorithm approaches employed for alignments, namely exhaustive (dynamic) approaches and heuristic approaches.

Most of the sequence alignments need to be done online to search for something associated with the sequence of interest to the researcher. This suggests that the search has to be done in all kinds of databases that contain the reference sequences, be it DNA, RNA, or protein sequences. To obtain the desired results, there are certain requirements that we must implement on algorithms for sequence database searching.

Sensitivity: Sensitivity refers to the ability of an algorithm to find as many correct hits as possible. These hits are generally known as true positives.

Selectivity: Selectivity refers to the ability of an algorithm to exclude incorrect hits. These hits are generally known as false positives.

Speed: Speed, as the name suggests, refers to the time an algorithm takes to obtain results from database searches.

Table 14.1 Comparison of the notations of the two disciplines of computer sciences and biology.

Computer sciences	Biology
String, word	Sequence
Substring (contiguous sequence of letters)	Subsequence
Subsequence	Noncontiguous segment of a sequence
Exact matching	N/A
Inexact matching	Alignment

For a desirable search, a researcher would prefer all three requirements to be met, but unfortunately as the sensitivity of a search increases so does the time it takes to find the results, while if sensitivity is reduced, selectivity increases giving rise to false positives. So, to obtain the desirable results, one ought to choose the algorithm that fits their needs.

Exhaustive type algorithms are the rigorous algorithms that use brute force to find the best and exact solution for a particular problem by examining all possible mathematical combinations for alignments; however, these methods are extremely slow and time consuming. These programs are also very taxing for computer memory. They keep the first two criteria of sensitivity and selectivity in account but at the cost of speed. This is why most of the current database searches do not use exhaustive-type algorithms because they do not give results in real time. Two of the methods used by exhaustive-type algorithms are the global pairwise alignment algorithm also known as the Needleman−Wunsch method and the local pairwise alignment algorithm also known as the Smith−Waterman method.

14.2.5 Dynamic programming

To perform a pairwise sequence alignment, this algorithm builds a 2D matrix with one sequence on the X-axis and the other on the Y-axis. Algorithmic rules are applied in each case (the local pairwise alignment algorithm is slightly different from the global pairwise alignment algorithm) giving a scoring scheme to fill the matrix. Scores are given to every possible combination of residues of two sequences and then the maximum scored region representing the best alignment by backtracking through the matrix is found (Fig. 14.6). One can imagine why this kind of algorithm can be extremely time consuming; the databases have sequences in the billions and will take a massive amount of computer power and a long time to find the best alignment. Global pairwise alignment, as the name suggests, would find the most accurate sequence alignment throughout the sequence; however, local pairwise alignment would only search for the most conserved region in between the two sequences (for further details, study the dynamic Needleman−Wunsch method and the Smith−Waterman method).

On the other hand, heuristic programming is a computational strategy and approach to find an empirical or near optimal solution by using certain thumb rules that make processing less time intensive. Heuristic algorithm approaches provide approximate solutions to a complex problem by taking shortcuts and circumventing the exhaustive brute force approach. Though the use of heuristics may not amount to the most accurate results, they do promise results in a realistic time frame without compromising the authenticity of the results. Sometimes, these results are not able to formally prove whether the solutions derived through heuristic approaches actually solve the problem at hand, yet these algorithm approaches work much faster than exact algorithms and since this is a software-based strategy, it is relatively cheaper and more widely available. Exhaustive models of sequence alignments are thus accurate but slow, whereas heuristic approaches run at a much faster rate,

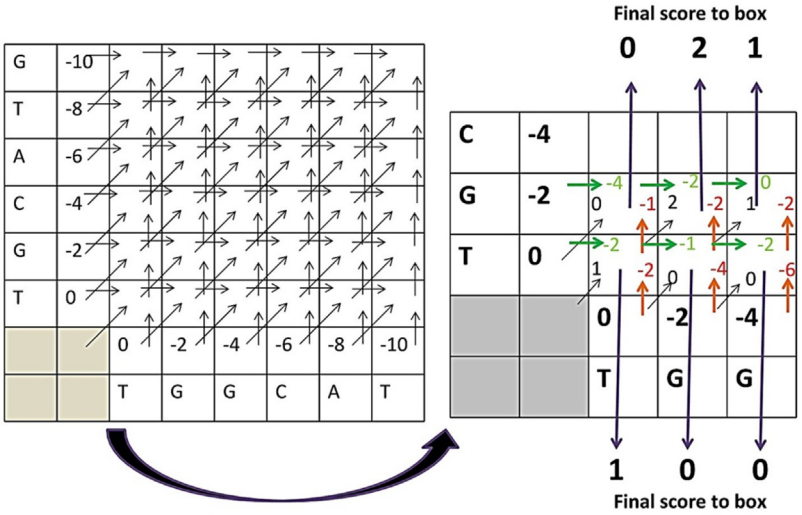


FIGURE 14.6

The exhaustive approach of pairwise alignment of one of the two methods (global alignment in this case). As the figure shows, every possible residue/nucleotide is compared with every possible residue/nucleotide of the other sequence. Score is given based on match, mismatch, and gap in the case of nucleotides, while the scoring matrix is used to score match and substitution in the case of amino acids. Each box is scored from three sides: below, side, and diagonal. Below and side give the score of the gap, that is −2, while diagonal gives the score as per match/mismatch/substitution between two residues associated with that box (in the case of a nucleotide, 1 for match, 0 for mismatch). Of the three values, the highest value is given to that box. Similarly, the entire matrix is filled with scores from below, adjacent to left, and adjacent diagonal box. Finally, from the top right of the box, backtracking is done to each box that has given the highest score to the said box until backtracking reaches the bottom left. Final residues of backtracking boxes are aligned, which gives the best possible alignment between the two sequences. This method is extremely time consuming and taxing on computer memory; however, it does give the most accurate of alignments. Local alignment is done in a similar way with slight changes in scores; where there are no negative scores, scores are either positive or zero (even if calculation gives a negative value, the matrix is filled with zero). For further clarification, please read the Needleman–Wunsch and Smith–Waterman alignment algorithms in detail.

but their shortcuts and complexity make them more difficult to implement at times. Nevertheless, researchers around the world have been completely content with the results obtained using heuristic methods as it has been quite useful so far, and double checks have revealed that the results are nearly perfect. Heuristic programming approaches for sequence alignment employ certain assumptions based on observations such as:

1. Substitutions occur much more likely than insertions or deletions. This is because organisms tend to prefer substitution mutation over indels since deletion of a residue entails deletion of three nucleotides, and similarly, insertion of an amino acid would mean insertion of three nucleotides. On the contrary, substitution may merely involve insertion/deletion of a single nucleotide; moreover, substitution mutation may occur due to frame shifts as well. All the possible ways of having substitution mutation are less taxing on the organism, while indels are more taxing on them, which is why they prefer substitution rather than indels. (Having said that, this does not mean that there are no indels. It does happen over the course of evolution; it is just that cells prefer it less.)
2. Homologous sequences contain many segments with matches or substitutions without gaps or insertions/deletions. These segments can thus be used to kick start the searching operation.

14.3 Basic local alignment search tool

By now, we should have an understanding of what biological sequences are, how they are aligned, how different types of scoring matrices are used to give scores to different alignments to get the most optimal alignment, and how an alignment of a sequence into the database can reveal structural, functional, and even evolutionary information regarding other sequences. We also learned how certain algorithms are best at giving the most accurate alignments but are not practical due to their time constraints, whereas other algorithms sacrifice some sensitivity to get closer to the best alignment in a very acceptable time run. One such program is known as BLAST. BLAST is merely a search tool, similar to the Google search engine, but for biological sequences. It is a tool that instead of keywords takes up biological sequences (in FASTA or GenBank format) and a weight matrix (PAM or BLOSUM), and searches them against other biological sequences on various online databases. The beauty of this structure lies in the fact that you can choose the databases and even the sequences you wish to search against among other biological, computational, and statistical parameters, which are entirely customizable. The biological sequences are then aligned against other sequences from the internet to produce an output that can be delivered in a variety of formats, ranging from HTML to plain text to XML formatting. Essentially, the output contains a list of search results (called hits), each containing alignment information regarding the extent to which the two sequences are perfectly aligned. This is expressed as a percentage and can be referred to as similarity or identity in a biological context. Homology, similarity, and identity were discussed earlier in this chapter.

BLAST is a heuristic local alignment-based algorithm that aligns genomic or proteomic sequences to discover regions of local similarity between the sequences. According to the National Center for Biotechnology Information (NCBI), BLAST is

the most widely used sequence similarity tool. The program compares nucleotide (DNA or RNA) and/or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

The BLAST algorithm facilitates comparison of sequences that are classified as follows:

1. **Query sequence:** This refers to the subject nucleotide or protein sequence that is to be searched.
2. **Target sequence/s or database sequences:** This refers to the library or sequence database to which the query sequences are aligned to identify target sequences from a database that resemble the query sequence above a certain statistically significant threshold.

BLAST was created by Stephen Altschul et al. at the National Institute of Health and was published in the *Journal of Molecular Biology* in 1990. It is known to be derived from the 1990 stochastic model of Samuel Karlin and Stephen Altschul, who "proposed a method for estimating similarities between the known DNA sequences of one organism with that of another," and their work is often attributed as the statistical backbone for BLAST.

This is known to be a significant improvement over the previously existing Smith−Waterman algorithm as it uses a much faster alignment approach and less computer power. It is much more time efficient than another heuristic approach, FASTA, as it searches only for the more significant blocks in sequences with comparative selectivity and sensitivity.

BLAST was developed to allow extremely fast searches through databases. The main criteria for these database searches as explained before were speed, sensitivity, and selectivity:

1. **Speed:** This refers to the number of computational alignments made while searching through a given database with respect to time. These databases grow exponentially, due to the enormous amounts of data being uploaded every day. Therefore the faster the algorithm, the more efficient its output would be. Speed thus contributes greatly for the success of BLAST as an alignment algorithm.
2. **Sensitivity:** This means that the algorithm needs to acquire all or most of the matches that are within a database. The higher the sensitivity of the algorithm, the better its efficiency would be. BLAST has proven to provide near perfect alignments, which has kept researchers satisfied around the globe.
3. **Selectivity:** This is another parameter that substantiates the efficiency of the BLAST algorithm and suggests that all or most of the matches that the algorithm finds in the database must be correct. The higher the selectivity of the algorithm, the better its efficiency.

14.4 How BLAST works

The BLAST algorithm works by taking the query sequence and breaking the entire sequence into "words." "Words" are created by breaking a DNA sequence into substrings of 11 nucleotides each, while protein sequences are broken into three substrings (explained in Fig. 14.7). This process is known as seeding and the list must include every possible word that can be extracted from the query sequence. After seeding, all created "words" are allowed to search for a match using pairwise alignment in a database for the occurrence of these words (since the size of "words" is extremely small, this part of the algorithm takes only a small amount of time and eliminates several million sequences as potential false positives at a time, thus saving time again). For an alignment to start, at least one "word" from the list of "words" should match the sequence in the database. The matching of words is scored by a given substitution matrix (PAM or BLOSUM) and a "word" is considered as a match if its score is above threshold. Once the "word[s]" identifies their match with the appropriate threshold, the next step involves extension of pairwise alignment from the words in both directions while counting the alignment score using the same substitution matrix that was used at the beginning of the program. The extension of pairwise alignment and addition of residues/nucleotides on either or both sides of the sequence continue until the score of the alignment drops below a specific threshold due to encountering mismatches. The dropping threshold for DNA is 20, while for protein it is 22. The resulting contiguously aligned segment pair without any gaps after pairwise alignment is called the *high-scoring segment pair* or HSP. The algorithm also looks for more than one HSP with the possibility of combining several together to make a larger HSP, which would suggest better sequence alignment.

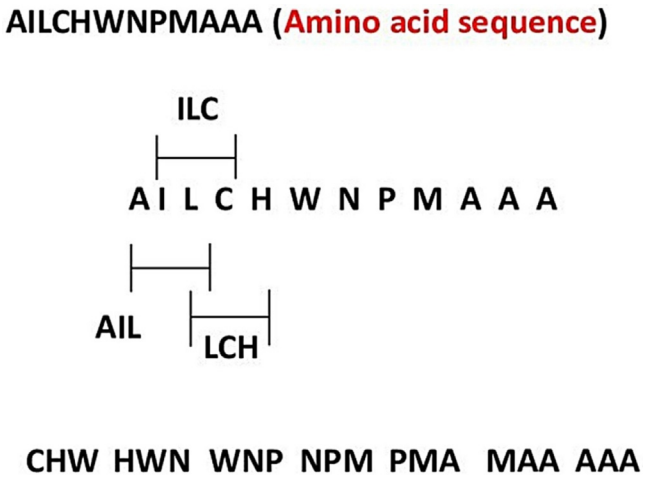


FIGURE 14.7

Seeding of a query amino acid sequence where the given sequence is broken into every possible "word." Triplet for an amino acid, 11 base pair for the nucleotide sequence.

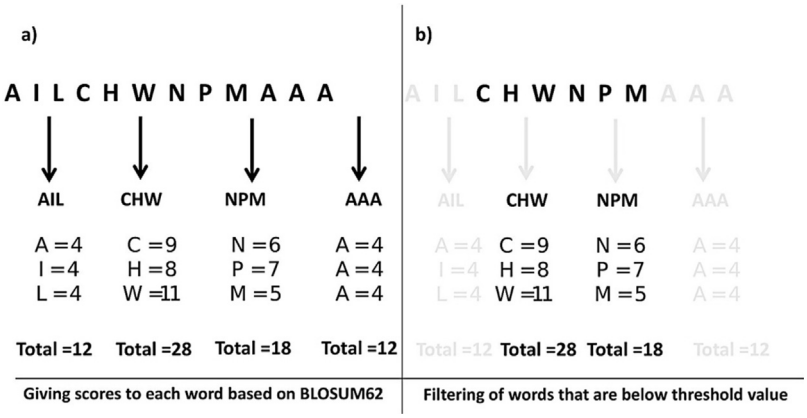


FIGURE 14.8

(A) Seeded "words" receiving the best possible scores they can achieve upon aligning with other sequences using the BLOSUM62 scoring matrix. (B) Filtering off the words whose scores are below the threshold value, while those words above the threshold score are allowed to go further and have more chances of achieving a significant match.

In the example given in Fig. 14.8A, all the possible "words" are identified and each individual word (all three amino acids) is given a score as per the substitution matrix (PAM or BLOSUM). According to the scoring matrices, the best score is always along the diagonal line of an amino acid to itself. For instance, according to the BLOSUM62 matrix, in the case of LIA, L matched to L would give 4, I aligned to I would give 4, and A aligned to A would give 4 as well. Thus the highest possible total for the word LIA in the best possible alignment case would be 12, which happens when LIA in the query sequence matches the LIA sequence in the database. Similarly, all possible "words" of the query amino acid will have the best possible score they can achieve with respect to the given substitution matrix, as shown in Fig. 14.8A.

"Words" that are above a certain given threshold are kept, while the rest whose scores fall below this given threshold are discarded (Fig. 14.8B). This threshold value can be specified in most BLAST algorithms by using the (-f) option given in the user interface, thereby removing unimportant or uninformative parts of the sequence that are not going to be useful while searching through a database, as the probability of them occurring is very high. Once the words with low scores are discarded, the next step is extending the words with higher scores than the threshold. The word is extended if the score is above a certain threshold or the score is above a certain value and is extended in both directions. If the alignment cannot be extended further, then the sufficiently high-scored segments are introduced by gaps; however, if the score falls below the specific threshold, extension is ended; otherwise, it is temporarily maintained until the score rises again. Once the extension has reached its limit, the final alignment is called the HSP; again, the substitution

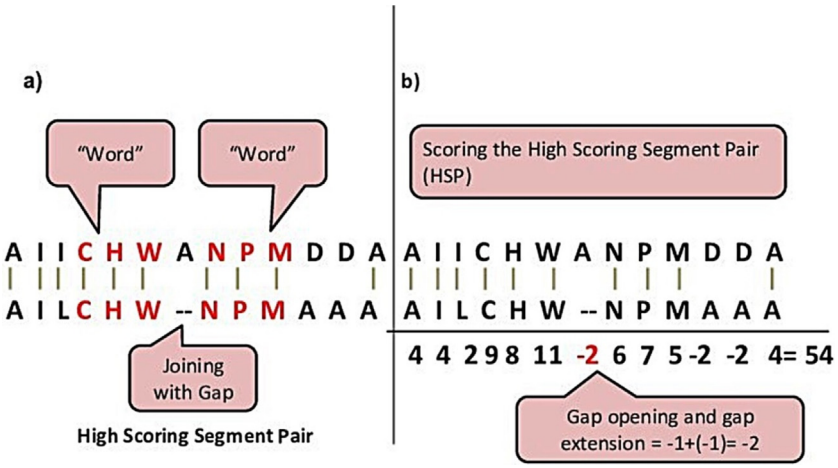


FIGURE 14.9

A) Introduction of a "gap" within the limits of threshold value giving rise to a high-scoring segment pair (HSP). (B) Scoring of HSP as per BLOSUM62 while applying the gap score of −1 for opening the gap and −1 score for gap extension. The final score, also known as "raw score," is 54.

matrix is used to score this HSP alignment (Fig. 14.9A). An HSP is a local alignment that has no gaps and achieves one of the highest alignment scores in a given search.

In the case of gaps, gap opening penalty and gap extension penalty are applied to the score. In our example, the total score is 54. The score is essentially a property of the alignment and the substitution matrix that was used to give a score to the alignment. As long the same scoring matrix is used for the same alignment, the score will always be 54. This score of 54 is called the "raw score," and it never changes (Fig. 14.9B).

However, here arises the question: Is this score significant, and if yes, what is it significantly different from?

It is of paramount importance to find out whether the score and the alignment given is a statistically significant match or whether the alignment is produced on a mere chance. The significance scores help to distinguish between the evolutionarily related sequences and the unrelated ones. Thus we start with a pairwise score and try to calculate whether the score is significant or not. For this, the initial amino acid sequence is jumbled and all the possible combinations of those amino acids at random are identified. Thus our choices have the exact same amino acid present in the sequence, but in a different order. Next, we calculate the entire possible anagram sequences, align them against the main sequence, and score these sequence alignments. The scores obtained from randomized shuffling and then alignment with the given sequence can be used to plot a graph by taking the score on the X-axis and the number of alignments on the Y-axis to obtain an inverse graph (Fig. 14.10). As the number of alignments decreases, the score increases and vice versa.

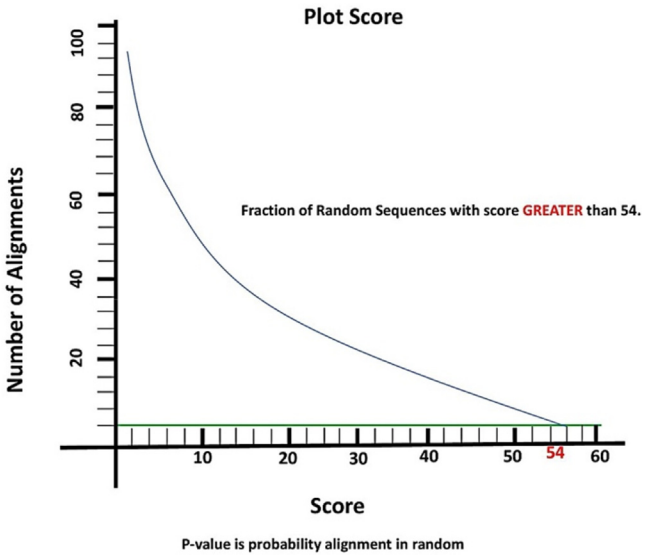


FIGURE 14.10

Inverse graph plotted between the number of alignments and the score obtained when sequence of interest was randomly jumbled and aligned with itself. The graph shows the fraction of random sequences with a score less than the raw score obtained from BLAST, which is extremely low, suggestive of a low *P*-value (i.e., result of a random chance) and a significant match.

The numbers of alignments are thus inversely proportional to the score. We now take this curve and on it we place our score of 54 (raw score from our sequence alignment). On plotting our raw score on this inverse graph we can get the probability that our alignment is random. The fraction of arbitrary sequences with a score greater than our raw score provides us with the likelihood that our alignment was just as random.

If the fraction is extremely low, it implies that the score is significant. However, if the fraction is very high, it would imply that there is a good probability that the alignment was purely because of chance. This fraction value is called the *P*-value and is defined as the probability of the sequence alignment being random. Canonically, the *P*-value is calculated by relating the observed alignment score, *S*, to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant *P*-values will be those close to zero. Table 14.2 explains a basic rule of thumb to generate inferences based on our *P*-values. These are only rules of thumb and these inferences or understandings can be modified based on our experience with sequence alignment to understand the results much better.

Based on this understanding, we can now completely understand the traditional definition of BLAST as described on the NCBI website.

Table 14.2 Different relevant inferences that can be derived on the basis of obtained *P*-value results.

P-values	Inference
10^{-100}	Identical sequences
1×10^{-50} to 1×10^{-100}	Nearly identical sequences
1×10^{-5} to 1×10^{-50}	Homologous sequences
1×10^{-1} to 1×10^{-5}	Distantly homologous sequences
$>1 \times 10^{-1}$	The sequence alignment may be random

BLAST (Altschul et al., 1990, 1997) is a sequence comparison algorithm optimized for speed and used to search sequence databases for optimal local alignments to a query. The initial search is done for a word of length "*W*" that scores at least "*T*" when compared to a query using a substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "*S*." The "*T*" parameter dictates the speed and sensitivity of the search.

BLAST is, however, a lot more complex than this. The reasons for this complexity include problems with sampling and oversampled statistics. Another problem with the algorithm is that longer sequences are more likely to find higher scoring pairs due to their lengths. Coupled with this, longer databases are more likely to find higher scoring pairs because they contain more sequences that our query sequence can match against.

Thus the developers of the BLAST algorithm decided to convert these *P*-values into expectation values or *E*-values (*E*). This value essentially gives us an idea about the expected frequency of a researcher finding this alignment in the database at random. The *E*-value is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. This value decreases exponentially as the *S* of the match increases. Essentially, the *E*-value describes the random background noise. For example, an *E*-value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see one match with a similar score simply by chance.

The lower the *E*-value, or the closer it is to zero, the more "significant" the match is. However, it should be kept in mind that virtually identical short alignments have relatively high *E*-values. This is because calculation of the *E*-value takes into account the length of the query sequence. These "high" *E*-values make sense because shorter sequences have a higher probability of occurring in the database purely by chance. The *E*-value can also be used as a convenient way to create a significance threshold for reporting results. You can change the *E*-value threshold on most BLAST search pages. When the *E*-value is increased from the default value of 10, a larger list with more low-scoring hits can be reported.

The basic formula for the *E*-value is as follows:

$$= \text{length of database}(m)\ x\ \text{length of sequence}(n)\ x\ \text{probability}(P)$$

$$E = mnP$$

The *E*-value (Expected value) can also be described as:

$$E = kmne^{-\lambda S}$$

where k and λ are scaling factors (these can be defined as parameters), m is the length of the database, n is the length of the query sequence, and S is the raw HSP score (e.g.,: 54).

Since, k and λ are scaling factors, we can simply eliminate these and convert our raw HSP score directly to a more refined score called the bit score (S').

$$S' = \frac{\lambda S - \ln k}{\ln k}$$

It is important to note that *E*-values are subject to change as new sequences are uploaded to databases, and they keep growing with each passing day. A higher bit score corresponds to a lower *E*-value. Table 14.3 suggests another basic rule of thumb for making certain inferences based on *E*-values shown at the end of a BLAST result.

The NCBI hosts a webpage at blast.ncbi.nlm.nih.gov as well as a network service. The algorithm can be run as a standalone application for researchers who intend to run it on their own machines or with their own personal sequence databases. A number of standalone latest generation applications can be downloaded by installing the BLAST+ package from the Web. NCBI also serves as a great portal for assisting researchers in carrying out alignments by providing them access to NIH Genetic and Proteomic Sequence databases that contain a large volume of nucleotide and protein sequence data. The algorithm can also be used in association with other bioinformatics algorithms that require approximate sequence matching.

Table 14.3 Different relevant inferences that can be derived on the basis of obtained *E*-value results.

Expected values (E-values)	Inference
$<1 \times 10^{-100}$	Identical sequences
1×10^{-50} to 1×10^{-100}	Nearly identical sequences
1×10^{-5} to 1×10^{-50}	Homologous sequences
1×10^{-1} to 1×10^{-5}	Distantly homologous sequences
$>1 \times 10^{-1}$	The sequence alignment may be random

14.5 Codons, reading frames, and open reading frames

In molecular sciences, a reading frame is a way of dividing nucleotide sequences (RNA or DNA) into triplets that are consecutive and nonoverlapping. These triplets code for amino acids during translation, and are thus called codons.

An open reading frame (ORF) is the segment of a reading frame that has the tendency to be translated. It is a register for reading the section of DNA that starts with ATG (start codon), uses three letters at a time to call for amino acids, and uses one or more amino acids to close the reading. It is a continuous stretch of codons that start with the start codon (AUG) and end at a stop codon (UGA, UAA, or UAG). The start codon within the ORF marks the point where the translation starts and the stop codon, in turn, marks the transcription termination site.

Since DNA is interpreted in nucleotide triplets called codons, a DNA strand has three distinct reading frames. However, we know that the DNA molecule exists as a double helix having two strands that runs in an antiparallel manner (i.e., if one strand runs in the $5' \rightarrow 3'$ direction, the other runs in the $3' \rightarrow 5'$ direction), having three reading frames each. Therefore there are six possible frames of translation.

There are three reading frames that can be read in the $5' \rightarrow 3'$ direction, each one beginning from a different nucleotide in the triplet, and an additional three reading frames that can be read from the other complementary strand.

NOTE: Not every ORF makes a protein. As a rule of thumb, researchers consider a minimum of 100 amino acids to produce a protein. ORFs are therefore essential in transcriptomics, metabolomics, bioinformatics, and chemoinformatics studies.

Now that we have a basic understanding of codons, reading frames, and ORFs, we can correlate that knowledge to understand different types or variants of BLAST that exist for providing researchers with different prospects to go further with their projects. The different variants or types of BLAST are discussed ahead. The next section of this chapter describes how four major BLAST programs function by utilizing different types of databases and query sequences. These specialized types of BLAST can be correlated with the applications of BLAST mentioned earlier in this chapter to curate specific research projects depending on various targeted research objectives.

The major types of BLAST are:

1. BLASTN: Compares a DNA query to a DNA database; searches for both strands automatically. This type of BLAST is optimized for speed over sensitivity.
2. BLASTP: Known to compare a protein query against a protein database.
3. BLASTX: Compares a DNA query sequence to a protein database by translating the former in six different ORFs, and then compares each of them against the database, which has three reading frames from each DNA strand.
4. TBLASTN: Compares a protein query to a DNA sequence database in all the six possible ORFs of the database.

5. TBLASTX: Compares the protein encoded in the DNA query to the protein encoded in the DNA database in the 6 × 6 possible frames of the database and query sequences.

Apart from the types of BLAST described here, there exist many other BLAST wrappers and derivatives. BLAST wrappers are specialized scripts that run BLAST in specific ways. Other variants include PHI-BLAST, PSI-BLAST, organism-specific BLAST, MegaBLAST, BLASTZ, XBLAST, HT-BLAST, GENE-BLAST, and MPBLAST.

Once researchers have a basic understanding of sequence alignments, homology, the BLAST algorithm, ORFs, and the traditional alignment approaches, they can extrapolate that information and knowledge to work with various specialized forms of BLAST mentioned earlier. These forms cater to specific research needs and require additional fields or parameters that can be mentioned in the user interface fields according to the requirement of the researcher.

14.6 Bioinformatics and drug design

Drug discovery involves the process of identifying new medications based on the knowledge provided by their targets; these targets are often biological molecules like membrane receptor proteins, membrane transport proteins, protein kinases, enzymes, transcription factors, and sometimes specific binding regions of DNA. The medications, however, also known as drug molecules, are often small chemical molecules that either activate or inhibit the function of biological molecules (drug targets), which in turn interfere with their metabolic or signaling pathway in disease. Thus, in layman's terms, a drug molecule should be designed in its shape and structure in such a way that it interferes with the normal pathway of a target molecule by binding to it. The binding of a drug molecule can be anywhere on the target molecule as long as its binding inhibits the target molecule's functions and qualifies for all the criteria, including absorption, distribution, metabolism, excretion, and toxicological (ADMET) profiles. Traditional drug discovery using the trial-and-error method is usually time consuming and economically taxing, thus a more robust method was required that would screen, identify, and validate drug targets and drug molecules on a larger scale while saving time and money. This type of modeling is often referred to as computer-aided drug design (CADD), which is a specialized discipline that uses computational methods to study the drug−receptor interactions. CADD methods are heavily dependent on bioinformatics tools, applications, and databases. The drug discovery process usually starts with an analysis of binding sites in target proteins and an identification of structural features common to active compounds (Fig. 14.11). Moreover, CADD is also used to predict the conformational changes in the target that may occur when the small molecule binds to it, allowing efficient screening of lead molecules.

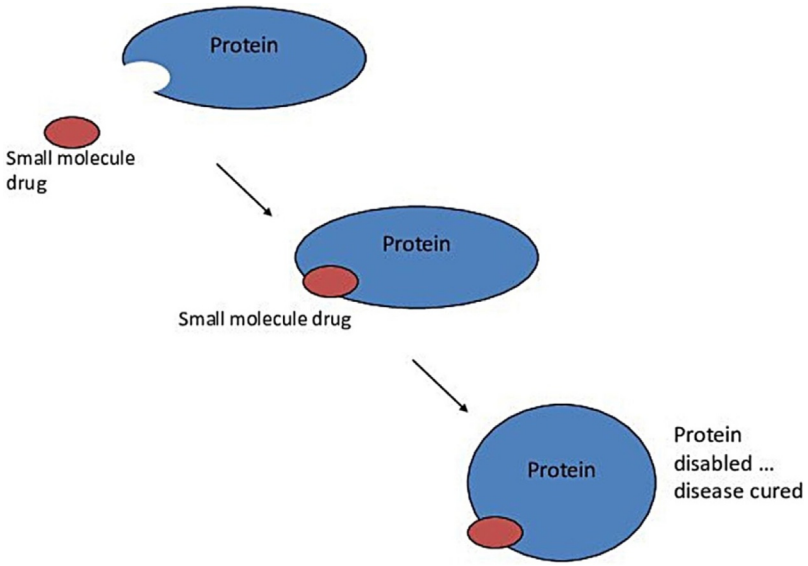


FIGURE 14.11

Drug–target interaction in its most oversimplified form. However, this concept is core to identifying putative drugs for any disease to identify the protein of interest (drug target) for which a modified lead/drug can be used. Bioinformatics plays a major role in identifying such targets, especially with the help of BLAST.

CADD indeed plays a major role in the rapid assessment of already existing chemical libraries to speed up the early-stage development of new active compounds, saving billions of dollars and a massive amount of time. One of the most important functions of CADD is to estimate the strength of intermolecular interaction between putative drug molecule and target molecule. This suggests, be it ligand-based strategy or structure-based strategy, there are two major pillars on which the efficiency and even the possibility of CADD to function is dependent: "lead molecule," also known as drug molecule, and "target molecule," which is mostly a protein (Fig. 14.12).

CADD entails a vast number of computational methodologies like virtual library design, virtual screening, lead optimization, and de novo design. Due to its proven ability of computational techniques to guide the selection of new hit compounds, chemoinformatics is still the scientific discipline that is in full bloom. While there are databases and libraries that contain lists of thousands of chemical compounds that can be used as lead molecules, they still need to be optimized and validated for each specific target protein. Here, BLAST plays a crucial role in identifying such drug targets by identifying those protein and/or DNA sequences that are similar to or to some extent share homology with already existing targets. So, if screening, identifying, and validating a lead molecule is one the pivotal functions of CADD,

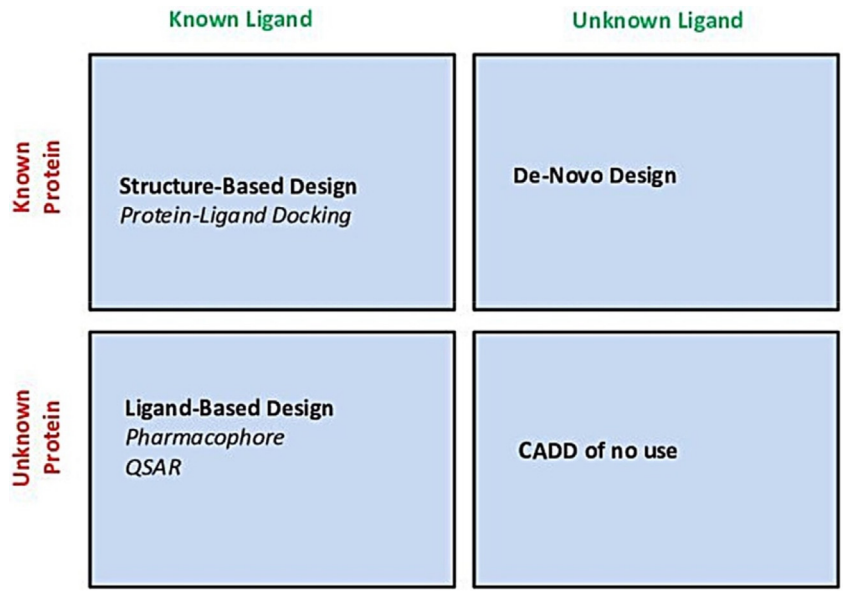


FIGURE 14.12

Computer-aided drug design (CADD) identifies the drug targets based on already existing knowledge of protein structures as well us by performing de novo synthesis of protein structures based on the chemical structure of already known drugs.
QSAR, Quantitative structure–activity relationship.

sequence analysis of the target molecule to be a perfect match for the lead molecule is equally important. BLAST is used to identify drug targets in two different ways: either identifying the drug target (protein, DNA, RNA, peptide) that is already available in protein and other databases or generating a 3D structure of the active site or whole protein via de novo synthesis. Most of the drugs exert their effects on target proteins by impairing structural and/or functional capabilities. The BLAST technique as we understand it is at the core of finding sequences that share similar structural and functional properties. Therefore if the structure of a drug molecule is known, regarding the target site it binds to, and if the structure of the target molecule (e.g., catalytic site of an enzyme protein, structural conformation of a receptor protein, subcellular localization of transport protein) is known, then BLAST can be utilized to identify other putative target proteins in a family that shares some similarity with the target molecule. On the other hand, if the structure of the target molecule (protein) is unknown, then ligand-based design, where pharmacophores are used to identify the putative target site of the unknown protein, is used to elucidate the structure of a "part" of the target protein. Once CADD provides the basic putative structure of a target site, homology-based 3D structuring of proteins will be carried out for which, again, BLAST plays a critical role by identifying the most promising backbone for the protein.

14.7 Applications of BLAST

1. BLAST plays a very pivotal role in chemoinformatics and drug discovery (explained earlier).
2. BLAST helps in the identification of secondary and tertiary structures of proteins from existing databases, which can be further studied as drug targets.
3. BLAST helps to kick start any kind of de novo 3D structure prediction, which can further be used to identify putative binding drug molecules.
4. BLAST can be used for identifying species when working with a DNA sequence from an unknown species or in finding homologous, orthologous, and paralogous sequences for a given query sequence. This is useful when a microbial, plant, or animal species has a protein that is related phylogenetically in lineage to a certain protein with a known amino acid sequence.
5. While working with proteins, a researcher can input a given protein sequence into the BLAST algorithm to find out various domains within the sequence of interest. This is useful to discover other genes that encode proteins that exhibit motifs such as those that have been determined in an experiment?
6. A BLAST output can construct a radial or linear phylogenetic tree and help establish relationships between the sequence under study and its ancestral sequences. If a particular protein causes a microbial infection in a host, the evolutionary ancestor of this protein can be determined and that information is used to establish links between vaccines or drugs developed for the ancestor and their effectiveness in combating the disease caused by the protein under study.
7. BLAST is also very useful in DNA mapping when an unknown region of DNA sequence is encountered while working with a known species. In this scenario, BLAST can be used to compare the region of interest to the relevant sequences in the database(s). NCBI has a "Magic-BLAST" tool built around BLAST for this purpose.
8. When working with genes, BLAST can locate common genes in two related species, and can be used to map annotations from one organism to another.
9. Different polymorphisms of any gene can be studied with the help of BLAST.
10. BLAST can identify novel genes by checking the structural and functional similarity between genes of interest and sequences in vast databases.

14.8 Understanding coronavirus: the menace of 2020

In December 2019, a novel coronavirus outbreak was discovered in Wuhan, China. This virus outbreak was labeled a Public Health Emergency of International Concern by the World Health Organization within 2 months of the discovery of the first few patients. Though the virus seems relatively less hazardous (with a case fatality rate of about 3%−4%), the number of deaths caused by the global

pandemic has already surpassed that of the 2002–03 severe acute respiratory syndrome (SARS) outbreak. The rapid increase in pandemic cases has resulted in more genomes being sequenced every day, in hopes that they may provide some clarity and evidence of viral mutations, specifically, the possibility of the introduction of different variants of the virus into the human population.

With an average rise of 160,000 global COVID-19 cases per day, as of June 21st, there are widespread concerns that the virus will acquire more substrains and variants, and the virulent viral strain is estimated to emerge with a much stronger toxicity, causing extreme outcomes for the global population. This makes it extremely crucial to track, model, and characterize the viral strains, variants, symptoms, patient profiles, treatment responses, and geographical locations. BLAST can be employed to carry out various sequence alignments and enable researchers to find new viral variants of SARS-CoV-2 from the comfort of their homes, in combination with various in silico approaches discussed in this book. Coupled with this, BLAST results can be employed to carry out phylogenetic analysis, and thereby conduct research for a better understanding of the origin of the viral genome. This can catapult biological and clinical research toward finding effective vaccine candidates based on sequence similarities found between SARS-CoV-2 and other related viruses. As of this writing, there is no slowdown in COVID-19 cases, and the number of infected patients seems to be rising exponentially. Our fight against this viral menace will be a long and tiring one, until we develop vaccines or effective approaches and treatments to tackle this unforeseen situation. We are still at an early stage of the global pandemic and have little information about the tendency of the virus to mutate into newer, more deadly forms. It is therefore incumbent upon bioinformatics and chemoinformatics researchers and scientists to employ their expertise, knowledge, and skills to develop better experimental approaches that can help accelerate the vaccine development process, among other research crusades. As of this writing, various online portals, including NCBI, EMBL, and DDBJ, are making efforts to release viral genomes in open-source, web repositories, enabling researchers to engage in variant analysis experiments.

14.8.1 BLAST simulation practical

1. Go to https://www.ncbi.nlm.nih.gov/sars-cov-2/ to retrieve SARS-CoV-2 sequences that have been uploaded to this specialized NCBI resource.
2. Access the NCBI sequence records through the website mentioned in 1. by clicking on the "Download Sequence List" button. This redirects you to a page that starts the download for a PDF file that has all the latest lists of SARS-CoV-2 nucleotide sequences. You can now query these IDs in GenBank.
3. You must now open the PDF file and select the accession numbers of the two coronavirus sequences you wish to align and draw results from. Accession numbers are unique numbers or fingerprints that are special to an entry in the NCBI database. Each accession number corresponds to only one biological

sequence, and is linked to a specific page, dedicated completely for that particular sequence.

4. You must now open a new tab and redirect to https://blast.ncbi.nlm.nih.gov/Blast.cgi This is the official BLAST page hosted by NCBI. You can now select Nucleotide BLAST on this website. This action will redirect you to the Nucleotide BLAST page hosted by NCBI.
5. On the Nucleotide BLAST page, you can now paste the selected accession numbers and enter the parameters of search or use the default BLAST parameters to carry out the alignment.
6. Once you have selected the Search Set and Program in the "Choose Search Set" and "Program Selection" parameters, you can finally click on BLAST. The input sequences now enter the algorithm for processing. This processing step might take a few minutes.
7. You will now be redirected to the BLAST Results page, and has access to the alignment results, which can be used to draw further inferences. These include a description tab that has information regarding scores, identity, query cover and many other parameters, a graphic summary tab that provides a graphical result of the alignment, an alignments tab that enables you to change the alignment view, and a taxonomy tab that enables you to work with information regarding lineage, organism, and taxonomy. The BLAST output also includes *E*-values for the alignments that will enable you to draw conclusions and inferences discussed in Table 14.3 in this chapter.

The following are five accession numbers to different SARS-CoV-2 sequences. Try aligning different combinations of the five sequences by taking varying pairs one at a time:

1. MT039890.1
2. MN994468.1
3. NC_045512.2
4. MT019532.1
5. MT044258.1

The PDF file downloaded from the NCBI website has the accession numbers to numerous SARS-CoV-2 variants, which can be used as inputs for BLAST. This opens a wide new horizon of research for students, researchers, and biology enthusiasts. Any and every scientific enthusiast who reads this chapter can now follow the foregoing instructions to access the BLAST portal and draw inferences based on these results from the comfort of their homes. Find a unique combination of two sequences, pay attention to the BLAST results, and you might be the next great scientist who discovers a mutation in the different viral genomes. Remember, when you change the way you look at things, the things you look at begin to change. These data might sometimes seem inconsequential, but if you dive deeper into the very basics of your understanding and start seeing things differently, these very data may lead to further inferences that the best of scientists might have somehow missed.

An alternative approach may include aligning any two sequences from the aforementioned resource, and carrying out a BLAST against the whole nucleotide database to see similarity with other viruses, and draw various inferences from them.

For instance, a BLAST between the parent/zeroth sequence of SARS-CoV-2 from Wuhan, China, shows a higher percentage identity with a Middle East respiratory syndrome (MERS) genomic sequence. An appropriate inference and course of action may be to start using the previously done research on MERS for drug discovery and for strategically curating the vaccine development process. This accelerates the vaccine development process, thereby saving months of time and thousands, if not, millions of lives. Research from the comfort of your home may now impact millions of lives. A few clicks and you might be the one to save the world from a global pandemic.

Learn. Think. Act.

14.9 Conclusions

In this chapter, we explained the very basics of BLAST to understand what BLAST is, how it is used, and the background on which BLAST was created. We briefly explained the core concepts that are required to understand the functionality of BLAST and more importantly for students to have an underlying understanding of sequence comparison, which is pivotal for drug discovery. BLAST in itself is a technique that has been used in several fields of science; here, we briefly explained how it is an integral part of drug discovery. We live in a world where technological advances from the past 70 years have transformed from using a computer that was as big as a room to exploiting artificial intelligence in gadgets that fit in the palm of a hand. There was a time when biology, chemistry, mathematics, and computer sciences where subjects of their own right; however, with the advent of bioinformatics all these subjects were amalgamated to give results in a time frame that was not possible before. One such area that benefited from the fruits of bioinformatics is the field of pharmacology in novel drug discovery. Traditionally, drug discovery was a tedious process, which for a single drug could take more than 20 years and millions of dollars, and was at risk of being disqualified for ADMET properties to pass through clinical trials at the last stage. Due to this slow progress and massive expenditure, market pressure to find new drugs in a short period of time along with bare minimum risks has fueled the interest of researchers to design drugs using bioinformatics. These challenges were overcome by the introduction of CADD that uses cost-effective and time-efficient procedures for the development of drugs. With the help of CADD, millions of drug molecules (virtual screening) and drug targets (protein receptors, enzymes, kinases, signaling proteins, etc.) are analyzed, screened, modeled, and predicted with the least amount of money and time spent, and moreover with added accuracy. Although experimentation and in vivo testing are the only ways to identify the authenticity of a drug, CADD eliminates millions of spurious candidates and targets at a time (read Chapter Chemoinformatics for a more detailed insight into CADD).

The identification of regions of interest in drug targets includes ORFs, conserved motifs, conserved domains, promoters, catalytic region, terminators, subcellular localization signal regions, and other structural and functional regions that are crucial to the protein involved in a specific disease. Without knowledge of these regions, no amount of efficient drug molecule is of any use unless the target that it binds to is properly identified, verified, and functionally understood. To validate the drug target, it is imperative to have the sequence analysis of the said candidate protein, and this is precisely why: BLAST is a decisive technique at the very core of pharmacology and chemoinformatics. In the grand scheme of things and the vast fields of chemoinformatics and complicated methodology of CADD, BLAST may seem to be a single isolated technique, but this technique is one of the integral pillars of the foundation on which modern drug discovery is laid. Comparative/homology modeling is one of the most straightforward approaches to predict a 3D structure of a protein molecule. BLAST helps in the identification of a related template (at least 30% sequence identity with target protein), which is further used to predict the unknown structure of the target protein. Once the structure of a target protein is identified, CADD, quantitative structure−activity relationship, virtual screening, and other programs can be used by researchers to tailor specific compounds that bind at a particular site for a given protein. The Protein Data Bank (PDB) is one of the most sought databases that have protein structures that are verified by nuclear magnetic resonance and X-ray crystallography; however, due to slow updating of the database, relatively low numbers of structures are available compared to primary and secondary structures available in a database like Uni-Prot. Apart from using PDB, there are several other approaches used to model the target protein if a homolog of an unknown protein is available, such as MODELLER, 3D-JIGSAW, and COMPOSER, among others. Nevertheless, no matter what software is used to design a drug for the protein pertaining to a disease, BLAST remains at the very center, or should we say at the very beginning, of each analysis.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403−410.

CHAPTER

Pseudoternary phase diagrams used in emulsion preparation

15

Vikas Jhawat, Monika Gulia, Anil Kumar Sharma

School of Medical and Allied Sciences, GD Goenka University, Gurugram, Haryana, India

15.1 Introduction

An emulsion is a biphasic dispersion of two immiscible liquids in which one phase is evenly distributed as a fine globule form (dispersed phase) into the other continuous phase (dispersion medium) (Mehta, 2002; Winfield and Richards, 2004). Emulsions are thermodynamically unstable, which tend to separate into two phases, i.e., aqueous and oily phases on standing undisturbed; therefore requires special additives named emulsifying agents to be stable for longer time (Goodarzi and Zendehboudi, 2019; Agarwal and Rajesh, 2007). Depending on the phases, emulsions may be classified as oil-in-water (O/W) emulsion and water-in-oil (W/O) emulsion. Complex forms of these two emulsions are called double emulsions or multiple emulsions. When an emulsion is emulsified with another phase, e.g., a W/O emulsion, and emulsified again with water, it produces a water-in-oil-in-water (W/O/W) emulsion (Khan et al., 2006). The globule size of the dispersed phase in an emulsion depends on many factors, such as the method of preparation, nature of oil used, and type of emulsifying agent. This categorizes the emulsion as a macroemulsion (0.1−10 μm), microemulsion (5−50 nm), and nanoemulsion (20−1000 nm) (Fuhrman, 2006; Binks, 1998), which further determine the stability of the emulsion (Dickinson, 1994). The larger the size of droplets of the dispersed phase, the smaller the stability because the large droplets tend to coalescence with each other, which leads to phase separation. Therefore macroemulsions are less stable than microemulsions over time (Ganguli and Ganguli, 2003). Addition of surfactants (emulsifying agents) reduces the interfacial tension at the oil and water interface and increases the miscibility of these two phases, which makes them stable for a prolonged time (Sharma et al., 2014). The emulsifiers are amphiphilic molecules, which contain both polar and nonpolar groups that form micelles around the dispersed phase at the liquid interface and hold both liquids tightly to make them miscible (Amashita et al., 2017).

In the area of pharmaceuticals, emulsions have broader applications in drug delivery systems as most of the drug molecules are hydrophobic in nature, which poses solubility and bioavailability problems in drug delivery to the body. So, the emulsions can be of great importance in the rapid delivery of both hydrophilic and lipophilic drugs due to the presence of the characteristics of both water and

Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. https://doi.org/10.1016/B978-0-12-821748-1.00011-7

lipids, which can be delivered either through the oral route or through the topical route (Ahmad et al., 2011; Kommuru, 2001). A pseudoternary phase diagram is a tool that optimizes the three components of any typical emulsion, i.e., water, oil, and surfactant, to obtain the concentration range of these components, which form a stable emulsion. The three components of a system are plotted on the three corners of the triangle. The sides of the triangle represent the binary system, and the middle part represents the possible equilibrium state between the three components at constant conditions of temperature and pressure (Richard et al., 2013). The phase diagram can also determine the pattern of arrangement of surfactant molecules at the water and oil interface. A ternary phase diagram helps to identify and determine the effect of component variabilities such as type of oil phase, surfactant ratio on the globule size, viscosity, pH, conductivity, refractive index, and transmittance properties of the prepared emulsion. The stability-indicating factors of the emulsion can be optimized based on the foregoing information, which further assists in the designing of the drug delivery system for different drug molecules (Kumar et al., 2016). The mutual interaction of the three components in different proportions produces different types of phases: either a completely miscible phase, which is stable for a longer time, or a partially miscible phase, which separates into two pure components, i.e., water and oil, when standing for some time. Regarding the weight percentage range of each component, making a stable emulsion can be identified from a phase diagram. Therefore a further detailed study of pseudoternary phase vdiagrams of emulsions is discussed in the following sections to understand, optimize, and develop emulsions into stable drug delivery systems.

15.2 Classification of emulsions

Pharmaceutical emulsions can be classified broadly in two ways (Fig. 15.1) (Fatima et al., 2014; Bari et al., 2019):

1. Based on the dispersed phase:
   - **(a)** Simple emulsion (Kempin et al., 2020; Wang et al., 2020):
     - (i) W/O emulsion
     - (ii) O/W emulsion
   - **(b)** Complex/multiple emulsion (Iqbal et al., 2020; Ji et al., 2020):
     - (i) W/O/W emulsion
     - (ii) Oil-in-water-in-oil (O/W/O) emulsion
2. Based on the globule size (Hazlett and Schechter, 1988; Das et al., 2020; Jiang et al., 2020):
   - **(a)** Macroemulsion
   - **(b)** Microemulsion
   - **(c)** Nanoemulsion

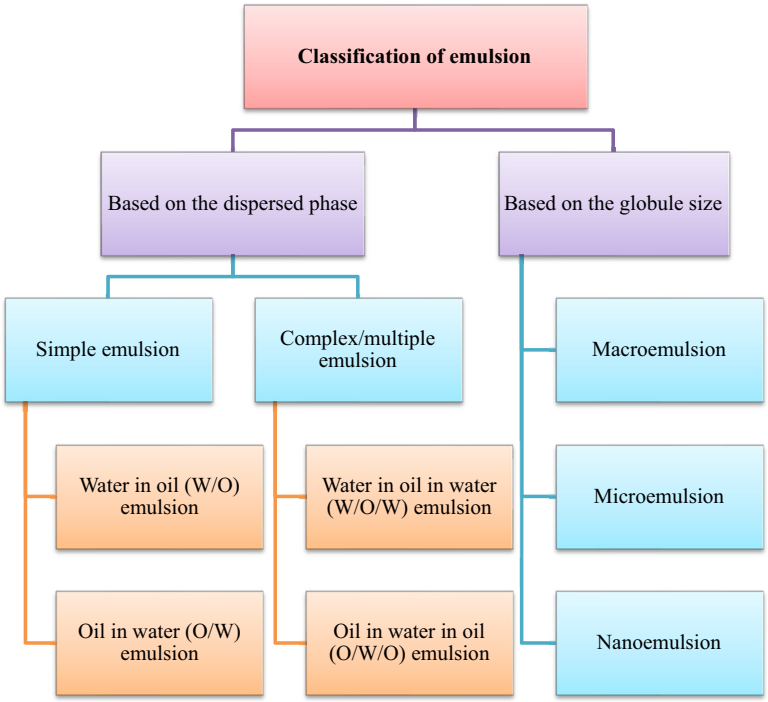


FIGURE 15.1

Classification of emulsions.

15.2.1 Simple emulsion

A simple emulsion is the mixture of two phases, i.e., water and oil (Fig. 15.2). W/O emulsion is produced when the water content is less than 45% of the total weight, and an oil-soluble surfactant is used during the preparation of the emulsion. In this condition, water droplets are suspended into the oil with the help of oil-soluble emulsifying agents such as wool fat, beeswax, fatty acids, and resins. W/O emulsion is generally meant for external application to the skin (Opawale and Burgess, 1998; Bokhout et al., 1981; Bobra, 1991). In the case of O/W emulsion, the water is in excess, i.e., more than 45% of the total weight of the emulsion. A suitable water-soluble emulsifying agent is used, which suspends the oil droplets in the water with the help of water-soluble emulsifying agents such as tragacanth, acacia, and cellulose derivatives. Addition of surfactant may also be desired to stabilize the emulsion for extended periods. These are prepared for internal use (Paulo et al., 2020; Stepisnik et al., 2019).

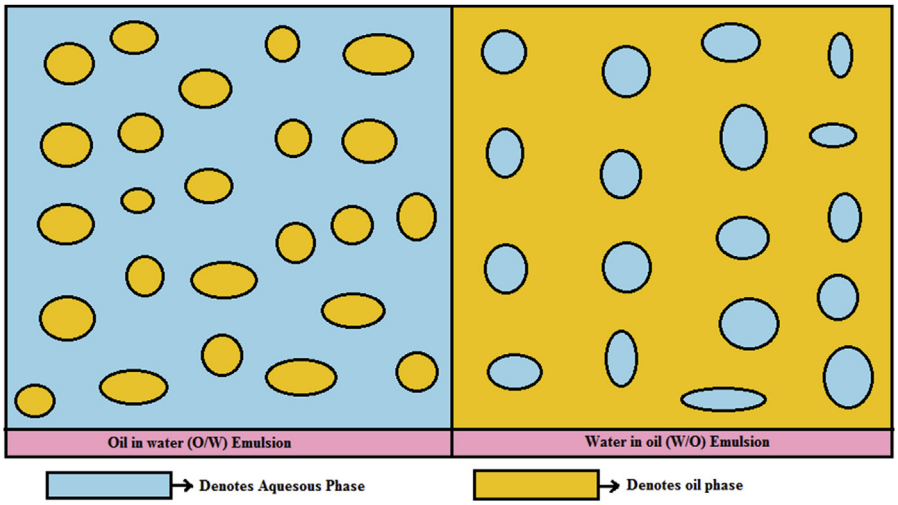


FIGURE 15.2

Representation of simple emulsion types.

15.2.2 Complex/multiple emulsion

A complex emulsion is one that is produced by the reemulsification of the simple emulsion, e.g., W/O emulsion is reemulsified with water to produce W/O/W emulsion by the addition of suitable emulsifiers (Fig. 15.3). Similarly, O/W emulsion is reemulsified with oil to produce O/W/O emulsion. Complex emulsions are very difficult to produce and maintain as these contain multiple surfactants of opposite nature, which may cause phase separation, cracking, or phase inversion of the emulsion (Matsumoto et al., 1976; Silva et al., 2016; Soriano-Ruiz et al., 2019; Liu et al., 2004).

15.2.3 Macroemulsion

Macroemulsions are the mixture of two immiscible liquids with a dispersed phase having a droplet size >0.1 μm. These are thermodynamically unstable and appear turbid or milky when two phases are mixed. Due to their larger size, the droplets of the dispersed phase coalesce with each other, and the two immiscible phases separate out on standing of the emulsion. However, use of a suitable emulsifier may increase the stability of macroemulsions to some extent (Sharma and Shah, 1985; Ruckenstein, 1999).

15.2.4 Microemulsion

In contrast to a macroemulsion, microemulsions are thermodynamically more stable with disperse-phase droplet sizes of 5−50 nm. These appear as a transparent and

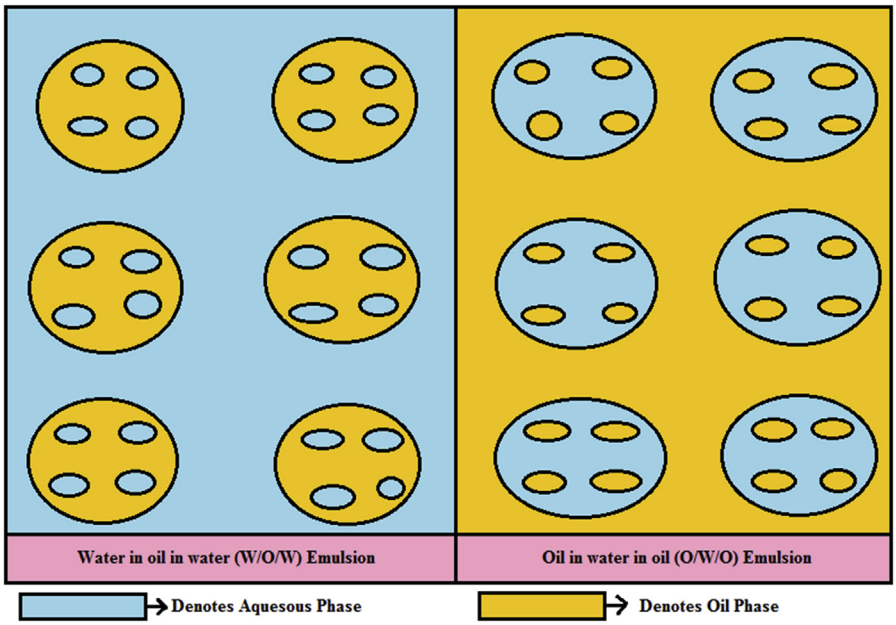


FIGURE 15.3

Representation of complex emulsion.

clear liquid when prepared by using suitable emulsifying agents (Fig. 15.4). The proportion of surfactant is higher in microemulsions compared to macroemulsions, which lowers the interfacial tension to a considerable extent and makes microemulsions quite stable (Kumar and Mittal, 1999; Lawrence and Rees, 2000).

15.2.5 Nanoemulsion

Nanoemulsion is a biphasic colloidal system with a droplet size of the dispersed phase in the submicron size range of 20−1000 nm. Nanoemulsions are thermodynamically most stable and transparent because of the presence of one or more amphiphilic surfactants. Nanoemulsion is the most advanced type of emulsion system, which has great potential in the area of drug delivery systems. It is used in targeted delivery, mucosal delivery, transdermal delivery, and site-specific delivery and as a diagnostic tool (Singh et al., 2017; Shah et al., 2010; Fryd and Mason, 2012).

15.3 Emulsifying agents (surfactants)

Emulsifying agents are a range of hydrophilic and lipophilic surfactants (low molecular weight chemicals), which are used in the preparation of different

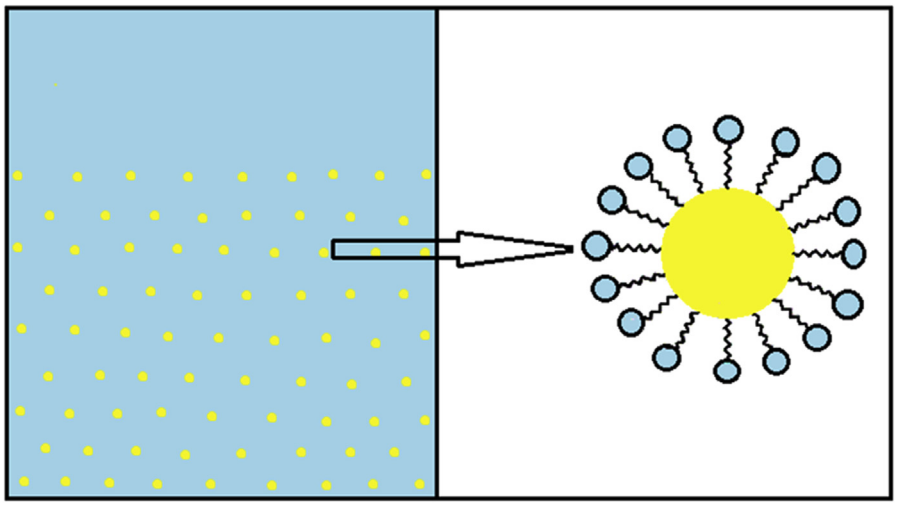


FIGURE 15.4

Representation of oil-in-water microemulsion.

types of emulsions. These are amphiphilic in nature and contain hydrophilic and lipophilic characteristics in a single chemical moiety. This moiety is distributed at the interface of two immiscible liquids and reduces the surface tension (Vijayakumar and Saravanan, 2015; Sobrinho et al., 2013). Emulsifying agents can be classified based on their solubilities, such as either water soluble or oil soluble, using the hydrophilic−lipophilic balance (HLB) scale proposed by Griffin (Yamashita and Sakamoto, 2016). The HLB scale classifies surfactants on an imaginary scale as values from 0 to 20, which are based on the relative proportion of polar-to-nonpolar groups in a nonionic surfactant molecule. For emulsifying agents, an HLB value of 2−6 indicates them as oil-soluble surfactants and 12−15 as water-soluble surfactants (Gadhave, 2014). This scale has now been extended to ionic surfactants too, which have higher HLB values up to 50 based on their ionization extent. Therefore, based on the nature and type of emulsion being prepared, a suitable emulsifying agent, either single or in combination, can be selected based on the HLB classification as shown in Table 15.1 (Rieger, 1987). If the continuous phase is oil, then the surfactants with HLB values of 2−6 are most suitable. For water, continuous-phase surfactants with an HLB value of 12−15 are suitable (Fig. 15.5) (Griffin, 1949). In the case of complex or multiple emulsions, the combination of surfactants can be utilized to achieve the desired stability of the emulsion (Fox, 1986).

When a single surfactant is unable to provide the desired *HLB* value, then the addition of another surfactant with the previous one may be chosen to obtain the desired *HLB* to prepare a stable emulsion. Two or more surfactants can be mixed based on the fraction of oil and fats used for the preparation of the emulsion. Suppose there is a fraction f of surfactant X and fraction $(1 - f)$ of surfactant Y to

Table 15.1 List of surfactants with their hydrophilic–lipophilic balance (HLB) values used for emulsion preparation.

Name	HLB	Water dispersibility
Ethylene glycol distearate	1.5	No dispersion
Sorbitan tristearate	2.1	
Propylene glycol monostearate	3.4	
Sorbitan sesquioleate	3.7	
Glyceryl monostearate, nonself-emulsifying	3.8	Poor dispersion
Propylene glycol monolaurate	4.5	
Sorbitan monostearate	4.7	
Diethylene glycol monostearate	4.7	
Glyceryl monostearate, self-emulsifying	5.5	
Diethylene glycol monolaurate	6.1	Milky dispersion (not stable)
Sorbitan monopalmitate	6.7	
Sucrose dioleate	7.1	
Propylene glycol (200) monooleate	8.0	
Sorbitan monolaurate	8.6	
Polyethylene (4) lauryl ether	9.5	Milky dispersion (stable)
Polyoxyethylene (4) sorbitan monostearate	9.6	
Polyoxyethylene (6) cetyl ether	10.3	
Polyoxyethylene (20) sorbitan tristearate	10.5	Translucent to clear dispersion
Polyoxyethylene glycol (400) monooleate	11.4	
Polyoxyethylene glycol (400) monostearate	11.6	
Polyoxyethylene (9) nonyl phenol	13.0	
Propylene glycol (400) monolaurate	13.1	Clear solution
Polyoxyethylene (4) sorbitan monolaurate	13.3	
Polyoxyethylene (20) sorbitan monooleate	15.0	
Polyoxyethylene (20) oleyl ether	15.4	
Polyoxyethylene (20) sorbitan monopalmitate	15.6	
Polyoxyethylene (20) cetyl ether	15.7	
Polyoxyethylene (40) stearate	16.9	
Sodium oleate	18.0	
Polyoxyethylene (100) stearate	18.8	
Potassium oleate	20.0	
Sodium lauryl sulfate	Approx. 40	

be included in the formulation, then the *HLB* value of the mixture would be calculated as per Eqs. (15.1)–(15.3):

$$\mathrm{HLB}_{\mathrm{mixture}} = \mathrm{f} \cdot \mathrm{HLB}_{\mathrm{X}} + (1 - \mathrm{f}) \cdot \mathrm{HLB}_{\mathrm{Y}} \tag{15.1}$$

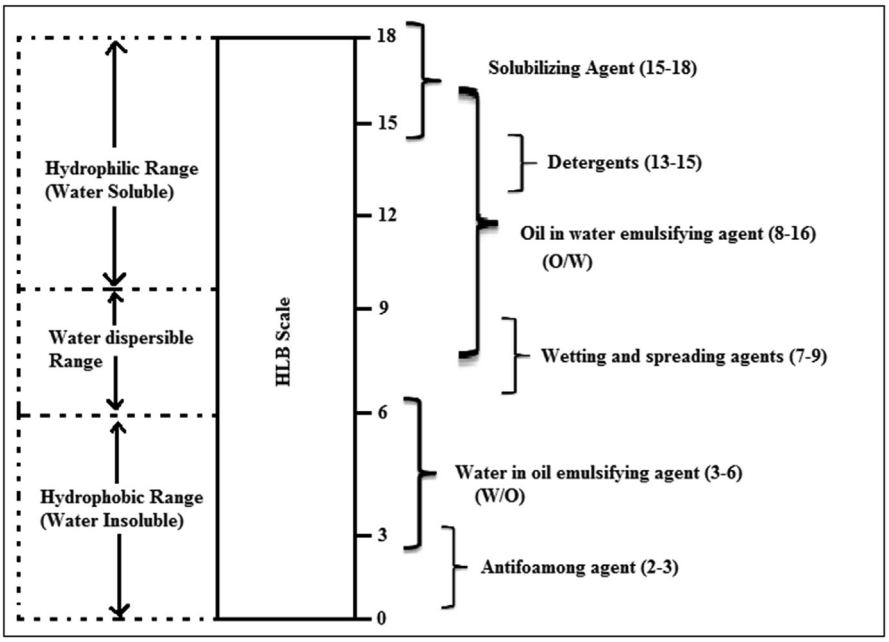


FIGURE 15.5

Hydrophilic–lipophilic balance (HLB) classification of surfactants.

The percentage amount of surfactant *X* and surfactant *Y* needed to produce the required *HLB* can be calculated by rearranging Eq. (15.1):

$$X = 100 \cdot (f - HLB_Y)/HLB_X - HLB_Y \tag{15.2}$$

$$Y = 100 - X \tag{15.3}$$

15.4 Pseudoternary phase diagrams

Any typical emulsion is formed by a suitable blend of oil, water, and surfactant/cosurfactant. So, a ternary phase diagram is the graphical representation of these three phases in the form of an imaginary triangle. This triangle determines the phase behavior and type of emulsion, droplet size, properties, and stability of formed emulsion (Ahmad et al., 2013). The apex of triangles represents the pure component, i.e., 100%, which reduces gradually to 0% on reaching another apex where another component is 100%, as shown in Fig. 15.6.

Pseudoternary phase diagrams are produced by the water titration method in which first S_{mix} (suitable fixed weight ratios of surfactant and cosurfactant) is prepared and then mixed with the oil phase in different w/w ratios. This mixture of

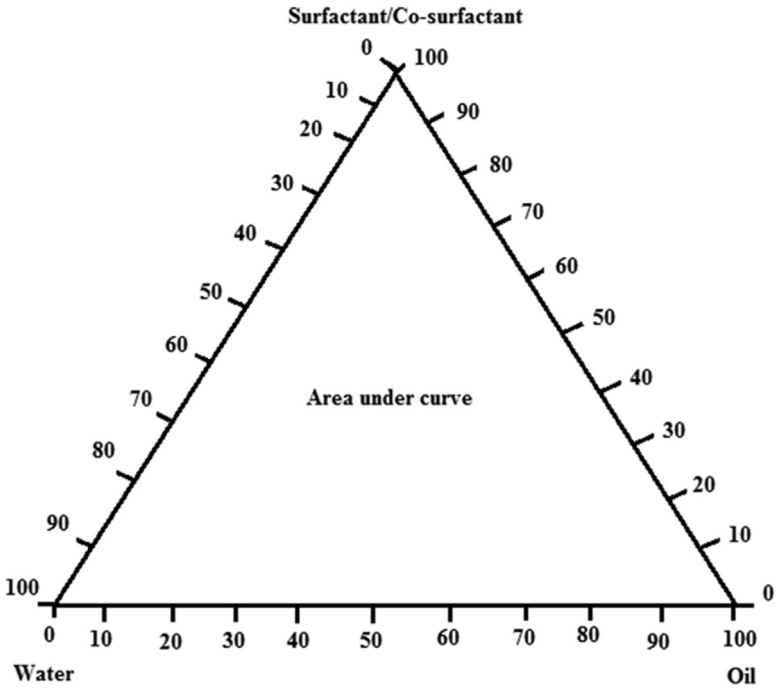


FIGURE 15.6

Representation of a pseudoternary phase diagram containing all three phases of an emulsion, i.e., water, oil, and surfactant/cosurfactant.

oil and surfactant (S_{mix}:oil) is titrated with water added dropwise. The resulting emulsions are observed for their physicochemical properties such as transparency, globule size, and stability. These ratio points are then plotted on the phase diagram and the area covered under these points gives the range of microemulsion existence (Chandra et al., 2014; Sabale and Vora, 2012). The phase behavior of emulsion components can easily be presented and studied with the help of a pseudoternary phase diagram. The components of the system must be mixed in a fixed ratio either in weight or in volume in such a manner that the concentration of one component decreases from 100 to 0 and that of other component gradually increases from 0 to 100. Generally, the phase diagram of emulsions is prepared by using the fixed ratios of water to surfactant or ratio of surfactant to cosurfactant (Nazzal et al., 2002; Rao and Shao, 2008; Zhang et al., 2008). Phase diagrams are the easiest way to identify the regions of existence of microemulsion, nanoemulsion, or coarse emulsion and their compositions, including water, oil, and surfactant (Elnaggar et al., 2009; Shafiq et al., 2007; Kang et al., 2004). A ternary phase diagram can predict the nature, probability, and type of emulsion formed using different compositions of water, oil, and surfactants, as shown in Fig. 15.7:

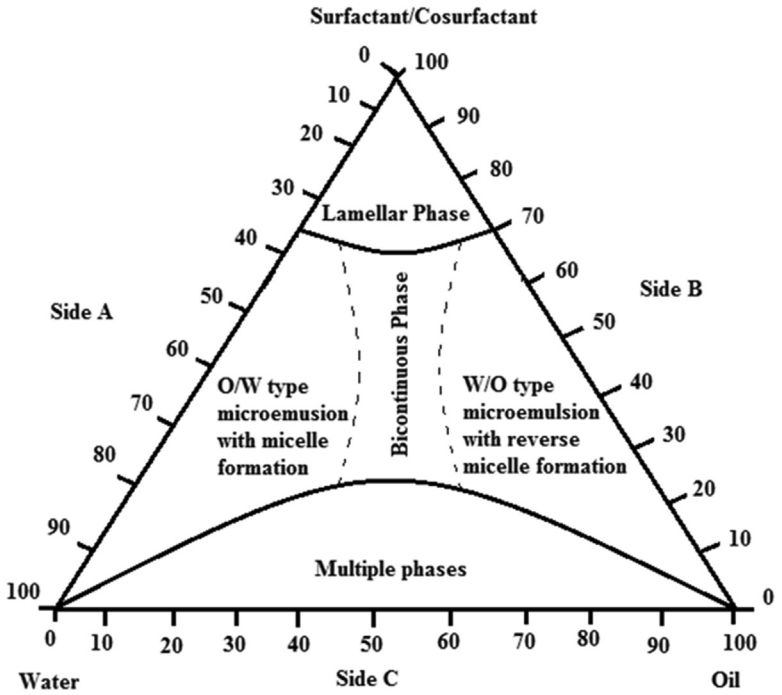


FIGURE 15.7

An imaginary ternary phase diagram showing the different regions with increased probability of a particular type of emulsion. *O/W*, Oil in water; *W/O*, water in oil.

As shown in Fig. 15.7, we will discuss each side of the triangle representing the binary components and their probable generalized interactions with each other to form either type of emulsion.

Side A: Side A of the triangle represents the water/surfactant/cosurfactant binary system. The probability of formation of O/W emulsion is at maximum because in this region the water is in excess with limited oil and surfactant concentration. The surfactant forms micelles with the hydrophilic group arranged toward the outside embedded in water.

Side B: Side B of the pseudoternary phase diagram represents the oil/surfactant binary system. Here, the oil phase will behave as the dispersion medium and maximum probability of formation of W/O emulsion. The surfactant forms reverse micelles with the hydrophobic group arranged toward the outside embedded in oil.

Side C: Side C of the pseudoternary phase diagram represents a water and oil binary system with a reduced amount of surfactant. In this region, multiple phases may exist either as separate components or any type of emulsion with reduced stability.

Upper apex: The upper apex of the triangle at the surfactant/cosurfactant side has low water- and low oil-phase concentration. In contrast, the surfactant/cosurfactant concentration is highest at this point, so this apex will have the surfactant molecules arranged in the form of a lamellar sheet and devoid of any type of emulsion (Lawrence and Rees, 2012).

Bicontinuous phase: The bicontinuous phase is the region where oil and water phases are almost equal in volume with a sufficient amount of surfactant. In this case, a particular emulsion is not formed, but continuous layers of water and oil are present bounded by the surfactant monolayers in between.

As previously mentioned, the pattern of the phase diagram is not applicable for every type of surfactant, cosurfactant, and every composition of different components. Many factors collectively determine which type of emulsion would result from a particular combination of components and further the stability of emulsion formed so far. These factors may include the HLB value, nature of hydrophobic chain and solubility of surfactant alone or in combination with cosurfactants, the pattern of micelle forming, nature and type of oil, and effect of temperature and pressure on the binary system. If these factors do not favor the particular type of emulsion, then phase inversion, phase separation, coalescence, or cracking of emulsion may take place.

So, in short, we cannot predict the behavior of emulsion. Still, in general, we can assume that more probably the component with reduced volume would form the droplets of the dispersed phase. On the other hand, the component with higher volume would form the continuous phase of the binary system provided that the added surfactant must favor the dispersion of the reduced volume component into the higher-volume component.

15.4.1 Phase behavior

Phase behavior studies are an important aspect of surfactant systems because they provide information about the different components' variability, temperature, and structural arrangement of surfactant molecules (Laughlin, 1976). Therefore the phase behavior of any microemulsion can be attributed to many factors, such as properties of lipids used, nature of surfactants and cosurfactants, temperature, pressure, and number of different components. Different types of lipids are available for the preparation of microemulsions, such as glycerides, long-chain fatty acids, medium-chain fatty acids, vegetable oils, and polyalcohols, and they also differ in hydrophobicity and hydrophilicity. Some are highly hydrophobic and have almost zero HLB value, while others may have a mixture of both hydrophobic and hydrophilic groups. Lipids with different hydrophobic and hydrophilic groups are difficult to process to formulate a microemulsion. On the other hand, high molecular weight lipids are very difficult to emulsify because these create problems in penetrating the interfacial surfactant film. Similarly, the choice of surfactant also depends on the type of lipids and type of emulsion , i.e., O/W emulsion or W/O emulsion. The surfactants should lower the interfacial tension between the oil and water interface to

the maximum extent to solubilize them into one phase by partitioning themselves between the two phases. To achieve the desired stability of the microemulsion, a cosurfactant could be added along with the surfactants, which increases the flexibility on the interface resulting in the formulation of nanodroplets of the dispersed phase. Regarding the type of surfactant used in the preparation of a microemulsion, generally nonionic surfactants are preferred over ionic because of reduced irritancy and toxicity and increased stability (Baroli et al., 2000). A surfactant with a lower HLB value favors W/O microemulsion and a surfactant with a high HLB value favors O/W microemulsion. Sometimes the HLB value of a surfactant is too high, so the addition of a cosurfactant is required to adjust the HLB value of the surfactant to the desired level. It should be kept in mind that every combination of different components may not result in a stable microemulsion. So, optimization is required when using different combinations of different lipids, surfactants, and cosurfactants. Even some time addition of the drug to either phase and range of operating temperature may also affect the behavior of the phase of an emulsion such as inversion of the phase from W/O to O/W or vice versa through the change in the interface behavior, i.e., micelle to reverse micelle, lamellar to bicontinuous phase (Fig. 15.8). In this way, depending on the conditions and concentration of surfactant,

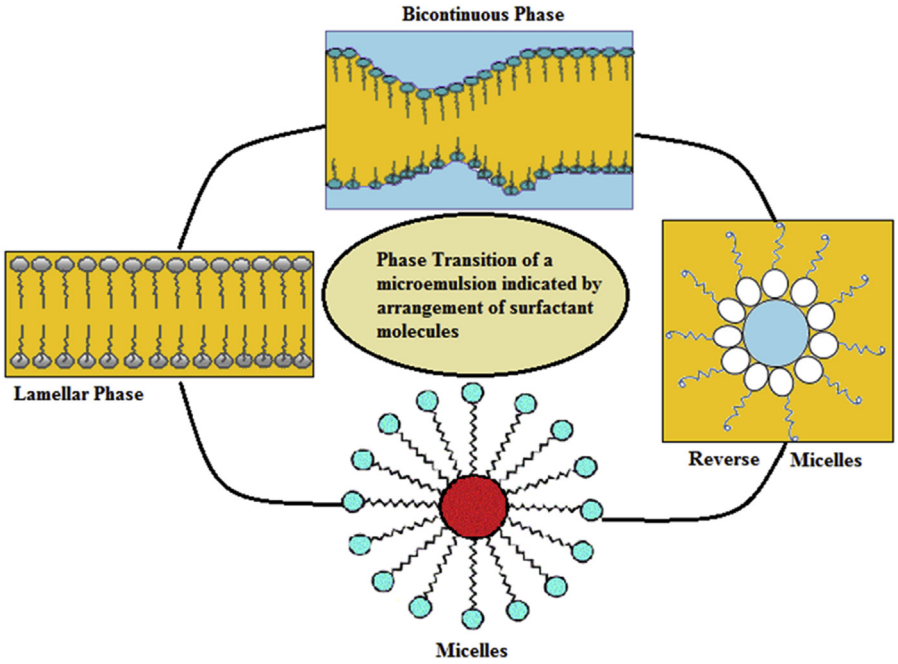


FIGURE 15.8

Representation of phase behavior of a microemulsion.

the arrangement of hydrophilic and hydrophobic groups of the surfactant may vary producing different types of phases, which also affect the nature and stability of an emulsion.

15.4.2 Understanding of the pseudoternary phase diagram

An emulsion is composed of three phases, i.e., water, oil, and surfactant/cosurfactant. Therefore a pseudoternary phase diagram for an emulsion can be represented by a triangle having three corners. Each corner of the triangle represents one component of the emulsion in pure form (100% of the concentration of this component) as shown in Fig. 15.9. Moving away from a corner, the concentration of that component starts decreasing and becomes 0% on the other corner where the concentration of other component is 100%. The sides of the triangle between any two corners represent increasing order of one component and decreasing order of the second component, for example, in Fig. 15.9, on the side of the triangle between component A and component B, by moving from A to B, the magnitude of A continues decreasing from 100% to 0%, and when moving from B to A, the magnitude of B continues decreasing from 100% to 0%. Similarly, the other sides of the triangle can be read for other pairs of components. The points at any location on a triangle side

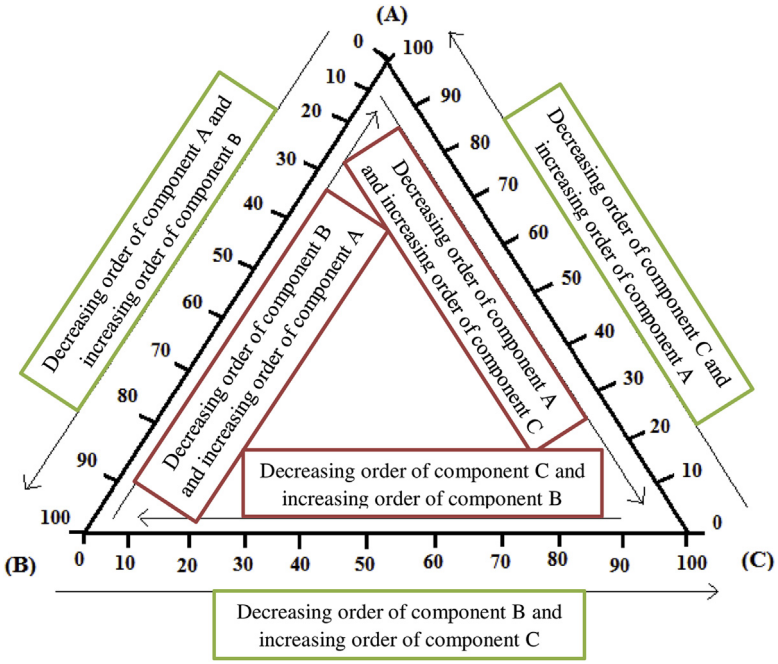


FIGURE 15.9

Understanding of the pseudoternary phase diagram.

give the ratio of component A and component B. Now the desired range of different components such as surfactant/cosurfactant, water, and oil can be easily identified for a stable emulsion by the area under the curve of the triangle.

For a better understanding refer to Fig. 15.7, where it is clearly seen that by changing the concentration of different components on the area under the curve of the triangle, different types of phase appear such as lamellar phase, O/W emulsion, W/O emulsion, bicontinuous phase, and mixture of multiple phases. Therefore from a ternary phase diagram, we can discover the concentration ranges of different components to form a stable emulsion of the desired droplet size. The effect of any additional components such as drug molecules, polymers, and other additives on the phase behavior of an emulsion can also be observed from a phase diagram.

15.4.3 How to plot values on the triangle of the pseudoternary phase diagram

To plot the value or identify the point value on the phase diagram, consider the three components A, B, and C on the triangle, as shown in Fig. 15.10. Each component is divided into an equal percentage fraction (each fraction of 10%) of components shown on the sides of the triangle in decreasing order of magnitude.

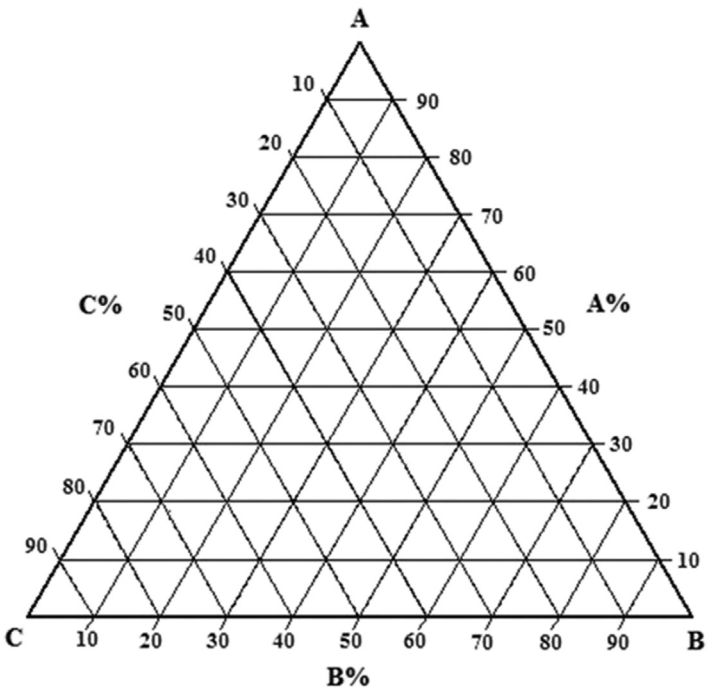


FIGURE 15.10

Marking and division of the triangle in percentage fractions.

Mark the percentage values in decreasing order of magnitude for every point, i.e., for component A, the decreasing order is from point A to B, for component B, it is from B to C, and for component C, it is from C to A. Percentage value of different components for any given point on the triangle is read by moving along the red lines lying parallel to the front side of that component, e.g., to read the value for component A, the red lines parallel to the CB side will be considered and the values on the AB side will give the value of A, for component B, the red lines parallel to AC will be read, and the values on the CB side will give the value of B, and for component C, the red lines parallel to AB will be read and the values on the AC side will give the value of C as shown by the red lines in Fig. 15.11.

Therefore for the combined form of the triangles in Fig. 15.11, look at Fig. 15.12, where an unknown point X (shown as a red dot) is given for which the percentage fractions of each component, i.e., A, B and C, are to be found. The value for component A is 30%, component B is 20%, and component C is 50%. The direction of reading the value for each component is indicated by the redlines from the unknown point X. So in this way, anyone can read a phase diagram or prepare it for emulsions by plotting the experimental values on this triangle to find out the suitable range of components for a stable emulsion.

15.4.4 Preparation of the pseudoternary phase diagram

Preparing a pseudoternary phase diagram is a very critical and time-consuming step where the first experiment is set up. Then, the data obtained from this experiment are reported on the phase diagram. Some important points are mentioned next for the preparation of a pseudoternary phase diagram (Siriporn, 2017).

15.4.4.1 Preparation of surfactant mix (S_{mix})

First, prepare the surfactant system required for the emulsion by mixing two or more surfactants in a predefined ratio, e.g., mixing of surfactant A and surfactant B in weight ratios of 1:1, 1:2, and 2:1. This ratio can be varied as per the need and type of surfactants used. Quantifying the different surfactants helps to identify and optimize the right surfactant in the right amount required for the preparation of stable emulsion.

15.4.4.2 Mixing of surfactant (S_{mix}) and oil in a defined ratio

Now, S_{mix} and the given oil are mixed in a fixed weight proportion such as 0:1, 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1, and 1:0 or in other ratios as per the requirement and types of oil used. This step helps in selecting the optimized ratio of surfactant mix to emulsify the given oil in a better way.

15.4.4.3 Determination of equilibrium point

The above-prepared ratios of surfactant and oil are titrated with distilled water until the endpoint has the appearance of a clear or milky solution. The volume of water consumed in titration and composition of all components converted in weight percent are noted.

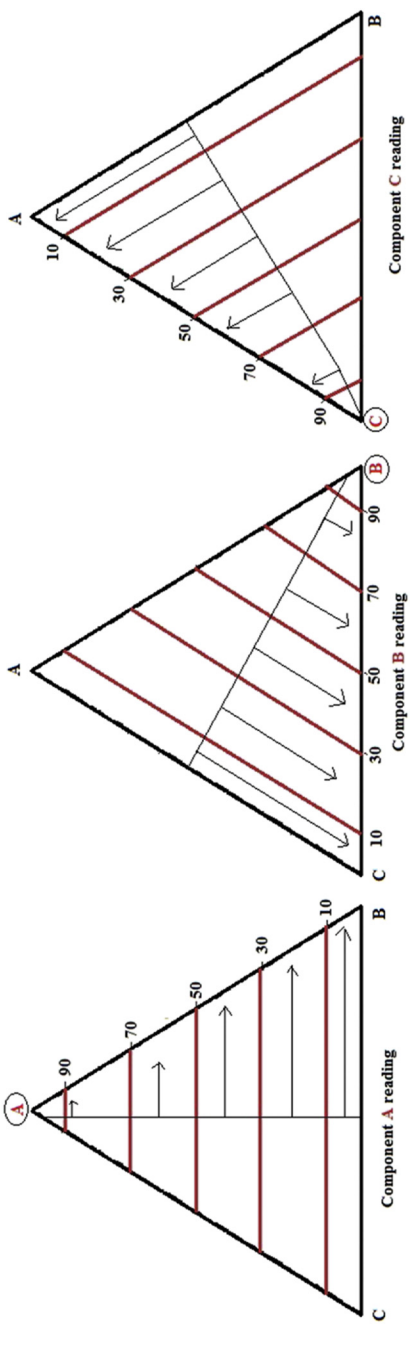


FIGURE 15.11

Understanding of way of reading the three components on a triangle.

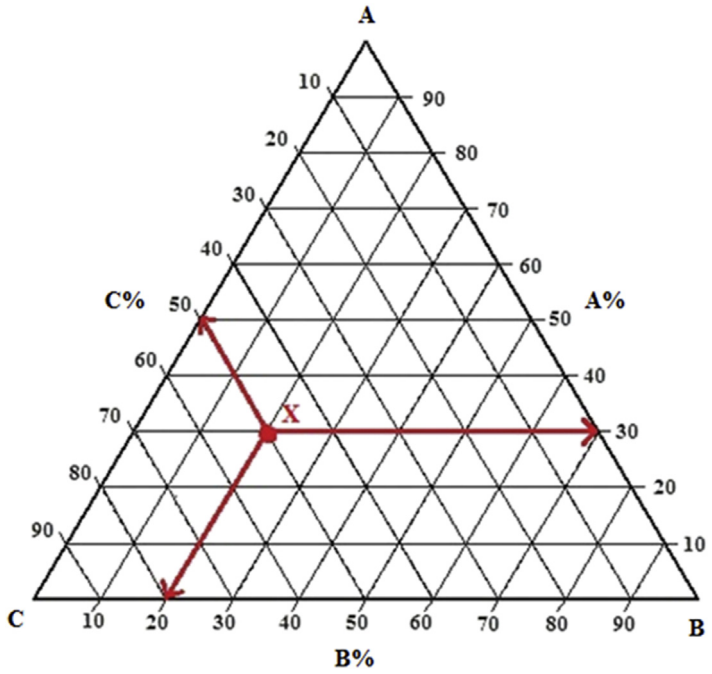


FIGURE 15.12

Spotting of a point on the triangle to find the percentage fraction of each component.

15.4.4.4 Preparation of ternary phase diagrams

Now the weight percent values of oil, water, and surfactant at an equilibrium point for each ratio of S_{mix} and oil are plotted on the ternary phase diagram. The phase boundary produced by plotting in such way can differentiate between single-phase region and two-phase region.

15.5 Software used for the preparation of pseudoternary phase diagrams

Advancement in science and technology makes our work very easy and less time consuming. So, preparing and reading a pseudoternary phase diagram using a computer program or software is now effortless. There are different software programs available either as freeware or with a subscription, which are very useful in the preparation of pseudoternary phase diagrams for emulsions. The basic principle or working of all software is quite similar, but they differ only in user interface modules. Therefore we will study only a few of them to understand the working and methodology of the software. Some of them are discussed next:

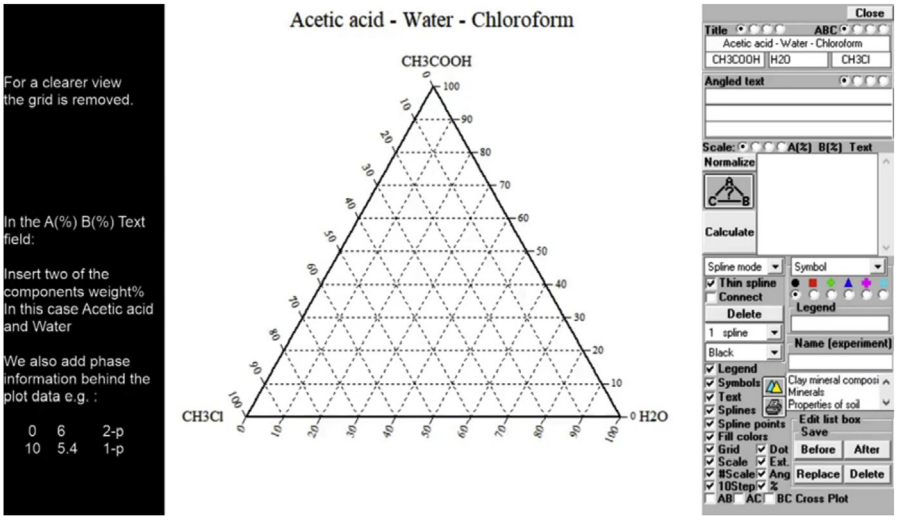
1. Chemix School
2. Design Expert

3. Minitab
4. SPSS
5. Sigma plot
6. Graph pad
7. Matlab
8. XL Stat
9. Delta plotware
10. Triplot
11. R
12. Lab Plot
13. ProSim

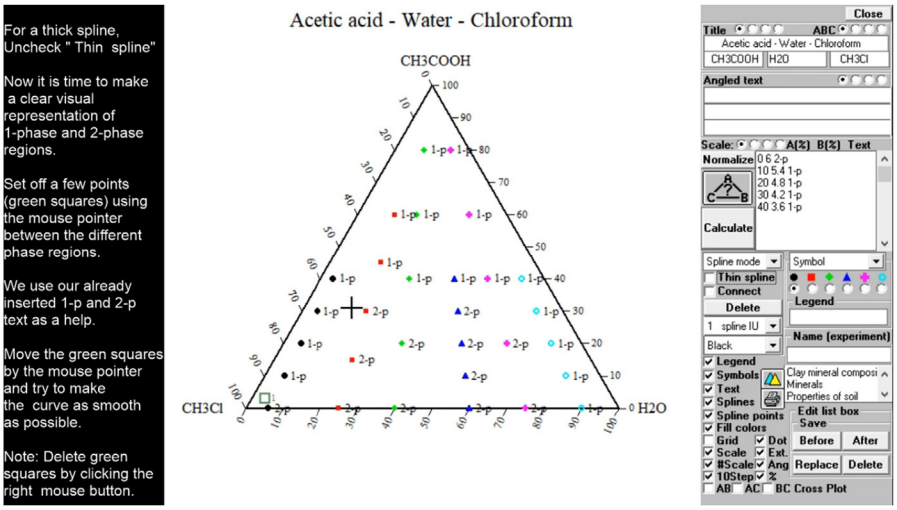
15.5.1 Chemix School

Chemix School is an interactive educational program covering a wide range of topics in the area of chemistry. It is utilized for calculation purposes in classrooms and laboratories for different types of studies, including pseudoternary phase diagrams for emulsions. It helps students to perform the repeated process of hit, trial, and error in an easy-going way to avoid manual calculation (Arne; Arne). This computer program is very useful for undergraduates, postgraduates, and research scholars in solving different problems. The procedure for the preparation of a phase diagram for an acetic acid/water/chloroform system is given next.

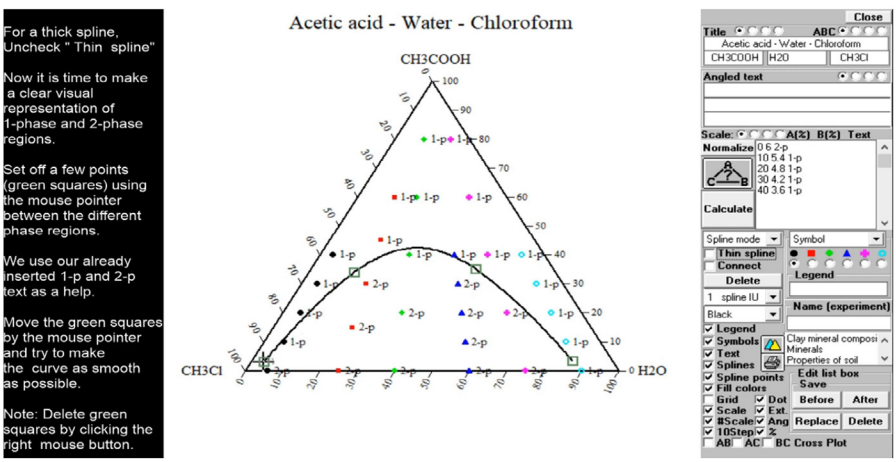
1. Specify the name/title of the ternary system that you are going to prepare, e.g., acetic acid/water/chloroform system.



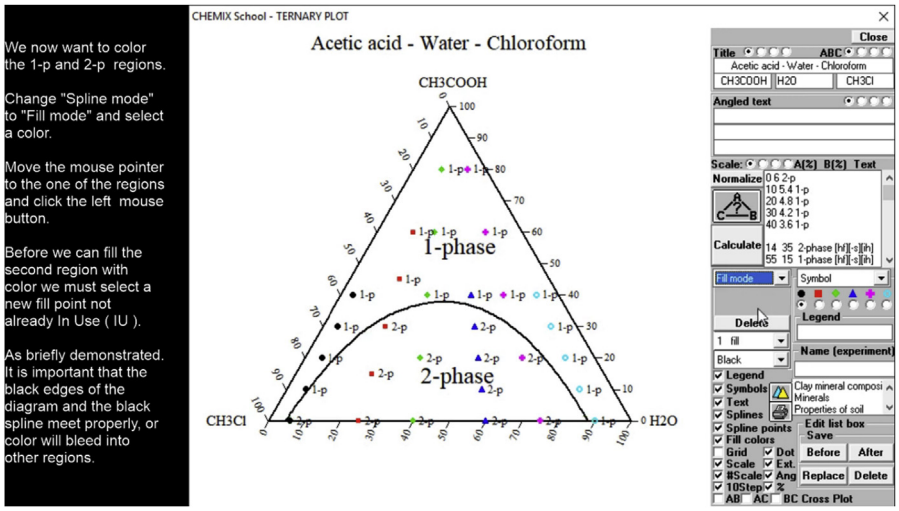
2. Add the weight percent of two components such as acetic acid and chloroform as optimized in the formula by putting the annotation 1-p for phase-1 and 2-p for phase-2 for all observations and click on the calculate button. The software will calculate the weight percent of the third component (water). As to the total 100% sum of three components, the weight percent of two are known, so for the third it will be calculated automatically by subtracting the sum of these two components from 100.



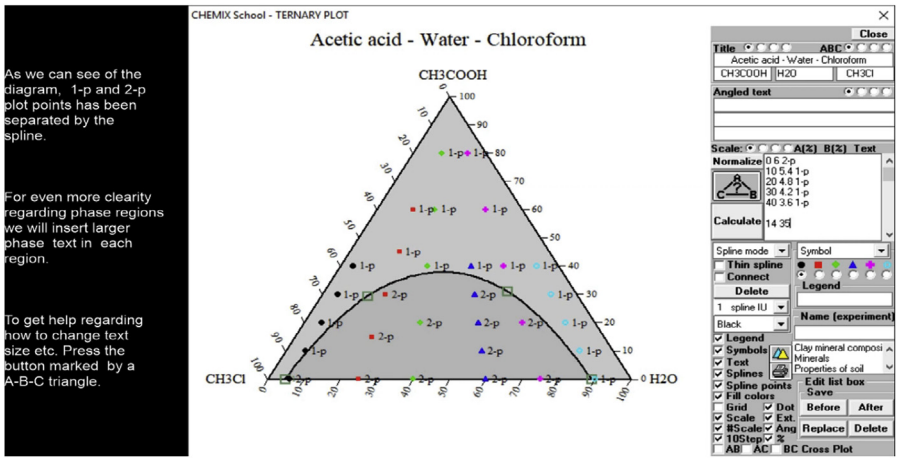
3. Set the boundary points between the 1-p and 2-p system by pointing the cursor and clicking the mouse at every point to separate the two phases.



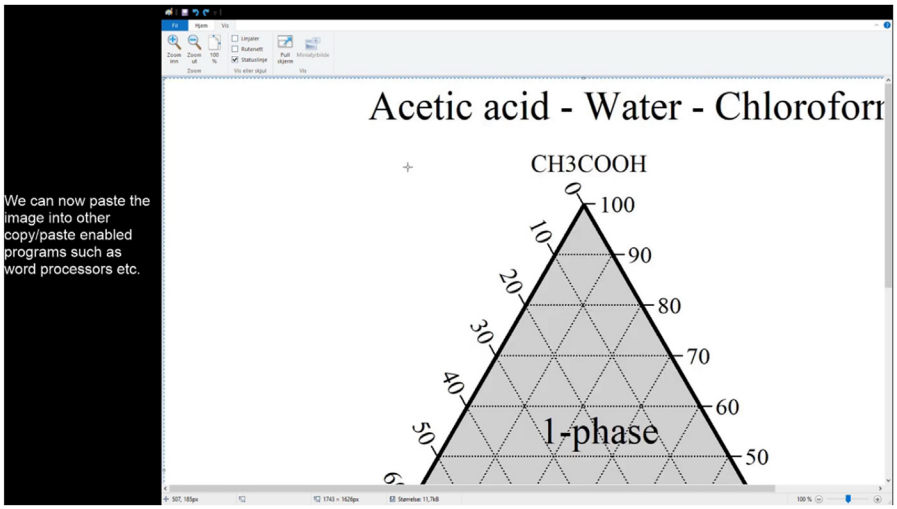
4. Now change the graph from "spline mode" to "fill mode" to visualize the two phases on the graph.



5. You can color the fill point region with different colors to distinguish between the two phases.



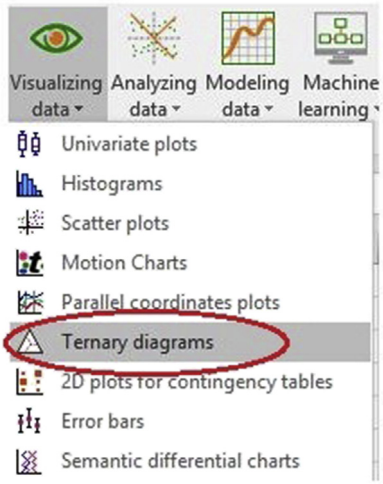
6. Now you can save the prepared phase diagram in high quality and convert to a suitable format such as JPG, JPEG, PNG, or PDF.



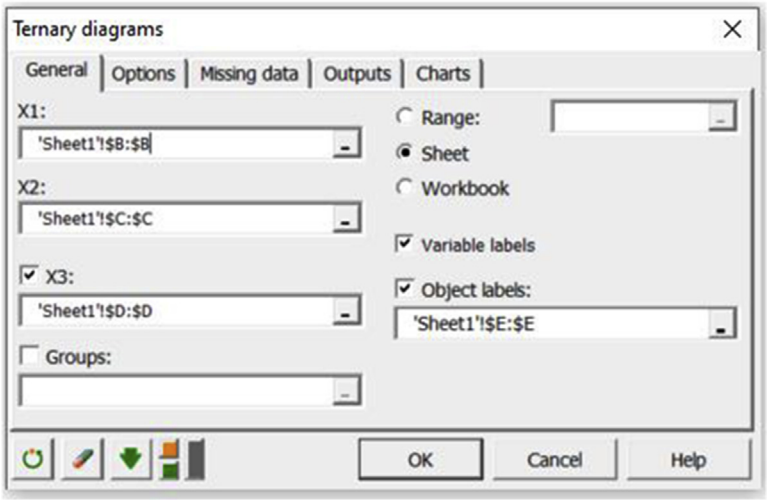
15.5.2 XL stat

Like Chemix School software, XL Stat is also a very useful tool to create a ternary phase diagram of three components. The required steps for this are discussed in a step-by-step manner next (Ternary diagram in Excel tutorial, XLSTAT Support Center):

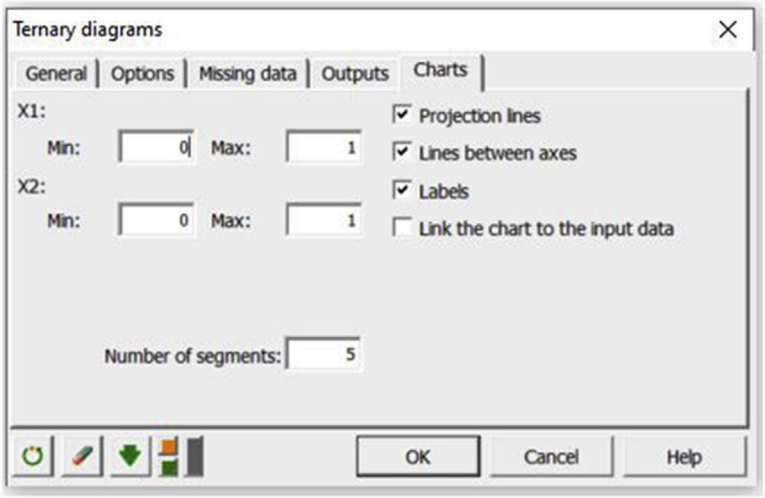
1. **Preparation of dataset:** The dataset for the given three component ternary systems is prepared in an Excel sheet before preparing the phase diagram. Any one set of data contains three values of weight percent, which represent the coordinates of three components in space. However, there is one requirement for this system: the sum of values of the three components will always be 100%.
2. **Preparation of ternary diagram:** Once you are ready with the data, open the XL Stat and under the Visualizing Data tab select Ternary Diagrams as shown in the red circle of the following image.



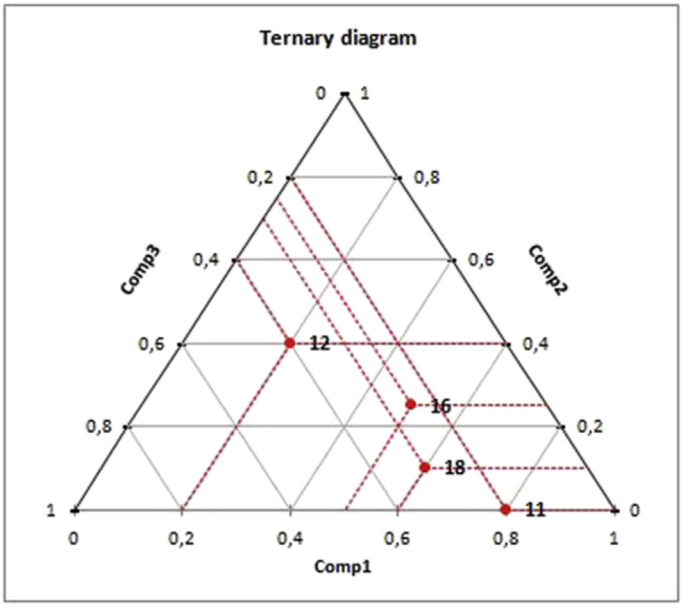
3. Once you click on Ternary Diagram, a new window will appear as shown in the following image. Under the General tab of this window enter the value of three components by selecting the data from the already prepared Excel sheet.



4. Under the Chart tab, you can customize the final chart that will be prepared in the next step as shown in the following image. Set the minimum values to 0 and maximum values to 1. Also, press the tick marks for the desired parameter that you need to be displayed on the final chart and click OK.



5. **Preparation of the phase diagram:** When you click OK after entering all the necessary information in the system, the ternary phase diagram will appear on the screen as shown in the following image:



So. in this manner, you can prepare a ternary phase diagram for any ternary system by using any dataset.

15.6 Conclusion

This chapter summarized the different classifications of emulsions, surfactants, HLB value of surfactant, and phase behavior of an emulsion based on the composition and concentrations of different components of an emulsion. The ratio of oil, water, and surfactant can be easily identified using pseudoternary phase diagrams to form a thermodynamically stable emulsion. A pseudoternary phase diagram suggests that the effectiveness of any surfactant can be optimized either by changing the concentration of the surfactant or by adding a secondary surfactant, i.e., cosurfactant, to the primary one. The solubility of drugs in the oil phase and aqueous phase and drug release pattern can be improved based on the data obtained from the phase diagram. In the later section of the chapter, the ternary phase diagram was explained in detail to enable readers to understand it easily. Plotting of the datasets on a phase triangle and reading of the phase diagram were also explained. With the advancement in science and technology, the application of a computer also extended the calculation and preparation of phase diagrams, which further make the process easy. Therefore the application and use of different computer software programs in the preparation of phase diagrams, e.g., Chemix School and XL Stat, were included and discussed for a better understanding of the method to prepare a ternary phase diagram.

References

Agarwal, S.P., Rajesh, K., 2007. Physical Pharmacy. CBS Publisher, Delhi, India, pp. 177–186.

Ahmad, J., Kohli, K., Mir, S.R., Amin, S., 2011. Formulation of self-nanoemulsifying drug delivery system for telmisartan with improved dissolution and oral bioavailability. J. Dispersion Sci. Technol. 32 (7), 958–968.

Ahmad, J., Amin, S., Kohli, K., Mir, S.R., 2013. Construction of pseudoternary phase diagram and its evaluation: development of self-dispersible oral formulation. Int. J. Drug Dev. Res. 5 (2), 84–90.

Amashita, Y.Y., Miyahara, R., Sakamoto, K., 2017. In: Sakamoto, K., Lochhead, R.Y., Maibach, H.I., Yamashita, Y. (Eds.), Emulsion and Emulsification Technology. Cosmetic Science and Technology Elsevier, pp. p489–506.

Arne, S. Chemix School Software. Microcontroller Programming; Available from: file:///D:/softwares/Stats/Chemix%20School/CHEMIX%20School%208_00/chemix-school-user-manual.html.

Arne, S. Chemix Tutorial. Microcontroller Programming; Available from: https://www.youtube.com/watch?v=YbOmiNclXLE.

Bari, V.D., Sullo, A., Norton, J., Norton, I., 2019. Material properties of cocoa butter emulsions: effect of dispersed phase droplet size and compression speed on mechanical response. Colloids. Surf. Physicochem. Eng. Asp 575, 292−298.

Baroli, B., López-Quintela, M.A., Delgado-Charro, M.B., Fadda, A.M., Blanco-Méndez, J., 2000. Microemulsions for topical delivery of 8-methoxsalen. J. Contr. Release 69 (1), 209−218.

Binks, B.P., 1998. Chapter 1: emulsions—recent advances in understanding. In: Modern Aspects of Emulsion Science, pp. 1−55.

Bobra, M., 1991. Water-in-oil emulsification: a physicochemical study. Int. Oil. Spill. Conf. Proc 1, 483−488.

Bokhout, B.A., van, Gaalen, C., van, der, H., 1981. A selected water-in-oil emulsion: composition and usefulness as an immunological adjuvant. Vet. Immunol. Immunopathol. 2 (5), 491−500.

Chandra, A., Sharma, P.K., Irchhiaya, R., 2014. Microemulsion-based hydrogel formulation for transdermal delivery of dexamethasone. Asian J. Pharm. 3 (1), 30−36.

Das, S., Lee, S.H., Chia, V.D., Chow, P.S., Macbeath, C., Liu, Y., Shlieout, G., 2020. Development of microemulsion based topical ivermectin formulations: pre-formulation and formulation studies. Coll. Surf. B Biointer. 189, 110823.

Dickinson, E., 1994. Emulsion stability. In: Food Hydrocolloids. Springer, Boston, MA, pp. 387−398.

Elnaggar, Y.S.R., El-Massik, M.A., Abdallah, O.Y., 2009. Self-nanoemulsifying drug delivery systems of tamoxifen citrate: design and optimization. Int. J. Pharm. 380 (1−2), 133−141.

Fatima, M., Sheraz, M.A., Ahmed, S., Kazi, S.H., Ahmad, I., 2014. Emulsion separation, classification and stability assessment. RADS J. Pharm. Pharm. Sci 2 (2), 56−62.

Fox, C., 1986. Rationale for the selection of emulsifying agents. Cosmet. Toilet. 101 (11), 25−44.

Fryd, M.M., Mason, T.G., 2012. Advanced nanoemulsions. Annu. Rev. Phys. Chem. 63 (1), 493−518.

Fuhrman, L.C., 2006. Ansel's pharmaceutical dosage forms and drug delivery systems, eighth ed. Am. J. Pharmaceut. Educ. 70 (3), 71.

Gadhave, A., 2014. Determination of hydrophilic-lipophilic balance value. Int. J. Sci. Res. 3 (4), 573−575.

Ganguli, D., Ganguli, M., 2003. Emulsions: a general introduction. In: Inorganic Particle Synthesis via Macro and Microemulsions. Springer, Boston, MA, pp. 1−19.

Goodarzi, F., Zendehboudi, S., 2019. A comprehensive review on emulsions and emulsion stability in chemical and energy industries. Can. J. Chem. Eng. 97 (1), 281−309.

Griffin, W.C., 1949. Classification of surface-active agents by HLB. J. Soc. Cosmet. Chem. 1 (5), 311−326.

Hazlett, R.D., Schechter, R.S., 1988. Stability of macroemulsions. Colloids Surf., A 29 (1), 53−69.

Iqbal, S., Chen, X.D., Kirk, T.V., Huang, H., 2020. Controlling the rheological properties of W1/O/W2 multiple emulsions using osmotic swelling: impact of WPI-pectin gelation in the internal and external aqueous phases. Coll. Surf. B Biointerf. 185, 110629.

Ji, J., Chao, D., Liu, X., Qin, J., 2020. Fabrication of porous polyimide hollow microspheres through O/W/O multiple emulsion. Coll. Surf. Physicochem. Eng. Asp 591, 124537.

Jiang, T., Liao, W., Charcosset, C., 2020. Recent advances in encapsulation of curcumin in nanoemulsions: a review of encapsulation technologies, bioaccessibility and applications. Food Res. Int. 132, 109035.

Kang, B.K., Lee, J.S., Chon, S.K., Jeong, S.Y., Yuk, S.H., Khang, G., et al., 2004. Development of self-microemulsifying drug delivery systems (SMEDDS) for oral bioavailability enhancement of simvastatin in beagle dogs. Int. J. Pharm. 274 (1−2), 65−73.

Kempin, M.V., Kraume, M., Drews, A., 2020. W/O Pickering emulsion preparation using a batch rotor-stator mixer − influence on rheology, drop size distribution and filtration behavior. J. Colloid Interface Sci. 573, 135−149.

Khan, A.Y., Talegaonkar, S., Iqbal, Z., Ahmed, F.J., Khar, R.K., 2006. Multiple emulsions: an overview. Curr. Drug Deliv. 3 (4), 429−443.

Kommuru, T.R., Gurley, B., Khan, M.A., Reddy, I.K., 2001. Self-emulsifying drug delivery systems (SEDDS) of coenzyme Q10: formulation development and bioavailability assessment. Int. J. Pharm. 212 (2), 233−246.

Kumar, R., Kumar, S., Sinha, V.R., 2016. Evaluation and optimization of water-in-oil microemulsion using ternary phase diagram and central composite design. J. Dispersion Sci. Technol. 37 (2), 166−172.

Kumar, P., Mittal, K.L., 1999. Handbook of Microemulsion Science and Technology. CRC Press, p. 872.

Laughlin, R.G., 1976. The aqueous phase behavior of surfactants. In: The Aqueous Phase Behavior of Surfactants. Academic Press Inc, London, p. 257.

Lawrence, M.J., Rees, G.D., 2000. Microemulsion-based media as novel drug delivery systems. Adv. Drug Deliv. Rev. 45 (1), 89−121.

Lawrence, M.J., Rees, G.D., 2012. Microemulsion-based media as novel drug delivery systems. Adv. Drug Deliv. Rev. 64, 175−193.

Liu, H., Ni, Y., Wang, F., Yin, G., Hong, J., Ma, Q., Xu, Z., 2004. Fabrication of submicron Cu2O hollow spheres in an O/W/O multiple emulsions. Coll. Surf. Physicochem. Eng. Asp. 235 (1−3), 79−82.

Matsumoto, S., Kita, Y., Yonezawa, D., 1976. An attempt at preparing water-in-oil-in-water multiple-phase emulsions. J. Colloid Interface Sci. 57 (2), 353−361.

Mehta, R.M., 2002. Dispensing Pharmacy, first ed. Vallabh Prakashan, Delhi, pp. 211−212 (reprint), Published by.

Nazzal, S., Smalyukh, I.I., Lavrentovich, O.D., Khan, M.A., 2002. Preparation and in vitro characterization of a eutectic based semisolid self-nanoemulsified drug delivery system (SNEDDS) of ubiquinone: mechanism and progress of emulsion formation. Int. J. Pharm. 235 (1−2), 247−265.

Opawale, F.O., Burgess, D.J., 1998. Influence of interfacial properties of lipophilic surfactants on water-in-oil emulsion stability. J. Colloid Interface Sci. 197 (1), 142−150.

Paulo, B.B., Alvim, I.D., Reineccius, G., Prata, A.S., 2020. Performance of oil-in-water emulsions stabilized by different types of surface-active components. Coll. Surf. B Biointerf. 190, 110939.

Rao, S.V.R., Shao, J., 2008. Self-nanoemulsifying drug delivery systems (SNEDDS) for oral delivery of protein drugs: I. Formulation development. Int. J. Pharm. 362 (1−2), 2−9.

Richard, S., Hiroshi, I., Cor, P., 2013. Introduction to Supercritical Fluids, first ed., vol. 4. Elsevier Science, p. 752.

Rieger, M.M., 1987. Emulsions. In: Lachman, L., Lieberman, H.A., Kanig, J.L. (Eds.), The Theory and Practice of Industrial Pharmacy, third ed. Varghese Publishing House, pp. 502−533. India Edition.

Ruckenstein, E., 1999. Thermodynamic insights on macroemulsion stability. Adv. Colloid Interface Sci. 79 (1), 59−76.

Sabale, V., Vora, S., 2012. Formulation and evaluation of microemulsion-based hydrogel for topical delivery. Int. J. Pharm. Investig 2 (3), 140–149.

Shafiq, S., Shakeel, F., Talegaonkar, S., Ahmad, F.J., Khar, R.K., Ali, M., 2007. Development and bioavailability assessment of ramipril nanoemulsion formulation. Eur. J. Pharm. Biopharm. 66 (2), 227–243.

Shah, P., Bhalodia, D., Shelat, P., 2010. Nanoemulsion: a pharmaceutical review. Sys. Rev. Pharm. 1, 24–32.

Sharma, M.K., Shah, D.O., 1985. Introduction to macro- and microemulsions. In: Shah, D.O. (Ed.), Macro- and Microemulsions. American Chemical Society, Washington, DC, pp. 1–18.

Sharma, S., Shukla, P., Misra, A., Misra, P.R., 2014. Interfacial and Colloidal Properties of Emulsified Systems: Pharmaceutical and Biological Perspective A2—Ohshima, Hiroyuki. Colloid and Interface Science in Pharmaceutical Research and Development. Makino K. Elsevier, Amsterdam, pp. 149–172.

Silva, B.F.B., Rodríguez-Abreu, C., Vilanova, N., 2016. Recent advances in multiple emulsions and their application as templates. Curr. Opin. Colloid Inter. Sci. 25, 98–108.

Singh, Y., Meher, J.G., Raval, K., Khan, F.A., Chaurasia, M., Jain, N.K., Chourasia, M.K., 2017. Nanoemulsion: concepts, development and applications in drug delivery. J. Contr. Release 252, 28–49.

Siriporn, O., 2017. Pseudoternary Phase Diagram Construction. protocols.Io. https://doi.org/10.17504/protocols.io.jyccpsw.

Sobrinho, H.B.S., Luna, J.M., Rufino, R.D., Porto, A.L.F., Sarubbo, L.A., 2013. Biosurfactants: classification, properties and environmental applications. In: Govil, J.N. (Ed.), Recent Developments in Biotechnology, first ed., vol. 11. Studium Press LLC, Houston, TX, USA, pp. 1–29.

Soriano-Ruiz, J.L., Suñer-Carbó, J., Calpena-Campmany, A.C., et al., 2019. Clotrimazole multiple W/O/W emulsion as anticandidal agent: characterization and evaluation on skin and mucosae. Coll. Surf. B Biointerf. 175, 166–174.

Stepisnik, P.T., Zupanc, M., Dular, M., 2019. Revision of the mechanisms behind oil-water (O/W) emulsion preparation by ultrasound and cavitation. Ultrason. Sonochem. 51, 298–304.

Ternary diagram in Excel tutorial | XLSTAT support center. Available from: https://help.xlstat.com/s/article/ternary-diagram-in-excel-tutorial?language=en_US.

Vijayakumar, S., Saravanan, V., 2015. Biosurfactants-types, sources and applications. Res. J. Microbiol. 10 (5), 181–192.

Wang, L., Li, Y., Xiang, D., Zhang, W., Bai, X., 2020. Stability of lutein in O/W emulsion prepared using xanthan and propylene glycol alginate. Int. J. Biol. Macromol. 152, 371–379.

Winfield, A.J., Richards, R.M.E., 2004. Pharmaceutical Practice, third ed. Churchill Livingstone Publisher, London, pp. 199–202.

Yamashita, Y., Sakamoto, K., 2016. Hydrophilic–lipophilic balance (HLB): classical indexation and novel indexation of surfactant. In: Encyclopedia of Biocolloid and Biointerface Science 2V Set. John Wiley & Sons, Ltd, pp. 570–574.

Zhang, P., Liu, Y., Feng, N., Xu, J., 2008. Preparation and evaluation of self-microemulsifying drug delivery system of oridonin. Int. J. Pharm. 355 (1–2), 269–276.

Index

Note: 'Page numbers followed by "f" indicate figures and "t" indicate tables.'

C

H

I

T

U