Methods in
Molecular Biology 1216

Springer Protocols

Valentin Köhler *Editor*

Protein Design

Methods and Applications

Second Edition

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Protein Design

Methods and Applications

Second Edition

Edited by

Valentin Köhler

Department of Chemistry, University of Basel, Switzerland

Humana Press

Editor
Valentin Köhler
Department of Chemistry
University of Basel
Switzerland

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-4939-1485-2 ISBN 978-1-4939-1486-9 (eBook)
DOI 10.1007/978-1-4939-1486-9
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014947803

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The second edition of protein design in the Methods in Molecular Biology series aims at providing the reader with practical guidance and general ideas on how to approach a potential protein design project. Considering the complexity of the subject and its attention in the scientific community it is apparent that only a selection of subjects, approaches, methods, studies, and ideas can be presented.

The design of well-folded peptide structures and the redesign of existing proteins serve multiple purposes from potentially unlimited and only just developing applications in medicine, material science, catalysis, the realization of systems chemistry, and synthetic biology to a deeper understanding of molecular evolution.

The book is roughly organized in increasing complexity of the systems studied. Additional emphasis is put on metals as structure-forming elements and functional sites of proteins towards the end.

A computational algorithm for the design of stable alpha helices is discussed in the first chapter and is accessible in the form of a web-based tool. An extensive review on monomeric β-hairpin and β-sheet peptides follows. In the design of these species any tendency to self-assemble has to be carefully considered. In contrast, Chapter 3 exploits just this phenomenon—peptides engineered to self-assemble into fibrils.

Subsequently, some possibilities and aspects resulting from the incorporation of unnatural amino acids are outlined. In the practical methods chapter on the redesign of RNase A, a variable α-helical fragment is reassembled with the remainder of the protein structure, generated by enzymatic cleavage. Chapter 5 discusses the design and characterization of fluorinated proteins, which are entirely synthetic. Comparisons to non-fluorinated analogous structures are included and practical advice is offered.

This is followed by an overview of considerations for the generation of binary-patterned protein libraries leading on to library-scale computational protein design for the engineering of improved protein variants. The latter is exemplified for cellobiohydrolase II and a study aimed at changing the co-substrate specificity of a ketol-acid reductoisomerase. Chapter 8 focuses on the elaboration of symmetric protein folds in an approach termed "top-down symmetric deconstruction," which prepares the folds for subsequent functional design studies.

The identification of a suitable scaffold for design purposes by means of the scaffold search program ScaffoldSelection is the topic of Chapter 9.

The computational design of novel enzymes without cofactor is demonstrated for a Diels-Alderase in Chapter 10.

The final four chapters deal with metal involvement in the designed or redesigned structures, either as structural elements or functional centers. The begin is made with a tutorial review that imparts general knowledge for the design of peptide scaffolds as novel pre-organized ligands for metal-ion coordination and then exemplifies these further in a respective case study. This is followed by an introduction on the computational design of metalloproteins, which encompasses metal incorporation into existing folds, fold design by exploiting symmetry, and fold design in asymmetric scaffolds. The potential power of cofactor exchange is addressed with the focus on a practical protocol for the preparation of apomyoglobin and the incorporation of zinc porphyrin in the penultimate chapter. The book concludes with a case study on the computational redesign of metalloenzymes carried out with the aim to assign a new enzymatic function.

This volume of Methods in Molecular Biology contains a number of practical protocols, but compared to other volumes of the series, a larger contribution of reviews or general introductions is provided. Those, however, are presented in a tutorial fashion to communicate principles that can be applied to individual research projects.

I sincerely do hope that the reader finds this edition of protein design helpful for devising their own experiments.

I warmly thank all the authors for their very valuable contributions, their dedication, and not least their patience.

Basel, Switzerland ***Valentin Köhler***

Contents

Contributors

Georgios Archontis • *Department of Physics, University of Cyprus, Nicosia, Cyprus*

Michael Blaber • *Department of Biomedical Sciences, College of Medicine, Florida State University, Tallahassee, FL, USA*

Luke H. Bradley • *Departments of Anatomy and Neurobiology, Molecular and Cellular Biochemistry, and the Center of Structural Biology, University of Kentucky College of Medicine, Lexington, KY, USA*

Benjamin C. Buer • *Department of Chemistry, University of Michigan, Ann Arbor, MI, USA*

Kaspar Feldmeier • *Max Planck Institute for Developmental Biology, Tübingen, Germany*

Aimee J. Gamble • *School of Chemistry, University of Birmingham, Birmingham, UK*

Maika Genz • *Faculty of Chemistry and Mineralogy, Center for Biotechnology and Biomedicine, Institute of Bioanalytical Chemistry, University of Leipzig, Leipzig, Germany*

Daniela Grabs-Röthlisberger • *Arzeda Corp., Seattle, WA, USA*

Per Jr. Greisen • *Department of Biochemistry, University of Washington, Seattle, WA, USA*

Takashi Hayashi • *Department of Applied Chemistry, Graduate School of Engineering, Osaka University, Suita, Osaka, Japan*

Birte Höcker • *Max Planck Institute for Developmental Biology, Tübingen, Germany*

Thaddaus R. Huber • *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO, USA*

M. Angeles Jiménez • *Consejo Superior de Investigaciones Científicas (CSIC), Instituto de Química Física Rocasolano (IQFR), Madrid, Spain*

Lucas B. Johnson • *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO, USA*

Emmanouil Kasotakis • *Department of Materials Science and Technology, University of Crete, Heraklion, Crete, Greece*

Sagar D. Khare • *Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Rutgers University, Piscataway, NJ, USA*

Liam M. Longo • *Department of Biomedical Sciences, College of Medicine, Florida State University, Tallahassee, FL, USA*

E. Neil G. Marsh • *Department of Chemistry, University of Michigan, Ann Arbor, MI, USA; Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, MI, USA*

Anna Mitraki • *Department of Materials Science and Technology, University of Crete, Heraklion, Crete, Greece; Institute for Electronic Structure and Laser, Foundation for Research and Technology-Hellas (IESL-FORTH), Heraklion, Crete, Greece*

Vikas Nanda • *Department of Biochemistry and Molecular Biology, Center for Advanced Biotechnology and Medicine, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ, USA*

Koji Oohora • *Department of Applied Chemistry, Graduate School of Engineering, Osaka University, Suita, Osaka, Japan*

AVANISH S. PARMAR • *Department of Biochemistry and Molecular Biology, Center for Advanced Biotechnology and Medicine, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ, USA*

ANNA F.A. PEACOCK • *School of Chemistry, University of Birmingham, Birmingham, UK*

MICHAEL PETUKHOV • *Department of Molecular and Radiation Biophysics, Petersburg Nuclear Physics Institute, NRC Kurchatov Institute, Gatchina, Russia; Saint Petersburg State Polytechnical University, Saint Petersburg, Russia*

DOUGLAS PIKE • *Department of Biochemistry and Molecular Biology, Center for Advanced Biotechnology and Medicine, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ, USA*

GEORGY RYCHKOV • *Department of Molecular and Radiation Biophysics, Petersburg Nuclear Physics Institute, NRC Kurchatov Institute, Gatchina, Russia; Saint Petersburg State Polytechnical University, Saint Petersburg, Russia*

MATTHEW D. SMITH • *Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA*

CHRISTOPHER D. SNOW • *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO, USA*

ANDRÉ C. STIEL • *Max Planck Institute for Developmental Biology, Tübingen, Germany*

NORBERT STRÄTER • *Faculty of Chemistry and Mineralogy, Center for Biotechnology and Biomedicine, Institute of Bioanalytical Chemistry, University of Leipzig, Leipzig, Germany*

PHANOURIOS TAMAMIS • *Department of Physics, University of Cyprus, Nicosia, Cyprus*

ALEXANDER YAKIMOV • *Department of Molecular and Radiation Biophysics, Petersburg Nuclear Physics Institute, NRC Kurchatov Institute, Gatchina, Russia; Saint Petersburg State Polytechnical University, Saint Petersburg, Russia*

ALEXANDRE ZANGHELLINI • *Arzeda Corp., Seattle, WA, USA*

Chapter 1

De Novo Design of Stable α-Helices

Alexander Yakimov, Georgy Rychkov, and Michael Petukhov

Abstract

Recent studies have elucidated key principles governing folding and stability of α-helices in short peptides and globular proteins. In this chapter we review briefly those principles and describe a protocol for the de novo design of highly stable α-helixes using the SEQOPT algorithm. This algorithm is based on AGADIR, the statistical mechanical theory for helix-coil transitions in monomeric peptides, and the tunneling algorithm for global sequence optimization.

Key words α-Helix, Stability, Sequence optimization, Solubility

1 Introduction

The α-helix is one of the most abundant elements of protein secondary structure. Numerous studies of α-helical peptides not only contributed to a better understanding of protein folding but also represent an increasing pharmacological interest in their practical utility for the development of novel therapeutics to modulate protein-protein interactions in vivo [1].

A large amount of information on α-helix folding and stability has been gathered since the early 1990s [2, 3]. The data show that sequences of protein helices are not, in general, optimized for high conformational stability. This may be an important factor in preventing the accumulation of nonnative intermediates in protein folding [4–6]. Nevertheless, designing short α-helical peptides and proteins with sufficient conformational stability under given environmental conditions (temperature, pH, and ionic strength) still remains an area of intense investigation in protein engineering [1].

Furthermore a large body of information has been accumulated regarding the factors which govern the stability of α-helices in proteins and the helical behavior of both isolated protein fragments and designed helical sequences in solution [4]. These factors include interactions between amino acid side chains [7–9], the helix macrodipole [10], and terminal capping [11].

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216, DOI 10.1007/978-1-4939-1486-9_1, © Springer Science+Business Media New York 2014

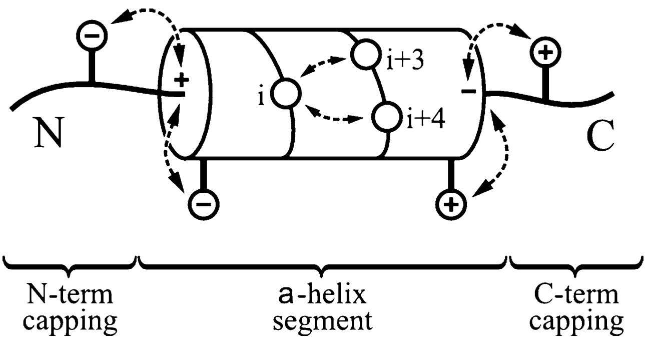


Fig. 1 Schematic view of the physical interactions stabilizing the α-helix segment

All these factors have been considered separately in attempts to increase the conformational stability of α-helices in peptides and in natural proteins [12, 13]. However, the design of peptide sequences with the optimal implementation of all these factors can often not be achieved even for short peptides, since they can be mutually exclusive. The stability of the α-helix is controlled by diverse and accurately balanced interactions. For example a positively charged amino acid at position i prefers that the $i+3$, $i+4$ and also the $i-3$, $i-4$ positions of the helix (Fig. 1) are occupied by negatively charged residues that may on the other hand be unfavorable for helix formation if they occur close to the carboxy-terminus where they lead to negative interactions with the helix macrodipole [10]. The problem increases rapidly with peptide length, since it determines the number of interactions to be considered.

Several de novo protein design methods, based on RosettaDesign [14], EGAD [15], Liang-Grishin [16], and RosettaDesign-SR [17] programs, have been developed during the past decade. These methods can also be applied for the design of α-helix-forming peptides [18]. Unlike these approaches, the AGADIR method is based on free energy contributions, obtained from experimental data.

The number of possible sequences of a peptide with N amino acid residues equals 20^N. Thus, it is computationally impossible to calculate the helical content for a complete permutation library even for short peptides as short as ten amino acids. To overcome this problem we used the tunneling algorithm for global optimization of multidimensional functions [19]. The main advantage of this approach is that it does not require an examination of all possible sequences to find a suitable solution for most practical purposes. The method is simple and robust and requires only the calculation of the first derivatives of the goal function. It has been reported that the method was successfully applied to identify global minima to many problems with many thousands of local minima [19]. However all available global optimization techniques can be described as random walkers which cover to a greater or lesser extent a significant region of phase space spanned by the task at hand. None of them can claim the true globality of a found solution. Besides taking into account imperfectness of theoretical approximations employed to predict helix stability, it is unlikely that the solution for any peptide sequence above a certain length (5–7 amino acids) can be globally optimized currently and in the near future. The inability of theoretical models to guarantee convergence to a globally optimized peptide sequence motivates the development of efficient tools for protein helix optimization, even if the inherent problem itself cannot be overcome. For protein engineering applications sufficiently optimized sequences are employed instead of truly globally optimized ones. Creating and testing such a tool on short peptide helices was the main goal of the work presented in the form of a practical method.

Recently we developed a new method for the design of α-helices in peptides and proteins using AGADIR (located at http://agadir.crg.es/) [20], the statistical mechanical theory for helix-coil transitions in monomeric peptides, and the tunneling algorithm of global optimization of multidimensional functions [19] for optimization of amino acid sequences [5]. Unlike traditional approaches that are often used to increase protein stability by adding a few favorable interactions to the protein structure, this method deals with all possible sequences of protein helices and selects a suitable one. Under certain conditions the method can be a powerful practical tool not only for the design of highly stable peptide helices but also for protein engineering purposes. In the study for the design of peptide helices we used an approach combining statistical mechanical calculations based on the AGADIR model [12] including several of its more recent modifications [21–27] and the global optimization algorithm [19].

In work [5] we used one sequence approximation of the AGADIR model (AGADIR1s) for helix-random coil transitions in monomeric peptides. As any other theoretical model it has its own simplifications and limitations. Most importantly it includes the AGADIR partition function physical interactions only within helical segments and those from a few flanking residues at both N- and C-termini (the so-called N- and C-capping interactions). The SEQOPT sequence optimization is not only applicable for short monomeric peptides in an aqueous environment but also for solvent-exposed parts of protein alpha-helices which show only intrahelical residue interactions. As another important simplification AGADIR1s ignores the possible existence of multiple helical segments in each peptide conformation. Multiple sequence approximation (AGADIRms) of the AGADIR model has also been developed [28] and its predictions of peptide conformational stability were compared with results of AGADIR1s as well as with Zimm-Bragg and Lifson-Roig classic models for helix-coil transition in peptides. It was shown that for all tested peptides having less than

56 residues the helical contents predicted by AGADIR1s are within 0.3 % error with those of AGADIRms. In addition AGADIR1s is computationally much faster.

1.1 α-Helix Structure and Stability

In the mid-1970s it was predicted by Finkelstein and Ptitsyn that short peptides consisting of amino acids with high α-helix propensity should have a fairly stable α-helical conformation in aqueous solution [29–33]. Later this theory has been verified experimentally by examining synthetic peptide sequences of ribonuclease A [34, 35]. The theoretical model developed by Finkelstein and Ptitsyn describes the probability of the formation of α-helices and β-structures and turns in short peptides and globular proteins based on the modified classical Zimm-Bragg model. It takes into account some additional physical interactions, including hydrophobic interactions of a number of amino acid side chains, electrostatic interactions between the charged side chains themselves, as well as the α-helix macrodipole. The computer program (ALB) based on this theoretical model was shown to successfully predict not only an approximate level of the conformational stability of α-helical peptides [2] but also, with a probability of ~65 %, the distribution of secondary structure elements in globular proteins.

Beginning in the late 1980s and increasing in the 1990s, a large number of experiments with amino acid substitutions in short synthetic peptides exploring different interactions in α-helices have been described in the literature [3]. We would like to point out the approach proposed by Scholtz and Baldwin, which enables the accumulation of sufficient experimental data to proceed to a quantitative description of the cooperative mechanisms of conformational transitions of α-helical conformations in peptides with random sequences.

Collected data allowed to establish the principle of intrinsic helical propensity of any amino acid to populate the α-helix formation. This propensity [22] has been attributed to changes of configurational entropy [36] and solvent electrostatic screening of amino acid side chains [37]. For instance methionine, alanine, leucine, uncharged glutamic acid, and Lys have high intrinsic helical propensities, whereas proline and glycine have poor ones. Proline residues either break or kink a helix, both because they cannot provide an amide hydrogen for hydrogen bonding (having no amide hydrogen), and also because its side chain interferes sterically with the backbone of the preceding turn; inside a helix, this forces a bend of about 30° in the helix axis [38]. Nevertheless due to its rigid structure proline is often found to be the first N-terminal residue in protein α-helices [39]. On the other hand glycine also tends to disrupt helices because its high conformational flexibility makes it entropically expensive to adopt the relatively constrained α-helical structure. Nevertheless it often plays a role as N- and C-cap residue of protein helices [40].

The intrinsic helical propensity of the amino acids has often been assumed to be independent of their position within the α-helix because the alpha-helical structure is highly symmetrical [2, 20, 41]. Later it has been shown that intrinsic helical propensities of some amino acids are different in the first and last α-helix turn as compared to central helix positions [25–27]. Additionally there are also side-chain:side-chain interactions in α-helices between residues at positions i and $i+3$ as well as i and $i+4$ interactions of charged or polar residues with the helix macrodipole and capping interactions between the residues flanking the α-helix and the free NH and CO groups at the first or last helical turn (for a review, *see* ref. 22). Furthermore, local motifs involving residues outside the helix that pack against helical residues have been described at both the N terminus (hydrophobic staple [42, 43]) and C terminus (Schellman motif [44, 45]). Several theoretical approaches have been developed to predict helical content of an arbitrary peptide sequence under given environmental conditions [20, 30, 41, 46, 47]. In work [5] we focus on the AGADIR model, which was tested to accurately predict the helical properties of several hundred short peptides in aqueous solution [20–22]. Short peptides do not possess a single stable conformation under typical environmental conditions. The AGADIR model accounts for free energy contributions from all possible helical segments in the peptide under consideration as follows: The difference in free energy between the random-coil and helical states for a given segment ($\Delta G_{\text{helical\_segment}}$) is calculated as the following summation:

$$\Delta G_{helical\_segment} = \Delta G_{int} + \Delta G_{hb} + \Delta G_{sc} + \Delta G_{el} + \Delta G_{nonH} + \Delta G_{macrodipole}$$

where ΔG_{int} is the summation of the intrinsic propensities of all residues in a given helical segment including its observed positional dependencies [25–27]; ΔG_{hb} is the sum of the main-chain:main-chain enthalpic contributions, which include the formation of i, $i+4$ hydrogen bonds; ΔG_{sc} sums the net contributions, with respect to the random-coil state, of all side-chain:side-chain interactions located at positions i, $i+3$ and i, $i+4$ in the helical region; ΔG_{el} includes all electrostatic interactions between two charged residues inside and outside the helical segment; ΔG_{nonH} represents the sum of all contributions to helix stability of a given segment from residues that are not in a helical conformation (N- and C-capping, Capping Box, hydrophobic staple motif, Schellman motif, etc.); and $\Delta G_{\text{macrodipole}}$ represents the interaction of charged groups with the helix macrodipole. All the free energy contributions are included with their respective dependencies on temperature, pH, and ionic strength as described in reference [21]. In the AGADIR model the helix content (HC) of a peptide under consideration is calculated as

$$HC = \frac{\sum e^{-\frac{\Delta G_{helical\_segment}}{RT}}}{1 + \sum e^{-\frac{\Delta G_{helical\_segment}}{RT}}}$$

where the sum includes all possible α-helical segments. In addition to the original AGADIR set of energy parameters [22] we incorporated several modifications of the parameter set of the theory published later [23, 24].

1.2 α-Helices with Optimized Sequences

Properties of peptides with optimized sequences were tested both theoretically and experimentally [5]. Despite the assignment of the highest α-helical propensity for Ala, only very few optimized sequences of short peptides contained this residue. Also the number of identified central salt bridges in the optimized sequences was quite low. The cause is probably associated with the influence of terminal positions in these peptides. It seems that hydrophobic residues (Leu) at central positions are more "tolerant" to the terminal requirements for accommodation of both positive charges from amino-termini and negative charges from carboxy-termini. Generally, the longer a peptide, the more complicated and difficult to rationalize are the patterns of sequential motifs that are found at the top of the list of the best peptide sequences.

The most stable peptide helices mainly consist of a few amino acid types (Leu, Met, Trp, Tyr, Glu, and Arg) having both high intrinsic helical propensities and high potential for other stabilizing interactions such as side-chain:side-chain interactions and N- and C-capping interactions. It is of interest that top positions of the peptide series are occupied by poly-Leu and poly-Trp motifs indicating that an accumulation of favorable hydrophobic side-chain:side-chain interactions can fully compensate for the loss of other helix-stabilizing factors such as beneficial N- and C-capping motifs and electrostatic interactions with the helix macrodipole and between the side chains. Certainly these homopolymeric motifs are not really useful due to their very low solubility. However, there are many soluble sequences that are only a little less stable than the homopolymer sequences. These sequences often have a few common motifs such as the "Capping Box", wherein side chains of the first (Thr) and the fourth (Glu) residue form a specific pattern of hydrogen bonding, with the amide protons of the main chain stabilizing the α-helix [23, 48] and where C-terminal positions are often occupied by positively charged amino acids that can stabilize an α-helix by charge–helix macrodipole interactions.

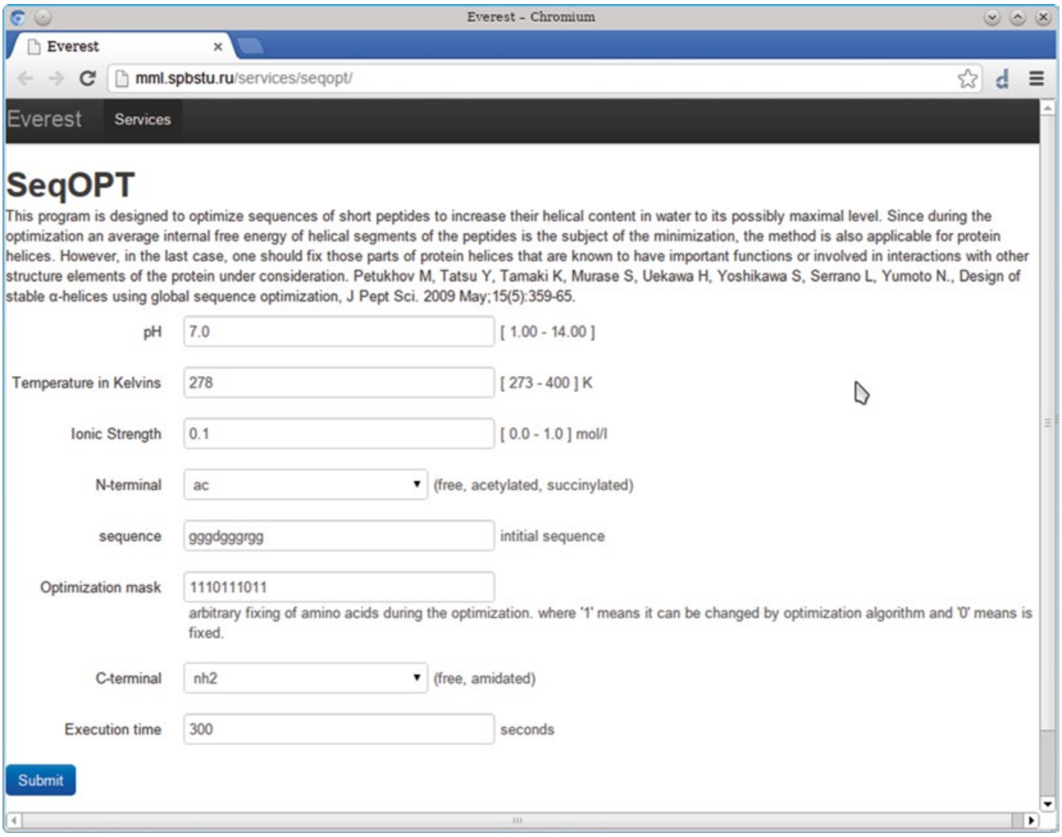
One of the important features of the proposed method is the ability to arbitrarily fix any functional segments of primary structure and to optimize just the nonfunctional elements. The usefulness of this feature can for instance be easily illustrated for the case of helix optimization in globular proteins with the aim of increasing their thermostability. In this case, only solvent-exposed amino acid positions of protein α-helices having local intrahelical contacts should be allowed to vary during the course of sequence optimization. These positions should be carefully selected based on the analysis of the protein 3D structure. All other amino acid positions of the helix should be fixed to their native sequence to preserve important tertiary interactions in the protein native structure.

2 Methods

2.1 SEQOPT Algorithm

The SEQOPT algorithm is based on the tunneling algorithm [19] of global optimization calculations. SEQOPT comprises two main phases: a local minimization phase and a tunneling phase. During the minimization phase, the target function of peptide helicity is minimized by the conjugate gradient method as implemented in the Fletcher–Reeves method [49]. During the tunneling phase, the algorithm starts from the vicinity of the sequence, which resulted from the previous phase and searches for a zero value of the auxiliary function by using the modified Newton method. Nonconvergence of the tunneling phase within 100 iterations of the algorithm was defined to be the stop condition of the optimization process. SEQOPT uses calculations of helical content as the target function for our global optimization procedure. These calculations are based on the sequence approximation AGADIR1s [21, 22]. In addition to the original AGADIR set of energy parameters [21] we incorporated several modifications of the parameter set of the theory published later [24]. Also the dependence of the intrinsic propensities of amino acids on their positions within helical segments was incorporated, as has been described [25–27]; besides, the energy parameters for those helical segments where formation of a capping box was possible were calculated as described [23]. The dependence of the energy parameters on temperature and pH was included according to Munoz and Serrano [22].

In order to use the tunneling algorithm for peptide sequence optimization, it is necessary to treat the amino acids of the primary structure as real variables. Therefore, we interpolated all the discrete energy parameters used in the statistical mechanical calculations of the goal function as follows: (a) integers from 1 to 20 were assigned to each type of amino acid; (b) the energy parameters of the AGADIR system were assigned to these integers on the real axis; (c) energy barriers of 2.5 kcal/mol were introduced at the midpoints between the integers assigned to the amino acids; and (d) the regular grids of the energy parameters and the barriers were used for one-dimensional and two-dimensional cubic spline interpolations [50]. The splines obtained by this procedure are continuously differentiable functions with well-separated energy minima at the integer points of



SeqOPT

This program is designed to optimize sequences of short peptides to increase their helical content in water to its possibly maximal level. Since during the optimization an average internal free energy of helical segments of the peptides is the subject of the minimization, the method is also applicable for protein helices. However, in the last case, one should fix those parts of protein helices that are known to have important functions or involved in interactions with other structure elements of the protein under consideration. Petukhov M, Tatsu Y, Tamaki K, Murase S, Uekawa H, Yoshikawa S, Serrano L, Yumoto N., Design of stable α-helices using global sequence optimization, J Pept Sci. 2009 May;15(5):359-65.

pH	7.0	[1.00 - 14.00]
Temperature in Kelvins	278	[273 - 400] K
Ionic Strength	0.1	[0.0 - 1.0] mol/l
N-terminal	ac ▾	(free, acetylated, succinylated)
sequence	gggdgggrgg	intitial sequence
Optimization mask	1110111011 arbitrary fixing of amino acids during the optimization. where '1' means it can be changed by optimization algorithm and '0' means is fixed.	
C-terminal	nh2 ▾	(free, amidated)
Execution time	300	seconds

Submit

Fig. 2 Screenshot of the web server main page containing an example of an initial setup of a SEQOPT calculation with mask fixing two amino acid residues

the real axis where they have both the true values of the AGADIR set of energy parameters and zero gradients.

To avoid the uncertainties that are associated with the tendency of the tunneling algorithm to escape from the permitted range of the real axis (from 1 to 20), the following periodical boundary conditions were employed for all points of the real axis:

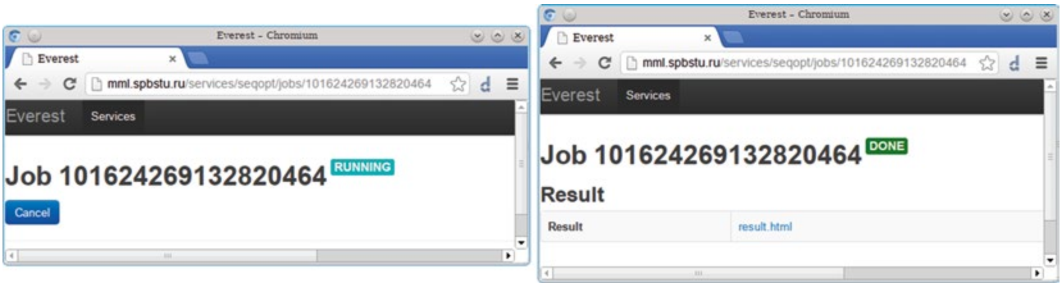
$$P_{int}\left(t_{aa}+n\times 20\right)=P_{int}\left(t_{aa}\right)$$

where P_{int} is the interpolation value of a parameter, t_{aa} is a variable type of amino acid, and n is an integer.

2.2 The SEQOPT Web Server

Using the publicly available SEQOPT web server [51] located at http://mml.spbstu.ru/services/seqopt/ one can optimize a peptide sequence with the option to define amino acids in desired positions. The server utilizes a web engine software called Everest (http://mathcloud.org/ project).

A SEQOPT session can be started from the initial web page shown in Fig. 2. This page provides a set of specified options


Everest - Chromium
Everest
mml.spbstu.ru/services/seqopt/jobs/101624269132820464
Everest
Services
RUNNING
Cancel
Everest - Chromium
Everest
mml.spbstu.ru/services/seqopt/jobs/101624269132820464
Everest
Services
DONE

Job 101624269132820464

Job 101624269132820464

Result

Result	result.html

Fig. 3 SEQOPT server screenshot during a job execution (*left panel*) and upon calculation completion (*right panel*)

including the choice of pH, temperature, ionic strength, and initial peptide sequence with an optimization mask, which prohibits selected residues to vary during the optimization, including N- and C-terminal blocking groups.

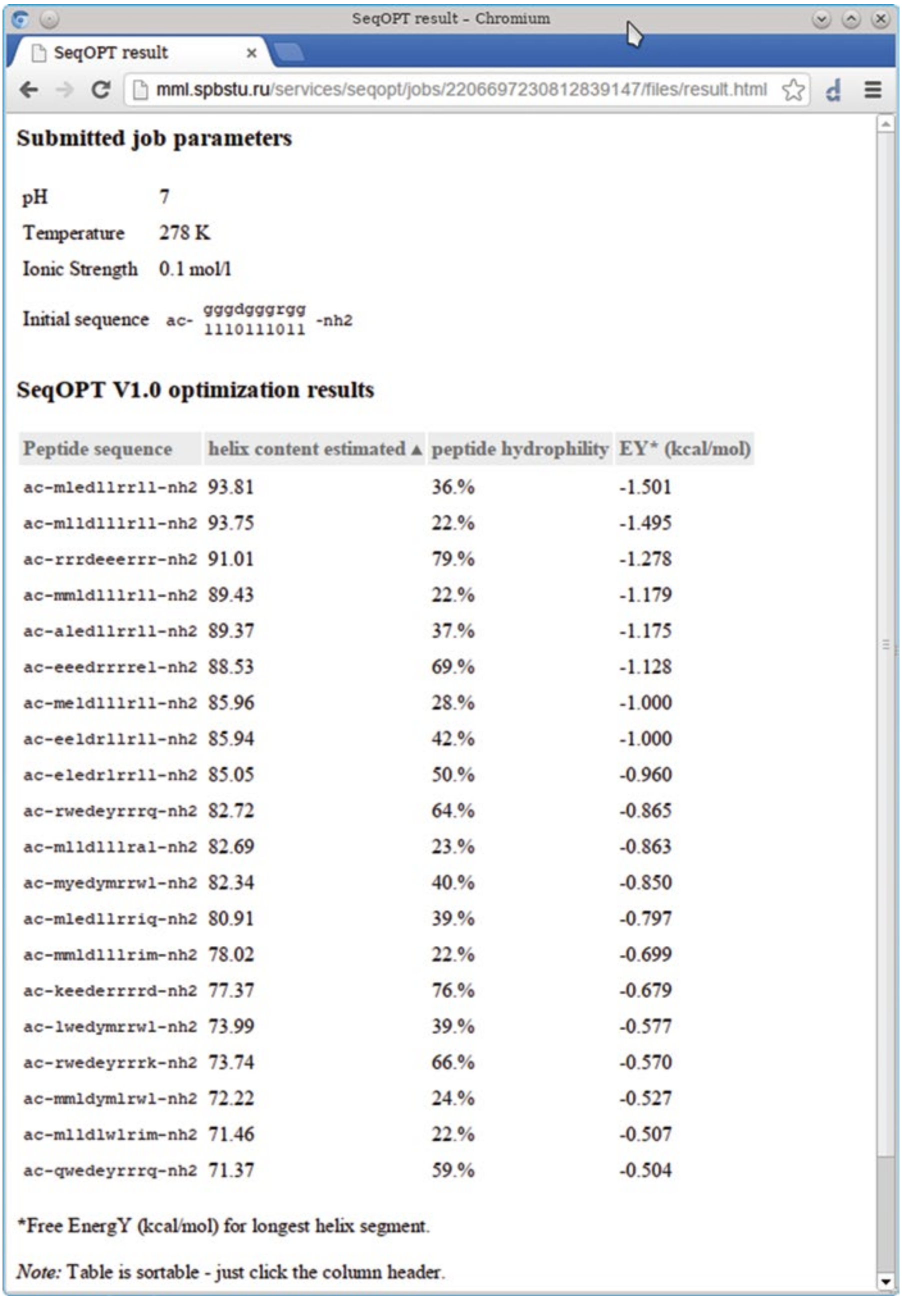
The buffer pH is set to 7.0 by default and can be changed according to experimental conditions. The default temperature setting in SEQOPT is 278 K. Since all energy contributions to free energies of peptide folding include their relevant temperature dependencies, the temperature can be set to any feasible value. Nevertheless it should be noted that the AGADIR parameter set was verified based on experimental data derived for peptides at around 5 °C and theoretical predictions are therefore preferably carried out at low temperatures. Experimental data showed that at high temperatures (80–90 °C) SEQOPT is expected to overestimate peptide helical content approximately by 10 %. Ionic strength is set to 0.1 M by default and can be changed according to needs. The sequence input data frame includes the initial peptide sequence and N- and C-terminal blocking groups. Note the necessity to set the mask of fixed residues to "0", otherwise use "1" for residues to be optimized. It is recommended to set the execution time according to the number of unfixed residues (*N*) using the formula

$$t[\text{seconds}] = 1.207e^{0.363N}.$$

After setting the specified parameters and submitting the job, the server runs the optimization process and displays the results available for download (*see* Fig. 3). One user can submit several jobs in one session and get access to the results using a provided digital JobID.

A SEQOPT job can be canceled during the execution (*see* Fig. 3. left panel). The successful accomplishment of the task submitted to SEQOPT is indicated by the generation of a result page shown in Fig. 3, right panel. Links to results are displayed in a table in HTML format as described below.

Since different interactions within α-helices tend to compensate each other, normally SEQOPT produces a number of diverse optimized sequences with similar helix stability values


SeqOPT result - Chromium
SeqOPT result
mml.spbstu.ru/services/seqopt/jobs/2206697230812839147/files/result.html

Submitted job parameters

pH	7
Temperature	278 K
Ionic Strength	0.1 mol/l
Initial sequence	ac- gggdgggrgg 1110111011 -nh2

SeqOPT V1.0 optimization results

Peptide sequence	helix content estimated ▲	peptide hydrophility	EY* (kcal/mol)
ac-mledllrrll-nh2	93.81	36.%	-1.501
ac-mlldlllrll-nh2	93.75	22.%	-1.495
ac-rrrdeeerrr-nh2	91.01	79.%	-1.278
ac-mmldlllrll-nh2	89.43	22.%	-1.179
ac-aledllrrll-nh2	89.37	37.%	-1.175
ac-eeedrrrrel-nh2	88.53	69.%	-1.128
ac-meldlllrll-nh2	85.96	28.%	-1.000
ac-eeldrllrll-nh2	85.94	42.%	-1.000
ac-eledrlrrll-nh2	85.05	50.%	-0.960
ac-rwedeyrrrq-nh2	82.72	64.%	-0.865
ac-mlldlllral-nh2	82.69	23.%	-0.863
ac-myedymrrwl-nh2	82.34	40.%	-0.850
ac-mledllrriq-nh2	80.91	39.%	-0.797
ac-mmldlllrim-nh2	78.02	22.%	-0.699
ac-keederrrrd-nh2	77.37	76.%	-0.679
ac-lwedymrrwl-nh2	73.99	39.%	-0.577
ac-rwedeyrrrk-nh2	73.74	66.%	-0.570
ac-mmldymlrwl-nh2	72.22	24.%	-0.527
ac-mlldlwlrim-nh2	71.46	22.%	-0.507
ac-qwedeyrrrq-nh2	71.37	59.%	-0.504

*Free EnergY (kcal/mol) for longest helix segment.

Note: Table is sortable - just click the column header.

Fig. 4 Sample result table of short peptide sequence optimization with fixed salt bridge in the central position of the α-helix

(within the expected approximation errors). One has to analyze the result table containing the most stable peptide sequences to select a suitable one, displaying the desired properties (Fig. 4).

The helix content (HC) of each peptide sequence is calculated as described above (*see* Subheading 1.1) and appears in the second column of the result table (Fig. 4). The table also lists the peptide hydrophilicity with typical values around ~40 % as well as the solubility of the peptides (columns 3 and 4). The prediction of peptide solubility is not an easy task. Solubility is usually estimated using one of several hydrophobicity scales reported in the literature [52–55]. For peptide solubility calculations SEQOPT utilizes the amino acid hydrophobicity scale described by Goldman and co-workers [56].

The last column of the result table (EY) lists free energies for the longest α-helical segment of a peptide as calculated using the modified AGADIR parameter set [23–27]. This is very useful for the design of α-helices in globular proteins where positions of helix ends are normally fixed [57]. Generally HC and EY are highly correlated.

2.3 In Silico Validation of α-Helix Stability

In protein crystal structures, α-helices can be assigned by the DSSP (Dictionary of Protein Secondary Structure) algorithm [58]. A variety of molecular modeling packages have been widely used to estimate the stability and energies of α-helical conformations in short peptides and globular proteins (ICM-Pro [59], AMBER [60], GROMACS [61], and APBS [62]) using different force fields. Given the recent increase in accessibility of supercomputer technology, molecular dynamics simulations (AMBER and GROMACS) of folding and unfolding processes in α-helices of short peptides and globular proteins on the microsecond time scale are now possible to simulate [63, 64]. MD simulations can provide *in silico* validation of high α-helical stability of peptides with optimized sequences without starting virtually long and expensive wet-lab experiments.

2.4 Experimental Validation of α-Helix Stability

Temperature-dependent circular dichroism (CD) spectroscopy is a standard method to experimentally characterize the stability of secondary structure elements in monomeric peptides and globular proteins [65, 66]. Characteristic ultraviolet CD spectra for α-helices exhibit minimum bands at approximately 222 and 208 nm and a maximum at approximately 192 nm. Providing accurate enough concentration measurements of proteins under investigation, the CD signal at 222 nm can be interpreted in terms of helical content using an empirical formula [67, 68]. Thus, CD spectroscopy provides a quick way to confirm whether or not a designed peptide adopts an α-helix structure at nearly native aqueous solution conditions (pH, ionic strength).

NMR spectroscopy is another powerful method of secondary structure determination in solution. To confirm the α-helix structure, it is important to obtain NMR-restraint characteristics of the peptide, like the nuclear Overhauser effect (NOE) pattern of $\alpha N(i,i+2)$, $\alpha N(i,i+4)$, and $\alpha\beta(i,i+3)$ atom interactions. $^3J_{HNH\alpha}$coupling constants should be in the range of 3–5 Hz [69]. However, extreme signal overlap within alanine-based peptides usually leads to a complication of the assignment task for nonlabeled peptides.

3 Conclusions

In this chapter we have presented the SEQOPT method for the rational design of α-helices based on proteinogenic amino acids, to achieve a high conformational stability by global optimization of the protein segment/peptide sequence. The method has three key characteristic properties: (1) only the 20 standard amino acids can be used, (2) it offers the possibility to arbitrarily fix any functionally important fragments of the primary structure, and (3) it offers accordingly the possibility to optimize the helical content of only those fragments that do not contain important functional groups of the protein. It has been shown that the proposed method is an effective tool for protein engineering [56]. In contrast to other methods for global energy optimization (molecular dynamics, Monte Carlo, etc.) that are often used to engineer the stability of the protein under investigation by altering only one or two amino acid residues and searching for advantageous physical interactions, the SEQOPT method deals with all possible sequences of protein α-helices and selects a suitable solution for most practical purposes.

Acknowledgments

This work was supported, in part, by grants from the Russian Ministry of Education and Science (grant No. 11.519.11.2002) and from the Russian Foundation of Basic Research (grant No 12-04-91444-NIH_a).

References

1. Estieu-Gionnet K, Guichard G (2011) Stabilized helical peptides: overview of the technologies and therapeutic promises. Expert Opin Drug Discov 6:937–963
2. Finkelstein AV, Badretdinov AY, Ptitsyn OB (1991) Physical reasons for secondary structure stability: alpha-helices in short peptides. Proteins 10:287–299
3. Scholtz JM, Baldwin RL (1992) The mechanism of alpha-helix formation by peptides. Annu Rev Biophys Biomol Struct 21:95–118
4. Errington N, Iqbalsyah T, Doig AJ (2006) Structure and stability of the alpha-helix: lessons for design. Methods Mol Biol 340:3–26
5. Petukhov M, Tatsu Y, Tamaki K, Murase S, Uekawa H, Yoshikawa S et al (2009) Design of stable alpha-helices using global sequence optimization. J Pept Sci 15:359–365
6. Azzarito V, Long K, Murphy NS, Wilson AJ (2013) Inhibition of [alpha]-helix-mediated protein-protein interactions using designed molecules. Nat Chem 5:161–173
7. Armstrong KM, Fairman R, Baldwin RL (1993) The (i, i+4) Phe-His interaction studied in an alanine-based alpha-helix. J Mol Biol 230:284–291
8. Huyghues-Despointes BM, Scholtz JM, Baldwin RL (1993) Helical peptides with three pairs of Asp-Arg and Glu-Arg residues in different orientations and spacings. Protein Sci 2:80–85
9. Padmanabhan S, Baldwin RL (1994) Tests for helix-stabilizing interactions between various nonpolar side chains in alanine-based peptides. Protein Sci 3:1992–1997
10. Lockhart DJ, Kim PS (1992) Internal stark effect measurement of the electric field at the amino terminus of an alpha helix. Science 257:947–951
11. Aurora R, Rose GD (1998) Helix capping. Protein Sci 7:21–38

12. Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT et al (1995) Protein design: a hierarchic approach. Science 270:935–941

13. Villegas V, Viguera AR, Avilés FX, Serrano L (1996) Stabilization of proteins by rational design of alpha-helix stability using helix/coil transition theory. Fold Des 1:29–34

14. Liu Y, Kuhlman B (2006) RosettaDesign server for protein design. Nucleic Acids Res 34(Web Server):W235–W238

15. Pokala N, Handel TM (2004) Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. Protein Sci 13:925–936

16. Liang S, Grishin NV (2003) Effective scoring function for protein sequence design. Proteins 54:271–281

17. Dai L, Yang Y, Kim HR, Zhou Y (2010) Improving computational protein design by using structure-derived sequence profile. Proteins 78:2338–2348

18. Li Z, Yang Y, Zhan J, Dai L, Zhou Y (2013) Energy functions in de novo protein design: current challenges and future prospects. Annu Rev Biophys 42:315–335

19. Levy A, Montalvo A (1985) The tunneling algorithm for the global minimization of functions. SIAM J Sci Comput 6:15–29

20. Muñoz V, Serrano L (1994) Elucidating the folding problem of helical peptides using empirical parameters. Nat Struct Biol 1:399–409

21. Muñoz V, Serrano L (1995) Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. J Mol Biol 245: 275–296

22. Muñoz V, Serrano L (1995) Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. J Mol Biol 245:297–308

23. Petukhov M, Yumoto N, Murase S, Onmura R, Yoshikawa S (1996) Factors that affect the stabilization of alpha-helices in short peptides by a capping box. Biochemistry 35:387–397

24. Lacroix E, Viguera AR, Serrano L (1998) Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. J Mol Biol 284:173–191

25. Petukhov M, Muñoz V, Yumoto N, Yoshikawa S, Serrano L (1998) Position dependence of non-polar amino acid intrinsic helical propensities. J Mol Biol 278:279–289

26. Petukhov M, Uegaki K, Yumoto N, Yoshikawa S, Serrano L (1999) Position dependence of amino acid intrinsic helical propensities II: non-charged polar residues: Ser, Thr, Asn, and Gln. Protein Sci 8:2144–2150

27. Petukhov M, Uegaki K, Yumoto N, Serrano L (2002) Amino acid intrinsic alpha-helical propensities III: positional dependence at several positions of C terminus. Protein Sci 11: 766–777

28. Muñoz V, Serrano L (1997) Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. Biopolymers 41:495–509

29. Finkelstein AV, Ptitsyn OB (1976) A theory of protein molecule self-organization. IV. Helical and irregular local structures of unfolded protein chains. J Mol Biol 103:15–24

30. Finkelstein AV (1977) Theory of protein molecule self-organization. III. A calculating method for the probabilities of the secondary structure formation in an unfolded polypeptide chain. Biopolymers 16:525–529

31. Finkelstein AV (1977) Electrostatic interactions of charged groups in water environment and their influence on the polypeptide chain secondary structure formation. Molek Biol (USSR) 10:811–819

32. Finkelstein AV, Ptitsyn OB (1977) Theory of protein molecule self-organization. I. Thermodynamic parameters of local secondary structures in the unfolded protein chain. Biopolymers 16:469–495

33. Finkelstein AV, Ptitsyn OB, Kozitsyn SA (1977) Theory of protein molecule self-organization. II. A comparison of calculated thermodynamic parameters of local secondary structures with experiments. Biopolymers 16:497–524

34. Bierzynski A, Kim PS, Baldwin RL (1982) A salt bridge stabilizes the helix formed by isolated C-peptide of RNase A. Proc Natl Acad Sci U S A 79:2470–2474

35. Kim PS, Baldwin RL (1984) A helix stop signal in the isolated S-peptide of ribonuclease A. Nature 307:329–334

36. Creamer TP, Rose GD (1994) Alpha-helix-forming propensities in peptides and proteins. Proteins 19:85–97

37. Avbelj F, Moult J (1995) Role of electrostatic screening in determining protein main chain conformational preferences. Biochemistry 34:755–764

38. Chang DK, Cheng SF, Trivedi VD, Lin KL (1999) Proline affects oligomerization of a coiled coil by inducing a kink in a long helix. J Struct Biol 128:270–279

39. Viguera AR, Serrano L (1999) Stable proline box motif at the N-terminal end of alpha-helices. Protein Sci 8:1733–1742

40. Strehlow KG, Baldwin RL (1989) Effect of the substitution Ala—Gly at each of five residue positions in the C-peptide helix. Biochemistry 28:2130–2133
41. Stapley BJ, Rohl CA, Doig AJ (1995) Addition of side chain interactions to modified Lifson-Roig helix-coil theory: application to energetics of phenylalanine-methionine interactions. Protein Sci 4:2383–2391
42. Seale JW, Srinivasan R, Rose GD (1994) Sequence determinants of the capping box, a stabilizing motif at the N-termini of α-helices. Protein Sci 3:1741–1745
43. Muñoz V, Blanco FJ, Serrano L (1995) The hydrophobic-staple motif and a role for loop-residues in alpha-helix stability and protein folding. Nat Struct Biol 2:380–385
44. Aurora R, Srinivasan R, Rose GD (1994) Rules for alpha-helix termination by glycine. Science 264:1126–1130
45. Viguera AR, Serrano L (1995) Experimental analysis of the Schellman motif. J Mol Biol 251:150–160
46. Zimm BH, Doty P, Iso K (1959) Determination of the parameters for helix formation in poly-gamma-benzyl-l-glutamate. Proc Natl Acad Sci U S A 45:1601–1607
47. Lifson S, Roig A (1961) On the theory of helix—coil transition in polypeptides. J Chem Phys 34:1963–1973
48. Harper ET, Rose GD (1993) Helix stop signals in proteins and peptides: the capping box. Biochemistry 32(30):7605–7609
49. Himmelblau DM (1972) Applied nonlinear programming. McGraw-Hill, New York
50. Bartenev OV (2000) FORTRAN for professionals 1. Dialog-MIPI, Moscow
51. Yakimov A, Rychkov G, Petukhov M (2013) SeqOPT: web based server for rational design of conformationally stable alpha-helices in monomeric peptides and globular proteins. FEBS J 280(Suppl s1):127–128
52. Rose GD, Wolfenden R (1993) Hydrogen bonding, hydrophobicity, packing, and protein folding. Annu Rev Biophys Biomol Struct 22:381–415
53. Eisenberg D, Weiss RM, Terwilliger TC (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci U S A 81:140–144
54. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132
55. Biswas KM, DeVido DR, Dorsey JG (2003) Evaluation of methods for measuring amino acid hydrophobicities and interactions. J Chromatogr A 1000:637–655
56. Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 15:321–353
57. Surzhik MA, Churkina SV, Shmidt AE, Shvetsov AV, Kozhina TN, Firsov DL, Firsov LM, Petukhov MG (2010) The effect of point amino acid substitutions in an internal alpha-helix on thermostability of *Aspergillus awamori* X100 glucoamylase. Prikl Biokhim Mikrobiol 46:221–227
58. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
59. 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
60. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM et al (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668–1688
61. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4:435–447
62. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A 98:10037–10041
63. Best RB, de Sancho D, Mittal J (2012) Residue-specific α-helix propensities from molecular simulation. Biophys J 102:1462–1467
64. Galzitskaya OV, Higo J, Finkelstein AV (2002) alpha-Helix and beta-hairpin folding from experiment, analytical theory and molecular dynamics simulations. Curr Protein Pept Sci 3:191–200
65. Dodero VI, Quirolo ZB, Sequeira MA (2011) Biomolecular studies by circular dichroism. Front Biosci 16:61–73
66. Kuwajima K (1995) Circular dichroism. Methods Mol Biol 40:115–136
67. Chen Y-H, Yang JT (1971) A new approach to the calculation of secondary structures of globular proteins by optical rotatory dispersion and circular dichroism. Biochem Biophys Res Commun 44:1285–1291
68. Luo P, Baldwin RL (1997) Mechanism of helix induction by trifluoroethanol: a framework for extrapolating the helix-forming properties of peptides from trifluoroethanol/water mixtures back to water. Biochemistry 36:8413–8421
69. Hinds MG, Norton RS (1995) NMR spectroscopy of peptides and proteins. Methods Mol Biol 36:131–154

Chapter 2

Design of Monomeric Water-Soluble β-Hairpin and β-Sheet Peptides

M. Angeles Jiménez

Abstract

Since the first report in 1993 (JACS 115, 5887–5888) of a peptide able to form a monomeric β-hairpin structure in aqueous solution, the design of peptides forming either β-hairpins (two-stranded antiparallel β-sheets) or three-stranded antiparallel β-sheets has become a field of growing interest and activity. These studies have yielded great insights into the principles governing the stability and folding of β-hairpins and antiparallel β-sheets. This chapter provides an overview of the reported β-hairpin/β-sheet peptides focussed on the applied design criteria, reviews briefly the factors contributing to β-hairpin/β-sheet stability, and describes a protocol for the de novo design of β-sheet-forming peptides based on them. Guidelines to select appropriate turn and strand residues and to avoid self-association are provided. The methods employed to check the success of new designed peptides are also summarized. Since NMR is the best technique to that end, NOEs and chemical shifts characteristic of β-hairpins and three-stranded antiparallel β-sheets are given.

Key words Antiparallel β-sheet, β-Hairpin, NMR, Peptide structure, β-Sheet propensities, Side chain/side chain interactions, Solubility, β-Turn prediction, β-Turn propensities

1 Introduction

Protein structures consist of a limited set of secondary structure elements, namely helices, β-strands, and turns, which are organized in different numbers and orientations to produce an extraordinary diversity of protein tertiary structures. A reasonable approach to understand protein folding and stability is the study of the conformational behavior of protein fragments and designed peptides. In this way, a large amount of information on α-helix folding and stability has been gathered since the early 1980s (*see* Chapter 1). In contrast, early efforts on studying β-sheet-forming peptides did not succeed, likely as a consequence of the strong tendency of sequences with high β-sheet propensity to self-associate. The first peptide able to adopt a monomeric β-hairpin in aqueous solution was reported in 1993 [1]. A β-hairpin is the simplest antiparallel

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_2, © Springer Science+Business Media New York 2014

β-sheet motif (Figs. 1 and 2). Mimicking parallel β-sheet motifs requires either the use of nonnatural scaffolds to join β-strands N-to-N-end or C-to-C-end [2–4] or long peptides in which the adjacent β-strands are connected by lengthy connectors, for example, an α-helix as in βαβ motifs in natural proteins, and in a designed 36-mer peptide [5]. These parallel β-sheet peptides as well as the use of β-turn or β-strand peptidomimetics to induce β-hairpin structures [6, 7] are beyond the scope of this chapter.

Since the report of the first β-hairpin peptide [1], the forces involved in the stability and folding of two- and three-stranded antiparallel β-sheets have been extensively investigated by several research groups (*for reviews see* [8–22]). Based on their conclusions, it is now possible to establish general guidelines for the design of new antiparallel β-sheet-forming peptides (Subheading 2). Previous to the description of the proposed design protocol, the structural characteristics of β-hairpins and three-stranded antiparallel β-sheets will be illustrated (Subheading 1.1). Next, the β-sheet-forming peptides reported up to now are overviewed focussed on the employed design strategies (Subheading 1.2), and the main conclusions derived from the extensive studies on β-hairpin and β-sheet stability using peptide models are summarized (Subheading 1.3).

1.1 Characteristics of β-Hairpin and Three-Stranded Antiparallel β-Sheet Structures

A β-hairpin consists of two antiparallel hydrogen-bonded β-strands linked by a loop region (Figs. 1 and 2). Characteristic average values for the ϕ and ψ angles of β-strand residues in antiparallel β-sheets are −139° and +135°, respectively [23]. β-Hairpin motifs differ in the length and shape of the loop and are classified according to the number of residues in the turn and the number of interstrand hydrogen bonds between the residues flanking the turn ($n-1$ and $c+1$ in Fig. 1). This β-hairpin classification uses a X:Y nomenclature [24], with X being the number of residues in the turn region and either Y = X if the CO and NH groups of the two residues that precede and follow the turn form two hydrogen bonds (for example, in 2:2 and 4:4 β-hairpins; Fig. 1a, d, respectively) or Y = X + 2 if these residues form only one hydrogen bond (as in 3:5 β-hairpins; Fig. 1c). The loops in 2:2, 3:5, and 4:4 protein β-hairpins are short, and very often their conformation corresponds to regular β-turns [24]. A β-turn (Fig. 3) consists of four residues and changes the direction of the protein main chain by having the first (i) and fourth (i + 3) residues spatially close (distance between their Cα atoms is less than 7 Å); in many cases the main chain CO of residue i is hydrogen-bonded to the amide NH of residue $i+3$ [25]. β-Turns are classified according to the ϕ and ψ dihedral angles of the two central residues (i + 1 and i + 2). The β-turns present in short β-hairpin loops are those with geometries adequate for the characteristic right-handed twist of antiparallel β-sheets (Fig. 2a). Thus, the most frequent β-turn in 2:2 β-hairpins is type I´ (Fig. 3), followed by type II´ (Fig. 3), whereas type I

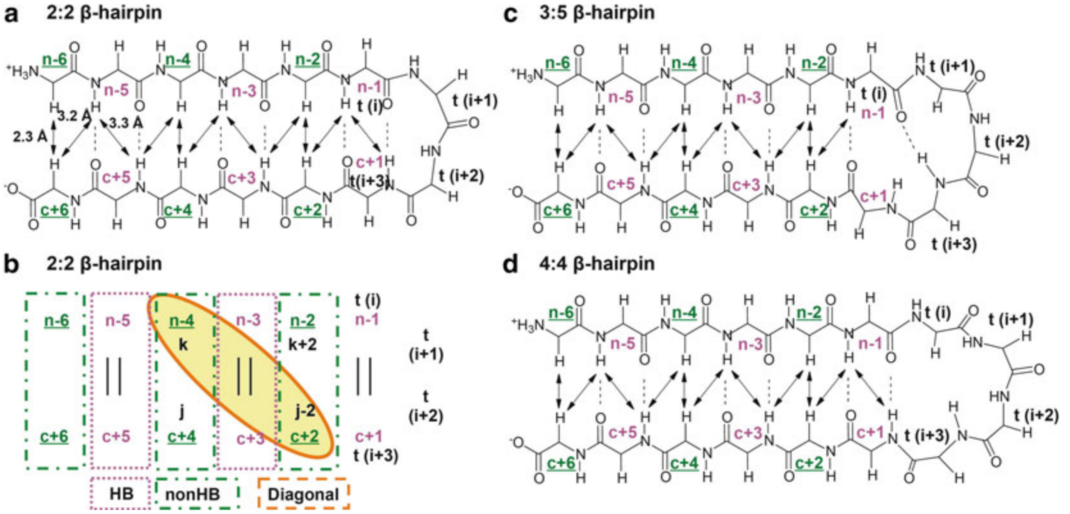


Fig. 1 Schematic representation of the peptide backbone conformation of 2:2 (**a**, **b**), 3:5 (**c**), and 4:4 (**d**) β-hairpins. Residues at the N-terminal β-strand, at the turn, and at the C-terminal strand are labelled as n, t, and c, respectively. Hydrogen bonds are indicated by *dotted lines* linking the NH proton and the acceptor CO oxygen in panels **a**, **c**, and **d**, and by *vertical lines* in panel **b**. *Black arrows* indicate the observable long-range NOEs involving Hα and NH backbone protons (Subheading 2.7.3). The corresponding average distances in protein antiparallel β-sheets are shown in panel **a**. Labels for residues at hydrogen-bonded sites (HB) are colored in *magenta*, and those at non-hydrogen-bonded sites (non-HB) in *green* and *underlined*. Side chains of residues in HB and non-HB sites point outwards at opposite faces of the β-sheet plane. In panel **b**, pairs of facing residues in HB and non-HB sites are indicated by *magenta* and *green rectangles*, respectively. Residues in a cluster of side chains in non-HB sites are labelled *k*, *k*+2, *j*−2, and *j*. A *yellow ellipse* highlights a diagonal pair interaction in non-HB sites

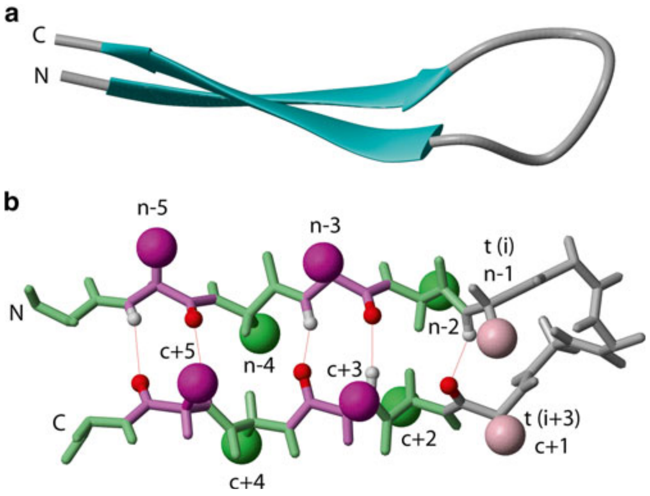


Fig. 2 β-Hairpin structure calculated for a designed 15-residue peptide [65]. (**a**) Ribbon representation where the β-sheet twist can be appreciated. N- and C-termini are indicated. (**b**) Backbone structure. Hydrogen-bonded oxygen atoms and NH protons are displayed as *red* and *white spheres*, respectively, and connected by a *red line*. Non-HB and HB sites are shown in *green* and *magenta*, respectively. The Cβ carbons of β-strand residues are shown as *spheres* and labelled as in Fig. 1c. Residues adjacent to the turn are colored in *light magenta*

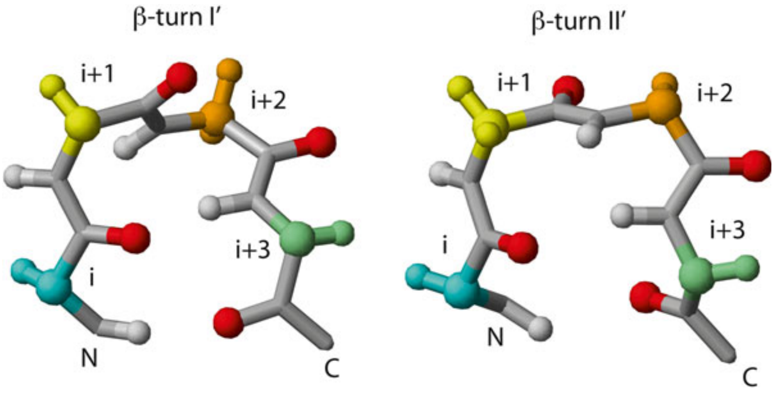


Fig. 3 Backbone structures of two types of β-turns. The type I′ β-turn (*left*) of sequence VSGV was taken from the 2:2 β-hairpin of protein TtCdnL from *T. thermophilus* (pdb code: 2LQK), and the type II′ (*right*) of sequence EGDL from the 2:2 β-hairpin of protein Ta0095 from *T. acidophilum* (pdb code: 2JOI). The Cα and Hα atoms and their bond are colored *cyan* for the first residue of the turn (*i*), *yellow* for the second residue (*i*+1), *orange* for the third residue (*i*+2), and *light green* for the last turn residue (*i*+4). Oxygen atoms and NH protons are displayed as *red* and *white spheres*, respectively. N- and C-termini are indicated

occurs less frequently. The ϕ and ψ dihedral angles characteristic of ideal β-turns of these types (I, I′, and II′) are listed in Table 1. The statistical occurrence found in 2:2 β-hairpins is explained by the fact that type I′ and II′ β-turns have a right-handed twist suitable for the β-strand pairing, while type I and II β-turns are left-handed twist, with the degree of twist being larger in types I and I′ β-turns than in type II and II′ ones. Type II β-turns are very frequent in proteins, but quite rare in β-hairpins. 3:5 β-hairpins normally exhibit a type I + G1 loop, that is, a type I β-turn with the residue at the i + 3 position, usually a Gly, forming a sort of bulge in the hairpin, whereas most of the 4:4 β-hairpins have a canonical type I β-turn.

Two kinds of β-strand positions can be distinguished for facing residues according to whether they form hydrogen bonds or not, i.e., hydrogen-bonded sites (HB) and non-hydrogen-bonded (non-HB) sites (Figs. 1 and 2b). In the β-hairpin, the side chains of consecutive residues in a strand point outwards opposite sides of the β-sheet plane, while the side chains of facing residues—corresponding to adjacent strands—are on the same face of the β-sheet (Figs. 1 and 2b). Since the averaged distances between the side chains of facing residues are 2.4 Å in non-HB sites, and 2.8 Å in HB sites [26], the contribution of a particular side chain/side chain interaction to β-hairpin stability depends on the site (Subheading 1.3.3). As a consequence of the right-handed twist of β-sheets, the side chains of residues in two consecutive non-HB sites (labelled as k and j − 2 in Fig. 1) are also quite close (3.0 Å; [26]). The interaction between these side chains, referred to as a diagonal interaction (Fig. 1b), also contributes to β-hairpin stability (Subheading 1.3.3).

Table 1
Average values of the ϕ and ψ dihedral angles and residues with high statistical probabilities in type I, type I′, and type II′ β-turns (Fig. 3). Residues are ordered from the most to the least favorable one [32]. The d-amino-acid dP is indicated by a lower case "p"

β-Turn	ϕ, ψ angles		Frequent residues				β-Hairpin
	i+1	*i*+2	*i*	*i*+1	*i*+2	*i*+3	
I	−60, −30	−90, 0	D > N > > H ≈ C ≈ S > P	P > > E ≈ S	D > N > > T > S ≈ W	G > > C ≈ T ≈ D ≈ R ≈ N	3:5, 4:4, 4:6
I′	+60, +30	+90, 0	Y > H > > I ≈ V	N > H > > D > G	G	K > > N ≈ R ≈ E ≈ Q	2:2
II′	+60, −120	−80, 0	Y > V ≈ S ≈ H ≈ F	p > G	N > S > D > H	T > G > N ≈ R ≈ F ≈ K	2:2

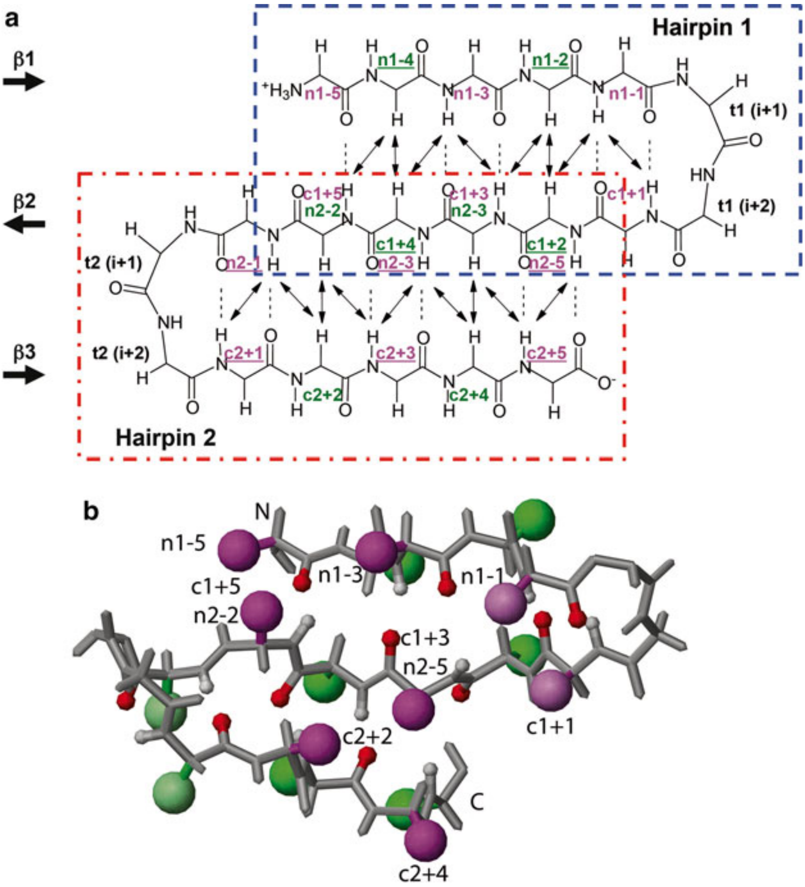


Fig. 4 Three-stranded antiparallel β-sheet motif. (**a**) Schematic representation of the peptide backbone. The meander β1–β2–β3 topology is indicated by *black arrows* (N to C direction) on the left site of the scheme. The two *large rectangles* surround the residues belonging to each of the 2:2 β-hairpins that compose this β-sheet motif. Residues at the N-terminal β-strand, at the turn, and at the C-terminal β-strand are labelled, respectively, as n1, t1, and c1, for hairpin 1, and n2, t2, and c2 for hairpin 2. *Dotted lines* link the NH proton and the acceptor CO oxygen of the β-sheet hydrogen bonds. Labels for residues at HB sites and non-HB sites are in *magenta* and *green*, respectively. Side chains of underlined strand residues are pointing outwards from the same β-sheet face, and those not underlined outwards from the other. *Double black arrows* indicate the observable long-range NOEs involving backbone Hα and NH protons (Subheading 2.7.3). (**b**) Backbone structure of a designed β-sheet peptide ([112]; Table 3). N- and C-termini are indicated. Hydrogen-bonded oxygen atoms and NH protons are displayed as *red* and *white spheres*, respectively. The Cβ carbons of β-strand residues are shown as *magenta spheres* for those side chains pointing upwards from the β-sheet plane, and as *green spheres* if pointing downwards. The *light green* and *light magenta* coloring indicates residues adjacent to the turn

After β-hairpins, the next simplest kind of β-sheet motifs are three-stranded antiparallel β-sheets with topology β1–β2–β3, sometimes denoted as meander β-sheets. They can be regarded as composed of two β-hairpins with a common β-strand (β2); that is, the C-terminal strand of hairpin 1 is the N-terminal strand of hairpin 2 (Fig. 4).

1.2 An Overview of Designed Antiparallel β-Sheet Peptides

The design of β-sheet peptides can be utilized to understand protein β-sheet folding and stability [8–22] and/or to achieve a biological functionality (Subheading 1.2.4). This section briefly reviews the peptides reported to form antiparallel β-sheets in aqueous solution (*see* **Note 1**) and highlights the employed design strategies. Tables 2 and 3 list the sequences of representative β-sheet-forming peptides.

1.2.1 Peptides Derived from Protein β-Sheets

Peptides that encompass the sequences of protein β-hairpins are mostly random coil in aqueous solution. The only reported protein fragments that adopt native-like β-hairpins in aqueous solution correspond to residues 41–56 of the domain B1 of protein G ($GB1_{41–56}$; [27]), to residues 46–61 and 48–59 of the domain B3 of protein G ($GB3_{46–61}$ [28] and $GB3_{48–59}$; [29]), and to residues 1–17 [30] and 4–14 [31] of ubiquitin (Table 2). These peptides have been taken as templates to investigate the factors contributing to β-hairpin stability (Subheading 1.3) or to design more stable β-hairpins. The strategies followed to achieve β-hairpin stabilization on peptides derived from protein β-hairpins are optimization of the β-turn sequence, optimization of inter-strand side chain interactions, statistical analysis within a protein family, and connection of the N- and C-termini of adjacent, non-consecutive, antiparallel β-strands via short loops.

The earliest successful strategy in the design of β-hairpin peptides consisted in substituting the native turn sequence for those residues with the highest intrinsic probability to occupy the corresponding β-turn positions [32]. Thus, the first reported β-hairpin peptide [1] was derived from residues 15–23 of Tendamistat, a 2:2 β-hairpin, by replacing the sequence of its native type I β-turn (SWRY) by NPDG, a sequence optimal for a type I β-turn (Table 1). This 9-mer peptide (Table 2) adopts a 3:5 β-hairpin with a type I + G1 loop and a non-native β-strand register. The peptide spanning the native sequence of Tendamistat is random coil [1].

Application of the "β-turn optimization" strategy to the N-terminal region of ubiquitin, a 3:5 β-hairpin with a type I + G1 loop, yielded peptides that adopt different β-hairpin conformations (Table 2). Thus, the substitution of the full-length native loop sequence TLTGK by NPDG rendered a 16-mer peptide [33] that forms a 3:5 β-hairpin with a type I + G1 loop, but with a nonnative strand register. In contrast, the native 3:5 β-hairpin was converted into 2:2 β-hairpins with native strand register by the replacement of the native loop TLTGK by four-residue sequences suitable for type I′ β-turns (FNGK; VNGK, TNGK, and GGGK; [34]), or by the substitution of the central loop residues LTG by two residues optimal for β-turns of either type I′ (DPDA) or type II′ (DPA, DPG; [35]). Also, a single-residue substitution (TLTGK by TLDGK) stabilizes the native 3:5 β-hairpin [36], but deletion of the G residue resulted in a random coil peptide [37].

Table 2
Sequences for some representative β-hairpin peptides[a]

Peptide system	Nº aa	Peptide sequence[a]			β-Hairpin	β-Turn
		N-terminal β-strand	Loop	C-terminal β-strand		
Tendamistat$_{15-23}$ [1]	9	YQN	PDG	SQA	3:5	I + G1[b, c]
Ubiquitin$_{1-17}$ [30]	17	MQIFVKT	LTG	KTITLEV	3:5	I + G1
	16	MQIFVKN	PDG	TITLEV	3:5	I + G1[c]
	16	MQIFVKT	pa	KTITKKV	2:2	I′
	16	MQIFVKT	pA	KTITKKV	2:2	I′
	16	MQIFVKT	pG	KTITKKV	2:2	II′
GB1$_{41-56}$ [27][d]	16	GEWTYD	DATK	TFTVTE	4:6	I
Fesinmeyer et al. [38]	16	KKWTYN	PATG	KFTVQE	4:6	I
Trpzip4 [41]	16	GEWTWD	DATK	TWTWTE	4:6	I
HP5W4 [38]	16	KKWTWN	PATG	KWTWQE	4:6	I
Chignolin [46]	10	GYD	PETG	TWG	4:6	I
CLN025 [47, 48]	10	YYD	PETG	TWY	4:6	I
Vammin$_{69-80}$ [44]	12	MRWN	PDRTQ	SWKM	4:6	I[e]
	12	MRCN	PDRTQ	SCKM	4:6	I[e]
Met repressor dimer [49]	16	KKYTVSI	NG	KKITVSI	2:2	I′
De Alba et al. [57]	10	IYSN	PDG	TWT	3:5	I + G1
De Alba et al. [60]	10	SYIN	SDG	TWT	3:5	I + G1
De Alba et al. [60]	10	YITN	SDG	TWT	3:5	I + G1
De Alba et al. [59]	10	IYS	AKAG	TWT	4:4	I

De Alba et al. [61]	15	**S**E**I**Y**S**N	PDG	T**W**T**V**T**E**	3:5	I + G1
Santiveri et al. [63, 64]	15	**S**E**S**Y**I**N	PDG	T**W**T**V**T**E**	3:5	I + G1
Santiveri et al. [63, 64]	14	**S**E**S**Y**I**Y	NG	K**W**T**V**T**E**	2:2	I′
BH8 [58]	12	_RG_I**T**V	NG	K**T**Y_GR_	2:2	I′
BH8KE [67]	14	_RG_**K**I**T**V	NG	K**T**Y**E**_GR_	2:2	I′
Stanger and Gellman [70]	12	R**Y**V**E**V	pG	O**K**I**L**Q-NH$_2$	2:2	II′
Stanger and Gellman [70]	12	R**Y**V**E**V	NG	O**K**I**L**Q-NH$_2$	2:2	I′
WKWK [114]	12	Ac-R**W**V**K**V	NG	O**W**I**K**Q	2:2	I′
Espinosa et al. [72]	12	R**W**Q**Y**V	pG	K**F**T**V**Q-NH$_2$	2:2	II′
CX$_8$C scaffold [77]	10	**C**T**W**E	GN	K**L**T**C**	2:2	II′
CX$_8$C scaffold [78]	10	**C**T**W**E	pN	K**L**T**C**	2:2	II′
CX$_8$C scaffold [78]	10	**C**T**W**E	pG	K**L**T**C**	2:2	II′
CX$_8$C scaffold [78]	10	**C**T**W**E	NG	K**L**T**C**	2:2	I′
CX$_8$C scaffold [80]	10	**C**T**W**E	PDG	K**L**T**C**	3:5	I + G1
Trpzip1 [41]	12	S**W**T**W**E	GN	K**W**T**W**K	2:2	II′
Trpzip2 [41]	12	S**W**T**W**E	NG	K**W**T**W**K	2:2	I′
Trpzip3 [41]	12	S**W**T**W**E	pG	K**W**T**W**K	2:2	II′
HP7 [85]	7	Ac-**WI**	NG	K**W**T-NH$_2$	2:2	I′

[a] Different peptide families are separated by dotted lines. Residues at non-HB sites are shown in bold. Residues added at the N- and C-termini of some sequences to increase solubility are shown in italics. Lower case "p" and "a" refer to the d-amino acids dP and dA, respectively. Ac stands for an acetylated amino group, and NH$_2$ indicates a terminal amidated carboxylate group

[b] Native protein β-hairpin is a 2:2 β-hairpin with a type I β-turn

[c] Nonnative strand register

[d] The sequences of the C-terminal β-hairpins of the domains B1 and B3 of protein G differ only by one residue at the first β-strand, which is E in domain B1 and V in domain B3

[e] This loop contains also a non-G bulge

Table 3
Sequences for representative three-stranded antiparallel β-sheet peptides

Citation	Nº residues	Peptide sequence[a]					β-Turn type at the loops
		N-terminal β-strand	Loop	Shared β-strand	Loop	N-terminal β-strand	
Kortemme et al. [100]	20	*R*GWSVQ	NG	KYTN	NG	KTTE*GR*	I′
Lopez de la Paz et al. [110]	20	*R*GWSLQ	NG	KYTL	NG	KTME*GR*	I′
Fernandez-Escamilla et al. [111]	24	*GSG*WSLY	NG	KYYL	NG	KTTWTDP*GK*	I′
Sharman and Searle [102]	24	KKFTLSI	NG	KKYTIS	NG	KTYITGR	I′
Griffiths-Jones and Searle [113]	24	KGEWTFV	NG	KYTVSI	NG	KKITVSI	I′
Schenck and Gellman [101]	20	VFITS	pG	KTYTEV	pG	OKILQ	II′
De Alba et al. [103]	20	TWIQN	GS	TKWYQN	GS	TKIYT	II′
Santiveri et al. [112]	20	TWIQN	pG	TKWYQN	pG	TKIYT	II′

[a] Residues with side chains pointing outwards from one of the β-sheet faces are underlined. Residues added at N- and C-termini to increase solubility are shown in italics. A “p” stands for dP

The stability of the 4:6 β-hairpins formed by peptide GB1$_{41–56}$ [27], and Trpzip4 (Table 2) was enhanced by the incorporation of a loop sequence (NPATGK) optimal according to amino acid frequencies at each loop position (the turn residues and the residues adjacent to the turn, *n* – 1 and *c* + 1 in Fig. 1d) in the GB family [38].

Further 16-residue peptides able to form 2:2 β-hairpins were derived from the C-terminal β-hairpin of the human YAP65 WW domain by turn sequence optimization and substitution of two neutral Q residues by two charged K residues [39]. A peptide derived from residues 18–35 of BPTI that has a modified turn sequence and a disulfide bond at the terminal HB site (Subheading 1.3.4) populates a native-like 3:5 β-hairpin, but also a nonnative-like 4:4 β-hairpin [40].

Since the finding that a facing W/W pair at a non-HB site is the most stabilizing cross-strand pair interaction, the incorporation of these pairs has proven to be a successful strategy for β-hairpin stabilization (*see* Subheading 1.3.3). Thus, 16-residue peptides that adopt native-like 4:6 β-hairpins more stable than the parent peptide GB1$_{41–56}$ ([27]; Table 2) were obtained by replacing the hydrophobic cluster formed by the side chains of the residues W/Y/F/V at the non-HB sites (*k*/*k*+2/*j*–2/*j* in Fig. 1b) by W-rich clusters, W/W/W/W (peptide Trpzip4 in Table 2), W/Y/F/W, and W/W/W/V [41]. These peptides belong to the tryptophan zipper family (*see* Subheading 1.2.2). The GB1$_{41–56}$ and Tripzip4 β-hairpins were additionally stabilized by incorporating a favorable ion pair interaction (K/E) between the N- and C-terminal residues [38, 42, 43]. Taking as template residues 69–80 of Vammin, a 4:6 β-hairpin in the native protein structure, the replacement of the non-HB-facing residues V/S by either a W/W pair or a disulfide bond yielded 12-residue peptides (Table 2) that adopt native-like β-hairpins [44]. The peptide that encompasses the native sequence is mainly random coil [44]. The stability of the N-terminal β-hairpin of ubiquitin has also been enhanced by incorporation of a hydrophobic cluster [45].

The sequence of a decapeptide (chignolin in Table 2) that adopts a 4:6 β-hairpin is the consensus found by statistical analysis for the central eight residues of structurally aligned homologues of GB1$_{41–56}$ plus a G/G pair at the terminal HB site [46]. Substitution of this G/G pair by salt bridges (E/K, E/R) or hydrophobic β-branched (T, I, V) and aromatic (F, Y, W) residues led to chignolin variants with increased β-hairpin stability [47], in particular, those with either a E/K, a I/I, or a Y/Y pair. This last variant (CLN025 in Table 2) was crystallized, and, based on CD studies, retains β-hairpin structure in the presence of urea and guanidinium chloride [48], losing β-hairpin conformation only in 8 M urea at high temperature (333 K).

β-Hairpin peptides have also been derived from protein adjacent, non-consecutive, antiparallel β-strands. Thus, the DNA-binding motif of the met repressor protein dimer has two identical β-strands, one from each subunit, that form an antiparallel β-sheet at the dimer interface. To mimic this binding motif, the C-terminus of one β-strand was linked to the N-terminus of the other by the two residue sequence NG, which is the most favorable for a type I′ β-turn (Table 1), appropriate for 2:2 β-hairpins (Subheading 1.1). Besides, an I residue in one of the two strands was replaced by an aromatic Y residue to facilitate NMR assignment (Subheading 2.7.3). As intended by design, the resulting 16-residue peptide adopts a 2:2 β-hairpin (Table 2; [49]). This peptide has been used as a model to study the effect of β-turn and side chain interactions on β-hairpin stability [50–55]. The same design strategy has been applied to peptides derived from two adjacent non-consecutive strands of protein CD2 ([56]; Subheading 1.2.4).

1.2.2 De Novo-Designed β-Hairpin Peptides

The first water-soluble (*see* **Note 1**) de novo-designed β-hairpins were reported in 1996 [57, 58]. These peptides and all other de novo-designed β-hairpins have been extensively studied to understand β-hairpin formation and used as prototypes to get either more stable or minimal β-hairpin peptides. They can be classified into a few families: the 10-mer [57, 59, 60] and 14/15-mer [61–65] peptides designed in our group; the 12-mer peptide denoted as BH8 [58] and those derived from it [66–69]; the 12-mer β-hairpin peptides reported by Gellman's group [70–75] and their longer 16- and 20-mer derivatives [74, 76]; the CX_8C scaffold, formed by a series of disulfide-cyclized 10-residue peptides [77–80]; the tryptophan zippers or Trpzip peptides [41] and their derivatives [81–84]; the numerous optimized and/or shortened peptides derived from Trpzip by Andersen and co-workers [38, 85–89]; and the 12- and 14-mer β-hairpin peptides designed by Water's group [90–98] using as templates those of Gellman's group.

In all of them (Table 2), the sequences were selected to have a good β-turn sequence at the loop, i.e., NPDG or PDG to have 3:5 β-hairpins with a type I + G1 loop [57, 59, 60]; GN, dPN, and dPG for 2:2 β-hairpins with a type II′ β-turn; NG for 2:2 β-hairpins with a type I′ β-turn; and PATG for 4:6 β-hairpins with type I β-turns [38]. The 2:2 β-hairpins formed by the CX_8C scaffold were converted into 3:5 β-hairpins with a type I + G1 loop by the substitution of the two central residues by a three-residue sequence, such as PDG [80]. Changes in the loop sequence transformed the 3:5 β-hairpin peptides reported by de Alba et al. [57] into 4:4 [59] and 2:2 β-hairpins [63, 64]. Some of them populated two β-hairpins, which differed in β-strand registers and type (3:5 and 4:4 [57, 59–61] or 2:2 with I′ β-turn and 2:2 with II′ β-turn [59, 61]).

The criteria to select β-strand residues differ among the different peptide systems. In the decapeptides reported by de Alba et al. [57], β-strand residues were chosen only by their high intrinsic β-sheet propensities (Subheading 2.3.1). Stability in these 3:5 β-hairpins is affected by the composition and order of β-strand residues [60].

Some of these short peptides were lengthened by the addition of good β-sheet former residues to the N- and C-termini to yield 14- and 15-mer peptides that, in general, adopted more stable β-hairpins [61–65]. In peptide BH8 [58], the β-strand residues were those statistically favorable in the corresponding β-strand positions deduced from the examination of the protein structure database included at the time in the WHATIF program [99]. At the strands, the CX_8C scaffold, which is formed by a series of disulfide-cyclized 10-residue peptides [77, 78], contains two HB sites and two non-HB sites. The disulfide-bonded C residues are placed in the terminal non-HB site (Fig. 1; Subheading 1.3.4). Aromatic residues (F, Y, W) and/or L were placed at the other non-HB site because these side chains provide the best packing with a disulfide bond that connect adjacent antiparallel strands in proteins. An E/K salt bridge was located at the HB site adjacent to the turn, and two good β-sheet-forming residues (Subheading 1.3.2) at the other HB site. Tryptophan zippers or Trpzip [41] are 12-residue peptides derived from the CX_8C scaffold peptides by removal of the disulfide bond and incorporation of a W/W/W/W cluster at the non-HB sites (*k*/*k*+2/*j*–2/*j* in Fig. 1b). The shortest, but still stable, β-hairpins designed by Andersen's group contain a stabilizing W/W pair [85, 86].

Some criteria to prevent aggregation and enhance water solubility were considered in all designs. Thus, the 12-mer peptides reported by Stanger and Gellman [70] have an overall charge ≥+3. In the case of peptide BH8 [58], positively charged R residues are placed at the N- and C-termini, and separated from the eight central "truly" β-hairpin-forming residues by flexible G residues. In the Trpzip peptides [41], a polar S and a positively charged K were added at the N- and at the C-termini, respectively.

1.2.3 Three-Stranded Antiparallel β-Sheets

After the success in designing β-hairpin peptides, several research groups addressed the design of three-stranded antiparallel β-sheets with a meander β1–β2–β3 topology (Fig. 4; *see* **Note 2**), the next step up in motif complexity. Almost simultaneously, four different peptides were reported to adopt monomeric meander three-stranded β-sheets in aqueous solution (Table 3; [100–103]). The β-sheet motif in the four peptides consists of two 2:2 β-hairpins (Fig. 4). Taking into account the crucial role of the turn in β-hairpin folding and stability (Subheading 1.3.1), the incorporation of sequences adequate to form either type I′ β-turns or type II′ β-turns (Table 1) was essential to achieve successful designs. Although the design strategies differ in the procedure for the selection of strand residues, all of them considered intrinsic β-sheet propensities (Subheading 1.3.2; [104–109]) and intended to have favorable side chain interactions (Table 4; Subheading 1.3.3). Criteria to prevent aggregation and to aid solubility were also important and consequently all of the designs incorporate from two to five positively charged residues with their side chains pointing outwards on

Table 4
Favorable cross-strand side chain/side chain interactions between facing and diagonal β-strand residues (Fig. 1)

	Non-HB sites		**HB sites**
	Cross-strand facing pairs	Diagonal pairs	Cross-strand facing pairs
Statistical data	**C/C**>>**E/K**>D/H>N/N>**W/W**>C/W≈D/G≈D/R>K/N≈N/S>H/P≈Q/R [143]	W/Y>K/E≈D/K≈N/N>E/R≈R/E≈D/N≈N/D [26]	C/C>**E/K**≈E/R>H/H≈Q/R≈D/N≈F/F≈C/H≈S/S≈D/K≈K/Q≈**N/T** [143]
Experimental[b]	**C/C** [77, 78] **K/E** [55, 67] F/F>**E/K** [92] Y/W [60]; F/F [90]; Y/F [134] I/W [62]; Y/L [73] **W/W**>>W/F>W/Y>W/L>W/M>W/I>W/V>>Y/L>M/L>F/L>L/L>I/L≈V/L [77, 78]	Y/K [73] W/R≈W/M>W/K>F/M>F/R≈W/S>F/K [91, 93, 94]	**N/T** [59]; T/T [60]; S/T [60, 62] Y/I≈Y/V>Y/F≈Y/Y>Y/W [69] V/V>H/V≈V/H>T/T [79] L/I>L/V≈I/I>S/K≈S/I≈E/I>K/V≈I/V≈K/I>V/I≈**E/K**≈S/T [135] **E/K**>I/I>Y/Y>Y/F≈W/F>F/F≈W/Y>F/Y≈V/V≈E/R≈V/Y≈V/I>V/F≈W/W≈I/V>>T/T [47]

Pair-wise interactions found to be favorable (statistically and experimentally) are shown in bold

[a] As deduced from experimental studies on different β-hairpin peptides (Subheading 1.3.3)

both sides of the β-sheet plane (*see* **Note 3**). In this way, the positive charge is distributed over the β-sheet and self-association is minimized. As previously in the design of β-hairpin peptide BH8 (Subheading 1.2.2), statistical analysis of β-sheet sequences in the protein databank (PDB) was considered to select the sequence of a 24-residue β-sheet (Table 3; [102]). The "truly" β-sheet-forming residues of the 20-mer peptide denoted "Betanova" (Table 3; [100]) count only 16, since the sequence RG was added at N- and C-termini to improve water solubility, as in peptide BH8 (Subheading 1.2.2). The residues of the three β-strands were chosen by evaluating the van der Waals energies of several sequences using a template backbone structure derived from two proteins with antiparallel β-sheets, a dehydrogenase fragment and a WW domain (*see* **Note 4**). The stability of the "Betanova" β-sheet was improved by triple-residue substitutions that enhance hydrophobic side chain packing, as indicated by computational evaluation (Table 3; [110]). Further stabilization of the resulting β-sheet peptide was achieved by structural sequence alignment with a WW domain (Table 3; [111]). In the 20-residue β-sheet model designed by de Alba et al. [61], β-strand residues were selected to have favorable cross-strand side chain interactions according to statistical data (Table 4) and previous results obtained from β-hairpin models. This β-sheet was strongly stabilized by changing the GS turn sequences to the more rigid DPG sequences (Table 3; [112]). Another β-sheet design consisted in extending the sequence of a β-hairpin peptide by adding a type I′ β-turn, an NG sequence, and a third strand to its C-terminus. This third strand contains F and W residues at non-HB sites which face Y and V residues in the second shared strand to give rise to a stabilizing hydrophobic cluster (Table 3; [113]).

1.2.4 "Functional" β-Hairpins

β-Hairpin structures can be used as scaffolds to get peptides with specific functions or activities (*see* [6] *for a review on β-hairpin peptidomimetics*; *see* **Note 5**), such as ligand binding, antimicrobial activity, and inhibitors of protein–protein interactions. Thus, the 12-residue β-hairpin designed by Gellman's group (Table 2; [70]) was converted into a β-hairpin able to bind nucleotides (ATP, GTP, CTP, and FMN) with high affinity by the substitution of the Y/E/K/L cluster at non-HB sites ($k/k+2/j-2/j$ in Fig. 1b) by W/K/W/K (Table 2; [114–116]). The W/K/W/K cluster was designed to contain a diagonal W/W pair (Fig. 1b) where a nucleobase could intercalate, since the two W diagonal residues of Trpzip motifs seem to form a cleft [41], and two K residues that might afford favorable electrostatic interactions with nucleotide phosphates. The peptide has a net charge of +4 to increase solubility. Taking this nucleotide-binding β-hairpin as a starting point, β-hairpin dimers that bind single- and double-stranded DNA and RNA have been designed [117–119]. Also, a three-stranded β-sheet peptide able to bind single-stranded DNA [120] has been obtained by combining the sequences of a WW domain (*see* **Note 4**) and the designed nucleotide-binding β-hairpins [114–116]. Using WKWK and Trpzip3 derivatives (Table 2), the cleft formed by the diagonal W/W pair in a W/W/W/W cluster, but not in a W/K/W/K cluster, was found to bind a polyproline peptide [121]. Two β-hairpin peptides that are stabilized upon metal binding have also been reported: a 17-mer peptide that binds Zn^{2+} to adequately located H residues [122], and 20-mer peptides that are stabilized upon As^{3+} binding when their sequences contain reduced C residues at appropriate positions [123]. In addition, a 14-mer peptide derived from a choline-binding protein has been shown to bind tri-methylammonium [124] and redox activity has been reported for an 18-mer β-hairpin peptide containing a cross-strand Y/H pair at a non-HB site [125].

Furthermore, β-hairpin peptides that mimic protein "hot spots" are being designed to disrupt protein/protein interactions. Thus, a 17-mer peptide spanning the receptor-binding region of placental growth factor adopts a native-like 3:5 β-hairpin and interacts with a domain of the vascular endothelial growth factor receptor 1 [126]. Also, a 10-residue peptide able to modulate protein–protein interactions between the cell adhesion protein CD2 and CD58, important in the immune response, has been obtained from the adjacent antiparallel β-strands c and f of protein CD2 [56]. The two N- and C-termini of the strands c and f were connected in the designed peptide, i.e., one via a two-residue sequence dPP, good for type I′ and II′ β-turns (Table 1), and the other by main-chain peptide bond formation.

1.3 Main Conclusions About Contributions to β-Sheet Folding and Stability

Analyses of the conformational behavior of peptides able to adopt β-sheet motifs in solution have provided information about their formation and stability [8, 9, 11–18, 20–22]. These conclusions are summarized below with emphasis on their applicability as design guidelines rather than on the physical-chemical basis of β-hairpin and β-sheet stability (*see* **Note 6**).

1.3.1 Essential Role of the Turn Sequence

The first evidence about the importance of the loop region in β-hairpin folding came from the early β-hairpin peptides designed by turn optimization (*see* Subheadings 1.2.1.and 1.2.2). Thus, the β-hairpin derived from Tendamistat by incorporating a NPDG turn sequence shows a nonnative β-strand register [1], and ubiquitin-derived peptides (Table 2) displayed various β-strand registers, native-like and nonnative, depending on the turn sequence [33–37]. The essential role played by the turn in directing β-hairpin formation [50, 52, 59, 61, 63] and in determining its final stability has been demonstrated in many β-hairpin peptide systems (*see reviews* [9–18, 20–22]). β-Hairpin population has been found to correlate with the statistical occurrence of the residues at the position $i+1$ of type I′ β-turns in protein 2:2 β-hairpins [66]. Also, the β-turn sequence has been shown to determine the β-strand register in three-stranded antiparallel β-sheet systems [127]. A dP as the $i+1$ turn residue greatly stabilizes β-hairpins and multi-stranded antiparallel β-sheets, whereas substitution of dP by P prevents β-sheet formation [70, 101, 110, 112, 127]. Even in the very stable Tripzip peptides, β-hairpin stability depends on the turn sequence [41, 84]. On the whole, a good β-turn sequence is a necessary but not a sufficient condition for a peptide sequence to adopt a β-hairpin structure [63].

The fact that, as a first approach, turns and strand–strand interactions appear to make independent and additive contributions to β-hairpin stability [43, 78] provides a solid basis for the design protocol described below where β-turn and β-strand residues are selected independently (Subheading 2).

1.3.2 Intrinsic β-Sheet Propensities

Strand residues with high intrinsic β-sheet propensities help to stabilize β-hairpins and three-stranded β-sheets, whereas those with low intrinsic β-sheet propensities destabilize them. For example, the incorporation of a G residue either in an edge strand or in the central strand led to a large destabilization of a three-stranded antiparallel β-sheet peptide [128]. However, the interaction of a G residue with a facing cross-strand aromatic side chain at an HB site has been reported to stabilize protein β-sheets [129]. The fact that residues with high intrinsic β-sheet propensities (Subheading 2.3.1) are generally hydrophobic accounts for the high tendency of β-sheet peptides to aggregate.

1.3.3 Side Chain Interactions Among β-Sheet Residues

In antiparallel β-sheets, the contributions to stability of side chain interactions depend on their nature, i.e., hydrophobic, electrostatic, and cation-pi, and also on the position that the interacting residues occupy within the β-sheet, i.e., whether they are facing residues in non-HB sites or in HB sites, show diagonal interactions, and on their distance to the loop (Fig. 1b; Table 4). The stabilization provided by favorable interactions increases with their proximity to the loop region, as found for the W/V/Y/F hydrophobic cluster [74], for the cross-strand pair I/W at a non-HB site [62], and for the facing pair S/T at an HB site [62]. The contribution of a particular pair of residues can also be asymmetric, that is, depends on which residue is located at the β-strand preceding and following the loop, for example, E/K is different from K/E ([67]).

β-Hairpin structures are stabilized by hydrophobic clusters, particularly when placed at non-HB sites, such as positions $k/k+2/j-2/j$ in Fig. 1b. This stabilizing effect was demonstrated by incorporating the W/V/Y/F hydrophobic cluster taken from the native β-hairpin of protein G B1 domain (GB1$_{41-56}$ in Table 2) into non-HB sites of designed 12-residue β-hairpin peptides [72, 75]. Side chain packing within the hydrophobic cluster and, hence its contribution to stability, is affected by turn flexibility, because of the steric limits imposed by a rigid turn (dPG) that can impede the optimal side chain contacts. However, a flexible turn (NG) allows side chains to optimize their interactions and packing [75, 84, 130]. The Trpzip peptides (Subheadings 1.2.1 and 1.2.2) which contain a cluster of four W residues at non-HB sites (*k*/*k*+2/*j*−2/*j* in Fig. 1b) have become the paradigm of stable β-hairpin peptides due to their remarkable stability [41]. The indole rings of each W/W pair at a non-HB site adopt an edge-to-face geometry [8, 41]. Replacement of W residues in the Trpzip motif (peptides Trpzip1 and Tripzip2; Table 2) decreases β-hairpin stability, but the degree of destabilization depends on the position of the replaced tryptophan/s, and the substituting aromatic (Y; [81, 83]) or hydrophobic residues (V; [82]). Thus, a single cross-strand W/W pair, but not a diagonal, suffices to stabilize the β-hairpin, and the W/W is slightly more stabilizing when closer to the turn. The rankings of the examined cross-strand interactions are W/aliphatic < W/Y < W/W, and Y/Y ≈ V/V (Table 4). The geometry of the aromatic rings in the cross-strand W/Y and Y/Y pairs is edge to face, as in the W/W pairs.

The contributions of cross-strand side chain/side chain interactions of facing residues at non-HB and HB sites differ due to their different geometry (Figs. 1 and 2b); that is, the distances between side chains of facing residues are different (Subheading 1.1).

As in α-helices, ionic interactions stabilize β-hairpin peptides. Thus, K/E ion pairs located at non-HB sites enhance β-hairpin stability in several peptide systems [42, 43, 55, 67], prevent fraying if located at the peptide ends [42, 43], and are more stabilizing than E/K pairs at the same position [67]. The contributions of these K/E salt bridges are pH dependent and decrease at low pH due to protonation of the carboxylate group of the E side chain [55]. Two K/E salt bridges contribute more to β-hairpin stability than the sum of the two individual K/E interactions, indicating that co-operativity (Subheading 1.3.6) plays a role [55]. The interaction between the N-terminal positively charged amino group and the negatively charged carboxylate group at the C-terminus also contributes to β-hairpin stability [131]. Furthermore, a 2:2 β-hairpin peptide is stabilized by a salt bridge interaction between the N-terminal, positively charged, K residue and the C-terminal carboxylate [15].

Aliphatic, aromatic, and mixed aliphatic/aromatic pairs favor β-hairpin formation (Table 4). The contributions of many cross-strand interactions involving mainly hydrophobic and aromatic residues have been examined using the CX_8C scaffold model ([77, 78]; Table 2; Subheading 1.2.2). The main finding was that an edge-to-face W/W pair at a non-HB site is the most β-hairpin-stabilizing interaction [77, 78], which has been confirmed in other β-hairpin peptides [41, 44, 85–87, 132]. In Vammin-derived peptides (Subheading 1.2.1), a W/W pair at a non-HB site led to a more stable and more native-like β-hairpin than an equally located covalent disulfide bond [44]. However, a W/W pair placed at an HB site did not stabilize a β-hairpin structure in a Vammin-derived peptide [133].

Other aromatic pairs also contribute to β-hairpin stability, such as a Y/W pair in a 10-mer peptide designed by de Alba et al. [60], a Y/F pair in peptide $GB1_{46–51}$ [134], and a F/F pair in the 12-residue peptide designed by Gellman's group [70, 71] that showed an edge-to-face geometry [90], and was more stabilizing than an equally positioned ionic E/K interaction [92].

In contrast to cation-π, anion-π interactions are unfavorable. Thus, the interactions between a W and negatively charged amino acids (E, or phosphorylated S, T, and Y) at non-HB sites have been shown to be repulsive and destabilize 2:2 β-hairpins [96]. In the same peptide system, the destabilizing effect of the W/phosphoserine interaction is context dependent [97].

Cross-strand interactions at the HB site (Fig. 1b) adjacent to the turn are important for β-hairpin stability. In particular, the polar pairs N/T [59], T/T [60], and S/T [60, 62], and the aromatic/aliphatic pairs Y/I and Y/V [69] were found to favor β-hairpin formation at that position (Table 4). The aromatic pairs Y/F, Y/Y, and Y/W were less stabilizing than the aromatic/aliphatic pairs [69].

Using Trpzip 4 as model (Table 2), the pair V/V in an inner HB site ($n-3/c+3$ in Fig. 1d) was found to be more stabilizing than the pairs V/H or H/V, and these better than T/T [79]. However, at the HB site adjacent to the disulfide bond in the CX_8C scaffold model (Table 2; Subheading 1.2.2), the residue V (and also I) is favored at the N-terminal strand, so that the pair V/H provides higher stabilization than the pair H/V [79].

Cross-strand amino acid pairs have also been examined at terminal HB sites in decapeptide chignolin ([47]; Table 2; Subheading 1.2.1) and in a very short 8-mer peptide model containing a nonnatural residue at the non-HB site [135]. In general, the preferred pairs (Table 4) are hydrophobic [47, 135], aromatic [47], hydrophobic-aromatic [47], a hydrophobic residue with a charged residue with a large aliphatic region (K or E; [135]), and salt bridges (K/E; [47, 135]), and the most destabilizing are pairs of residues bearing charges of the same sign [135], such as D/D. Besides, most pairs containing a D residue are destabilizing [135].

Diagonal side chain/side chain interactions (Fig. 1b) also contribute to β-hairpin stability. A Y/K pair placed at diagonal non-HB sites in a 12-residue β-hairpin peptide was the first diagonal pair reported to be β-hairpin stabilizing [73]. Afterwards, the diagonal cation-π interactions between F and W with either K or R have been extensively investigated (*see* [8] and references therein), being more stabilizing with W than F, and R than K [91]. Methylated R and K lead to more favorable cation-π interactions than R [136] and K [137, 138]. In the same peptide model, the diagonal interaction between an aromatic residue, W or F, and M was found to be as stabilizing as an equivalent cation-π interaction [93].

A stabilizing capping motif has been reported by Andersen and co-workers [88, 89]. This motif, designated as β-cap, consists of an Ac-W at the N-terminus and a WTG extension at the C-end. The two W residues have to face each other at a non-HB site. This motif reduces the fraying at the N- and C-termini generally observed in β-hairpin peptides.

1.3.4 Disulfide Bonds and Other Covalent Cross-Links

Stabilization of protein structures by disulfide bonds is explained mainly by entropic contributions (the unfolded state becomes more rigid, and hence the loss of entropy upon folding decreases), but enthalpy also affects the contribution to stability of disulfide bonds [139–141]. In the case of β-hairpin structures, some natural antimicrobial peptides are β-hairpins stabilized by several disulfide bonds [142]. Furthermore, disulfide-cyclized peptides, as well as N-to-C-backbone-cyclized peptides, have been used as references for the folded state of designed β-hairpin peptides [71], and a disulfide bond links the terminal residues in the CX_8C scaffold (Table 2; [77, 78]). However, the entropic effect of disulfide bonds, and any other covalent cross-links can be counterbalanced by (1) unfavorable enthalpic effect due to steric strain caused by inadequate geometry and/or (2) a decrease in the contributions of other inter-strand interactions that can be impeded to adopt their optimal arrangement by the rigidity imposed by the cross-link. In fact, disulfide bonds in protein β-hairpins are more abundant at non-HB sites than at HB sites (Fig. 1), where they occur very rarely [143–145]. Accordingly, a disulfide bond stabilizes designed β-hairpin peptides at non-HB sites, but not at HB sites [65]. In searching mimics of VEGF loop 3, a disulfide bond as well as amide bridges placed at an HB site led to β-turn stabilization, but the peptide did not populate any β-hairpin structure, either native or nonnative [146]. More recently, tri-azole bridges have been shown to trigger β-hairpin stability in both non-HB and HB sites [147], but the stabilized β-hairpins might display some distortion. Their stabilizing capacity depends on the number of methylene groups present in the bridge [148].

1.3.5 β-Sheet Twist

The right-handed twist characteristic of β-sheets (Fig. 2a; Subheading 1.1) seems to be related to β-sheet stability. The most twisted β-hairpins are usually the most stable ones. This is the case for the 3:5 β-hairpins that are more twisted and also more stable than the 4:4 β-hairpins [61]. The existence of a correlation between the degree of twist and the buried hydrophobic surface has been found in a three-stranded β-sheet [112]. Thus, β-sheet twist appears to contribute to β-sheet stability by increasing hydrophobic surface burial.

1.3.6 Co-operativity

The question of whether the folding of β-hairpin and β-sheet peptides is co-operative or not still remains open. Two types of co-operativity can be distinguished in antiparallel β-sheet peptides, one being longitudinal or parallel to the strand axis, and the other perpendicular to the strand direction. The increase of β-hairpin stability observed upon strand lengthening indicates the existence of longitudinal co-operativity [76]. In regard to perpendicular co-operativity, it has been shown that adding a fourth strand to a three-stranded β-sheet peptide leads to further β-sheet stability [130].

1.3.7 Hydrogen Bonds

The contribution of hydrogen bonds to β-hairpin stability remains a subject of discussion. Based on the amide I region of infrared spectra, a 16-mer β-hairpin peptide showed no features suggestive of inter-strand hydrogen bonds [54], whereas in another 16-mer β-hairpin peptide an amide I band at ~1,617 cm^{-1} was attributed to hydrogen bonding across the strands [149]. The fact that the rare protein 2:2 β-hairpins with the unsuitable type I β-turn usually exhibit longer strands than those with the appropriate type I′ β-turns might be explained by the additional hydrogen bonds compensating for the unfavorable type I β-turn [145].

1.3.8 Solvent, Salt, and pH Effects

Stabilization of β-hairpin and β-sheet structures by alcoholic co-solvents, such as methanol [30, 33, 42, 45, 50, 52, 55, 102, 113, 132, 150, 151] or trifluoroethanol (TFE; [28, 38, 42, 57, 58, 61, 64, 66, 67, 112, 152]), has been reported in many peptide systems, even though TFE is generally considered a helix-inducing co-solvent. A 20-mer peptide encompassing the N-terminal region of ferredoxin I which is not structured in aqueous solution adopts a native-like β-hairpin in the presence of either methanol or TFE [153]. This β-hairpin is also stabilized in sodium dodecyl sulfate micelles [154]. This is the only peptide reported to adopt a 2:2 β-hairpin with a type I β-turn, which are rare in 2:2 protein β-hairpins because of the inadequacy of the geometric parameters of type I β-turns to properly align the antiparallel strands.

A very short random coil peptide of sequence Ac-GAN**PN**AAG with its N- and C-termini modified by long alkyl tails has been shown to be stabilized in the presence of liposomes by insertion of the two aliphatic tails into the liposome [155].

β-Hairpin stability has been reported to be pH dependent in several β-hairpin peptides [42, 53, 57, 67, 131, 151]. In all cases, the pH dependence was caused by pH effects on electrostatics interactions (Subheading 1.3.3).

Salt additives can also affect β-hairpin stability as shown for ionic [67, 92] and/or indole interactions in W/W pairs [156].

2 Methods

The protocol proposed here for the rational design of monomeric water-soluble β-sheet peptides consists of the following steps:

Step 1. Selection of goal β-sheet and peptide length (Subheading 2.1).

Step 2. Selection of β-turn sequence (Subheading 2.2).

Step 3. Selection of β-strand residues and application of solubility criteria (Subheadings 2.3 and 2.4).

Step 4. "In silico" validation of the sequence resulting from **steps 1** to **3** (Subheading 2.6).

Step 5. Peptide preparation.

Step 6. Experimental validation of the designed β-sheet (Subheading 2.7).

The two structurally different regions that constitute a β-hairpin (Figs. 1 and 2; Subheading 1.1), the β-turn (**step 2**) and the two antiparallel β-strands (**step 3**), are considered independently. **Step 4** was not performed for most reported β-sheet-forming peptides. Methods to carry out peptide preparation (**step 5**) either by chemical synthesis or by biotechnological methods are beyond the scope of the current review. From the design perspective, the procedure employed for the preparation of the designed peptide is important only in respect to the possible protection of peptide ends or the incorporation of nonnatural amino acids.

2.1 Selection of the Goal β-Sheet Structure and Peptide Length

The design of a protein or peptide structure aims to get a sequence able to adopt a selected goal structure. The choice of this structure is the starting point in any design project. The principles described in the following sections are applicable for β-hairpins with short loops and can be extended to three-stranded antiparallel β-sheets with β1–β2–β3 topology (Fig. 4). To redesign a protein domain or fragment of known β-sheet structure (for example, to improve certain structural characteristics of a natural biologically active sequence; Subheading 1.2.4), the guidelines given in Subheadings 2.2–2.4 can be applied by maintaining the residues important for the biological activity, and substituting some other residues so as to improve its structural stability. Sequence alignment procedures are very useful in these redesign cases (Subheading 2.5).

The length of a β-hairpin peptide is $n + c + t$, where n and c are the number of residues at N- and C-terminal strands, respectively, and t the number of residues in the turn (Fig. 1). This last value depends on the type of β-hairpin, being two in 2:2 β-hairpins, three in 3:5 β-hairpins, and four in 4:4 and 4:6 β-hairpins (Subheading 1.1; Fig. 1). If n and c are different, residues at the end of the longest strand are not paired. Strand length in the reported β-hairpin peptides ranges from two up to nine, most of them having three to seven strand residues. In terms of stability, the strand length increases stability up to seven residues, but no stability increment is observed beyond ([76]; Subheading 1.3.6). Statistical studies indicate that protein β-hairpins with even numbers of strand residues are more frequent than with odd numbers, and that most of them terminate at a non-HB site [157]. Among the reported β-hairpin peptides, some fulfil these preferences, but others do not (*see* Table 2).

To design three-stranded antiparallel β-sheets with a meander β1–β2–β3 topology (Fig 4; Subheadings 1.1 and 1.2.3), **steps 2** (Subheading 2.2) and **3** (Subheading 2.3) of the proposed protocol are applied twice, once for the N-terminal hairpin and then again for the C-terminal hairpin (hairpins 1 and 2 in Fig. 4). The two β-hairpins can be of the same type and thus have the same number of turn residues, or of a different type. It is necessary to bear in mind that, in this meander antiparallel β-sheet motif (Fig. 4), the sequences of the C-terminal strand of hairpin 1 and the N-terminal strand of hairpin 2 are the same and that every residue at this middle strand, which is located at an HB site in hairpin 1, is at a non-HB site in hairpin 2, and vice versa (Fig. 3). The designed meander three-stranded β-sheets reported up to now (Table 3; Subheading 1.2.3) have four to six residues per strand.

2.2 Selection of β-Turn Residues

Since the turn region plays an essential role in determining β-hairpin conformation and its stability (Subheading 1.3), the selection of an adequate β-turn sequence is crucial to ensure that the designed peptide will adopt the target β-hairpin. Table 1, which has been built using β-turn positional potentials statistically derived from protein structures [32], is very useful to select the β-turn residues. The optimal turn sequence depends on the type of the desired β-hairpin. Thus, the preferred β-turns are types I′ or II′ (Fig. 3) for 2:2 β-hairpins, and type I for 3:5, 4:4, and 4:6 β-hairpins. The optimal sequences (Table 1), and also the most commonly present in designed 2:2 β-hairpins (Tables 2 and 3), are NG and DG for type I′ β-turns (NG better than DG; *see* **Note** 7), and dPG (the best), GN, and GS for type II′. In the case of 3:5 β-hairpins, it is convenient to have a G at the $i+2$ residue to promote a G1 β-bulge, since the most suitable loop for 3:5 β-hairpins is a type I β-turn with a G1 β-bulge (I+G1). Indeed, all reported 3:5 β-hairpins contain this G residue at their turn (Table 2); their most common sequences are PDG and SDG. The loop sequences in the reported 4:4 and 4:6 β-hairpins are quite diverse (Table 2), but statistical analysis of the GB family indicated the sequence NPATGK as an optimal loop sequence (Subheading 1.2.1). Residues N and K correspond to the positions adjacent to the turn ($n-1$ and $c+1$ in Fig. 1d), and P and G, respectively, at positions i and $i+4$ of the turn were considered the most important [38]. The final complete sequence of a designed β-hairpin must not only have an optimal β-turn sequence, but the design should also ensure the lack of sequences likely to form alternative β-turns.

2.3 Selection of β-Strand Residues

The following guidelines have to be considered to select β-strand residues:

2.3.1 Intrinsic β-Sheet Propensities

As a general rule, β-strands should contain residues with high intrinsic β-sheet propensities and lack residues with no β-sheet-forming tendency. Since intrinsic β-sheet propensities seem to be context dependent [106], the reported scales of β-sheet propensities show differences. Nevertheless, they offer useful guidance for the classification of residues into good and bad β-sheet formers. Thus, aromatic (Y, F, and W) and β-branched residues (V, I, and T) are good β-sheet formers in all the reported scales, both statistically [104] and experimentally [105–109]. Residues L, M, and S are slightly beneficial or neutral (neither good nor bad β-sheet formers), R and C are neutral, Q is neutral or slightly detrimental, and residues D, G, P, A, E, K, N, and H have a negative effect on β-sheet formation. The different scales differ in the rank order of the residues in each group (*see* **Note 8**).

2.3.2 Residues Adjacent to the Turn

Apart from having high intrinsic β-sheet propensities, the best residues to precede ($n-1$ in Fig. 1a, b) and follow ($c+1$ in Fig. 1a, b) the β-turn in 2:2 β-hairpins should also have high intrinsic probability to be at the positions i and $i+3$, respectively, of either a I′ or a II′ β-turn (Table 1). In the case of 3:5 β-hairpins (Fig. 1c), the best residue to precede the β-turn should have high intrinsic probability to be at the position i of a type I β-turn (Table 1). For 4:4 and 4:6 hairpins (Fig. 1d), pair residues found to be favorable at an HB site adjacent to the turn can also be taken into account (Subheadings 1.3.3 and 2.2).

2.3.3 Pair-Wise Cross-Strand Interactions

Facing residues at two antiparallel β-strands should correspond to pairs with favorable side chain/side chain interactions. The inclusion of pairs with unfavorable interactions should be avoided, except if essential for the target biological activity of the peptide (Subheading 1.2.4). The statistically most favorable pair-wise interactions [143] listed in Table 2 are a useful guide for selection, though the statistical data do not always coincide with the experimental results regarding β-sheet stability. Furthermore statistical analyses reported by different authors show some discrepancies. Other statistical data on pair-wise interactions, which are not included in Table 2, might also be used [26, 144, 145]. The cross-strand side chain/side chain interactions found to be stabilizing in model β-hairpin peptides are included in Table 2. It is convenient to have the most stabilizing pairs close to the turn [62, 92]. Experimentally, the facing pair W/W is the most β-hairpin stabilizing at non-HB sites (Subheading 1.3.3.2). A disulfide bond can also be incorporated into non-HB sites to stabilize a β-hairpin peptide (*see* Subheading 1.3.4 for other covalent linkers).

2.3.4 Diagonal Interactions

Diagonal interactions (Fig. 1c) also contribute to β-hairpin stability (Subheading 1.3.3). Table 2 lists preferred diagonal interactions according to statistical data [26] as well as those found experimentally and can be used as an aid to select favorable diagonal interactions.

2.3.5 Hydrophobic Clusters

Hydrophobic clusters at non-HB sites can also be incorporated to stabilize β-hairpin structures. It should be taken into account that their contribution to stability increases with their closeness to the turn region [74]. Also, residues that compose a good stabilizing hydrophobic cluster should have favorable pair-wise facing (Subheading 2.3.3) and diagonal interactions (Subheading 2.3.4). Up to now, the most prominent β-hairpin-stabilizing hydrophobic cluster is the Trpzip motif (W/W/W/W; Subheading 1.3.3).

2.3.6 Peptide Ends

Peptide ends can be free or protected by acetylation of the N-terminal amino group and by amidation of the C-terminal carboxylate group. If they are free and the two terminal amino acids face each other, the electrostatic interaction between the positively charged amino group and the negatively charged carboxylate makes a favorable contribution to β-hairpin stability [131]. To increase solubility, it can be advisable to protect one of the peptide ends (Subheading 2.4). The incorporation of a β-cap motif ([89]; Subheading 1.3.3) can also be considered to reduce fraying at the N- and C-termini.

Linking of the N- and C-termini to form cyclic peptides can lead to very stable β-hairpins [71–73, 75, 76]. The linkers have to be good β-turn-forming sequences (dPG has been mostly used). It has to be noted that the size of the cycle is important for β-hairpin stability [158]. Cyclic peptides with type I′ or II′ turns (two residues at each turn) and an even number of strand residues [1, 3, 5] adopt stable β-hairpins, but the β-hairpins in those with an odd number of strand residues [2, 4, 6] are unstable and distorted [158].

The options available to modify the peptide ends depend on the peptide preparation method. Thus, acetylation, amidation, and cyclization are easily achievable by peptide chemical synthesis, but not by cloning and expression of the peptide. However, to get ^{13}C- and/or ^{15}N-labelled peptide is less expensive by peptide expression. In this case, the peptides usually have some additional residues at the N-terminus, denoted as the cloning tag. This cloning tag is required to express peptides by the currently available Molecular Biology techniques [159].

2.4 Solubility Criteria (See Note 9)

Aggregation and solubility problems in peptides and proteins seem to be higher close to the pI (isoelectric point). A peptide with either a net positive charge or a net negative charge will likely be more water soluble. Incorporation of an electrostatic interaction in a β-hairpin resulted in peptide aggregation in one case [61]. Therefore, it can be advisable to protect the N-termini in D/E-containing peptides, and the C-termini in those containing positively charged residues (K, R, O). The distribution of the charged and polar residues also plays an important role for the solubility. Since amphipathic sheets with a hydrophilic face and a hydrophobic one are prone to aggregate, the charged and polar side chains should be pointing outwards from both faces of the antiparallel β-sheet. Another strategy to avoid self-association in designed β-hairpin peptides consists in the incorporation of charged residues at the peptide ends separated from the hairpin by spacing linkers consisting of G residues [58].

2.5 Alignment to Sequences That Adopt the Target β-Sheet Motif

Naturally occurring peptides or protein domains having the desired β-sheet motif or designed peptides previously reported to adopt the target structure can be used as the starting point for the design. In fact, many designed β-hairpin-forming peptides were derived from protein β-hairpins (Subheading 1.2.1). A consensus sequence [46] can be obtained from the alignment of peptides or protein domains that have the target structure and/or function.

2.6 "In Silico" Validation of Designed β-Sheet Sequences (See Note 10)

A β-hairpin prediction program developed in our laboratory (*see* **Note 11**) can be applied to evaluate the probability of the peptide sequences designed following the above guidelines to adopt the target β-sheet structure (Subheadings 2.1–2.5). The program contains two principal subroutines: One predicts the β-turn residues and the other deals with the β-strand residues. In a first step a normalized version of the β-turn positional potentials published by Hutchinson and Thornton [32] is employed to predict the residues most likely to form a β-turn and the type of turn they will adopt. In a second independent step, the prediction of the most favorable residues in β-sheet conformation is determined by a linear combination of terms derived from intrinsic β-sheet propensities [108], cross-strand pair interactions [143], and number of hydrogen bonds formed. Finally, the prediction of the β-hairpin type adopted by the amino acid sequence is a combination of both results, which are ranked numerically together according to rules based on the most favorable type of protein β-hairpins.

2.7 Experimental Check of the Design Success (See Note 12)

2.7.1 State of Association

The monomeric state of β-hairpin peptides is usually confirmed by analytical ultracentrifugation and by dilution experiments monitored by CD and/or NMR.

2.7.2 Fast Test of β-Hairpin Formation: Circular Dichroism Spectroscopy (CD) (See **Note 13**)

CD spectroscopy provides a quick way to confirm whether or not a designed peptide adopts a β-sheet structure, but no information about β-strand register. The characteristic far-UV CD spectra for β-sheets exhibit a minimum at ~216 nm and a maximum at ~195 nm [160]. In β-hairpins that contain facing edge-to-face W/W pairs at non-HB sites, the far-UV CD spectrum has a maximum at 227–229 nm and a minimum at 213–215 nm, and the near-UV CD presents well-defined bands around 290 nm (*see* [8] and references therein).

2.7.3 Determination of the β-Hairpin/β-Sheet Structure: NMR Spectroscopy (See ***Note 14****)*

NMR is the best technique to demonstrate that a particular peptide adopts its target β-sheet structure. Analysis of chemical shifts provides a fast way to qualitatively identify the formation of β-hairpins and three-stranded β-sheets, and NOEs provide unambiguous evidence about the strand register, and hence the type of β-hairpin, and also about the type of β-turn. The three-dimensional structure can also be determined from NMR parameters.

Thus, once the NMR signals are assigned (*see* **Note 15**), β-hairpin and β-sheet formation can be confirmed on the basis of the patterns of ^{1}Hα, ^{1}HN, ^{13}Cα, and ^{13}Cβ conformational shifts ($\Delta\delta = \delta^{\text{observed}} - \delta^{\text{random coil}}$, ppm; *see* **Note 16**). Thus, β-strands are delineated by the stretches of at least two consecutive residues having positive $\Delta\delta_{\text{H}}\alpha$, $\Delta\delta_{\text{HN}}$, and $\Delta\delta 13_{\text{C}}\beta$ values and negative $\Delta\delta 13_{\text{C}}\alpha$ values (*see* **Note 17**). Two of such stretches are observed in β-hairpins and three in three-stranded β-sheets. These stretches are separated by two to four residues that display a $\Delta\delta_{\text{H}}\alpha$ which is negative or very small in absolute value, and at least one residue with a positive $\Delta\delta 13_{\text{C}}\alpha$ value and a negative $\Delta\delta 13_{\text{C}}\beta$ value [161]. The particular type of β-hairpin and β-turn can be identified on the basis of the distinctive characteristic patterns of $\Delta\delta 13_{\text{C}}\alpha$ and $\Delta\delta 13_{\text{C}}\beta$ [161], and/or $\Delta\delta_{\text{HN}}$ [152] at the turn region.

Nevertheless, the NMR parameters that provide the strongest structural evidences are the ^{1}H-^{1}H NOEs (nuclear Overhauser effect), since an NOE correlation between two protons is observed only if they are spatially close (approximately less than 5.5 Å). The intensity of an NOE is inversely proportional to the sixth power of the proton-proton distance ($1/r^6$, where r is the proton–proton distance [162, 163]). In antiparallel β-sheets (Fig. 1a), the backbone protons that are close enough to give rise to NOEs are (1) the Hα protons of residues facing each other in a non-HB site (2.3 Å), (2) the NH amide protons of facing residues in HB sites (3.3 Å), and (3) the Hα protons in a non-HB site and NH protons in an HB site of residues in adjacent strands, when these two sites are consecutive (3.2 Å). Some Hα signals may be obscured by the residual water signal in the usually employed aqueous solution conditions (H_2O/D_2O 9:1 in volume; needed to observe backbone NH amide protons, and so required for assignment). Therefore, to observe Hα-Hα NOEs it is convenient to dissolve the peptide in D_2O, instead of the usual conditions. Further details on the structural features of the β-hairpin adopted can be gathered from the NOEs between protons of side chains located on the same β-sheet face.

In meander three-stranded antiparallel β-sheets (Fig. 4), the set of NOEs involving backbone protons is also compatible with the independent formation of β-hairpin 1 and β-hairpin 2. Only the observation of at least one long-range NOE involving side chain protons of residues at the N-terminal strand and at the C-terminal one demonstrates the formation of the three-stranded β-sheet motif [112].

In those peptides that adopt stable β-hairpins or β-sheets, the three-dimensional structure can be calculated from distance restraints derived from the complete set of observed NOEs by protocols similar to those used in proteins (*see* **Note 18**).

2.7.4 Quantification of β-Hairpin and β-Sheet Populations

In contrast to helical peptides where CD spectroscopy provides a good method for quantifying helix population, there is not a well-established method to quantify β-sheet populations. Assuming a two-state behavior for β-hairpin or β-sheet formation, populations can be evaluated from different NMR parameters, such as the intensity of Hα-Hα NOEs [57–60, 67, 103, 164], ^{1}Hα, ^{13}Cα, and ^{13}Cβ chemical shifts [58, 67–71, 73, 75, 76, 112, 161, 164], and the Gly splitting (chemical shift difference between the two Hα protons of a G residue at the turn region; [45, 85, 90, 91, 93–97, 113, 136, 137, 148, 165]). Apart from the validity of the two-state assumption, the absence of accurate reference values for the completely folded and random coil states limits the precision and accuracy of the quantification of β-sheet populations (*for details on this question see* [15, 18, 20, 21, 71, 152, 161]).

2.7.5 Determination of β-Hairpin Stability (See **Note 19***)*

In principle, the methods applicable to determine protein structure stability can be applied to measure β-hairpin stability. But, measurements and data analysis can be troublesome because the plateaux corresponding to fully folded or unfolded states cannot be reached at available experimental conditions, and also because β-hairpin unfolding can deviate from the two-state assumption. Despite this, thermal unfolding has been followed in many β-hairpin peptides by differential scanning calorimetry (DSC; [29, 166]), CD [38, 41, 46, 55, 79, 80, 85–91, 93–95, 113, 120, 147, 167–172]), FTIR [167–172]), and NMR chemical shifts [38, 42, 43, 46, 50, 52, 53, 63, 85, 87–91, 93–95, 112, 134, 136, 137, 165, 166]. In disulfide-cyclized peptides, as the CX$_8$C scaffold [77–79], β-hairpin stability can be measured by a non-spectroscopic method based on the changes in the thiol-disulfide equilibrium constant upon residue substitutions.

3 Notes

1. Non-water-soluble peptides that adopt β-hairpin structures in chloroform, benzene, dimethylsulfoxide (DMSO), and alcoholic solvents have also been designed. As in water-soluble peptides, the β-turn sequence and interactions between facing aromatic residues are important for β-hairpin stability. Thus, an apolar octapeptide with the β-turn sequence DPG adopts a 2:2

β-hairpin [173] that is destabilized by a dP-to-P change [174]. Substitution of the two-residue turn sequence by the three-residue sequence dPGdA led to a nonapeptide that forms a 3:5 β-hairpin with an altered strand register [175]. Cross-strand interactions between pairs of facing aromatic residues (Y/Y and Y/W) at non-HB sites stabilize β-hairpins also in organic solvents [176]. The stabilizing effect of these aromatic pairs is weaker at HB sites [176]. Some of these short and very hydrophobic β-hairpin peptides have been crystallized [177].

2. A designed three-stranded antiparallel β-sheet with a β2–β1–β3 topology has also been reported [178]. Several antiparallel β-sheets with more than three strands have also been designed: two four-stranded antiparallel β-sheet peptides, i.e., a 26-residue peptide that adopts the β-sheet in either pure methanol or in water–methanol solutions [179, 180] and a 50-residue molecule composed of two BPTI-derived β-hairpin modules that are connected by a cross-link between two Lys residues in the inner strands [181]; a 34-residue peptide that forms a five-stranded β-sheet and contains a metal-binding site [182]; and an eight β-stranded antiparallel β-sheet formed by connecting two four-stranded β-sheets with a disulfide bond [183]. Dimeric and tetrameric β-sheets [184] and a trimer composed of three β-hairpin modules [185] have also been designed.

3. *N*-methyl amino acids were incorporated in a designed three-stranded β-sheet to prevent aggregation [186]. A non-water-soluble three-stranded antiparallel β-sheet containing dPG sequences in the two turns has also been designed [179].

4. WW domains are small stable naturally occurring three-stranded β1–β2–β3 antiparallel β-sheets. The conclusions about β-sheet folding and stability obtained from studies on the WW domains [187] agree with those described here at Subheading 1.3. Thus, β-turn sequences play an essential role in stability and kinetics [188], and a W/W pair at a non-HB site increases stability in a 34-mer WW domain [189].

5. Designed β-hairpin peptides that self-assemble to form hydrogel materials are being intensely studied by Schneider's group [190].

6. For deeper discussion on the physical-chemical origin of the contributions to β-hairpin and β-sheet stability see reviews [8–10, 12–22] and references therein and here.

7. The N residue in the NG sequence is relatively prone to be de-amidated yielding D and isoAspartate (isoD). The β-hairpin formed by the 12-mer peptide derived from met repressor (Table 2) is strongly destabilized by the substitution of N for isoD [151].

8. According to statistical data [104], the residues with high intrinsic β-sheet propensities are V > I > T > Y > W > F > L, while C > M > Q > S > R are more or less neutral to adopt β-sheet

conformations, and in the experimental data reported by Smith et al. [108], the β-sheet favorable residues are Y > T > I > F > W > V and the neutral ones are S > M > C > L > R.

9. Programs that predict the tendency of a given sequence to aggregate may be used to screen out sequences with the strongest tendency to self-associate.

10. "In silico" validation of the designed sequences could also be performed by applying methods developed for predicting β-turns in a protein from its amino acid sequence, such as that available at web server http://www.imtech.res.in/raghava/betatpred2 [191], or for the recognition of β-hairpins in proteins [192]. However, so far these methods have not been used for the design of any reported β-hairpin or β-sheet peptide.

11. BEHAIRPRED (D. Pantoja-Uceda and M. A. Jiménez, unpublished) is accessible at web server http://triton.rmn.iqfr.csic.es/software/behairpredv1.0/behairpred.htm. It accepts the nonnatural residues O (ornithine) and dP as input. In addition, it is able to identify type II β-turns, even though they are quite uncommon in β-hairpins.

12. β-Hairpin stability has been shown to confer resistance to enzymatic degradation [193].

13. Other spectroscopic techniques such as Fourier-transform infrared (FT-IR) and fluorescence are being used to study β-hairpin structures in peptides [54, 81–84, 149, 168, 169, 194].

14. Crystal structures have been reported for a water-soluble 10-mer β-hairpin peptide [47] and for apolar peptides (*see* **Note 1**).

15. NMR signals of peptides are assigned by following the standard sequential strategy developed by Wüthrich [162, 163]. NMR assignment is complicated by repetitive sequences.

16. ^{1}Hα, ^{1}HN, ^{13}Cα, and ^{13}Cβ conformational shifts or chemical shift deviations ($\Delta\delta_{H}\alpha$, $\Delta\delta_{HN}$, $\Delta\delta_{C}\alpha$, and $\Delta\delta_{C}\beta$, respectively) are defined as the deviation of the experimentally measured chemical shift ($\delta^{observed}$, ppm) from reference δ values for the random coil state ($\delta^{random\ coil}$, ppm; [152, 161]). The best way to obtain $\Delta\delta_{HN}$ values is by using the CSDb2 program (http://andersenlab.chem.washington.edu/CSDb/; [152]).

17. The pattern of $\Delta\delta_{H}\alpha$ is sometimes distorted because the chemical shifts of Hα protons are affected not only by the ϕ and ψ angles, but also by context-dependent effects related to the occupied position: edge versus central strand in three-stranded antiparallel β-sheet motifs, and HB versus non-HB in β-hairpins or edge strands [195]. Also, the Hα proton of a residue that faces an aromatic residue (W, Y, F, H) can be shifted upfield (negative $\Delta\delta_{H}\alpha$ value instead of the positive value characteristic of β-strands; [61, 62, 64, 65]), because of the anisotropy

effects of the aromatic ring currents. Concerning $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts, they reflect the diagnostic $\Delta\delta 13_C\alpha$ and $\Delta\delta 13_C\beta$ values better at HB than at non-HB sites [196], and can also be affected by $\chi 1$ angles [197].

18. For details, see the procedures for the structure calculation in recently reported β-hairpin peptides [44, 65, 133].

19. Folding kinetics of several β-hairpin and β-sheet peptides have also been studied by fitting NMR signal broadening [103, 198, 199], and by temperature-jump coupled to IR [167–171, 200] and fluorescence [200–202].

Acknowledgements

Financial support from Spanish MINECO (CTQ2011-22514) and Comunidad de Madrid (S2010/BMD-2305) projects is acknowledged.

References

1. Blanco FJ, Jimenez MA, Herranz J, Rico M, Santoro J, Nieto JL (1993) NMR evidence of a short linear peptide that folds into a beta-hairpin in aqueous solution. J Am Chem Soc 115:5887–5888
2. Nowick JS (2008) Exploring beta-sheet structure and interactions with chemical model systems. Acc Chem Res 41:1319–1330
3. Khakshoor O, Nowick JS (2008) Artificial beta-sheets: chemical models of beta-sheets. Curr Opin Chem Biol 12:722–729
4. Freire F, Gellman SH (2009) Macrocyclic design strategies for small, stable parallel beta-sheet scaffolds. J Am Chem Soc 131:7970–7972
5. Liang H, Chen H, Fan K, Wei P, Guo X, Jin C et al (2009) De novo design of a beta alpha beta motif. Angew Chem Int Ed Engl 48: 3301–3303
6. Robinson JA (2008) Beta-hairpin peptidomimetics: design, structures and biological activities. Acc Chem Res 41:1278–1288
7. Fuller AA, Du D, Liu F, Davoren JE, Bhabha G, Kroon G et al (2009) Evaluating beta-turn mimics as beta-sheet folding nucleators. Proc Natl Acad Sci U S A 106:11067–11072
8. Santiveri CM, Jimenez MA (2010) Tryptophan residues: scarce in proteins but strong stabilizers of beta-hairpin peptides. Biopolymers 94: 779–790
9. Lewandowska A, Oldziej S, Liwo A, Scheraga HA (2010) beta-hairpin-forming peptides; models of early stages of protein folding. Biophys Chem 151:1–9
10. Hughes RM, Waters ML (2006) Model systems for beta-hairpins and beta-sheets. Curr Opin Struct Biol 16:514–512
11. Pantoja-Uceda D, Santiveri CM, Jimenez MA (2006) De novo design of monomeric beta-hairpin and beta-sheet peptides. Methods Mol Biol 340:27–51
12. Searle MS, Ciani B (2004) Design of β-sheet systems for understanding the thermodynamics and kinetics of protein folding. Curr Opin Struct Biol 14:458–464
13. Searle MS (2004) Insights into stabilizing weak interactions in designed peptide beta-hairpins. Biopolymers 76:185–195
14. Stotz CE, Topp EM (2004) Applications of model beta-hairpin peptides. J Pharm Sci 93: 2881–2894
15. Searle MS (2001) Peptide models of protein β-sheets: design, folding and insights into stabilising weak interactions. J Chem Soc Perkin Trans 2:1011–1020
16. Venkatraman J, Shankaramma SC, Balaram P (2001) Design of folded peptides. Chem Rev 101:3131–3152
17. Serrano L (2000) The relationship between sequence and structure in elementary folding units. Adv Protein Chem 53:49–85
18. Lacroix E, Kortemme T, Lopez de la Paz M, Serrano L (1999) The design of linear peptides that fold as monomeric beta-sheet structures. Curr Opin Struct Biol 9:487–493
19. Ramirez-Alvarado M, Kortemme T, Blanco FJ, Serrano L (1999) Beta-hairpin and beta-sheet

formation in designed linear peptides. Bioorg Med Chem 7:93–103

20. Blanco F, Ramirez-Alvarado M, Serrano L (1998) Formation and stability of beta-hairpin structures in polypeptides. Curr Opin Struct Biol 8:107–111
21. Gellman SH (1998) Minimal model systems for beta sheet secondary structure in proteins. Curr Opin Chem Biol 2:717–725
22. Smith CK, Regan L (1997) Construction and design of beta sheets. Acc Chem Res 30: 153–161
23. Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167–339
24. Sibanda BL, Blundell TL, Thornton JM (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. J Mol Biol 206:759–777
25. Rose GD, Gierasch LM, Smith JA (1985) Turns in peptides and proteins. Adv Protein Chem 37:1–109
26. Cootes AP, Curmi PM, Cunningham R, Donnelly C, Torda AE (1998) The dependence of amino acid pair correlations on structural environment. Proteins 32:175–189
27. Blanco FJ, Rivas G, Serrano L (1994) A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. Nat Struct Biol 1:584–590
28. Skwierawska A, Rodziewicz-Motowidlo S, Oldziej S, Liwo A, Scheraga HA (2008) Conformational studies of the alpha-helical 28-43 fragment of the B3 domain of the immunoglobulin binding protein G from Streptococcus. Biopolymers 89:1032–1044
29. Skwierawska A, Makowska J, Oldziej S, Liwo A, Scheraga HA (2009) Mechanism of formation of the C-terminal beta-hairpin of the B3 domain of the immunoglobulin binding protein G from Streptococcus. I. Importance of hydrophobic interactions in stabilization of beta-hairpin structure. Proteins 75:931–953
30. Zerella R, Evans PA, Ionides JM, Packman LC, Trotter BW, Mackay JP et al (1999) Autonomous folding of a peptide corresponding to the N-terminal beta-hairpin from ubiquitin. Protein Sci 8:1320–1331
31. Mei CG, Jahr N, Singer D, Berger S (2011) Hairpin conformation of an 11-mer peptide. Bioorg Med Chem 19:3497–3501
32. Hutchinson EG, Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. Protein Sci 3:2207–2216
33. Searle MS, Williams DH, Packman LC (1995) A short linear peptide derived from the N-terminal sequence of ubiquitin folds into a water-stable non-native beta-hairpin. Nat Struct Biol 2:999–1006
34. Simpson ER, Meldrum JK, Bofill R, Crespo MD, Holmes E, Searle MS (2005) Engineering enhanced protein stability through beta-turn optimization: insights for the design of stable peptide beta-hairpin systems. Angew Chem Int Ed 44:4939–4944
35. Haque TS, Gellman SH (1997) Insights into β-hairpin stability in aqueous solution from peptides with enforced type I′ and type II′ β-turns. J Am Chem Soc 119:2303–2304
36. Zerella R, Chen PY, Evans PA, Raine A, Williams DH (2000) Structural characterization of a mutant peptide derived from ubiquitin: implications for protein folding. Protein Sci 9:2142–2150
37. Chen PY, Gopalacushina BG, Yang CC, Chan SI, Evans PA (2001) The role of a beta-bulge in the folding of the beta-hairpin structure in ubiquitin. Protein Sci 10:2063–2074
38. Fesinmeyer RM, Hudson FM, Andersen NH (2004) Enhanced hairpin stability through loop design: the case of the protein G B1 domain hairpin. J Am Chem Soc 126:7238–7243
39. Espinosa JF, Syud FA, Gellman SH (2005) An autonomously folding beta-hairpin derived from the human YAP65 WW domain: attempts to define a minimum ligand-binding motif. Biopolymers 80:303–311
40. Carulla N, Woodward C, Barany G (2000) Synthesis and characterization of a beta-hairpin peptide that represents a ‘core module’ of bovine pancreatic trypsin inhibitor (BPTI). Biochemistry 39:7927–7937
41. Cochran AG, Skelton NJ, Starovasnik MA (2001) Tryptophan zippers: stable, monomeric beta-hairpins. Proc Natl Acad Sci U S A 98: 5578–5583
42. Huyghues-Despointes BM, Qu X, Tsai J, Scholtz JM (2006) Terminal ion pairs stabilize the second beta-hairpin of the B1 domain of protein G. Proteins 63:1005–1017
43. Wei Y, Huyghues-Despointes BM, Tsai J, Scholtz JM (2007) NMR study and molecular dynamics simulations of optimized beta-hairpin fragments of protein G. Proteins 69:285–296
44. Mirassou Y, Santiveri CM, Perez de Vega MJ, Gonzalez-Muniz R, Jimenez MA (2009) Disulfide bonds versus TrpTrp pairs in irregular beta-hairpins: NMR structure of vammin loop 3-derived peptides as a case study. Chembiochem 10:902–910
45. Riemen AJ, Waters ML (2008) Stabilization of the N-terminal beta-hairpin of ubiquitin by a terminal hydrophobic cluster. Biopolymers 90: 394–398

46. Honda S, Yamasaki K, Sawada Y, Morii H (2004) 10 residue folded peptide designed by segment statistics. Structure 12:1507–1518
47. Honda S, Akiba T, Kato YS, Sawada Y, Sekijima M, Ishimura M et al (2008) Crystal structure of a ten-amino acid protein. J Am Chem Soc 130:15327–15331
48. Hatfield MP, Murphy RF, Lovas S (2011) The CLN025 decapeptide retains a beta-hairpin conformation in urea and guanidinium chloride. J Phys Chem B 115:4971–4981
49. Maynard AJ, Searle MS (1997) NMR structural analysis of a β-hairpin peptide designed for DNA binding. Chem Commun 19:1297–1298
50. Maynard AJ, Sharman GJ, Searle MS (1998) Origin of β-hairpin stability in solution: structural and thermodynamic analysis of the folding of a model peptide supports hydrophobic stabilization in water. J Am Chem Soc 120: 1996–2007
51. Griffiths-Jones SR, Sharman GJ, Maynard AJ, Searle MS (1998) Modulation of intrinsic phi, psi propensities of amino acids by neighbouring residues in the coil regions of protein structures: NMR analysis and dissection of a beta-hairpin peptide. J Mol Biol 284:1597–1609
52. Griffiths-Jones SR, Maynard AJ, Searle MS (1999) Dissecting the stability of a beta-hairpin peptide that folds in water: NMR and molecular dynamics analysis of the beta-turn and beta-strand contributions to folding. J Mol Biol 292:1051–1069
53. Searle MS, Griffiths-Jones SR, Skinner-Smith H (1999) Energetics of weak interactions in a β-hairpin peptide: electrostatic and hydrophobic contributions to stability from lysine salt bridges. J Am Chem Soc 121:11615–11620
54. Colley CS, Griffiths-Jones SR, George MW, Searle MS (2000) Do interstrand hydrogen bonds contribute to b-hairpin stability in solution? IR analysis of peptide folding in water. Chem Commun 7:593–594
55. Ciani B, Jourdan M, Searle MS (2003) Stabilization of beta-hairpin peptides by salt bridges: role of preorganization in the energetic contribution of weak interactions. J Am Chem Soc 125:9038–9047
56. Gokhale A, Weldeghiorghis TK, Taneja V, Satyanarayanajois SD (2011) Conformationally constrained peptides from CD2 to modulate protein–protein interactions between CD2 and CD58. J Med Chem 54:5307–5319
57. de Alba E, Jimenez MA, Rico M, Nieto JL (1996) Conformational investigation of designed short linear peptides able to fold into beta-hairpin structures in aqueous solution. Fold Des 1:133–144
58. Ramirez-Alvarado M, Blanco FJ, Serrano L (1996) De novo design and structural analysis of a model beta-hairpin peptide system. Nat Struct Biol 3:604–612
59. de Alba E, Jimenez MA, Rico M (1997) Turn residue sequence determines beta-hairpin conformation in designed peptides. J Am Chem Soc 119:175–183
60. de Alba E, Rico M, Jimenez MA (1997) Cross-strand side-chain interactions versus turn conformation in beta-hairpins. Protein Sci 6: 2548–2560
61. de Alba E, Rico M, Jimenez MA (1999) The turn sequence directs beta-strand alignment in designed beta-hairpins. Protein Sci 8: 2234–2244
62. Santiveri CM, Rico M, Jimenez MA (2000) Position effect of cross-strand side-chain interactions on β-hairpin formation. Protein Sci 9: 2151–2160
63. Santiveri CM, Santoro J, Rico M, Jimenez MA (2002) Thermodynamic analysis of β-hairpin-forming peptides from the thermal dependence of ^{1}H NMR chemical shifts. J Am Chem Soc 124:14903–14909
64. Santiveri CM, Pantoja-Uceda D, Rico M, Jimenez MA (2005) Beta-hairpin formation in aqueous solution and in the presence of trifluoroethanol: a ^{1}H and ^{13}C nuclear magnetic resonance conformational study of designed peptides. Biopolymers 79:150–162
65. Santiveri CM, Leon E, Rico M, Jimenez MA (2008) Context-dependence of the contribution of disulfide bonds to beta-hairpin stability. Chemistry 14:488–499
66. Ramirez-Alvarado M, Blanco FJ, Niemann H, Serrano L (1997) Role of beta-turn residues in beta-hairpin formation and stability in designed peptides. J Mol Biol 273:898–912
67. Ramirez-Alvarado M, Blanco FJ, Serrano L (2001) Elongation of the BH8 beta-hairpin peptide: electrostatic interactions in beta-hairpin formation and stability. Protein Sci 10: 1381–1392
68. Pastor MT, Lopez de la Paz M, Lacroix E, Serrano L, Perez-Paya E (2002) Combinatorial approaches: a new tool to search for highly structured beta-hairpin peptides. Proc Natl Acad Sci U S A 99:614–619
69. Pastor MT, Gimenez-Giner A, Perez-Paya E (2005) The role of an aliphatic-aromatic interaction in the stabilization of a model beta-hairpin peptide. Chembiochem 6:1753–1756
70. Stanger HE, Gellman SH (1998) Rules for antiparallel β-sheet design:d-Pro-Gly is superior to l-Asn-Gly for β-hairpin nucleation. J Am Chem Soc 120:4236–4237

71. Syud FA, Espinosa JF, Gellman SH (1999) NMR-based quantification of beta-sheet populations in aqueous solution through use of reference peptides for the folded and unfolded states. J Am Chem Soc 121:11578–11579

72. Espinosa JF, Gellman SH (2000) A designed beta-hairpin containing a natural hydrophobic cluster. Angew Chem Int Ed 39:2330–2333

73. Syud FA, Stanger HE, Gellman SH (2001) Interstrand side chain–side chain interactions in a designed beta-hairpin: significance of both lateral and diagonal pairings. J Am Chem Soc 123:8667–8677

74. Espinosa JF, Munoz V, Gellman SH (2001) Interplay between hydrophobic cluster and loop propensity in β-hairpin formation. J Mol Biol 306:397–402

75. Espinosa JF, Syud FA, Gellman SH (2002) Analysis of the factors that stabilize a designed two-stranded antiparallel beta-sheet. Protein Sci 11:1492–1505

76. Stanger HE, Syud FA, Espinosa JF, Giriat I, Muir T, Gellman SH (2001) Length-dependent stability and strand length limits in antiparallel β-sheet secondary structure. Proc Natl Acad Sci U S A 98:12015–12020

77. Russell S, Cochran AG (2000) Designing stable β-hairpins: energetics contributions from cross-strand residues. J Am Chem Soc 122: 12600–12601

78. Cochran AG, Tong RT, Starovasnik MA, Park EJ, McDowell RS, Theaker JE et al (2001) A minimal peptide scaffold for beta-turn display: optimizing a strand position in disulfide-cyclized beta-hairpins. J Am Chem Soc 123:625–632

79. Russell SJ, Blandl T, Skelton NJ, Cochran AG (2003) Stability of cyclic beta-hairpins: asymmetric contributions from side chains of a hydrogen-bonded cross-strand residue pair. J Am Chem Soc 125:388–395

80. Blandl T, Cochran AG, Skelton NJ (2003) Turn stability in beta-hairpin peptides: investigation of peptides containing 3:5 type I G1 bulge turns. Protein Sci 12:237–247

81. Takekiyo T, Wu L, Yoshimura Y, Shimizu A, Keiderling TA (2009) Relationship between hydrophobic interactions and secondary structure stability for Trpzip beta-hairpin peptides. Biochemistry 48:1543–1552

82. Wu L, McElheny D, Huang R, Keiderling TA (2009) Role of tryptophan-tryptophan interactions in Trpzip beta-hairpin formation, structure, and stability. Biochemistry 48:10362–10371

83. Wu L, McElheny D, Takekiyo T, Keiderling TA (2010) Geometry and efficacy of cross-strand Trp/Trp, Trp/Tyr, and Tyr/Tyr aromatic interaction in a beta-hairpin peptide. Biochemistry 49:4705–4714

84. Wu L, McElheny D, Setnicka V, Hilario J, Keiderling TA (2012) Role of different beta-turns in beta-hairpin conformation and stability studied by optical spectroscopy. Proteins 80:44–60

85. Andersen NH, Olsen KA, Fesinmeyer RM, Tan X, Hudson FM, Eidenschink LA et al (2006) Minimization and optimization of designed beta-hairpin folds. J Am Chem Soc 128: 6101–6110

86. Kier BL, Andersen NH (2008) Probing the lower size limit for protein-like fold stability: ten-residue microproteins with specific, rigid structures in water. J Am Chem Soc 130: 14675–14683

87. Eidenschink L, Kier BL, Huggins KN, Andersen NH (2009) Very short peptides with stable folds: building on the interrelationship of Trp/Trp, Trp/cation, and Trp/backbone-amide interaction geometries. Proteins 75:308–322

88. Eidenschink L, Crabbe E, Andersen NH (2009) Terminal sidechain packing of a designed beta-hairpin influences conformation and stability. Biopolymers 91:557–564

89. Kier BL, Shu I, Eidenschink LA, Andersen NH (2010) Stabilizing capping motif for beta-hairpins and sheets. Proc Natl Acad Sci U S A 107:10466–10471

90. Tatko CD, Waters ML (2002) Selective aromatic interactions in beta-hairpin peptides. J Am Chem Soc 124:9372–9373

91. Tatko CD, Waters ML (2003) The geometry and efficacy of cation-pi interactions in a diagonal position of a designed beta-hairpin. Protein Sci 12:2443–2452

92. Kiehna SE, Waters ML (2003) Sequence dependence of beta-hairpin structure: comparison of a salt bridge and an aromatic interaction. Protein Sci 12:2657–2667

93. Tatko CD, Waters ML (2004) Investigation of the nature of the methionine-pi interaction in beta-hairpin peptide model systems. Protein Sci 13:2515–2522

94. Tatko CD, Waters ML (2004) Comparison of C-H…pi and hydrophobic interactions in a beta-hairpin peptide: impact on stability and specificity. J Am Chem Soc 126:2028–2034

95. Tatko CD, Waters ML (2004) Effect of halogenation on edge-face aromatic interactions in a beta-hairpin peptide: enhanced affinity with iodo-substituents. Org Lett 6:3969–3972

96. Riemen AJ, Waters ML (2009) Controlling peptide folding with repulsive interactions between phosphorylated amino acids and tryptophan. J Am Chem Soc 131:14081–14087

97. Riemen AJ, Waters ML (2010) Positional effects of phosphoserine on beta-hairpin stability. Org Biomol Chem 8:5411–5417

98. Meyer D, Mutschler C, Robertson I, Batt A, Tatko C (2013) Aromatic interactions with naphthylalanine in a beta-hairpin peptide. J Pept Sci 19:277–282

99. Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8:52–56

100. Kortemme T, Ramirez-Alvarado M, Serrano L (1998) Design of a 20-amino acid, three-stranded beta-sheet protein. Science 281: 253–256

101. Schenck HL, Gellman SH (1998) Use of a designed triple-stranded antiparallel β-sheet to probe β-sheet cooperativity in aqueous solution. J Am Chem Soc 120:4869–4870

102. Sharman GJ, Searle MS (1998) Cooperative interaction between the three strands of a designed antiparallel beta-sheet. J Am Chem Soc 120:5291–5300

103. de Alba E, Santoro J, Rico M, Jimenez MA (1999) De novo design of a monomeric three-stranded antiparallel beta-sheet. Protein Sci 8:854–865

104. Fasman GD (1989) The development of the prediction of protein structure. In: Fasman GD (ed) Prediction of protein structure and the principles of protein conformation. Plenum, New York, NY, pp 193–316

105. Kim CA, Berg JM (1993) Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide. Nature 362:267–270

106. Minor DL Jr, Kim PS (1994) Context is a major determinant of beta-sheet propensity. Nature 371:264–267

107. Minor DL Jr, Kim PS (1994) Measurement of the beta-sheet-forming propensities of amino acids. Nature 367:660–663

108. Smith CK, Withka JM, Regan L (1994) A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. Biochemistry 33:5510–5517

109. Street AG, Mayo SL (1999) Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. Proc Natl Acad Sci U S A 96: 9074–9076

110. Lopez de la Paz M, Lacroix E, Ramirez-Alvarado M, Serrano L (2001) Computer-aided design of beta-sheet peptides. J Mol Biol 312:229–246

111. Fernandez-Escamilla AM, Ventura S, Serrano L, Jimenez MA (2006) Design and NMR conformational study of a beta-sheet peptide based on Betanova and WW domains. Protein Sci 15:2278–2289

112. Santiveri CM, Santoro J, Rico M, Jimenez MA (2004) Factors involved in the stability of isolated β-sheets: turn sequence, β-sheet twisting, and hydrophobic surface burial. Protein Sci 13:1134–1147

113. Griffiths-Jones SR, Searle MS (2000) Structure, folding, and energetics of cooperative interactions between beta-strands of a de novo designed three-stranded antiparallel beta sheet peptide. J Am Chem Soc 122: 8350–8356

114. Butterfield SM, Waters ML (2003) A designed beta-hairpin peptide for molecular recognition of ATP in water. J Am Chem Soc 125: 9580–9581

115. Butterfield SM, Sweeney MM, Waters ML (2005) The recognition of nucleotides with model beta-hairpin receptors: investigation of critical contacts and nucleotide selectivity. J Org Chem 70:1105–1114

116. Butterfield SM, Goodman CM, Rotello VM, Waters ML (2004) A peptide flavoprotein mimic: flavin recognition and redox potential modulation in water by a designed beta hairpin. Angew Chem Int Ed 43:724–727

117. Butterfield SM, Cooper WJ, Waters ML (2005) Minimalist protein design: a beta-hairpin peptide that binds ssDNA. J Am Chem Soc 127:24–25

118. Stewart AL, Waters ML (2009) Structural effects on ss- and dsDNA recognition by a beta-hairpin peptide. Chembiochem 10:539–544

119. Cline LL, Waters ML (2009) Design of a beta-hairpin peptide-intercalator conjugate for simultaneous recognition of single stranded and double stranded regions of RNA. Org Biomol Chem 7:4622–4630

120. Stewart AL, Park JH, Waters ML (2011) Redesign of a WW domain peptide for selective recognition of single-stranded DNA. Biochemistry 50:2575–2584

121. Wilger DJ, Park JH, Hughes RM, Cuellar ME, Waters ML (2011) Induced-fit binding of a polyproline helix by a beta-hairpin peptide. Angew Chem Int Ed 50:12201–12204

122. Platt GW, Chung C-W, Searle MS (2001) Design of histidin-Zn^{2+} binding sites within a β-hairpin peptide: enhancement of β-sheet stability through metal complexation. Chem Commun 13:1162–1163

123. Ramadan D, Cline DJ, Bai S, Thorpe C, Schneider JP (2007) Effects of As(III) binding on beta-hairpin structure. J Am Chem Soc 129:2981–2988

124. Maestro B, Santiveri CM, Jimenez MA, Sanz JM (2011) Structural autonomy of a beta-hairpin peptide derived from the pneumococcal choline-binding protein LytA. Protein Eng Des Sel 24:113–122

125. Sibert RS, Josowicz M, Barry BA (2010) Control of proton and electron transfer in de novo designed, biomimetic beta hairpins. ACS Chem Biol 5:1157–1168

126. Diana D, Basile A, De Rosa L, Di Stasi R, Auriemma S, Arra C et al (2011) beta-hairpin peptide that targets vascular endothelial growth factor (VEGF) receptors: design, NMR characterization, and biological activity. J Biol Chem 286:41680–41691

127. Chen PY, Lin CK, Lee CT, Jan H, Chan SI (2001) Effects of turn residues in directing the formation of the beta-sheet and in the stability of the beta-sheet. Protein Sci 10: 1794–1800

128. Santiveri CM, Rico M, Jimenez MA, Pastor MT, Perez-Paya E (2003) Insights into the determinants of β-sheet stability: ^{1}H and ^{13}C NMR conformational investigation of three-stranded antiparallel β-sheet-forming peptides. J Pept Res 61:177–188

129. Merkel JS, Regan L (1998) Aromatic rescue of glycine in beta sheets. Fold Des 3: 449–455

130. Syud FA, Stanger HE, Mortell HS, Espinosa JF, Fisk JD, Fry CG et al (2003) Influence of strand number on antiparallel beta-sheet stability in designed three- and four-stranded beta-sheets. J Mol Biol 326:553–568

131. de Alba E, Blanco FJ, Jimenez MA, Rico M, Nieto JL (1995) Interactions responsible for the pH dependence of the beta-hairpin conformational population formed by a designed linear peptide. Eur J Biochem 233:283–292

132. Dhanasekaran M, Prakash O, Gong YX, Baures PW (2004) Expected and unexpected results from combined beta-hairpin design elements. Org Biomol Chem 2:2071–2082

133. Santiveri CM, Perez de Vega MJ, Gonzalez-Muniz R, Jimenez MA (2011) Trp-Trp pairs as beta-hairpin stabilisers: hydrogen-bonded versus non-hydrogen-bonded sites. Org Biomol Chem 9:5487–5492

134. Kobayashi N, Honda S, Yoshii H, Munekata E (2000) Role of side-chains in the cooperative beta-hairpin folding of the short C-terminal fragment derived from streptococcal protein G. Biochemistry 39:6564–6571

135. Phillips ST, Piersanti G, Bartlett PA (2005) Quantifying amino acid conformational preferences and side-chain-side-chain interactions in beta-hairpins. Proc Natl Acad Sci U S A 102:13737–13742

136. Hughes RM, Waters ML (2006) Arginine methylation in a beta-hairpin peptide: implications for Arg-pi interactions, DeltaCp(o), and the cold denatured state. J Am Chem Soc 128:12735–12742

137. Riemen AJ, Waters ML (2009) Design of highly stabilized beta-hairpin peptides through cation-pi interactions of lysine and n-methyllysine with an aromatic pocket. Biochemistry 48:1525–1531

138. Hughes RM, Waters ML (2005) Influence of N-methylation on a cation-pi interaction produces a remarkably stable beta-hairpin peptide. J Am Chem Soc 127:6518–6519

139. Betz SF (1993) Disulfide bonds and the stability of globular proteins. Protein Sci 2:1551–1558

140. Pace C (1990) Conformational stability of globular proteins. Trends Biochem Sci 15: 14–17

141. Pace C, Grimsley G, Thomson B, Barnett J (1988) Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. J Biol Chem 263: 11820–11825

142. Andreu D, Rivas L (1998) Animal antimicrobial peptides: an overview. Biopolymers 47: 415–433

143. Wouters MA, Curmi PM (1995) An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. Proteins 22:119–131

144. Hutchinson EG, Sessions RB, Thornton JM, Woolfson DN (1998) Determinants of strand register in antiparallel beta-sheets of proteins. Protein Sci 7:2287–2300

145. Gunasekaran K, Ramakrishnan C, Balaram P (1997) β-Hairpins in proteins revisited: lessons for de novo design. Protein Eng 10: 1131–1141

146. Garcia-Aranda MI, Mirassou Y, Gautier B, Martin-Martinez M, Inguimbert N, Vidal M et al (2011) Disulfide and amide-bridged cyclic peptide analogues of the VEGF(81–91) fragment: synthesis, conformational analysis and biological evaluation. Bioorg Med Chem 19:7526–7533

147. Park JH, Waters ML (2013) Positional effects of click cyclization on beta-hairpin structure, stability, and function. Org Biomol Chem 11: 69–77

148. Celentano V, Diana D, De Rosa L, Romanelli A, Fattorusso R, D'Andrea LD (2012) beta-Hairpin stabilization through an interstrand triazole bridge. Chem Commun 48:762–764

149. Arrondo JL, Blanco FJ, Serrano L, Goni FM (1996) Infrared evidence of a beta-hairpin peptide structure in solution. FEBS Lett 384: 35–37

150. Jourdan M, Griffiths-Jones SR, Searle MS (2000) Folding of a beta-hairpin peptide derived from the N-terminus of ubiquitin.

Conformational preferences of beta-turn residues dictate non-native beta-strand interactions. Eur J Biochem 267:3539–3548

151. Stotz CE, Borchardt RT, Middaugh CR, Siahaan TJ, Vander Velde D, Topp EM (2004) Secondary structure of a dynamic type I′ beta-hairpin peptide. J Pept Res 63:371–382

152. Fesinmeyer RM, Hudson FM, Olsen KA, White GW, Euser A, Andersen NH (2005) Chemical shifts provide fold populations and register of beta hairpins and beta sheets. J Biomol NMR 33:213–231

153. Searle MS, Zerella R, Williams DH, Packman LC (1996) Native-like beta-hairpin structure in an isolated fragment from ferredoxin: NMR and CD studies of solvent effects on the N-terminal 20 residues. Protein Eng 9:559–565

154. Searle MS, Jourdan M (2000) Templating peptide folding on the surface of a micelle: nucleating the formation of a beta-hairpin. Bioorg Med Chem Lett 10:1139–1142

155. Lowik DW, Linhardt JG, Adams PJ, van Hest JC (2003) Non-covalent stabilization of a beta-hairpin peptide into liposomes. Org Biomol Chem 1:1827–1829

156. Dempsey CE, Mason PE (2006) Insight into indole interactions from alkali metal chloride effects on a tryptophan zipper beta-hairpin peptide. J Am Chem Soc 128:2762–2763

157. Penel S, Morrison RG, Dobson PD, Mortishire-Smith RJ, Doig AJ (2003) Length preferences and periodicity in beta-strands. Antiparallel edge beta-sheets are more likely to finish in non-hydrogen bonded rings. Protein Eng 16:957–961

158. Gibbs AC, Kondejewski LH, Gronwald W, Nip AM, Hodges RS, Sykes BD et al (1998) Unusual beta-sheet periodicity in small cyclic peptides. Nat Struct Biol 5:284–288

159. Li Y (2011) Recombinant production of antimicrobial peptides in Escherichia coli: a review. Protein Expr Purif 80:260–267

160. Johnson WCJ (1988) Secondary structure of proteins through circular dichroism. Annu Rev Biophys Biophys Chem 17:145–166

161. Santiveri CM, Rico M, Jimenez MA (2001) ^{13}Cα and ^{13}Cβ chemical shifts as a tool to delineate β-hairpin structures in peptides. J Biomol NMR 19:331–345

162. Wuthrich K, Billeter M, Braun W (1984) Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. J Mol Biol 180:715–740

163. Wuthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York, NY

164. Ramirez-Alvarado M, Serrano L, Blanco FJ (1997) Conformational analysis of peptides corresponding to all the secondary structure elements of protein L B1 domain: secondary structure propensities are not conserved in proteins with the same fold. Protein Sci 6: 162–174

165. Hughes RM, Waters ML (2006) Effects of lysine acetylation in a beta-hairpin peptide: comparison of an amide-pi and a cation-pi interaction. J Am Chem Soc 128:13586–13591

166. Honda S, Kobayashi N, Munekata E (2000) Thermodynamics of a beta-hairpin structure: evidence for cooperative formation of folding nucleus. J Mol Biol 295:269–278

167. Huang R, Wu L, McElheny D, Bour P, Roy A, Keiderling TA (2009) Cross-strand coupling and site-specific unfolding thermodynamics of a trpzip beta-hairpin peptide using 13C isotopic labeling and IR spectroscopy. J Phys Chem B 113:5661–5674

168. Xu Y, Du D, Oyola R (2011) Infrared study of the stability and folding kinetics of a series of beta-hairpin peptides with a common NPDG turn. J Phys Chem B 115:15332–15338

169. Xu Y, Oyola R, Gai F (2003) Infrared study of the stability and folding kinetics of a 15-residue beta-hairpin. J Am Chem Soc 125: 15388–15394

170. Du D, Zhu Y, Huang CY, Gai F (2004) Understanding the key factors that control the rate of beta-hairpin folding. Proc Natl Acad Sci U S A 101:15915–15920

171. Hauser K, Krejtschi C, Huang R, Wu L, Keiderling TA (2008) Site-specific relaxation kinetics of a tryptophan zipper hairpin peptide using temperature-jump IR spectroscopy and isotopic labeling. J Am Chem Soc 130: 2984–2992

172. Kuznetsov SV, Hilario J, Keiderling TA, Ansari A (2003) Spectroscopic studies of structural changes in two beta-sheet-forming peptides show an ensemble of structures that unfold noncooperatively. Biochemistry 42:4321–4332

173. Awasthi SK, Raghothama S, Balaram P (1995) A designed beta-hairpin peptide. Biochem Biophys Res Commun 216:375–381

174. Raghothama SR, Awasthi SK, Balaram P (1998) β-Hairpin nucleation by Pro-Gly β-turns. Comparison of dPro-Gly and lPro-Gly sequences in an apolar octapeptide. J Chem Soc Perkin Trans 2:137–143

175. Rai R, Raghothama S, Balaram P (2006) Design of a peptide hairpin containing a central three-residue loop. J Am Chem Soc 128: 2675–2681

176. Mahalakshmi R, Raghothama S, Balaram P (2006) NMR analysis of aromatic interactions in designed peptide beta-hairpins. J Am Chem Soc 128:1125–1138

177. Karle IL, Awasthi SK, Balaram P (1996) A designed beta-hairpin peptide in crystals. Proc Natl Acad Sci U S A 93:8189–8193

178. Ottesen JJ, Imperiali B (2001) Design of a discretely folded mini-protein motif with predominantly beta-structure. Nat Struct Biol 8: 535–539

179. Das C, Raghothama S, Balaram P (1998) A designed three stranded β-sheet peptide as a multiple β-hairpin model. J Am Chem Soc 120:5812–5813

180. Das C, Nayak V, Raghothama S, Balaram P (2000) Synthetic protein design: construction of a four-stranded beta-sheet structure and evaluation of its integrity in methanol-water systems. J Pept Res 56:307–317

181. Carulla N, Woodward C, Barany G (2002) BetaCore, a designed water soluble four-stranded antiparallel beta-sheet protein. Protein Sci 11:1539–1551

182. Venkatraman J, Naganagowda GA, Sudha R, Balaram P (2001) De novo design of a five-stranded β-sheet anchoring a metal-ion binding site. Chem Commun 24:2660–2661

183. Venkatraman J, Nagana Gowda GA, Balaram P (2002) Design and construction of an open multistranded beta-sheet polypeptide stabilized by a disulfide bridge. J Am Chem Soc 124:4987–4994

184. Mayo KH, Ilyina E (1998) A folding pathway for betapep-4 peptide 33mer: from unfolded monomers and beta-sheet sandwich dimers to well-structured tetramers. Protein Sci 7: 358–368

185. Meier S, Guthe S, Kiefhaber T, Grzesiek S (2004) Foldon, the natural trimerization domain of T4 fibritin, dissociates into a monomeric A-state form containing a stable beta-hairpin: atomic details of trimer dissociation and local beta-hairpin stability from residual dipolar couplings. J Mol Biol 344:1051–1069

186. Doig AJ (1997) A three-stranded beta-sheet peptide in aqueous solution containing N-methyl amino acids to prevent aggregation. Chem Commun 22:2153–2154

187. Jager M, Deechongkit S, Koepf EK, Nguyen H, Gao J, Powers ET et al (2008) Understanding the mechanism of beta-sheet folding from a chemical and biological perspective. Biopolymers 90:751–758

188. Kaul R, Angeles AR, Jager M, Powers ET, Kelly JW (2001) Incorporating beta-turns and a turn mimetic out of context in loop 1 of the WW domain affords cooperatively folded beta-sheets. J Am Chem Soc 123:5206–5212

189. Jager M, Dendle M, Fuller AA, Kelly JW (2007) A cross-strand Trp Trp pair stabilizes the hPin1 WW domain at the expense of function. Protein Sci 16:2306–2313

190. Rajagopal K, Lamm MS, Haines-Butterick LA, Pochan DJ, Schneider JP (2009) Tuning the pH responsiveness of beta-hairpin peptide folding, self-assembly, and hydrogel material formation. Biomacromolecules 10:2619–2625

191. Kaur H, Raghava GP (2002) BetaTPred: prediction of beta-TURNS in a protein using statistical algorithms. Bioinformatics 18:498–499

192. de la Cruz X, Hutchinson EG, Shepherd A, Thornton JM (2002) Toward predicting protein topology: an approach to identifying beta hairpins. Proc Natl Acad Sci U S A 99: 11157–11162

193. Cline LL, Waters ML (2009) The structure of well-folded beta-hairpin peptides promotes resistance to peptidase degradation. Biopolymers 92:502–507

194. Hilario J, Kubelka J, Syud FA, Gellman SH, Keiderling TA (2002) Spectroscopic characterization of selected beta- sheet hairpin models. Biopolymers 67:233–236

195. Sharman GJ, Griffiths-Jones SR, Jourdan M, Searle MS (2001) Effects of amino acid phi, psi propensities and secondary structure interactions in modulating H alpha chemical shifts in peptide and protein beta-sheet. J Am Chem Soc 123:12318–12324

196. Shu I, Stewart JM, Scian M, Kier BL, Andersen NH (2011) beta-Sheet 13C structuring shifts appear only at the H-bonded sites of hairpins. J Am Chem Soc 133:1196–1199

197. Villegas ME, Vila JA, Scheraga HA (2007) Effects of side-chain orientation on the 13C chemical shifts of antiparallel beta-sheet model peptides. J Biomol NMR 37:137–146

198. Olsen KA, Fesinmeyer RM, Stewart JM, Andersen NH (2005) Hairpin folding rates reflect mutations within and remote from the turn region. Proc Natl Acad Sci U S A 102: 15483–15487

199. Scian M, Shu I, Olsen KA, Hassam K, Andersen NH (2013) Mutational effects on the folding dynamics of a minimized hairpin. Biochemistry 52:2556–2564

200. Davis CM, Xiao S, Raleigh DP, Dyer RB (2012) Raising the speed limit for beta-hairpin formation. J Am Chem Soc 134: 14476–14482

201. Munoz V, Thompson PA, Hofrichter J, Eaton WA (1997) Folding dynamics and mechanism of beta-hairpin formation. Nature 390: 196–199

202. McMillan AW, Kier BL, Shu I, Byrne A, Andersen NH, Parson WW (2013) Fluorescence of tryptophan in designed hairpin and Trp-cage miniproteins: measurements of fluorescence yields and calculations by quantum mechanical molecular dynamics simulations. J Phys Chem B 117:1790–1809

Chapter 3

Combination of Theoretical and Experimental Approaches for the Design and Study of Fibril-Forming Peptides

Phanourios Tamamis, Emmanouil Kasotakis, Georgios Archontis, and Anna Mitraki

Abstract

Self-assembling peptides that can form supramolecular structures such as fibrils, ribbons, and nanotubes are of particular interest to modern bionanotechnology and materials science. Their ability to form biocompatible nanostructures under mild conditions through non-covalent interactions offers a big biofabrication advantage. Structural motifs extracted from natural proteins are an important source of inspiration for the rational design of such peptides. Examples include designer self-assembling peptides that correspond to natural coiled-coil motifs, amyloid-forming proteins, and natural fibrous proteins. In this chapter, we focus on the exploitation of structural information from beta-structured natural fibers. We review a case study of short peptides that correspond to sequences from the adenovirus fiber shaft. We describe both theoretical methods for the study of their self-assembly potential and basic experimental protocols for the assessment of fibril-forming assembly.

Key words Peptides, Self-assembly, Amyloid fibrils, Beta-structure, Molecular dynamics, Implicit solvent, Replica-exchange, Adenovirus

1 Introduction

Natural fibrous proteins are usually built up from repetitive sequences and can adopt either helical or beta-structural folds [1]. As of today, many structural folds from intracellular and extracellular fibrous proteins, virus and bacteriophage fibers, and amyloid are accessible in the Protein Data Bank. These repetitive sequences can serve as useful building blocks for designing short peptides, or recombinant proteins made up from "concatamers" of these sequences [2–6]. Small building blocks such as short peptides that self-assemble into fibrils are particularly attractive since they can be synthesized under mild, physiological conditions, they are biocompatible, and they can withstand harsh physical and chemical conditions once formed [7, 8]. Structural information provided by the single-crystal structures of fibrous proteins is not sufficient to

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_3, © Springer Science+Business Media New York 2014

predict the conformation that will be adopted by small peptides corresponding to their sequences. Many fibrous proteins contain "registration domains" or "capping elements" that ensure that staggered assembly of the repetitive sequences will be avoided [9, 10]. As a result, in the absence of these domains, short peptide sequences may self-assemble into a structure different than the one assumed within the native protein context.

Molecular dynamics simulations can provide precious insight into the organization and interactions of the structures formed by the self-assembly of such peptides. We review in this chapter such a case study of self-assembling peptide rational design. The design is based on insight provided from our molecular dynamics simulations of natural sequences from the adenovirus fiber protein [11] (PDB entry: 1QIU). The adenovirus fibers are trimeric and consist of the following parts: an N-terminal capsid-binding domain, a fibrous shaft domain, and a globular, C-terminal receptor-binding domain; the C-terminal domain is necessary for the correct registration and folding of the fibrous shaft. The fiber shaft domain consists of repeats of the triple beta-spiral fold. Every repeat contains a beta-strand running parallel to the fiber axis, followed by a type II beta-turn. The turn is followed by another beta-strand that runs "backward" at an angle of 45° relative to the fiber axis [11]. The repeats are connected with a solvent-exposed loop of variable length.

Peptides ranging from 6 to 41 amino acids were initially designed on the basis of these repeats and were found to self-assemble into amyloid-type fibrils [3]. From these sequences, we were seeking to identify minimal peptide building blocks that could comprise a central self-assembling "core" and yet could carry positions amenable to modification. Such positions are essential for the incorporation of amino residues that could target binding of inorganic nanoparticles, cells, etc. This kind of approach is very useful for the rational design of biologically inspired nanomaterials [12].

In this chapter, we describe a combination of theoretical and experimental approaches towards the identification of such building blocks. We first describe a general computational protocol, based on molecular dynamics, for the investigation of the early self-assembly stage of peptide-based nanostructures. Application of this protocol to the self-assembly of amyloidogenic octapeptide and dodecapeptide sequences from the adenovirus fiber shaft is described in refs. [13, 14]. Additional studies investigate (i) an undecapeptide residue sequence (segment 155–165 of the adenovirus shaft, with sequence: LSGSDSDTLTV), and (ii) a 33-residue sequence (segment 360–392 of the adenovirus shaft, with sequence: KIGSGIDYNENGAMITKLGAGLSFDNSGAITIG); the findings of the latter study will be reported elsewhere. Insights from this work assisted in the design of fibril-forming sequences with the capacity to bind metals [12]. We further review two key experimental protocols (sample preparation for electron microscopy and X-ray fiber diffraction) for verification of the fibril-forming potential of the designed sequences.

2 Materials

2.1 Peptides

In order to give a practical guide towards the theoretical and experimental investigation of fibril-forming peptides, we will focus on a case study using the dodecapeptide LSFDNSGAITIG (Leucine-Serine-Phenylalanine-Aspartate-Asparagine-Serine-Glycine-Alanine-Isoleucine-Threonine-Isoleucine-Glycine) and its truncated octapeptide variant, NSGAITIG (Asparagine-Serine-Glycine-Alanine-Isoleucine-Threonine-Isoleucine-Glycine). Both are subsequences of the 15-residue repeating motif in the fiber shaft. The dodecapeptide corresponds to a strand-loop-strand region within the natural protein (residues 381–392) while the octapeptide corresponds to a short loop-and-strand region (residues 385–392) [3, 11]. The peptides (free N-termini, amidated C-termini) were purchased from Eurogentec (Belgium) and had a degree of purity higher than 95 %. Lyophilized peptide powders were dissolved and studied in ultrapure water. The same theoretical protocol has been applied to simulations of monomeric and trimeric sequences of the adenovirus shaft segment 360–392 (sequence KIGSGIDYNENGAMITKLGAGLSFDNSGAITIG, unpublished) and the segment 155–165 (LSGSDSDTLTV).

2.2 Transmission Electron Microscopy

1. Square mesh grids: 200 up to 400 MESH, Copper or Nickel, 3.05 mm, Formvar, Carbon, Formvar-Carbon (Agar Scientific, UK).
2. Dumont tweezers No. 5a Stainless, Dumoxel (Non-Magnetic).
3. Staining solutions: 1 %(w/v) uranyl acetate or sodium phosphotungstate, depending on solution pH (*see* **Note 1**).
4. JEOL JEM-100C transmission electron microscope operating at 80 kV, or equivalent.
5. Gatan Digital Micrograph software for the analysis of the images (http://www.gatan.com/).

2.3 Fiber Diffraction Samples

1. Glass rods (approx. 0.5 mm diameter).
2. Microscope glass slides.
3. Plasticine.
4. Zeiss Stemi 2000-C Stereoscope, or equivalent.
5. Bunsen burner (or equivalent).

3 Methods

3.1 Computational Protocols: Setup, Execution, and Analysis of MD Simulations of Fibril-Forming Peptides

MD simulations have emerged as a valuable tool for the investigation of the structural and dynamical properties of biomolecular complexes ([15–19] and references within), and can provide important insights into the structural organization, interactions, and stability of peptide-based nanostructures [13, 14, 20–31]. In planning peptide self-assembly simulations, an important initial modelling choice concerns the resolution of the molecular representation: (1) The atomic-detail representation of peptides and solvent provides the most accurate description [19]. However, atomic-detail simulations have a considerable computational cost, and explore mostly conformations in the vicinity of the initial structure. Hence, this representation is optimal when one or more plausible structural models of the nanostructures already exist [e.g., from experiments, previous modelling, or low-resolution (coarse-grained) simulations], and can be used as a starting point. The simulations can then investigate the robustness of the models and provide a set of representative conformations for further computational studies. For example, relative stabilities of the structural models and key interactions can be quantified by free-energy calculations (e.g., MM-PBSA or MM-GBSA [32, 33]) and free-energy component analyses [34–41] of the MD conformations. (2) In the absence of structural models, a principal goal of the simulations is to explore the early stages of aggregation and to identify interaction patterns that may be present in the nanostructures [13, 14, 20–23, 27, 29–31]. Such patterns could emerge in complex conformations of high symmetry (e.g., well-ordered multi-stranded beta-sheets), which appear infrequently. Therefore, it is important that the simulations are efficient and sufficiently long, so that they sample a large portion of the conformational space. The simulation efficiency can be increased by a multi-step computational protocol, in which the molecular representation is refined as more information becomes available. In early-phase ("structure-exploration") simulations, the protein and/or solvent can be represented by "coarse-grained" models, where selected groups of atoms are replaced by a single interaction center [15, 42–44]; alternatively, solvent effects can be included implicitly by suitable terms in the energy function [45–51]. Such reduced representations lower significantly the computational cost, and accelerate peptide conformational transitions. At the same time, they provide a compromise between accuracy and efficiency and need to be conducted with caution [50]. The resulting structural models can be assessed in a following, "validation" stage via atomic-detail simulations.

In what follows, we describe the execution and analysis of "structure exploration" simulations. A typical methodology consists of the following steps:

1. The peptide sequence, including terminal ends, is defined to match the actual experimental system.
2. Molecular representations and force fields are chosen for the peptide and solvent interactions. In the two test cases of the present section, sequences NSGAITIG [13], and LSFDNSG AITIG [14], as well as in additional studies of the segments 360–392 and 155–165 of the adenovirus fiber shaft, we used the CHARMM22 [52] or polar-hydrogen CHARMM19 force field [53] for the peptide interactions in conjunction with the FACTS [51] continuum electrostatics model for the aqueous solvent. In atomic-detail simulations we employed the TIP3P model for water and the all-atom CHARMM22 force field [52] for the peptide interactions, with a CMAP correction for backbone torsional interactions [54]. All simulations were conducted with the CHARMM program [19].
3. At the beginning it is recommended to simulate the fundamental nanostructure-building block at infinite-dilution conditions (e.g., a single peptide in implicit solvent). These simulations are relatively rapid and can compare the predictions of several implicit-solvent models (in conjunction with the corresponding peptide force fields). Comparison with available experimental data at the same conditions (e.g., structural data on the same peptide by NMR, CD) can help identify the optimum implicit-solvent model for the particular system. Fine-tuning of specific model parameters (e.g., the surface tension coefficient, the protein dielectric constant) might improve agreement with the experiment. We note though that such modifications should be done with caution, and their impact should ideally be tested on a range of sequences of variable sizes and structural motifs.

   In the absence of experimental information, the implicit-solvent results can be compared against explicit-solvent simulations. The explicit-solvent runs should be executed with a high-efficiency protocol, such as the Replica Exchange Molecular Dynamics (REMD) method [55], to ensure that they explore a large portion of the peptide conformational space.
4. Once the implicit-solvent model has been selected, representative conformations from the infinite-dilution run can be used as starting points for simulations in finite dilution. The results of the infinite- and finite-dilution runs with the optimum implicit model should be eventually compared. Conformational features emerging only in finite dilution may help identify key intermolecular interactions in the nanostructures.

5. The finite-dilution simulation system consists of a number of identical peptide copies in implicit solvent. The peptides are restrained to move in a suitably defined "container" (e.g., the interior of a sphere or a periodically replicated box). The container volume is adjusted to model a solution with peptide concentration in the range of experimental self-assembly concentrations.

6. The peptide copies are initially placed in random positions and orientations inside the container; typically a representative conformation from the infinite-dilution runs is chosen for each peptide; the initial distance between any two peptides is usually set to a smaller value than the cutoff of non-bonded interactions, to facilitate the formation of an initial aggregate. However, if the non-bonded cutoff is too small relative to the box size, peptides may spend a large portion of the simulation at remote parts of the container, without interacting with the rest of the system; in such a case, it is recommended to decrease the container size.

7. The simulations can be conducted with a high-efficiency protocol, such as the Replica Exchange Molecular Dynamics (REMD) method [56–61]. In the REMD scheme, several identical copies (replicas) of the entire system are simulated concurrently at different temperatures. The copies communicate at periodic intervals and exchange coordinates with a Metropolis criterion. This exchange allows low-temperature replicas to escape local minima and to borrow the sampling efficiency of high-temperature replicas. The Metropolis criterion ensures that energetically inaccessible conformations (at a given temperature) are not allowed, and that all replicas gain access to conformations with the proper thermodynamic weight. Furthermore, the high-temperature runs permit the frequent dissolution and reformation of aggregates and the sampling of a wide variety of intermolecular structures. This is an essential feature of the REMD simulations, which a traditional run at room temperature would lack.

8. For optimum use of the REMD method, the replica temperatures need to satisfy several criteria [60–64]: (a) The range of temperatures should ensure the rapid and extensive sampling of conformations compatible with the temperature of experimental conditions [62] (*see* **Note 1**). (b) The individual replica temperatures are chosen so as to achieve a uniform exchange probability between neighboring replicas [57, 60]; a typical target value of the exchange probability is in the range of 20–30 %. Easily implementable temperature optimization methods have been described in ref. 57, 60 (*see* **Note 2**). (c) The number of MD steps between exchange attempts should be reasonably large, to ensure equilibration and sufficient

sampling between exchanges [61] (*see* **Note 3**). (d) The temperature control of the simulations should be achieved with thermostats that produce a canonical ensemble [61, 63].

9. In the early stages of the REMD simulation it is important to fine-tune the simulation conditions with the aid of several tests: (a) it is helpful to analyze the sizes of the observed aggregates at various temperatures. Near the experimental temperature the observed aggregates should be relatively stable and contain most of the peptides, to ensure the exploration of intermolecular structures involving several peptides. At higher temperatures the aggregates should undergo extensive reorganization (with the frequent detachment and reattachment of peptides), to ensure sampling of a large variety of intermolecular structures. (b) The replica-exchange probabilities should be fairly constant throughout the temperature space. (c) Each replica should execute random walks covering the entire temperature space. (d) The radius of gyration (rgyr) provides information on the average distance among peptides. Examination of the rgyr temperature dependence may reveal a transition temperature in the range [T_{min}, T_{max}], beyond which the aggregates are not stable. Often, the replicas tend to stay either below or above this transition temperature and cannot complete random walks in the entire temperature space (*see* [64] for a discussion). In such cases, a denser arrangement or dynamical adaptation [65] of replicas near this transition temperature may eliminate the problem.

10. As simulation progresses, the formation of increasingly complex intermolecular structures should become more probable. For example, amyloidogenic peptides have an increased propensity for intermolecular beta-sheets. At the early stage of the simulations, the observed beta-sheets will involve a small number of peptides (typically two or three). Gradually, more complex sheets will be stabilized, involving several or all peptides in the system. The intermolecular structures can be detected by visualization of the trajectories, or analysis with secondary-structure algorithms (e.g., STRIDE [66], DSSP [67]).

11. Comparison of the trajectories from infinite- and finite-dilution simulations illuminates the impact of intermolecular interactions on the conformation of individual residues. This comparison can be achieved by merging the coordinates of individual peptides from the infinite- and finite-dilution simulations into a single (one-peptide) trajectory, and subject it to a clustering analysis. If the impact of finite-dilution conditions is significant, the clustering will partition the conformations from the trajectories of these two conditions into distinct clusters.

12. The main goal of the simulations is to identify inter- and intramolecular interaction patterns, which hint to the peptide organization in the nanostructures. Because the simulation system includes only a small number of peptides, and the solvent is represented implicitly, the peptides will mostly collapse into a globular aggregate; within this aggregate the peptides will tend to form nonspecific intermolecular interactions; hence, the emergence of patterns hinting to high-symmetry structures will require long simulations and careful analysis. In what follows, we describe our analysis of simulations targeting two peptides from the adenovirus fiber shaft with sequences NSGAITIG [13] and LSFDNSGAITIG [14], respectively. Both peptides have been shown experimentally to self-assemble in amyloid-like fibrils [3]. Hence, we target our analysis on the identification of intermolecular beta-sheets.

13. Raw data on intermolecular beta-sheet content of the aggregate are first obtained by post-processing the trajectories with STRIDE or DSSP. The output of these programs can be analyzed with a text-manipulation language (e.g., AWK or PERL), or a programming language. Information on residue pairs in intermolecular beta-bridges can be tabulated in suitably defined matrices. We use raw data from DSSP or STRIDE to construct five-dimensional matrices of the form F(pept1,*i*,pept2,*j*,*t*). The indices "pept1" and "pept2" run from 1 to *k* (the total number of peptides in the system); the indices "*i*" and "*j*" run from 1 to *N* (the number of residues in each peptide), and "*t*" runs from 1 to *M* (the maximum number of snapshots to be analyzed). Suppose at time step *t*, residues (*i*, …, *i* + *n*) of pept1 form contiguous beta-bridges with residues (*j*, …, *j* + *n*) of pept2. (a) In a parallel beta-sheet the interacting pairs are [*i*:*j*], [*i* + 1:*j* + 1], …, [*i* + *n*:*j* + *n*], and elements F(pept1,*i*,pept2,*j*,*t*), F(pept1,*i* + 1,pept2,*j* + 1,*t*), …, F(pept1,*i* + *n*,pept2,*j* + *n*,*t*) are set to 1 [equivalently, F(pept2,*j*,pept1,*i*,*t*), F(pept2,*j* + 1, pept1,*i* + 1,*t*), …, F(pept2,*j* + *n*,pept1,*i* + *n*,*t*) are set to 1]. (b) In an antiparallel sheet, the interacting pairs are [*i*:*j* + *n*], [*i* + 1:*j* + n−1], …, [*i* + *n*:*j*] and elements F(pept1,*i*,pept2,*j* + *n*,*t*), F(pept1,*i* + 1,pept2,*j* + *n*−1,*t*), …, F(pept1,*i* + *n*,pept2,*j*,*t*) are set to −1 [equivalently F(pept2,*j*,pept1,*i* + *n*,*t*), F(pept2,*j* + 1, pept1,*i* + *n*−1,*t*), …, F(pept2,*j* + *n*,pept1,*i*,*t*) are set to −1]. If any residue pair [*i*:*j*] of peptide pair [pept1:pept2] does not participate in a â-bridge at time *t*, the corresponding elements F(pept1,*i*,pept2,*j*,*t*) are set to 0.

14. Once matrix "F" is filled in, a FORTRAN loop identifies and classifies beta-sheet structures in order of decreasing complexity. A simulation system of three peptides can have the following states (in terms of beta-sheet content): one 3-stranded beta-sheet, one 2-stranded beta-sheet, and no intermolecular

beta-sheet. With the aid of matrix "f," 3-stranded beta-sheets can be labeled as antiparallel (A3), mixed (M3), and parallel (P3). Similarly, 2-stranded beta-sheets can be labeled as antiparallel (A2), or parallel (P2).

15. Running averages of the various states (A3, M3, P3, A2, P2, "no-sheet") demonstrate whether the simulations have reached convergence by the end of the simulation, and rank the beta-sheet families with respect to their formation probability. The early, equilibration phase, where these averages are far from their converged values, can be excluded from the analysis.

16. More detailed information on the sequence content of intermolecular beta-sheets can be obtained from two-dimensional probability maps of intermolecular beta-bridges. For a three-stranded sheet, it is convenient to assign the residues of the central peptide to one of the map axes, and the residues of the edge peptides to the second axis (results are averaged over the two edge peptides). Analysis of the maps provides information on the type of "in-" or "off-" register beta-sheets in states A3, M3, and P3. When analyzing beta-sheets with more than three peptides, it is convenient to record the interactions within each pair of adjacent strands; the final map will involve an average over all such pairs.

17. Probability maps of intermolecular side-chain contacts provide complementary information on key sheet-stabilizing side-chain interactions formed by beta-sheet interacting peptides. A useful contact criterion checks whether the distance between the geometric centers of two side chains is smaller than a user-defined cutoff (e.g., 6–7 Å).

18. In some systems the peptide arrangement in the nanostructures involves a combination of intra- and intermolecular interactions. An example is peptide LSFDNSGAITIG [14], which formed intermolecular beta-sheets with individual strands bent into a "U" shape. In such cases, additional descriptors are needed for a complete description of the beta-sheets. For example, matrix F can be augmented by one extra dimension [elements: F(pept1,i,pept2,j,q,t)]. The new variable q distinguishes between different peptide shapes. For example, if a strand is mainly observed in two shapes (e.g., a linear shape "I" and a bent shape "U"), the values q = 1–4 can label, respectively, the shape combinations pept1:pept2 = I:I, U:U, I:U, and U:I (symbol ":" denotes a beta-sheet interaction). Intermolecular beta-sheets can be classified in terms of the number and shape of their constituent strands (e.g., for three-stranded sheet, state "U3" denotes a U:U:U arrangement, state "U2I" denotes U:I:U or U:U:I). It is worth noting that the simultaneous classification of strand shape and orientation (parallel/antiparallel) could result in a large number of states;

accurate sampling of the corresponding formation probabilities may require long simulations.

19. A running-average analysis of the formation probabilities will identify the most populated highest-complexity configuration. In the NSGAITIG simulations [13], parallel 3-stranded sheets were more probable than antiparallel 3-stranded; in the LSFDNSGAITIG case [14], structures with all three peptides in a bent (U-shape) conformation were more probable than structures with all three peptides in a linear (I-shape) conformation. In general, non-symmetric sheets (containing strands in different shapes and orientations, such as a combination of I and U shapes in mixed parallel/antiparallel arrangements) are favored entropically. However, the peptides should be arranged in more symmetric patterns within the nanoscale structures (e.g., in one or more layers of fully antiparallel or parallel sheets, alternating U patterns); therefore, when constructing a model of the nanostructure it may be more relevant to employ lower probability, higher symmetry structural elements. In the case of LSFDNSGAITIG, beta-sheets with all strands in a U shape were the most probable high-symmetry structures [14].

20. A very useful analysis of MD trajectories involves clustering of the structures with respect to their atomic or internal coordinates [68]. Such analysis is not straightforward in systems of many identical molecules, such as the finite-dilution peptide solutions discussed here. Since the peptides of the simulation system are equivalent, conformations in which two peptides simply swap positions should be considered identical. For example, in a three-stranded sheet of peptides A, B, and C, all 3! possible strand arrangements (ABC, ACB, BAC, …) correspond to the same conformation, even though the atomic coordinates of interchanged peptides change between arrangements. A convenient way to overcome the previous issue is to classify conformations by suitably defined order parameters, which describe global structural features of the system. Frequently used order parameters are the polar order-parameter P_1 and the nematic order-parameter P_2 (Eq. 1), which reflect the orientation and degree of strand order in the aggregate. These parameters are widely used in the structural characterization of liquid crystals [69–71], and have been employed successfully in simulation studies of peptide aggregation [13, 14, 21]:

$$P_1 = \frac{1}{N}\sum_{i=1}^{N}\vec{z}_i\vec{d}, \quad P_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{3}{2}\left(\vec{z}_i\vec{d}\right)^2 - \frac{1}{2}. \tag{1}$$

In Eq. 1, N is the number of molecules in the simulation and $\vec{z}_i$ is a unit vector along a suitably defined molecular direction; $\vec{d}$ is a unit vector along a preferred direction of alignment, which emerges from the properties of the system.

21. By definition, P_1 can differentiate between parallel or antiparallel/mixed molecular vectors in the beta-sheets, whereas P_2 can differentiate between perfectly ordered (P_2, ~1) and disordered conformations ($P_2 < \sqrt{\frac{81}{40\pi N}}$), for a system of N peptides; for $N = 3$, $P_2 = 0.46$. P_1 and P_2 parameters can be conveniently computed with the program WORDOM [72].

22. A key step in the use of P_1 and P_2 is the choice of the molecular vector $\vec{z_i}$ entering in Eq. 1. If the strands are extended, the entire peptide backbone can be used as a vector; otherwise, it is best to select a suitable backbone fragment, which best describes the predominant direction (*see* **Note 4**).

23. Free-energy landscapes (FEL) can be constructed by the two-dimensional probability $P(P_1, P_2)$ as

$$G\left(P_1,P_2\right) = -k_B T \ln\left[P\left(P_1,P_2\right)\right] \quad (2)$$

The construction of meaningful FELs using polar and order nematic parameter(s) becomes more complex when the peptides within intermolecular beta-sheets are in a nonlinear conformation, as in the case of LSFDNSGAITIG. In such cases, a combined set of molecular vectors corresponding to two sets of parameters (P_1, P_1*|P_2, P_2*) may be required to describe and differentiate between different configurations (*see* ref. 14 for more details).

24. Representative structures can be extracted from the FEL minima and examined in more detail. Some of the resulting elementary beta-sheet patterns can constitute elementary building blocks of the naturally occurring beta-sheets in the nanostructures.

25. In the case of LSFDNSGAITIG, the global minimum indicated a U-shape folding of individual strands in the intermolecular sheets (Fig. 1a), in agreement with the experimental width of the fibril [3, 14].

26. In the case of NSGAITIG, the global minimum of the FES of three-stranded parallel sheets contained conformations with the first two residues of each peptide (Asn1 and Ser2) in a disordered, solvent-exposed state (Fig. 1b) [13]. This gave impetus for the design of new amyloidogenic peptides with cysteine substitutions at positions 1 and 2. The substituted cysteines were capable of binding to silver, gold, and platinum nanoparticles [12], in line with the availability of the N-terminal ends [13].

A similar analysis was applied to simulation studies of the self-assembly of the sequence 155–165 of the adenovirus shaft (LSGSDSDTLTV). This fragment is rich in aspartates and ser-

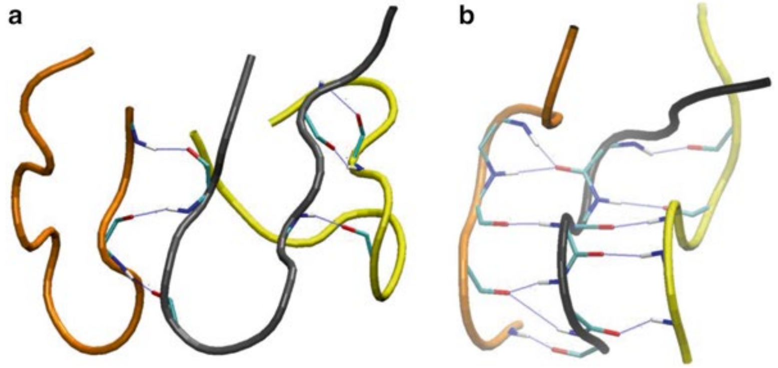


Fig. 1 Representative conformations of (**a**) U:U:U intermolecular beta-sheets are observed in the LSFDNSGAITIG simulations, and (**b**) parallel intermolecular beta-sheets observed in the NSGAITIG simulations. The backbone tube representation of each interacting peptide is shown in different color. Beta-sheet hydrogen bonds are presented with *dashed lines*

ines, both of which are naturally displayed on the surface of proteins with a role in calcium nucleation and hard tissue formation [73].

3.2 Preparation of Specimens for TEM Analysis

Transmission electron microscopy is a key method for the assessment of fibril-forming capacity of peptides. 8–10 μl of the peptide solution at the desired concentration are placed on a 300 mesh formvar-coated grid and after 2 min the excess fluid is removed with a filter paper. The samples are subsequently negatively stained with 8 μl staining solution 1 % for 2 min (*see* **Note 5**). When amyloid fibrils are formed, the following morphology should be expected: unbranched fibrils with diameters around 100 Å that could arrive at the order of microns in length (Fig. 2). The fibrils could be straight or twisted.

3.3 Preparation of Specimens for X-Ray Fiber Diffraction Analysis

Fiber diffraction is also a key method for the diagnosis of structural signature of fibrous biomolecules. Very often it is difficult to have in-house access and a collaboration with fiber diffraction specialists is necessary for X-ray data collection and analysis. We describe here a protocol for the preparation of specimens that can be used for fiber diffraction. The protocol is based on the formation of fibrous stalks between two glass rods. Two glass rods are passed in a Bunsen burner (or equivalent) to form round ends and are subsequently stabilized and aligned with plasticine on a glass microscope slide, as shown in Fig. 3. The distance between the round ends should be around 2–3 mm. The peptide solutions should be visually examined for an increase in viscosity, gel formation, or the appearance of precipitates. Once a change in viscosity or a gel/precipitate is observed, a droplet of 8–10 μl of peptide solution is placed between the rods and left to dry. As the drops dry, a fibrous stalk will form between the rods if fibrils are present (Fig. 4). If the

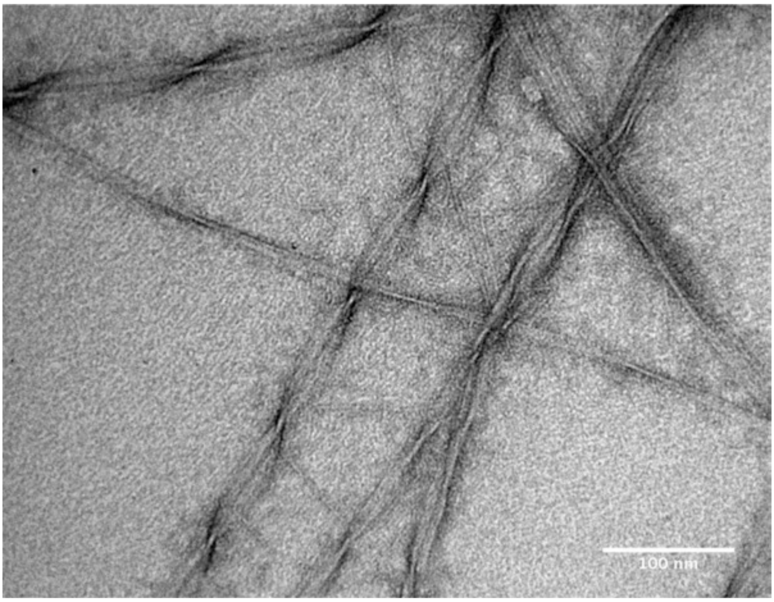


Fig. 2 Electron micrograph showing amyloid-type fibrils formed by the peptide NSGAITIG, negatively stained with uranyl acetate

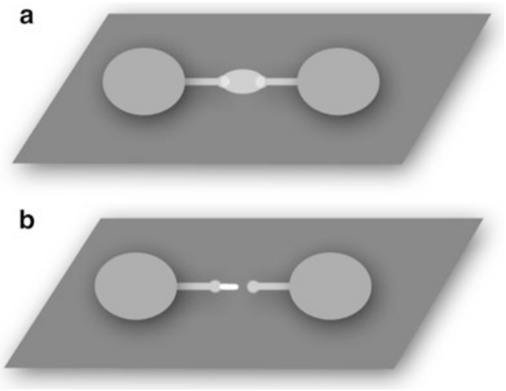


Fig. 3 Experimental setup for the formation of fibrous stalks between glass rods. Glass rods are fixed and aligned with plasticine on a microscope glass slide. A droplet of peptide solution is deposited between the rods (**a**) and allowed to dry until stalk is formed (**b**)

fibrils are well aligned, the stalks will be birefringent when viewed under crossed polars in a stereoscope. For amyloid-type fibrils, the following diagnostic pattern should be observed: a meridional reflection at 4.8 Å that corresponds to the distance between beta-strands that are perpendicular to the fibril axis, and an equatorial reflection at approximately 10 Å that corresponds to the distance between beta-sheets (Fig. 5).

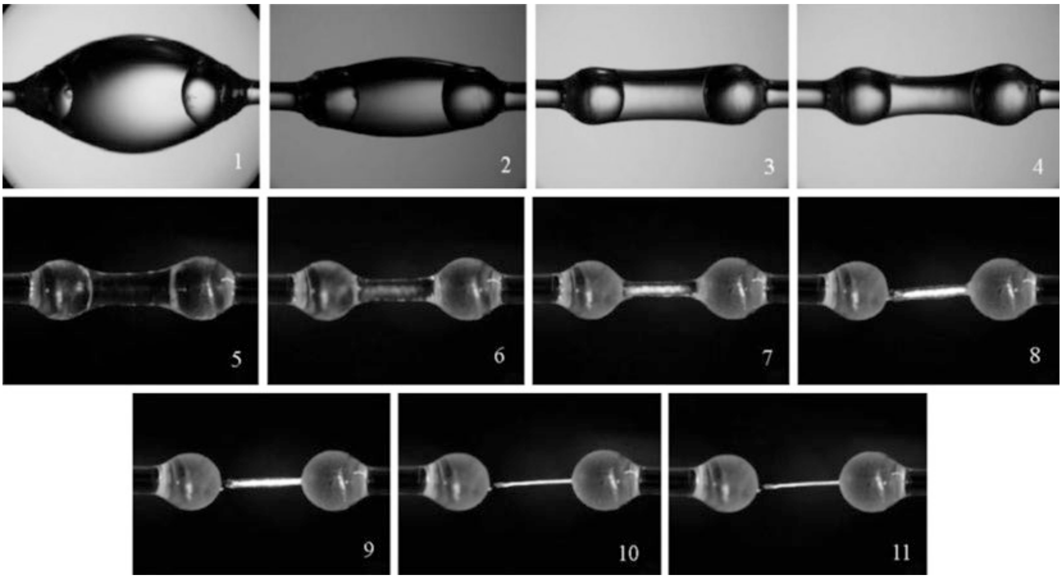


Fig. 4 Droplet drying steps and fibrous stalk formation as watched in a stereoscope. Clichés 1–4 are taken in bright field, while clichés 5–11 are taken with crossed polars. The observed birefringence indicates a good alignment of the sample

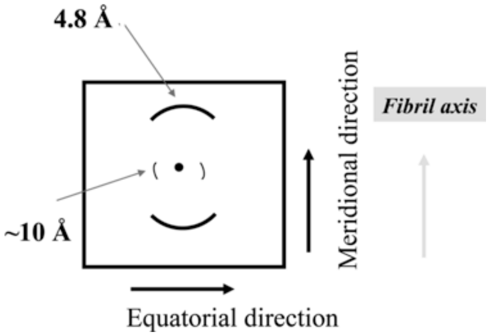


Fig. 5 Schematic drawing of the typical fiber diffraction pattern recorded when the fibrous stalks are placed at right angle to the X-ray beam. The two reflections, 4.8 Å in the meridional direction, and ~10 Å in the equatorial direction, are a diagnostic signature for the amyloid structure

4 Notes

1. Choose minimum (T_{min}) and maximum (T_{max}) temperatures to bracket the temperature of the self-assembly experiments (T_{exp}). Even though T_{min} can be set equal to T_{exp}, it is advisable that a sufficiently low value is selected, so that a few replicas are simulated at temperatures lower than T_{exp}. This facilitates the stabilization of conformations at $T = T_{\mathrm{exp}}$. The maximum temperature T_{max} depends on the total number of available processors.

Nevertheless, a sufficiently high value is desirable, so that individual peptides frequently disengage from the aggregate and undergo extensive conformational changes.

2. Conduct preliminary standard (non-replica) runs at temperatures spanning the anticipated temperature range of the replica-exchange simulations. Post-process the obtained trajectories, to determine the average potential energy of the simulated system as a function of temperature. Insert this temperature dependence in the replica exchange-probability equation [57, 60] to determine replica temperatures that yield a specific value for the replica-exchange probability. Iterate this procedure a few times, until the chosen temperatures yield replica-exchange probabilities near the target value (~20–30 %).
3. In our systems, we attempted replica exchanges every 10 ps.
4. For example, if intermolecular beta-sheets are in register and involve fragments 4–7 of individual strands, then a useful molecular vector would start from nitrogen N of residue 4 and end at atom C of residue 7 [13].
5. Uranyl acetate is used for negative staining in the pH range of 4–6, while sodium phosphotungstate is used in the pH range of 5–8.

Acknowledgments

GA and PhT acknowledge financial support from the University of Cyprus program "*Computational Investigation of Peptide Sequences from the Adenovirus and Reovirus Fiber Shaft*: *Insights on the Self-assembly of Peptide-based Nanostructures and the Stability of the Triple-beta Spiral Fold*," and the program UPGRADE/0609/11, "*Self-Assembly and Folding of Biologically-inspired Peptide Sequences From Natural Fibrous Proteins*: *Insights From Highly-Intensive Computational Studies*" that is co-funded by Desmi 2009–2010 of the Cyprus Research Promotion Foundation, the Republic of Cyprus, and the European Regional Development Fund. Simulations were conducted in Linux clusters of the Biophysics group, partly financed through program UPGRADE/0609/11. E.K. and A.M. acknowledge funding form the European Union (STREP NMP-CT-2006-033256, "BeNatural").

References

1. Cohen C (1998) Why fibrous proteins are romantic. J Struct Biol 122:3–16
2. Iconomidou VA, Chryssikos GD, Gionis V, Vriend G, Hoenger A, Hamodrakas S (2001) Amyloid-like fibrils from an 18-residue peptide analogue of a part of the central domain of the B-family of silkmoth chorion proteins. FEBS Lett 499:268–273
3. Papanikolopoulou K, Schoehn G, Forge V, Forsyth VT, Riekel C, Hernandez JF et al (2005) Amyloid fibril formation from sequences of a natural beta-structured fibrous

protein, the adenovirus fiber. J Biol Chem 280:2481–2490

4. Spiess K, Lammel A, Scheibel T (2010) Recombinant spider silk proteins for applications in biomaterials. Macromol Biosci 10:998–1007
5. Ryadnov MG, Woolfson DN (2007) Self-assembled templates for polypeptide synthesis. J Am Chem Soc 129:14074–14081
6. Girotti A, Fernandez-Colino A, Lopez IM, Rodriguez-Cabello JC, Arias FJ (2011) Elastin-like recombinamers: biosynthetic strategies and biotechnological applications. Biotechnol J 6: 1174–1186
7. Reches M, Gazit E (2006) Molecular self-assembly of peptide nanostructures: mechanism of association and potential uses. Curr Nanosci 2:105–111
8. Zhang SG (2003) Fabrication of novel biomaterials through molecular self-assembly. Nat Biotechnol 21:1171–1178
9. Mitraki A, Miller S, van Raaij MJ (2002) Review: conformation and folding of novel Beta-structural elements in viral fiber proteins—the triple Beta-spiral and triple Beta-helix. J Struct Biol 137:236–247
10. Mitraki A, Papanikolopoulou K, van Raaij MJ (2006) Natural triple beta-stranded fibrous folds. Adv Protein Chem 74:97–124
11. van Raaij MJ, Mitraki A, Lavigne G, Cusack S (1999) A triple beta-spiral in the adenovirus fibre shaft reveals a new structural motif for a fibrous protein. Nature 401:935–938
12. Kasotakis E, Mossou E, Adler-Abramovich L, Mitchell EP, Forsyth VT, Gazit E et al (2009) Design of metal-binding sites onto self-assembled peptide fibrils. Biopolymers 92: 164–172
13. Tamamis P, Kasotakis E, Mitraki A, Archontis G (2009) Amyloid-like self-assembly of peptide sequences from the adenovirus fiber shaft: insights from replica exchange MD simulations. J Phys Chem B 113:15639–15647
14. Tamamis P, Archontis G (2011) Amyloid-like self-assembly of a dodecapeptide sequence from the adenovirus fiber shaft: perspectives from molecular dynamics simulations. J Non-Cryst Solids 357:717–722
15. van Gunsteren WF, Dolenc J (2008) Biomolecular simulation: historical picture and future perspectives. Biochem Soc Trans 36:11–15
16. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9:646–652
17. Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. Proc Natl Acad Sci U S A 102:6679–6685
18. Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106: 1589–1615
19. Brooks BR, Brooks CL III, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B et al (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30:1545–1614
20. Tamamis P, Adler-Abramovich L, Reches M, Marshall K, Sikorski P, Serpell L et al (2009) Self-assembly of phenylalanine oligopeptides: insights from experiments and simulations. Biophys J 96:5020–5029
21. Cecchini M, Rao F, Seeber M, Caflisch A (2004) Replica exchange molecular dynamics simulations of amyloid peptide aggregation. J Chem Phys 121:10748–10756
22. Paci E, Gsponer J, Salvatella X, Vendruscolo M (2004) Molecular dynamics studies of the process of amyloid aggregation of peptide fragments of transthyretin. J Mol Biol 340: 555–569
23. Mousseau N, Derreumaux P (2005) Exploring the early steps of amyloid peptide aggregation by computers. Acc Chem Res 38:885–891
24. Ma BY, Nussinov R (2006) Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. Curr Opin Chem Biol 10:445–452
25. Baumketner A, Shea JE (2007) The structure of the Alzheimer amyloid â 10–35 peptide, probed through replica-exchange molecular dynamics simulations in explicit solvent. J Mol Biol 366:275–285
26. Hall CK, Wagoner VA (2007) Computational approaches to fibril structure and formation. Methods Enzymol 412:338–365
27. Hills RD, Brooks CL III (2007) Hydrophobic cooperativity as a mechanism for amyloid nucleation. J Mol Biol 368:894–901
28. Knowles TP, Fitzpatrick AW, Meehan S, Mott HR, Vendruscolo M, Dobson CM et al (2007) Role of intermolecular forces in defining material properties of protein nanofibrils. Science 318:1900–1903
29. Nguyen PH, Li MS, Stock G, Straub JE, Thirumalai D (2007) Monomer adds to preformed structured oligomers of Aâ peptides by a two-stage dock-lock mechanism. Proc Natl Acad Sci U S A 104:111–116
30. Song W, Wei G, Mousseau N, Derreumaux P (2008) Self-assembly of the b2-microglobulin NHVTLSQ peptide using a coarse-grained protein model reveals a b-barrel species. J Phys Chem B 112:4410–4418
31. Tarus B, Straub JE, Thirumalai D (2008) Structures and free energy landscapes of the

wild type and mutants of the Abeta (21–30) peptide are determined by an interplay between intrapeptide electrostatic and hydrophobic interactions. J Mol Biol 379:815–829

32. Kollman PA (1993) Free energy calculations: applications to chemical and biochemical phenomena. Chem Rev 93:2395–2417
33. Massova I, Kollman PA (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. Perspect Drug Discov Des 18: 113–135
34. Archontis G, Simonson T, Moras D, Karplus M (1998) Specific amino acid recognition by aspartyl-tRNA synthetase studied by free energy simulations. J Mol Biol 275:823–846
35. Archontis G, Simonson T, Karplus M (2001) Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase. J Mol Biol 306:307–327
36. Archontis G, Watson KÁ, Xie Q, Andreou G, Chrysina E, Zographos SE et al (2005) Molecular recognition and relative binding of glucopyranose spirohydantoin analogues to glycogen phosphorylase: a free energy perturbation study. Proteins 61:984–998
37. Tamamis P, Morikis D, Floudas CA, Archontis G (2010) Species specificity of the complement inhibitor compstatin investigated by all-atom molecular dynamics simulations. Proteins 78: 2655–2667
38. Tamamis P, Pierou P, Mytidou S, Floudas CA, Morikis D, Archontis G (2011) Design of a modified mouse protein with ligand binding properties of its human analog by molecular dynamics simulations: the case of C3 inhibition by compstatin. Proteins 79:3166–3179
39. Kieslich C, Tamamis P, Gorham RD Jr, López de Victoria A, Sausman N, Archontis G et al (2012) Exploring protein-protein and protein-ligand interactions in the immune system using molecular dynamics and continuum electrostatics. Curr Phys Chem 2:324–343
40. Tamamis P, López de Victoria A, Gorham RD Jr, Bellows ML, Pierou P, Floudas CA et al (2012) Molecular dynamics simulations in drug design: new generations of compstatin analogs. Chem Biol Drug Des 79:703–718
41. López de Victoria A, Tamamis P, Kieslich CA, Morikis D (2012) Insights into the structure, correlated motions, and electrostatic properties of two HIV-1 gp120 V3 loops. PLoS One 7:e49925
42. Derreumaux P, Mousseau N (2007) Coarse-grained protein molecular dynamics simulations. J Chem Phys 126:025101
43. Melquiond A, Dong X, Mousseau N, Derreumaux P (2008) Role of the region 23–28 in Abeta fibril formation: insights from simulations of the monomers and dimers of Alzheimer's peptides Abeta 40 and Abeta 42. Curr Alzheimer Res 5:244–250
44. Han W, Schulten K (2012) Further optimization of a hybrid united-atom and coarse-grained force field for folding simulations: improved backbone hydration and interactions between charged side chains. J Chem Theory Comput 8:4413–4424
45. Still WC, Tempczyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 112:6127–6129
46. Bashford D, Case DA (2000) Generalized born models of macromolecular solvation effects. Annu Rev Phys Chem 51:129–152
47. Im W, Lee MS, Brooks CL III (2003) Generalized born model with a simple smoothing function. J Comput Chem 24:1691–1702
48. Feig M, Im W, Brooks CL III (2004) Implicit solvation based on generalized born theory in different dielectric environments. J Chem Phys 190:903
49. Chen J, Im W, Brooks CL III (2006) Balancing solvation and intramolecular interactions: towards a self-consistent generalized born force field. J Am Chem Soc 128:3728–3736
50. Chen J, Brooks CL III, Khandogin J (2008) Recent advances in implicit solvent-based methods for biomolecular simulations. Curr Opin Struct Biol 2:140–148
51. Haberthür U, Caflisch AJ (2008) FACTS: fast analytical continuum treatment of solvation. Comput Chem 29:701–715
52. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ et al (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616
53. Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. J Chem Phys 105:1902–1921
54. Buck M, Bouguet-Bonnet S, Pastor RW, MacKerell AD (2005) Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of Hen Lysozyme. Biophys J 90:L36–L38
55. Pieridou G, Avgousti-Menelaou C, Tamamis P, Archontis G, Hayes SC (2011) UV resonance Raman study of TTR (105–115) structural evolution as a function of temperature. J Phys Chem B 115:4088–4098
56. Swendsen R, Wang J (1987) Non-universal critical dynamics in Monte Carlo simulations. Phys Rev Lett 57:2607–2609

57. Hukushima K, Nemoto K (1996) Exchange Monte Carlo method and application to spin glass simulation. J Phys Soc Jpn 65:1604–1608

58. Hansmann U (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 281:140–150

59. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151

60. Sanbonmatsu KY, Garcia AE (2002) Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. Proteins 46:225–234

61. Nymeyer H, Gnanakaran S, Garcia A (2004) Atomic simulations of protein folding, using the replica exchange algorithm. Methods Enzymol 30:119–149

62. Rao F, Caflisch A (2003) Replica exchange molecular dynamics simulations of reversible folding. J Chem Phys 119:4035

63. Rosta E, Buchete N-V, Hummer G (2009) Thermostat artifacts in replica exchange molecular dynamics simulations. J Chem Theory Comput 5:1393–1399

64. Kim J, Keyes T, Straub J (2010) Generalized replica exchange method. J Chem Phys 132: 224107

65. Lee MS, Olson MA (2011) Comparison of two adaptive temperature-based replica exchange methods applied to a sharp phase transition of protein unfolding-folding. J Chem Phys 134: 244111

66. Frishman D, Argos P (1995) Knowledge-based secondary structure assignment. Proteins 23: 566–579

67. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

68. Karpen ME, Tobias DT, Brooks CL III (1993) Statistical clustering techniques for analysis of long molecular dynamics trajectories. I: analysis of 2.2 ns trajectories of YPGDV. Biochemistry 32:412–420

69. Chandrasekhar S (1992) Liquid crystals. Cambridge University Press, Cambridge

70. de Gennes PG, Prost J (1993) The physics of liquid crystals, 2nd edn. Oxford University Press, Oxford

71. Berardi R, Muccioli L, Zannoni C (2004) Can nematic transitions be predicted by atomistic simulations? A computational study of the odd-even effect. Chem Phys Chem 5: 104–111

72. Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A (2007) WORDOM: a program for efficient analysis of molecular dynamics simulations. Bioinformatics 23:2625–2627

73. Tamamis P, Terzaki K, Kassinopoulos M, Mastrogiannis L, Mossou E, Forsyth VT et al (2014) Self-assembly of an aspartate-rich sequence from the adenovirus fibre shaft: insights from molecular dynamics simulations and experiments. J Phys Chem B 118:1765–1774

Chapter 4

Posttranslational Incorporation of Noncanonical Amino Acids in the RNase S System by Semisynthetic Protein Assembly

Maika Genz and Norbert Sträter

Abstract

The unique ribonuclease S (RNase S) system, derived from proteolytic cleavage of bovine ribonuclease A (RNase A), consists of a tight complex formed by a peptide (amino acids 1–20) and a protein (21–124) part. These fragments, designated as S-peptide and S-protein, can be separated by two purification steps. By addition of synthetic S-peptide derivatives to the S-protein, semisynthetic RNase S is reassembled with high efficiency. Based on this peptide–protein complementation noncanonical amino acids can be easily introduced into a protein host. Here we describe the preparation of the S-protein from RNase A as well as the characterization of the reassembled semisynthetic RNase S complex. Complex formation can be monitored by RNase activity, circular dichroism, or fluorescence polarization. Structure-based enzyme design of the RNase S scaffold is possible based on high-resolution crystal structures of RNase S and its semisynthetic variants.

Key words Ribonuclease, RNase, Assembly, Peptide–protein-complementation, S-protein, Fluorescence polarization, Circular dichroism, Crystallization

1 Introduction

The commercial availability and extraordinary stability of ribonuclease A (RNase A) (EC 3.1.27.5) made (and still makes) the protein one of the most attractive objects for studying enzymes and developing new methods in protein biochemistry. For a detailed overview of RNase A the reviews of Raines [1] and Marshall [2] are highly recommended.

RNase A is a pancreatic enzyme, which hydrolyzes ribonucleic acid. It consists of 124 amino acids and contains eight cysteines which are all involved in disulfide bridges (Fig. 1b). Residues H12, K41, H119, and the peptide backbone of F120 are probably involved in stabilizing the transition state during the catalytic reaction [1]. Herein we focus on one of the most attractive features of RNase A—the RNase S system (Fig. 1a) [3]. RNase A can be cleaved

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_4, © Springer Science+Business Media New York 2014

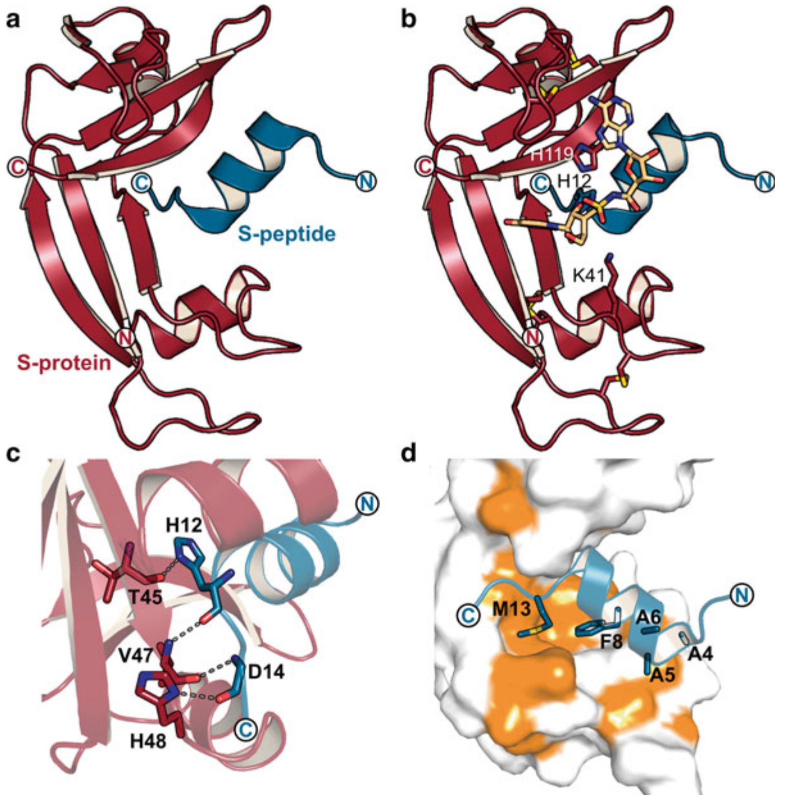


Fig. 1 RNase S wild-type structure (pdb 2RNS). (**a**) Overall structure of the RNase S complex containing S-protein (*red*) and S-peptide (*blue*). (**b**) Biochemical characteristics of the RNase S structure. The four disulfide bridges are displayed as *sticks* (between the cysteines at positions: 26 + 85, 40 + 95, 58 + 110, and 65 + 72). The active center comprising H12 (*blue sticks*) and H119 (*red sticks*) is shown in complex with a dinucleotide inhibitor (H119 and the inhibitor are superimposed from the RNase A structure 2XOG). K41 is essential for ribonuclease activity by stabilization of the transition state. (**c**) Hydrophobic interactions of the RNase S complex. The molecular surface of the S-protein (*white*) is shown such that hydrophobic surface areas (generated by amino acids V, L, I, M, or F) are depicted in *orange*. Hydrophobic amino acids of the S-peptide (*blue*) are displayed as *sticks* (A4, A5, A6, F8, M13). The main hydrophobic interaction occurs via F8 and M13 (sulfur in *yellow*). (**d**) Interactions of the C-terminal residues of the S-peptide with the edge of a β-sheet of the S-protein. From the S-peptide (*blue*) H12 and the main chain of D14 are involved in reassembly and displayed as *sticks*. Hydrogen bond distances are listed in Table 1

with the protease subtilisin gaining RNase S (Fig. 2a). The cleavage site is between the amino acids 20 and 21, generating an N-terminal fragment designated as S-peptide (1–20) and a C-terminal part termed the S-protein (21–124). The S-protein part can be purified from RNase S [4]. Afterwards the S-protein is able to reassemble with a synthetic peptide related to the S-peptide sequence (Fig. 2b).

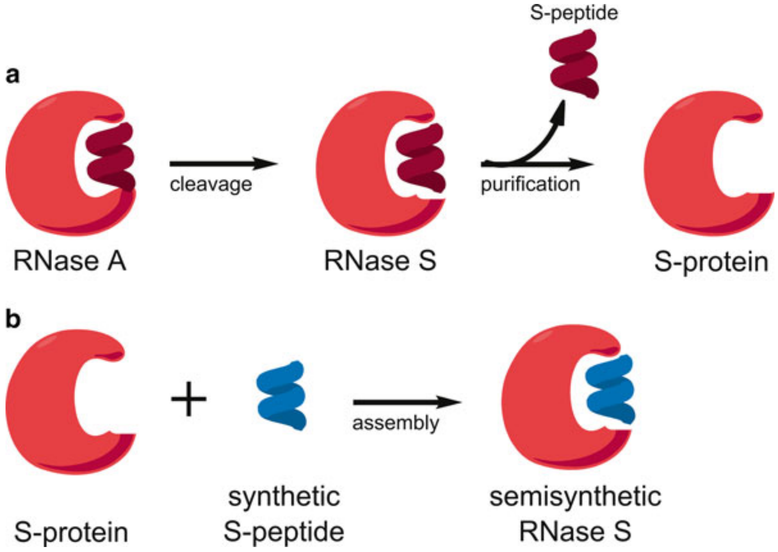


Fig. 2 Overview of the production of semisynthetic RNase S. (**a**) Cleavage of RNase A results in RNase S (step 1). The S-protein is isolated from the RNase S by removal of the S-peptide (step 2). (**b**) Reassembly of the S-protein (*red*) and a synthetic S-peptide (*blue*) results in a semisynthetic RNase S variant

For reassembling the RNase S complex only amino acids 1–15 of the S-peptide are essential [5]. The binding of the wild-type S-peptide to the S-protein is very specific and occurs tightly with a dissociation constant of 49 nM at 30 °C [6].

Usage of the RNase S system holds large potential for exploring new fields in protein biochemistry, in particular in enzyme design based on the introduction of noncanonical amino acids via a synthetic S-peptide. Here we share our knowledge in the production and characterization of semisynthetic RNase S variants. The structure of RNase S revealed that out of the 15 amino acids of the S-peptide part amino acids 3–13 form an α-helix (Fig. 1a) [5]. One face of the S-peptide helix is in contact with the S-protein, and the other side is exposed to the solvent. A hydrogen bonding network (Table 1 and Fig. 1c) and interactions with a hydrophobic pocket of the S-protein (Fig. 1d) contribute to the high affinity of the S-peptide. The C-terminal part of the S-peptide, mainly H12 and D14, forms hydrogen bonds to the edge of one of the two β-sheets of RNase S (Fig. 1c). Furthermore, the S-peptide interacts with a hydrophobic surface area of the S-protein via F8 and M13 (Fig. 1d). Exposure of this hydrophobic surface patch in the isolated S-protein probably contributes to the more hydrophobic behavior of this protein (our observations concerning the tendency of the S-protein to stick to vessel surfaces and to membranes). Separation of S-peptide and S-protein leads to a loss of ribonuclease activity [1]. None of the fragments have any hydrolase activity

Table 1
Hydrogen bonding interactions between S-peptide and S-protein in wild-type RNase S

S-peptide	S-protein	Distance [Å]
Arg 10 O	Arg 33 Nη1	2.79
Gln 11 O	Asn 44 Nδ	2.88
His 12 O	Val 47 N	2.77*
His 12 Nδ	Asn 44 Oδ	3.32
His 12 Nδ	Thr 45 O	2.85*
Met 13 O	Arg 33 Nη1	2.84
Met 13 O	Arg 33 Nη2	2.91
Asp 14 Oδ2	Tyr 25 OH	2.64
Asp 14 O	His 48 Nδ	2.95*
Asp 14 N	Val 47 O	2.82*
Ser 15 Oγ	Gln 49 O	2.80

Hydrogen bonds between S-peptide and S-protein of the RNase S wild-type complex (pdb code: 2RNS). C-terminal hydrogen bonds forming a pseudo-β-sheet-like structure are marked with an *asterisk* (*see* Fig. 1b)

as the active center is separated, with H12 on the S-peptide and H119 on the S-protein part (Fig. 1b). With reassembling S-peptide and S-protein to RNase S, the ribonuclease activity is restored. Although the separated S-peptide comprises the first 20 N-terminal amino acids, only amino acids 1–15 are required for reconstitution of complete ribonuclease functionality [5]. Before the development of modern molecular biology methods for mutational analysis, variants of the S-peptide were prepared by chemical synthesis and introduced into RNase S [7]. Later, non-proteinogenic amino acids were introduced into the RNase S system for various approaches, e.g., pyridoxamine phosphate-based amino acids for new catalytic activities [8], iminodiacetic acid groups for metal-induced regulation of ribonuclease activity [9], or phenylazophenylalanine for a photoswitchable RNase S variant [10]. Today, the peptide–protein complementation of the RNase S system is still a straightforward method to introduce non-proteinogenic amino acids into a protein scaffold.

Here we describe the preparation of the S-protein fragment starting from RNase A based on the early work of Richards and Vithayathil [4]. RNase A is cleaved by subtilisin to RNase S. From RNase S the S-protein is isolated via TCA precipitation and cation-exchange chromatography (Fig. 3a, b). By mixing the purified S-protein and a synthetic S-peptide, the RNase S complex can be restored.

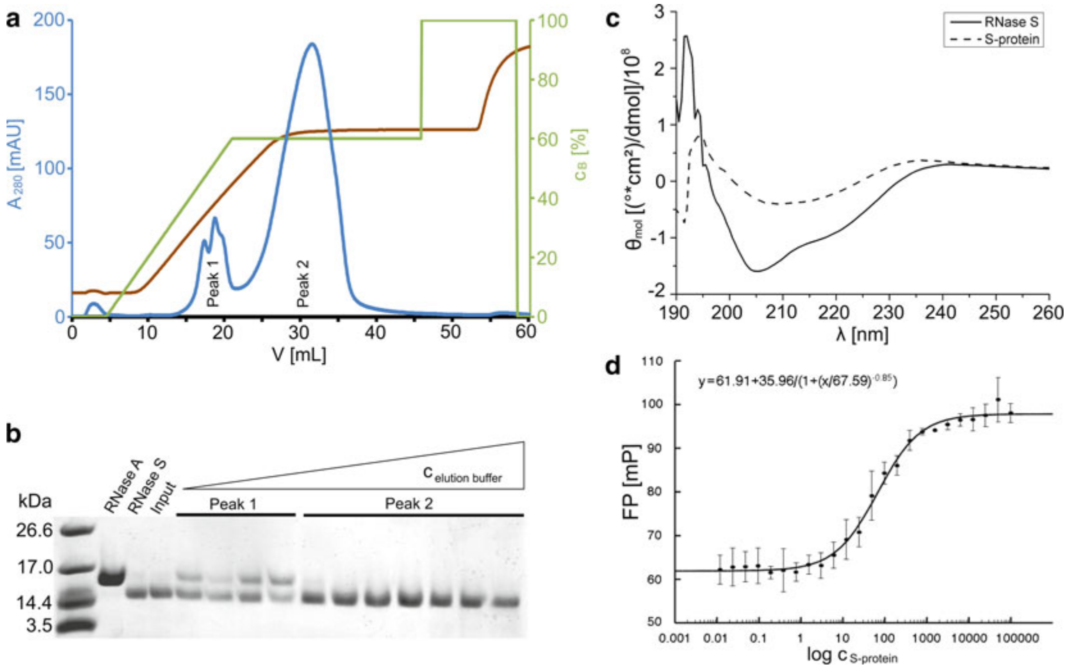


Fig. 3 S-protein purification and RNase S reassembly. (**a**) FLPC chromatogram on SP Sepharose (cation exchange) of cleaved RNase S after TCA precipitation and dialysis. The protein absorption at 280 nm is shown in *blue*. Peak 1 contains RNase A and S, Peak 2 contains purified S-protein (see SDS-PAGE in (**b**)). The percentile concentration of the elution buffer (c_B) is depicted as a *green line*; the corresponding conductivity is displayed in *brown*. (**b**) SDS-PAGE depicting RNase A cleavage (line 1 and 2) as well as identification of S-protein purification on cation-exchange chromatography. (**c**) CD spectra of S-protein (*dashed line*) and reassembled RNase S wild type with a synthetic S-peptide of residues 1–15 (*full line*). (**d**) FP titration curve of the S-protein in complex with a synthetic S-peptide (residues 1–15) (*black dots*). K_D was calculated by using the nonlinear dose–response fit (*solid line*), determined with 67.59 nM as calculated in the equation

Furthermore we highlight assays for monitoring RNase S complex formation. The first, probably simplest, method is checking S-peptide–S-protein assembly with the help of the restored ribonuclease activity. If the active center of the RNase S ($\text{H12}_{\text{peptide}}+\text{H119}_{\text{protein}}$) is correctly formed by peptide–protein complementation, RNase activity can be measured in various activity assays [11–14]. Here, we describe the methylene blue assay [15]. If the ribonuclease activity is not restored, other biophysical methods for RNase S complex formation can be employed. Via circular dichroism (CD) spectroscopy the change in secondary structure from S-protein to RNase S upon peptide binding can be monitored (Fig. 3c) [16]. The CD spectrum of the S-protein shows mainly β-sheet characteristics, whereas the isolated S-peptide is predominantly unfolded. In complex with the S-protein, the S-peptide adopts mainly an α-helical fold, which results in a significant change in the CD spectrum [17]. Alternatively, the dissociation constant (K_D) of the RNase S complex can be determined with the help of fluorescence polarization [18]. For this assay the S-peptide needs to

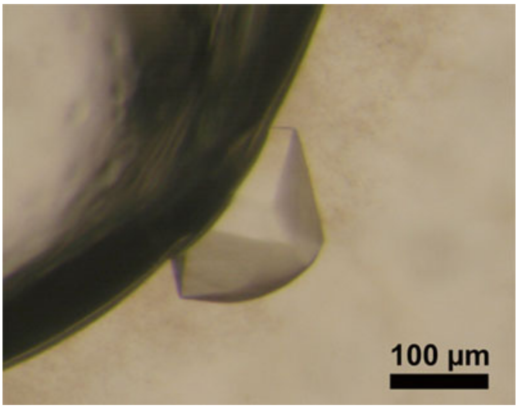


Fig. 4 Crystal of a semisynthetic RNase S construct. Crystals usually appear after 3–7 days

be labeled, e.g., at the N-terminus with 5(6)-carboxyfluorescein. Upon binding of the S-peptide to the S-protein the mobility of the S-peptide in solution is changed, resulting in a measurable shift in fluorescence polarization (Fig. 3d) [1]. Furthermore we describe a procedure for crystallization of semisynthetic RNase S constructs for X-ray crystallographic studies (Fig. 4) [6, 19].

2 Materials

All solutions are prepared with ultrapure water and analytical grade reagents. Solutions and reagents are stored at room temperature unless indicated otherwise.

2.1 RNase A Cleavage and S-Protein Purification

The following materials are needed for the preparation of S-protein from ~10 mg RNase A.

1. Cleavage buffer (50 mL): 25 mM Tris–HCl pH 8.0, 50 mM KCl (*see* **Note 1**).
2. Subtilisin stock solution (500 µL): 10 mg/mL subtilisin in water. Keep the solution on ice during use (*see* **Note 2**).
3. Hydrochloric acid (20 mL): 1 M solution in water (*see* **Note 3**).
4. Potassium hydroxide (20 mL): 1 M KOH (*see* **Note 3**).
5. TCA solution (50 mL): Prepare 20 % (w/v) TCA solution in water. Store at 4 °C.
6. Binding buffer (1 L): 20 mM Na-phosphate pH 5.0 (*see* **Note 4**).
7. Elution buffer (500 mL): 200 mM Na-phosphate pH 6.6, 20 mM Na-acetate.
8. Storage buffer (2 L): 20 mM Bis-Tris–HCl pH 6.5 (*see* **Note 5**).
9. Lyophilization buffer (2 L): 0.01 % (v/v) formic acid (pH ~3) (*see* **Note 6**).

10. pH universal indicator paper ranging from pH 0–14.
11. Dialysis membrane (MWCO 3,500).
12. Centrifugal concentrators (MWCO 5,000).
13. FPLC System including HiTrap SP HP 1 mL (material: SP sepharose high performance).

2.2 RNase S Assembly Assays

2.2.1 Methylene Blue Ribonuclease Assay

1. Reassembly buffer (50 mL): 25 mM Tris–HCl pH 8.0, 25 mM KCl.
2. Assay buffer (1 L): 0.1 M MOPS-NaOH pH 7.5, 2 mM EDTA. Store in a dark bottle at 4 °C (*see* **Note 7**).
3. RNA solution (50 mL): 10 mg/mL RNA in assay buffer. Store in a dark bottle at −20 °C.
4. RNase S (35 μL): 6 μM RNase S (1:1 mixture of S-protein and S-peptide) in reassembly buffer (*see* **Note 8**).
5. Cuvettes (we use disposable 1 mL cuvettes with 10 mm path length).

2.2.2 Circular Dichroism Spectroscopy (CD)

1. CD buffer (20 mL): 15 mM Tris–HCl pH 8.0.
2. Protein solution (400 μL): 15 μM S-protein in CD buffer.
3. Peptide solution (10 μL): 0.3 mM S-peptide in CD buffer.
4. Quartz cuvette (we use 1 mm path length Quartz cuvette).
5. CD-spectrometer setup: Range 260–190 nm, 0.5 nm per step, 5 s time point, 2 repeats, 1.0 nm bandwidth, T=25 °C (*see* **Note 9**).

2.2.3 Fluorescence Polarization Assay (FP)

1. FP buffer (5 mL): 15 mM Tris–HCl pH 8.0, 15 mM KCl.
2. S-protein solution (130 μL): 5 μM S-protein in FP buffer (*see* **Notes 10** and **11**).
3. N-terminally 5(6)-carboxyfluorescein-labeled peptide solution (250 μL): 20 nM S-peptide in FP buffer (*see* **Note 12**).
4. N-terminally 5(6)-carboxyfluorescein labeled peptide solution (3,000 μL): 10 nM S-peptide in FP buffer (*see* **Note 13**).
5. Black 384 well plates (F-shaped, Greiner Bio-One GmbH).
6. PARADIGM™ (Beckman, Coulter) containing the fluorescence polarization cartridge.

2.3 Crystallization

1. Reassembly buffer (50 mL): 25 mM Tris–HCl pH 8.0, 25 mM KCl (same as Subheading 2.2.1, **item 1**, *see* **Note 14**).
2. Buffer stock solutions (50 mL): 1 M Na-citrate pH 3.8, pH 3.9, pH 4.0, and pH 4.1.
3. Precipitant stock solution (250 mL): 3.8 M ammonium sulfate $(NH_4)_2SO_4$.
4. Peptide stock solution: 10 mg/mL S-peptide.

3 Methods

3.1 Preparation of the S-Protein

1. Weigh in 10 mg commercial RNase A in a 1.5 mL tube. Dissolve the RNase A in 1 mL cleavage buffer and place it on ice. Add 25 μL of subtilisin stock solution (10 mg/mL) and incubate for at least 2 h on ice (*see* **Note 15**).
2. To inactivate the subtilisin the pH of the protein solution is titrated to pH 3 with 1 M HCl and incubated on ice for 10 min. Afterwards the pH value is titrated to pH 8 with 1 M KOH (*see* **Note 16**).
3. Add 1/5 volume (~200 μL) 20 % ice-cold TCA solution and incubate overnight at room temperature to precipitate the S-protein (*see* **Note 17**).
4. Centrifuge the solution at 9,600 × *g* for 10 min at 4 °C. Discard the supernatant.
5. Resuspend the protein pellet in 1.4 mL water by using a thermomixer at 20 °C and 1,400 rpm (*see* **Note 18**).
6. Dialyze the resuspended protein against 2 L binding buffer for at least 2 h. Afterwards centrifuge the dialyzed solution at 16,800 × *g* to remove undissolved material which may affect the chromatography material in the following FPLC step (*see* **Note 19**).
7. Before applying the protein to the chromatography column, equilibrate the HiTrap SP HP column with the binding buffer and flush all tubings and the loop with the binding buffer.
8. Run the FPLC with the protocol specified in Table 2. Fractionate in RNase-free tubes in 1 mL aliquots. *See* Fig. 3a for an exemplary elution profile (*see* **Note 20**).
9. Concentrate the pooled S-protein-containing fractions by using centrifugal concentrators with a cutoff of 5,000 Da. Before applying the protein solution to the concentrator, wash the concentrator membrane with a mixture of 1:1 binding buffer:elution buffer with a volume of at least 5 mL (flow through). Centrifuge at 5,000 × *g* (*see* **Note 21**).

Table 2
S-protein purification protocol on HiTrap SP HP (1 mL)

Purification step	Column wash	Segment 1	Segment 2	Segment 3
Binding buffer in %	100	100 → 40	40	0
Elution buffer in %	0	0 → 60	60	100
Column volumes	2–5	20	30	10
Gradient type	Isocratic	Linear	Isocratic	Isocratic

10. Depending on further use, i.e., storage as frozen solution at −80 °C (a) or as lyophilized powder (b), dialyze the concentrated S-protein against 2 L storage buffer (a) or 2 L lyophilization buffer (b) (*see* **Note 22**).
11. (a) After dialysis overnight (or at least 3 h) against storage buffer at room temperature, prepare aliquots of 20–50 μL S-protein for storage at −80 °C (*see* **Note 23**). (b) After 2-h dialysis against lyophilization buffer at room temperature, prepare aliquots of 3 mg S-protein per tube. Freeze the aliquots for at least 30 min at −80 °C. Afterwards lyophilize the S-protein overnight (*see* **Note 24**).

3.2 Assays for RNase S Assembly

3.2.1 Methylene Blue Assay (Ribonuclease Activity)

1. Produce a methylene blue solution by dissolving a tip of a spatula methylene blue in assay buffer and dilute it with assay buffer until the OD at 688 nm is 0.5 (*see* **Note 25**).
2. Prepare at least 35 μL of RNase S from S-protein and S-peptide as a 1:1 mixture at a concentration of 6 μM (*see* **Note 26**).
3. Determine a calibration curve of RNA in methylene blue buffer ranking from 0 to 1,500 μg/mL RNA at OD_{688}. Calculate the dependency of RNA concentration on methylene blue absorption at OD_{688} with a standard hyperbolic function (*see* **Note 27**).
4. Prepare a solution of 0.8 mg/mL RNA in methylene blue buffer and incubate it at 30 °C. 3 mL is needed for one measurement in triplicate (*see* **Note 28**).
5. Pipette 10 μL of 6 μM RNase S solution to a cuvette. Start the reaction by adding 990 μL 0.8 mg/mL preheated RNA-methylene blue solution. Record immediately the change in adsorption at 688 nm for at least 90 s. Determine the activity from the linear slope (*see* **Note 29**).

3.2.2 Monitoring S-Peptide Binding via Circular Dichroism (CD)

1. Prepare 200 μL of S-protein solution in CD buffer by mixing 190 μL 15 μM S-protein and 10 μL CD buffer. Record a CD spectrum of the S-protein without S-peptide (*see* **Note 30**).
2. Prepare 200 μL of the RNase S solution with a molar ratio of 1:1 S-peptide:S-protein by mixing 190 μL 15 μM S-protein and 10 μL 0.3 mM S-peptide ($c_{\text{final}} = 14.25$ μM RNase S). Incubate the solution for 10 min at room temperature to allow RNase S complex formation. Record a CD spectrum using the same parameters as used for the S-protein spectrum (*see* **Note 31**).

3.2.3 Determination of S-Peptide Binding via Fluorescence Polarization (FP)

1. Pipette 40 μL 20 nM S-peptide solution in the first (left) well of each row of the multiwell plate. For the control fill 3 wells with 40 μL 20 nM S-peptide in the row that contains the control measurements (*see* **Notes 32** and **33**).
2. Pipette 40 μL 10 nM S-peptide in the following 23 wells for each row. For the control fill 3 wells with 40 μL 10 nM S-peptide in the row that contains the control measurements (*see* **Note 33**).

3. Add 40 μL S-protein stock solution ($c_{\text{start}} = 5$ μM) to the first (left) well of each row of the multiwell plate. Mix the solution with the 20 nM S-peptide solution by pipetting up and down. For the serial dilution, next transfer 40 μL from well 1 to the neighboring well 2 and again mix the solutions by pipetting up and down. Continue with the serial dilution procedure for wells 2–3, 3–4, and so on until 40 μL have been transferred to the last well of the row. From this well 40 μL are discarded to the waste after the mixing procedure, so that an equal volume is present in all wells. Be sure of sufficient mixing, as this will influence the precision of the results dramatically (*see* **Note 33**).
4. Incubate the FP plate for 30 min at room temperature in a dark place (*see* **Note 34**).
5. Measure the fluorescence anisotropy with an excitation wavelength of $\lambda_{\text{exc}} = 485$ nm at the emission wavelength of $\lambda_{\text{em}} = 535$ nm and at a temperature of 28–30 °C (*see* **Note 35**).
6. After calculation of the fluorescence polarization, the resulting data are plotted half logarithmically and fitted sigmoidal to determine the dissociation constant (K_D) of the peptides (*see* **Note 36** and Fig. 3d).

3.3 Crystallization of RNase S

1. Assemble the RNase S complex at a molar ratio of 1:3 (S-protein: S-peptide) at a final S-protein concentration of 4–5 mg/mL in 25 mM Tris–HCl pH 8.0 buffer for 30 min (*see* **Note 37**).
2. Prepare the crystallization plate with buffer conditions in a range of 100 mM Na-citrate pH 3.8–4.1 and precipitant concentration ranging from 2.2 to 3.2 M $(NH_4)_2SO_4$ (*see* **Note 38**).
3. Mix 1 μL RNase S solution with 1 μL reservoir buffer and incubate at 19 °C by using the hanging drop vapor diffusion technique. Under these conditions crystals are usually obtained after 3–7 days.
4. Transfer a crystal to the cryobuffer containing 1 M lithium chloride in the reservoir solution and incubate for ca. 3 min. Afterwards the crystal is flash-cooled in liquid nitrogen (*see* **Note 39**).

4 Notes

1. Because most of the buffers are based on tris(hydroxymethyl) aminomethane (Tris) and potassium chloride (KCl), two stock solutions may be prepared: 500 mL 1 M Tris–HCl pH 8.0 and 500 mL 2 M KCl.
2. The stock solution can be prepared in either water or cleavage buffer. For storage, place the subtilisin stock solution in a

−20 °C freezer (shock-freezing is not necessary). For further usage thaw the frozen subtilisin solution by placing the tube on the top of ice (in the ice, the thawing process will take much longer).

3. For preparation of 10 mg RNase A, at most 500 μL of each solution is needed. Concerning the number of preparations, adjust the volume of the HCl or KOH solution.
4. Prepare 500 mL 1 M stock solution of sodium phosphate (Na-phosphate) at pH 5.0. This stock solution is needed for dialysis and FPLC.
5. As the storage buffer will be used for dialysis, prepare a stock solution of 500 mL 1 M bis(2-hydroxyethyl)-amino-tris (hydroxymethyl)-methane (Bis-tris) and titrate with HCl to pH 6.5.
6. Fill 2 L of water in a graduated flask and add 200 μL formic acid (HCOOH).
7. Prepare a stock solution of 100 mL of 500 mM ethylenediaminetetraacetic acid (EDTA). Adjust the pH to 8.0 with sodium hydroxide (add slowly as solid or ~10 M solution). Be aware that EDTA at this concentration completely dissolves only during the addition of the base.
8. For each measurement 10 μL of 6 μM RNase S is needed (30 μL for triplicate measurements).
9. Setup parameters for CD spectroscopy may vary slightly between different equipment; thus the parameters are only guidance values.
10. For one data row of the titration curve 40 μL S-protein solution is needed. Thus for triplicate measurements 120 μL S-protein is needed to determine the K_D value of one S-peptide variant.
11. With the specified concentration of 5 μM of the S-protein solution, the highest concentration of the S-protein will be 2.5 μM. If the K_D value is around or above this concentration, higher concentrations of the S-protein solution are needed to achieve binding. At very high protein concentration, the fluorescence polarization changes nonspecifically due to the increased viscosity of the solution.
12. As starting point 40 μL 20 nM S-peptide is needed. For triplicate measurements and control measurements at least 240 μL 20 nM S-peptide will be needed.
13. For a triplicate measurement of one titration curve ~3,000 μL 10 nM S-peptide solution is needed (*see* **Note 33** for the pipetting scheme).

14. For crystallization, also other compositions of the reassembly buffer (or protein buffer) may work. Make sure that the pH range is around 8. Additionally avoid high salt concentration (>200 mM) or buffer components that may influence RNase S complex formation.

15. For the cleavage we use 1:40 (w/w) subtilisin to RNase A. The reaction can run longer but usually almost all of the RNase A is cleaved after 2 h. Make sure that the solution is on ice during the entire reaction time.

16. About 50–100 μL HCl is needed to adjust the pH to 3. Check the pH of the protein solution with the pH paper. Use a small amount of protein solution (less than 1 μL) for the pH measurements to minimize the loss of protein. An almost equal amount of KOH is needed to titrate the pH back to 8. Be careful not to overtitrate as protein denaturation may result at pH < 3 or pH > 9.

17. Depending on the volume of HCl and KOH, the amount of TCA solution may vary between 200 and 250 μL. The S-protein mostly precipitates directly after addition of the TCA solution. However, for a higher yield incubate the solution overnight at room temperature.

18. Resolubilization of the protein pellet can be initiated by carefully pipetting the solution up and down, until small particles of the precipitant are obtained in the solution. The solubilization by using a thermomixer can take up to 2 h. For resolubilization of the precipitated protein use at least 1.4 mL (in a 1.5 mL tube). Up to 1.9 mL may be used in a 2 mL tube.

19. As an alternative to the dialysis step it is possible to dilute the protein solution in binding buffer or water to a final volume of at least 10 mL. This is necessary to keep the TCA concentration as low as possible since it may affect protein binding to the cation-exchange chromatography material. We obtain a better yield when dialyzing the protein against binding buffer.

20. The S-protein fraction is usually obtained in the isocratic segment 2. Depending on the amount of S-protein binding to the SP Sepharose, the S-protein may additionally elude in segment 4. In segment 1 one or two small peaks often appear. These arise from contaminations of RNase A or RNase S in the S-protein preparation.

21. It is important that the centrifugal force does not exceed 5,000 × *g*. The S-protein tends to clog the ultrafiltration membrane which may break at higher centrifugal forces. Always check if the flow through contains protein due to a broken membrane. If the S-protein will be lyophilized in the end, the pH may be titrated to 5.0 with HCOOH before the concentration step, as this will fasten the concentration procedure.

22. The storage buffer should be slightly acidic if possible as this will minimize protein loss resulting from S-protein binding to

surfaces and membranes. We usually store the S-protein in 20 mM Bis-tris buffer pH 6.5. S-protein solutions should be stored at −20 °C (or better at −80 °C) for prolonged storage. For short-term storage (e.g., the next day), it is preferable to store it at 4 °C to avoid too many freezing-thawing cycles.

23. Determine the protein concentration after the dialysis step using the dialysis buffer as blank. After aliquotation directly store the tubes at −80 °C. No shock-freezing in liquid nitrogen is necessary.

24. Do not exceed the dialysis time of 2 h against the lyophilization buffer. 3 mg S-protein per tube is a reasonable protein amount for crystallization studies. The lyophilized S-protein is stored at −20 °C until usage. Check the pH value of S-protein solution obtained from lyophilized powder, as it might be acidic from the remaining HCOOH.

25. Use the assay buffer as the blank. Mix the methylene blue and the buffer with a magnetic stir until a homogenous blue solution is observed. After reaching an absorption of $OD_{688} = 0.5$, let the reaction stir for at least ten more minutes vigorously (caution!) and check the OD_{688} again. If OD_{688} is constant at 0.5 the solution is ready to use and should be stored in dark bottles at 4 °C. Do not store the methylene blue solution for longer than 4 weeks.

26. For each measurement 10 μL RNase S solution (6 μM) is needed. When using the wild-type S-protein with a molecular weight of 11,542 Da the concentration of 6 μM corresponds to 0.07 mg/mL. The amount of 35 μL has been calculated for triplicate measurements. Store the RNase S solution at room temperature until usage.

27. For each new methylene blue and/or RNA solution a new calibration curve needs to be determined. We use the following concentrations: 0, 50, 75, 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800, 900, 1,000, 1,100, 1,250, and 1,500 μg/mL. First pipet the 10 mg/mL RNA solution into the cuvette, i.e., 0, 5, 7.5, …, 150 μL. Afterwards add the methylene blue solution. Adapt the volume of methylene blue solution to fill the cuvette to a volume of 1 mL. We neglect the resulting variations in methylene blue concentration. We determine the calibration curve in triplicate measurements.

28. Prepare a new RNA-methylene blue solution each day, e.g., 50 mL for 50 measurements. Due to inhomogeneity of the commercial RNA, try to avoid using different RNA stock solutions when data from different mutants or conditions need to be compared. If different RNA stock solutions have to be used, make sure that they have almost equal OD_{688} after mixing with methylene blue buffer (±0.05). Keep the RNA-methylene blue solution in a dark tube.

29. Be aware to avoid the formation of air bubbles in the cuvette and discard measurements where air bubbles occur. Determine the initial linear decrease of the OD_{688} value. It is not mandatory to use the complete measurement of 90 s. In fact, due to the nonlinear relationship between absorbance change and changes in substrate concentration, deviations from linearity in $\Delta OD_{688}/\Delta t$ may arise from larger changes in RNA concentrations. Make sure that the values to determine $\Delta OD_{688}/\Delta t$ are taken from a range with linear OD_{688} change. From the negative slope at OD_{688} the rate of RNA hydrolysis and the specific activity can be calculated. We obtained a specific activity of ~30 U/mg for wild-type RNase S. Greiner-Stöffele et al. define 1 unit of enzyme activity as the amount of RNase that leads to an absorbance change of 0.01 min^{-1} at 25 °C at an RNA concentration of 0.8 mg/mL [15].

30. Make sure that no air bubbles occur when transferring the solutions to the cuvette. Use a small tip and pipette the solution slowly along the inner wall of the cuvette. Depending on the path length of the quartz cuvette it may be necessary to adapt the protein concentration. Here all concentrations are calculated for using a cuvette with 1 mm path length, which usually contains a volume of 100–200 μL. The recorded spectrum should reflect the typical behavior of β-sheets with one (relatively wide) minimum at 218 nm ($\Pi \rightarrow \Pi^*$) and one maximum at 196 nm ($n \rightarrow \Pi^*$); *see* Fig. 3c.

31. For the wild-type S-peptide with residues 1–15 prepare a 0.53 mg/mL solution (MW = 1,749 Da; c = 0.3 mM) in CD buffer. Add 10 μL S-peptide solution to the 190 μL S-protein solution for a molar ratio of 1:1. (It is also possible to reassemble the RNase S complex in the cuvette. Be sure to adapt the volume of S-peptide solution to achieve a molar ratio of 1:1.) The spectrum of RNase S reveals two minima at 209 nm ($\Pi \rightarrow \Pi^*$) and 222 nm ($n \rightarrow \Pi^*$) as well as a maximum at 192 nm ($\Pi \rightarrow \Pi^*$), resulting from the larger content of α-helices in RNase S compared to the S-protein (*see* Fig. 3c). If no complex formation occurs, the CD spectrum of the mixture of S-protein and S-peptide resembles that of the S-protein as the unstructured S-peptide does not give rise to the maximum at 192 nm.

32. Because the S-protein tends to adhere to surfaces, we recommend blocking the FP plates before use. We use the following buffer for blocking: 0.5 % (w/v) casein, 10 mM sodium phosphate buffer pH 7.4, 0.3 M NaCl, and 0.05 % (v/v) Tween20. The plates are incubated for 1 h at room temperature and afterwards washed three times with the same buffer without casein (10 mM sodium phosphate buffer pH 7.4, 0.3 M NaCl, and 0.05 % (v/v) Tween20).

33. A typical pipetting scheme for FP titration of one S-peptide variant is shown in Table 3.

Table 3
Serial dilution pipetting scheme for a 384-well format

		1	2	3	4	…	24	
A	S-peptide S-protein	X	40 µL 20 nM –	40 µL 20 nM –	40 µL 20 nM –	…	X	
B	S-peptide S-protein	40 µL 20 nM 40 µL 5 µM	40 µL 10 nM → 40 µL	40 µL 10 nM → 40 µL	40 µL 10 nM → 40 µL	…	40 µL 10 nM → 40 µL	→ 40 µL
C	S-peptide S-protein	40 µL 20 nM 40 µL 5 µM	40 µL 10 nM → 40 µL	40 µL 10 nM → 40 µL	40 µL 10 nM → 40 µL	…	40 µL 10 nM → 40 µL	→ 40 µL
D	S-peptide S-protein	40 µL 20 nM 40 µL 5 µM	40 µL 10 nM → 40 µL	40 µL 10 nM → 40 µL	40 µL 10 nM → 40 µL	…	40 µL 10 nM → 40 µL	→ 40 µL
…	**…**	…	…	…	…	…	…	
P	S-peptide S-protein	X	40 µL 10 nM –	40 µL 10 nM –	40 µL 10 nM –	…	X	

Rows A and P contain the solutions for the control measurements and the empty wells used for calibration measurements (marked X). Each row (B to O) contains the solutions for one titration curve. If each data point is measured in triplicate, rows B to D are identical for a given column and are averaged to obtain the fluorescence polarization of the corresponding S-protein concentration. Note that 40 µl are discarded after the completion of the serial dilution for each row

Table 4
Pipetting scheme for crystallization plate

$(NH_4)_2SO_4$ [M] Na-citrate pH	2.2	2.4	2.6	2.8	3.0	3.2
3.8						
3.9						
4.0						
4.1						

34. Shorter or longer incubation times may work fine.
35. We measure the FP on a PARADIGM™ (Beckman Coulter) using the fluorescence polarization detection cartridge in top read position.
36. We determine the dissociation constants with the program *SlideWrite* by using the dose–response logistic transition function [$y = a_0 + a_1/(1 + (x/a_2)^{a3})$, with $x = \lg (c_{\text{protein}})$ and $y = FP$]. The dissociation constant is represented by the a_2 coefficient (SlideWrite, Encinitas).
37. The RNase S complex formation between S-protein and S-peptide happens very fast, so 30 min should be sufficient time for reassembly.
38. For a typical crystallization plate scheme see below (Table 4).
39. At higher $(NH_4)_2SO_4$ concentration, lower LiCl concentration may be used. At around 3.2 M $(NH_4)_2SO_4$ no additional cryoprotectant is needed, as ammonium sulfate provides sufficient cryoprotection at this concentration. Glycerol at ~10 % is also a suitable cyroprotective for these crystallization conditions.

Acknowledgment

This work was supported by the European Fund for Regional Structure Development (EFRE, European Unit). The authors like to thank the group of Prof. Dr. Ralf Hoffmann, especially Dr. David Singer for peptide synthesis as well as Dr. Daniel Knappe and Nicole Berthold, for introduction into fluorescence polarization.

References

1. Raines RT (1998) Ribonuclease A. Chem Rev 98:1045–1065
2. Marshall GR, Feng J, Kuster DJ (2008) Back to the future: ribonuclease A. Biopolymers 90: 259–277
3. Richards FM (1955) Titration of amino groups released during the digestion of ribonuclease by subtilisin. C R Trav Lab Carlsberg Chim 29:322–328
4. Richards FM, Vithayathil PJ (1959) The preparation of subtilisin-modified ribonuclease and separation of the peptide and protein components. J Biol Chem 234:1459–1465
5. Taylor HC, Komoriya A, Chaiken IM (1985) Crystallographic structure of an active, sequence-engineered ribonuclease. Proc Natl Acad Sci U S A 82:6423–6426
6. Genz M, Singer D, Hey-Hawkins E, Hoffmann R, Sträter N (2013) Crystal structure of Apo- and metalated thiolate containing RNase S as structural basis for the design of artificial metalloenzymes by peptide-protein complementation. Z Anorg Allg Chem 639: 2395–2400
7. Barnard EA (1969) Ribonucleases. Ann Rev Biochem 38:677–732
8. Imperiali B, Roy RS (1994) Coenzyme: amino acid chimeras—new residues for the assembly of functional proteins. J Am Chem Soc 116: 12083–12084
9. Hamachi I, Yamada Y, Matsugi T, Shinkai S (1999) Single- or dual-mode switching of semisynthetic ribonuclease S' with an iminodiacetic acid moiety in response to the copper(II) concentration. Chem Eur J 5:1503–1511
10. Hamachi I, Hiraoka T, Yamada Y, Shinkai S (1998) Photoswitching of the enzymatic activity of semisynthetic ribonuclease S' bearing phenylazophenylalanine at a specific site. Chem Lett 6:537–538
11. Dickman SR, Trupin K (1959) Ribonuclease assay based on uridine phosphate determination. Arch Biochem Biophys 82:355–361
12. Korn K, Greiner-Stoeffele T, Hahn U (2001) Ribonuclease assays utilizing toluidine blue indicator plates, methylene blue, or fluorescence correlation spectroscopy. Methods Enzymol 341:142–153
13. Lee CC, Trotman CNA, Tate WP (1983) A ribonuclease assay: separation of product from undegraded RNA substrate. Anal Biochem 135:64–68
14. Postek KM, LaDue T, Nelson C, Sandwick RK (1992) Spectrophotometric ribonuclease assays using dinucleoside monophosphate substrates. Anal Biochem 203:47–52
15. Greiner-Stoeffele T, Grunow M, Hahn U (1996) A general ribonuclease assay using methylene blue. Anal Biochem 240:24–28
16. Simons ER, Blout ER (1968) Circular dichroism of ribonuclease A, ribonuclease S, and some fragments. J Biol Chem 243:218–221
17. Bastos M, Pease JH, Wemmer DE, Murphy KP, Connelly PR (2001) Thermodynamics of the helix-coil transition: binding of S15 and a hybrid sequence, disulfide stabilized peptide to the S-protein. Proteins 42:523–530
18. Moerke NJ (2009) Fluorescence polarization (FP) assays for monitoring peptide-protein or nucleic acid: protein binding. Curr Protoc Chem Biol 1:1–15
19. Kim EE, Varadarajan R, Wyckoff HW, Richards FM (1992) Refinement of the crystal structure of ribonuclease S. Comparison with and between the various ribonuclease A structures. Biochemistry 31:12304–12314

Chapter 5

Design, Synthesis, and Study of Fluorinated Proteins

Benjamin C. Buer and E. Neil G. Marsh

Abstract

Highly fluorinated analogs of hydrophobic amino acids have proven to be generally effective in increasing the thermodynamic stability of proteins. These non-proteogenic amino acids can be incorporated into both α-helix and β-sheet structural motifs and generally enhance protein stability towards unfolding by heat and chemical denaturants, and retard their degradation by proteases. Recent detailed structural and thermodynamic studies have demonstrated that the increase in buried hydrophobic surface area that accompanies fluorination is primarily responsible for the stabilizing properties of fluorinated side chains. Fluorination appears to be a particularly useful strategy for increasing protein stability because fluorinated amino acids closely retain the shape of the side chain, and are thus minimally perturbing to protein structure and function. The first part of this chapter discusses some examples of highly fluorinated model proteins designed by our laboratory and protocols for their synthesis. In the second part, methods for determining their thermodynamic stability, along with conditions that have proven to be useful for crystallizing these proteins, are presented.

Key words Fluorinated proteins, Protein stabilization, Hexafluoroleucine, Protein design, Hydrophobic effect, Coiled-coil proteins, Peptide synthesis, Fluorinated amino acids

1 Introduction

Although fluorine is essentially absent from biology, fluorine has proven to be a remarkably useful element to probe the underlying principles of biological molecules. For example, fluorinated substrates have been extensively used to investigate enzyme mechanisms, and ^{19}F NMR has proven to be a valuable tool for studying structure, dynamics, and interactions of fluorine-labeled proteins, peptides, lipids, and nucleic acids [1–11]. Fluorinated molecules also have important medical applications, as illustrated by the fact that ~20 % of all pharmaceuticals contain fluorine, which improves pharmacokinetic properties [12].

The development of various methods that allow a wide variety of unnatural amino acids to be incorporated into proteins has greatly expanded the possibilities for modifying protein structure. Indeed, it is now possible to introduce a diverse range of chemical

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_5, © Springer Science+Business Media New York 2014

functionality that is not seen in Nature into proteins. Over the last decade, our laboratory has focused on the incorporation of fluorinated amino acids into de novo-designed proteins [13–19] and bioactive peptides [1, 8, 20–22]. We have been particularly interested in understanding how fluorination stabilizes proteins against unfolding and in exploiting the NMR activity of fluorine to report on peptide-membrane interactions, and, most recently, on the pathways by which amyloid-forming peptides aggregate [10, 11]. Other laboratories, of course, have also made important contributions to this field. Thus the ability of a wide variety of fluorinated amino acids to modulate the properties of various proteins and peptides has been explored [23–54] and this field continues to expand as conveyed in numerous review articles [55–60].

Although fluorinated amino acids have largely been of academic interest, their potential to increase the thermodynamic stability of proteins could have industrial applications. Enhancing the stability of protein-based therapeutics could, for example, decrease their susceptibility to proteases, thereby increasing their potency. Likewise many industrial processes that use enzymatic transformations would also benefit from enzymes that have increased resistance towards heat and chemical denaturation.

1.1 Origin of the Stabilizing Effects of Fluorinated Amino Acid Residues

Extensively fluorinated analogs of hydrophobic amino acids have often been found to stabilize protein structure while being minimally perturbing to protein function (Fig. 1). However, to effectively utilize fluorinated amino acids in protein design it is important to understand how they stabilize proteins. Soluble proteins primarily derive their stability from the packing of nonpolar residues into the hydrophobic protein interior, i.e., the hydrophobic effect. This is primarily an entropic effect that is associated with the release of ordered water molecules that form clathrates around the exposed hydrophobic residues in the unfolded state. The change in free energy of folding due to the hydrophobic effect is proportional to the change in buried hydrophobic surface area on folding and has been quantified through various studies on proteins and hydrophobic small molecules. The general consensus is that the hydrophobic effect contributes 25–30 cal/mol/Å^2 of buried surface area to the stability of a globular protein [61–64].

Initially there was some thought that extensively fluorinated amino acids might behave differently from natural hydrocarbon amino acids [18, 19, 23–26, 32, 44, 47, 48, 51, 65]. Perfluorinated organic molecules often have significantly different physical properties than their hydrocarbon counterparts, a phenomenon that has been described as the “polar hydrophobic” effect [66]. Perfluorinated solvents exhibit unusual phase-segregating properties that have been effectively exploited in “fluorous” synthesis methodology [67]. However, as discussed below, our studies indicate

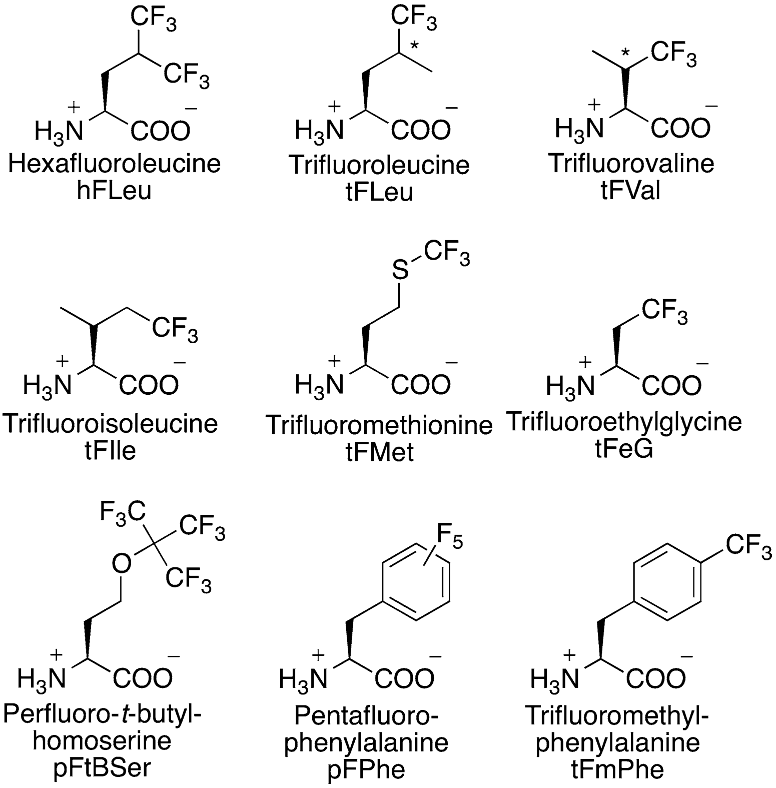


Fig. 1 Highly fluorinated amino acids that have been incorporated into proteins. The abbreviations are those that are used in this chapter (*asterisk* denotes a racemic stereocenter)

that, in the context of protein folding, extensively fluorinated amino acids such as hexafluoroleucine can be considered as conventional hydrophobic residues.

In the first part of this chapter we present some pertinent findings from our laboratory and summarize important contributions of other researchers relating to the design of proteins that incorporate extensively fluorinated amino acids. Our intention is to provide the reader with an overview of the state of the field, and present some general principles for consideration when designing fluorinated proteins. In the second part of the chapter we detail some of the methods commonly used in our laboratory for the preparation and characterization of fluorinated peptides.

1.2 Designing hFLeu-Containing α-Helical Proteins

As a model system with which to investigate the effects of incorporating fluorinated amino side chains we designed a small protein that we call α_4 [19]. α_4 comprises a 27-residue peptide chain that folds into a tetrameric antiparallel, four-helix bundle (Fig. 2). Such α-helical coiled-coil proteins provide simple model systems with well-defined hydrophobic cores created by contacts between

a

		abcdefg	abcdefg	abcdefg	a
α_4H	Ac-GN	ADELYKE	LEDLQER	LRKLRKK	LRSG-NH2
α_4Ht	Ac-GN	ADEXYKE	XEDXQER	XRKXRKK	XRSG-NH2
α_4F_2(6,24)	Ac-GN	ADEXYKE	LEDLQER	LRKLRKK	XRSG-NH2
α_4F_2(10,20)	Ac-GN	ADELYKE	XEDLQER	LRKXRKK	LRSG-NH2
α_4F_2(13,17)	Ac-GN	ADELYKE	LEDXQER	XRKLRKK	LRSG-NH2
α_4F_3(6-13)	Ac-GN	ADEXYKE	XEDXQER	LRKLRKK	LRSG-NH2
α_4F_3(17-24)	Ac-GN	ADELYKE	LEDLQER	XRKXRKK	XRSG-NH2
α_4F_3a	Ac-GN	ADELYKE	XEDLQER	XRKLRKK	XRSG-NH2
α_4F_3d	Ac-GN	ADEXYKE	LEDXQER	LRKXRKK	LRSG-NH2
$\alpha_4F_3$at	Ac-GN	ADEXYKE	XEDXQER	XRKXRKK	XRSG-NH2
$\alpha_4F_3af_3d$	Ac-GN	ADEXYKE	XEDXQER	XRKXRKK	XRSG-NH2
α_4F_6	Ac-GN	ADEXYKE	XEDXQER	XRKXRKK	XRSG-NH2

L = Leu
X = tBAla
X = hFLeu
X = tFeG

b

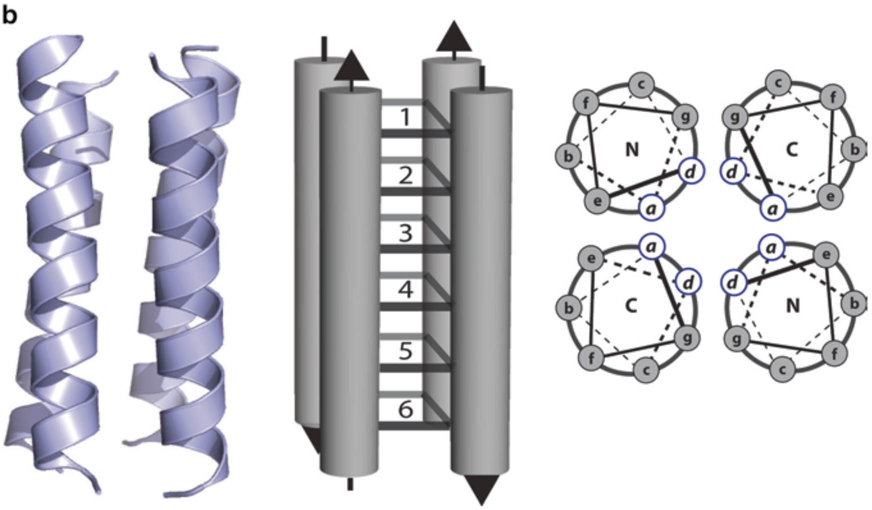


Fig. 2 Sequences and structures of α_4 proteins. (**a**) The 27-residue sequence of α_4 proteins designed in our studies, with noncanonical amino acid substitutions at *a* and *d* positions denoted by "X." (**b**) *Left*: Cartoon representation of a tetrameric, antiparallel, 4-helix bundle protein. *Middle*: Cartoon displaying the hydrophobic layers formed by two *a* and two *d* residues. *Right*: Helical wheel diagram illustrating residue placement for α_4 proteins

residues of adjacent helices. The geometry of an α-helix is defined by hydrogen bonding between every fourth residue (*i* to *i*+4); supercoiling of α-helices around each other results in a repeat pattern of seven residues for every two turns of the helix. These repeating positions of amino acids are referred to as the "heptad repeat," in which residues at the *a* and *d* positions point into the hydrophobic core, whereas residues at the *b*, *c*, *e*, and *g* positions are generally polar and form stabilizing salt bridges and hydrogen bonding interactions. α_4 was designed with two distinct polar interfaces that are formed between adjacent helices through interactions between residues in the *b* and *e* and *c* and *g* positions to enforce the four-helix antiparallel topology. We note that the antiparallel topology is more structurally robust than the parallel topology: parallel α-helix bundles are known to form different

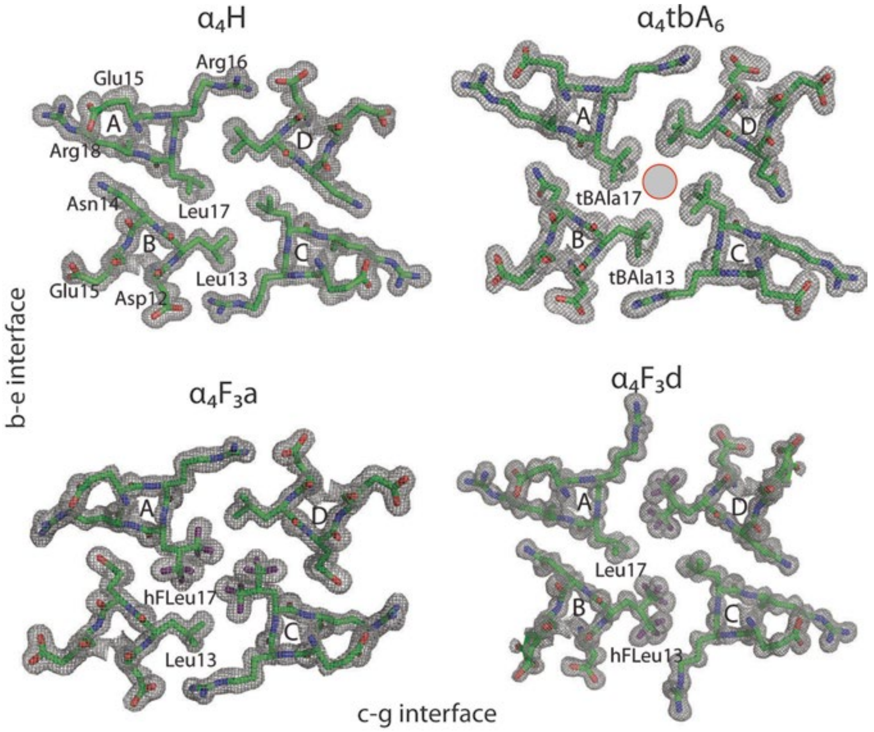


Fig. 3 X-ray structures of 4 α_4 proteins packed with either fluorinated or non-fluorinated amino acid residues. The figure displays a cross section of one layer of the hydrophobic core packing (the same layer in each protein). α_4H is the parent protein, packed with Leu at *a* and *d* positions. hFLeu is well accommodated in the structures of α_4F_3a and α_4F_3d which have hFLeu at the *a* and *d* positions, respectively. However the introduction of tBAla into the structure of $\alpha_4$tbA$_6$ reorganizes the protein core so that a void runs through the center of the protein, represented by the circle. Electron density maps ($2F_o - F_c$) are contoured at 1.0σ

oligomeric states in response to even subtle changes to the residues at *a* and *d* [25, 26, 68].

We have incorporated various numbers of hFLeu residues into the hydrophobic *a* and *d* positions of α_4 in a variety of combinations and in all cases this results in a more stable protein that retains its intended structure. The per-residue increase in stability depends upon the context. Introducing hFLeu at the terminal *a* and *d* positions results in a relatively small increase in stability of ~0.1 kcal/mol hFLeu/residue, presumably because these positions remain partially solvent exposed in the folded protein. Incorporating hFLeu into fully buried *a* and *d* positions results in a larger increase in stability of ~0.25 kcal/mol hFLeu/residue. In general, increasing the number of hFLeu residues results in a monotonic increase in stability. However, within the context of the α_4 four-helix bundle, an alternating pattern of two Leu residues and two hFleu residues per layer of the hydrophobic core (Fig. 3) that results from introducing hFLeu at either only *a* positions (α_4F_3a) or only *d* positions (α_4F_3d) appears to give the greatest per-residue stability of, respectively, 0.80 and 0.72 kcal/mol hFLeu/residue.

To investigate the origin of the apparently general stabilization of proteins by fluorination, we recently conducted an extensive thermodynamic analysis of a series of 12 α_4 variants incorporating different numbers of fluorinated residues at different positions within the protein core. For each of these we determined $\Delta G°$, $\Delta H°$, $\Delta S°$, and $\Delta C_p°$ for unfolding. Our analysis demonstrated that the increase in $\Delta G°_{\mathrm{unfold}}$ correlates well with the increase in buried hydrophobic surface area that occurs when a methyl group is substituted for a trifluoromethyl group, which results in an increase in volume of ~16 Å^3 [14]. Furthermore, most of this increase is accounted for by changes in entropy, as expected for the hydrophobic effect. From these experiments we calculated a value of ~28 cal/mol/Å^2 which is associated with burying a fluorinated side chain in the protein core, a value very similar to that for natural hydrophobic amino acids.

The question therefore arises: Why are fluorinated amino acids generally effective at stabilizing protein structure? Significant increases in protein stability have been obtained by mutating amino acids within the hydrophobic core to increase hydrophobicity, van der Waals contacts, and packing efficiency [62, 69]. However, attempts to stabilize proteins by point mutations invariably require replacing one residue for a larger and differently shaped hydrophobic side chain. Very often this disrupts the local structure and may lead to misfolding, and/or impairment of the protein's biological function.

We have used X-ray crystallography to solve the structures for a number of the α_4 variants containing either Leu, hFLeu, or β-t-butylalanine (tBAla) within their hydrophobic cores [15, 16]. From a detailed analysis of their structures it appears that fluorination is uniquely suited to stabilizing proteins because, although the replacement of hydrogen atoms by fluorine atoms increases the volume of the amino acid side chain, it very closely preserves the side chain's shape. This allows the fluorinated residues to integrate into the protein core with minimal perturbation to its structure. This is illustrated by a comparison of the structures of the parent α_4 protein, two variants containing hFLeu at either the *a* or the *d* positions, and a variant containing β-*t*-butylalanine at both *a* and *d* positions, which are shown in Fig. 3. The larger hFLeu residue induces minimal changes to the proteins' overall structure while contributing as much as ~0.8 kcal/mol/residue to the protein's stability. Although β-*t*-butylalanine does stabilize the protein by ~4.3 kcal/mol/residue due to its increased size and hydrophobicity, the change in shape imparted by the additional methyl group does not preserve the van der Waals contacts. This causes the hydrophobic core to reorganize so that a destabilizing void now runs through the center of the protein.

1.3 Studies on the Effects of Fluorination in Other Structural Contexts

Our studies on α_4 indicate that hFLeu is well tolerated and stabilizing within the context of a four-helix bundle. As many proteins adopt α-helical structures this suggests that, in this context, fluorination should be generally stabilizing. However, the context is clearly important, as the following examples from other small protein motifs demonstrate. Thus it was found that introducing hFLeu into a two-stranded parallel coiled-coil resulted in a change in oligomerization state to a four-stranded coiled-coil [25, 26]. In another case, substitution of pFPhe for one of the three Phe residues in the small villin headpiece protein was only stabilizing at one position and disruptive at the other two [34]. Lastly, the introduction of a hFLeu:hFleu cross-strand interaction in a β-hairpin-forming peptide was found to be destabilizing with respect to the equivalent Leu:Leu interaction [40].

Cheng and co-workers have attempted to evaluate the intrinsic α-helix and β-sheet-stabilizing propensities of various fluorinated amino acids by incorporating them at solvent-exposed positions (to minimize tertiary packing effects) in a monomeric α-helix or in a β-strand of a protein GB1 domain [38, 39]. In the β-sheet context, fluorinated analogs of Phe, Leu, and ethylglycine stabilized the protein fold. In contrast, for the α-helical context the effects were reversed: fluorinated analogs of Phe, Leu, and ethylglycine were all destabilizing with respect to their non-fluorinated counterparts. Koksch and co-workers investigated the stabilizing effects of increasing the fluorine content in a series of ethylglycine analogs incorporated into a model coiled-coil protein. They concluded that in this case both steric effects and polarity effects (due to the strong electron-withdrawing effect of fluorine) were important in modulating the stability of the protein [47].

From the above discussion it is evident that whereas fluorinated amino acids can be incorporated into a wide variety of protein structures, increases in stability are not guaranteed. α-Helical structures appear to be more tolerant of fluorinated residues than β-sheet structures, but fluorinated residues are not perfectly isosteric with the corresponding hydrocarbon residues (and would not lend stability if they were!) and so the site(s) for incorporating fluorinating residues needs to be carefully considered before proceeding with synthesis.

1.4 Synthesis of Fluorinated Proteins

Most of the work on fluorinated proteins, including all of our studies, has used solid-phase peptide synthesis (SPPS) to produce peptides of interest, as this allows the greatest control over the introduction of noncanonical amino acids into the sequence. For larger proteins, Tirrell and co-workers have developed methods for in vivo incorporation of fluorinated amino acids such as tFLeu, tFIle, tFVal, and hFLeu that can be activated by endogenous tRNA synthetases [28–30, 32, 33, 46]. One drawback is that protein expression does not result in 100 % incorporation of fluorinated analogs due to the presence of natural amino acid substrate derived from cellular proteins; efficiencies of 70–90 % are typical. In vivo protein incorporation also results in global substitution of a particular amino acid, which limits some applications [31–33, 46]. The use of an orthogonal tRNA synthetase/amber-suppressing tRNA pair, pioneered by Shultz and co-workers [70, 71], provides a further potential route for site-specific incorporation of fluorinated amino acids. However, although more than 100 nonnatural amino acids have been site-specifically incorporated into numerous proteins, to our knowledge, no highly fluorinated amino acid analogs have been incorporated by this method so far. Currently, expressed protein ligation techniques [72, 73] probably offer the best way to incorporate fluorinated amino acids at a specific position within a larger protein (longer than ~50 residues), although, again, we are not aware of a specific example where this technique has been used for the production of extensively fluorinated proteins.

Our laboratory has used both Fmoc- and Boc-protection strategies (Fmoc = fluorenylmethyloxycarbonyl, Boc = *tert-butyloxycarbonyl*) for the synthesis of various fluorinated peptides that incorporate the following fluorinated amino acids (Table 1): tFeG, tFmPhe, tFMet, pFtBSer, and hFLeu. Fmoc-protected peptide synthesis is the most popular SPPS method due to the mild basic Fmoc deprotection using 20 % piperidine, its amenability to automation, and the ease with which peptides can be cleaved from the resin and side chains deprotected at the end of the synthesis using trifluoroacetic acid (TFA). We have found that synthesis of peptides containing limited numbers of tFeG, tFmPhe, tFMet, or pFtBSer is readily accomplished by standard Fmoc SPPS protocols.

Table 1
Highly fluorinated amino acids that have been incorporated into proteins. Commercial suppliers are listed, along with protocols to resolve enantiomers of racemic mixtures. Amino acids listed that are not commercially available include references to synthetic protocols

Amino acid	Source	Racemic	Reference
Hexafluoroleucine (hFLeu)	Synthesized	–	[82–84]
Trifluoroleucine (tFLeu)	Synthesized	–	[88]
Trifluorovaline (tFVal)	Oakwood Chemical	Yes	[88, 89]
Trifluoroisoleucine (tFIle)	Synthesized	–	[33]
Trifluoromethylmethionine (tFMet)	Synthesized	–	[11]
Trifluoroethylglycine (tFeG)	SynQuest, Oakwood Chemical	Yes	[85]
Perfluoro-*t*-butylhomoserine (pFtBSer)	Synthesized	–	[20]
Pentafluorophenylalanine (pFPhe)	SynQuest, Sigma	No	
Trifluoromethylphenylalanine (tFmPhe)	Chem-Impex International	No	[10]

However, attempts to synthesize peptides containing multiple hFLeu residues using Fmoc-hFLeu encountered large decreases in coupling efficiency following incorporation of hFLeu residues. The reason for this remains unclear, but we have found the manual Boc SPPS protocol to give reliable results with all the peptides we have attempted to synthesize. The disadvantage of Boc synthesis is that cleavage from the resin is accomplished with HF [74], which requires specialized equipment to handle. However various peptide synthesis companies offer HF cleavage as a service, which is the route we use to cleave our peptides. The efficient side-chain deprotection afforded by Boc SPPS, in our experience, results in clean peptides that are easy to purify and justifies the additional time and expense of HF cleavage.

1.5 Measuring the Thermodynamic Stability of Fluorinated Proteins by Circular Dichroism

To measure the change in $\Delta G^{\circ}_{\text{unfold}}$ imparted by fluorinated residue we routinely employ guanidinium hydrochloride (GuHCl) titrations; we find that this method is generally superior to thermal denaturation as many coiled-coil proteins have very broad thermal transitions. To determine simple free energies of unfolding ($\Delta G^{\circ}_{\text{unfold}}$) a single titration at fixed protein concentration and temperature can be performed, with unfolding transitions followed by circular dichroism spectroscopy to monitor the change in secondary structure [75] (Fig. 4a). For monomeric and dimeric proteins, analysis of the unfolding transitions is straightforward (assuming it is a two-state process) and standard equations can be used to fit the data [76]. For oligomeric proteins, exact solutions to the unfolding curves can, in principle, be obtained for trimeric and tetrameric proteins but in practice are too cumbersome to fit to experimental data. Approximate methods work better and we provide a detailed description of this approach below and have appended the MATLAB (MathWorks Inc.) routine that we use to fit the data.

1.5.1 Determining ΔG° from GuHCl-Induced Protein Unfolding

For the analysis to be valid, GuHCl-induced unfolding of the proteins must follow a two-state equilibrium between the folded oligomeric protein (F) and the unfolded monomeric peptide (U), as shown in Eq. 1. This is characterized by an equilibrium constant $K([\text{GuHCl}])$ that is dependent on GuHCl concentration:

$$\text{F} \Leftrightarrow n\text{U} \tag{1}$$

Equation 2 relates $K([\text{GuHCl}])$ to [F], [U], and [P], which are the concentrations of folded tetramer, unfolded monomer, and total protein, respectively, so that $[\text{P}] = n[\text{F}] + [\text{U}]$:

$$K\left([\text{GuHCl}]\right) = \frac{[\text{U}]^{n}}{[\text{F}]} = \frac{n[\text{U}]^{n}}{[\text{P}]-[\text{U}]} \tag{2}$$

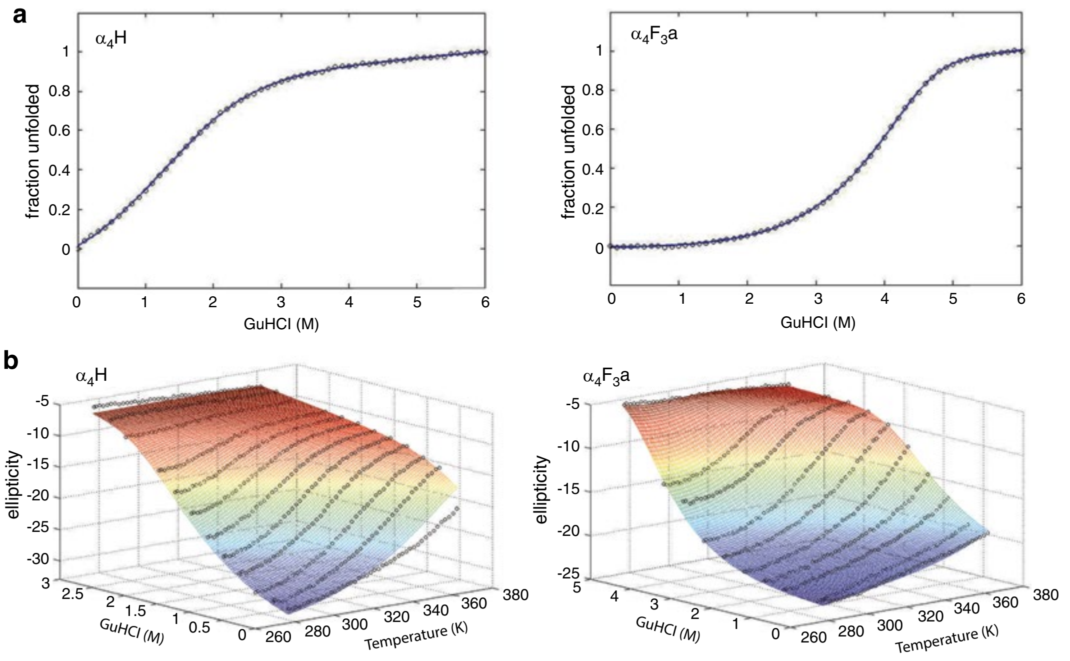


Fig. 4 Unfolding of α_4H and α_4F_3a monitored by changes in ellipticity at 222 nm. (**a**) Denaturation using GuHCl with experimental data represented by *black circles* and data fitting to determine $\Delta G°_u$ and *m* as a *blue line*. (**b**) Denaturation using heat and GuHCl with experimental data represented by *black circles* and data fitting to determine $\Delta H°$, $\Delta S°$, and $\Delta C_p°$ as colored surface

Rearrangement of Eq. 2 results in the polynomial expression (Eq. 3):

$$[\mathrm{U}]^n + \frac{K([\mathrm{GuHCl}])}{n}[\mathrm{U}] - \frac{K([\mathrm{GuHCl}])}{n}[\mathrm{P}] = 0 \qquad (3)$$

For fixed [P] (in this case 40 μM), given any nonnegative value of K([GuHCl]), Eq. 3 has a unique solution for [U] between 0 and [P]. Equation 3 can be solved numerically, which allows K([GuHCl]) to be calculated at each GuHCl concentration. Protein stability as a function of GuHCl concentration is modeled by the following relationship:

$$\Delta G^{\circ}([\mathrm{GuHCl}]) = \Delta G^{\circ}(0\mathrm{M\,GuHCl}) - m^{*}[\mathrm{GuHCl}] \qquad (4)$$

$$K([\mathrm{GuHCl}]) = \exp\left(\frac{-\left(\Delta G^{\circ}(0\mathrm{M\,GuHCl}) - m^{*}[\mathrm{GuHCl}]\right)}{RT}\right) \qquad (5)$$

where $\Delta G°$(0 M GuHCl) is the stability of a protein in the absence of GuHCl and m is the dependence of stability on GuHCl concentration. $K([\text{GuHCl}])$ is then given by Eq. 5 and global fitting of $K([\text{GuHCl}])$ as a function of [GuHCl] allows the values of $\Delta G°$ and m to be calculated.

1.5.2 Treatment of Baselines

Plotting the ellipticity of α_4 proteins as a function of GuHCl concentration results in a sigmoidal curve with the pre- and post-transition baselines corresponding to the ellipticity of folded protein (θ_f) and unfolded protein (θ_u). The ellipticity of the unfolded and folded proteins is assumed to vary linearly with [GuHCl] and is modeled using Eqs. 6 and 7, where the parameters a and d are the baseline intercept at 0 M GuHCl while c and f describe the baseline slope for unfolded and folded protein, respectively:

$$\theta_\text{u}\left(\left[\text{GuHCl}\right]\right) = a + c^{*}\left[\text{GuHCl}\right] \tag{6}$$

$$\theta_\text{f}\left(\left[\text{GuHCl}\right]\right) = d + f^{*}\left[\text{GuHCl}\right] \tag{7}$$

The observed ellipticity is the sum of the contributions from the unfolded and folded fractions of protein and is described by Eq. 8:

$$\theta_\text{Obsd} = \theta_\text{u}\left(\left[\text{GuHCl}\right]\right)\frac{[\text{U}]}{[\text{P}]} + \theta_\text{f}\left(\left[\text{GuHCl}\right]\right)\frac{[\text{P}]-[\text{U}]}{[\text{P}]} \tag{8}$$

Equations 3 and 5–7 are substituted implicitly into Eq. 8 which can be fitted to the data using the program MATLAB; *see* Subheading 2 for MATLAB code to calculate values for a, c, d, f, $\Delta G°$, and m.

1.5.3 Determining $\Delta H°'$, $\Delta S°'$, and $\Delta C_p°$ from Heat and GuHCl-Induced Protein Unfolding

A more detailed thermodynamic analysis that allows the values of $\Delta H°'$, $\Delta S°'$, and $\Delta C_\text{p}°'$ to be determined can be performed by fitting denaturation profiles to the Gibbs-Helmholtz equation. These thermodynamic parameters are helpful to diagnose what types of physical interactions are contributing to protein stability. This requires a two-dimensional approach in which the protein is thermally denatured at different GuHCl concentrations (Fig. 4b). The unfolding surface can then be fitted to the Gibbs-Helmholtz equation by approximate methods as detailed below.

To assess an individual protein, temperature-unfolding data from each GuHCl concentration experiment is combined into a single spreadsheet with columns corresponding to temperature (K), CD output (ellipticity), and GuHCl concentration (M). The thermal unfolding of the peptides is modeled by the unfolding of an n-mer to monomer as in Eq. 1 with the equilibrium constant $K(T,[\text{GuHCl}])$ similar to that in Eqs. 2 and 3, but now being dependent on both temperature and denaturant concentration.

To calculate the values ΔH°, ΔS°, and ΔC_{p}° associated with protein unfolding, $K(T,[\text{GuHCl}])$ is fitted to the Gibbs-Helmholtz equation (Eq. 9), modified by assuming that the Gibbs free energy, ΔG°, varies linearly with GuHCl concentration as described by Eq. 4, to give Eq. 10:

$$\Delta G^{\circ}(T) = \Delta H^{\circ} - T\Delta S^{\circ} + \Delta C_{p}^{\circ *}\left(T - T_{0} + T\ln\frac{T_{0}}{T}\right) \tag{9}$$

$$\Delta G^{\circ}(T,[\text{GuHCl}]) = \Delta H^{\circ} - T\Delta S^{\circ} + \Delta C_{p}^{\circ *}\left(T - T_{0} + T\ln\frac{T_{0}}{T}\right) - m^{*}[\text{GuHCl}] \tag{10}$$

In these equations T is temperature, T_0 is the reference temperature of 25 °C, ΔH° is the change in enthalpy, ΔS° is the change in entropy, and ΔC_{p}° is the change in heat capacity, each at the reference temperature T_0. It has been observed that ΔC_{p}° and m change little over the measured range of denaturant concentration and temperature and are assumed to be constant [77,78]. $K(T,[\text{GuHCl}])$ is then given by Eq. 11 and global fitting of $K(T,[\text{GuHCl}])$ as a function of T and [GuHCl] allows the values of ΔH°, ΔS°, ΔC_{p}°, ΔG°, and m to be calculated [77]:

$$K(T,[\text{GuHCl}]) = \exp\left(\frac{-\left(\Delta H^{\circ} - T\Delta S^{\circ} + \Delta C_{p}^{\circ *}\left(T - T_{0} + T\ln\frac{T_{0}}{T}\right) - m^{*}[\text{GuHCl}]\right)}{RT}\right) \tag{11}$$

1.5.4 Treatment of Base Planes

Plotting the ellipticity of α_4 proteins as a function of GuHCl concentration and temperature results in a two-dimensional surface with the pre- and post-transition base planes corresponding to the ellipticity of folded protein (θ_f) and unfolded protein (θ_u). The ellipticity of the unfolded and folded proteins is assumed to vary linearly with T and [GuHCl] and is modeled using Eqs. 12 and 13, where the parameters a, b, c, d, e, and f describe the ellipticity of the folded and unfolded states at various temperatures and GuHCl concentrations:

$$\theta_{u}(T,[\text{GuHCl}]) = a + b^{*}T + c^{*}[\text{GuHCl}] \tag{12}$$

$$\theta_{f}(T,[\text{GuHCl}]) = d + e^{*}T + f^{*}[\text{GuHCl}] \tag{13}$$

The observed ellipticity is the sum of the contributions from the unfolded and folded fractions of protein and is described by Eq. 14:

$$\theta_{\text{Obsd}} = \theta_{u}(T,[\text{GuHCl}])\frac{[\text{U}]}{[\text{P}]} + \theta_{f}(T,[\text{GuHCl}])\frac{[\text{P}]-[\text{U}]}{[\text{P}]} \tag{14}$$

Equations 3 and 11–13 are substituted implicitly into Eq. 14 which can be fitted to the data using the program MATLAB; *see* Subheading 2 for MATLAB code to calculate values for *a*, *b*, *c*, *d*, *e*, *f*, $\Delta H°$, $\Delta S°$, $\Delta C_p°$, and *m*. Data sets with a large number of data points, in our experiments between 430 and 512, allow a more robust fit. If reliable values for *e* and *f* cannot be obtained for marginally stable proteins because there are insufficient data points to define the folded base plane, *e* and *f* are set to zero. We have found that this approximation does not usually introduce significant errors into the analysis.

1.6 Structure Determination via X-Ray Crystallography

Although the determination of the three-dimensional structure of a protein is a significant undertaking, it represents the "gold standard" for evaluating how the introduction of noncanonical residues alters the molecular structure of a protein. With the rapid development of automated methods in X-ray crystallography it is increasingly feasible to obtain X-ray structures for proteins. Fluorinated proteins with their increased thermodynamic stability are generally good candidates for crystallography. And if the X-ray structure of the parent protein has been solved, this greatly aids in both crystallization trials and solving the structure by molecular replacement. The largest barrier to obtaining a structure is increasingly often obtaining crystals that diffract sufficiently well for X-ray crystallography. Therefore, we describe the conditions used in the crystallization of the α_4 proteins with the hope that this will provide some general guidance. Examples of protein crystals obtain in our studies are shown in Fig. 5.

The formation of three-dimensional protein crystals relies on *inter*-molecular interactions, usually between external polar residues of adjacent proteins. For α_4, the relatively flat sides of α-helical bundle proteins and the number of solvent-exposed residues increase the number of potential intermolecular interactions. The thermodynamic stability and structural homogeneity of the α_4 proteins, as indicated by circular dichroism and analytical ultracentrifugation, also made them strong candidates for protein crystallization.

1.7 Concluding Remarks

Highly fluorinated analogues of hydrophobic amino acids have shown great potential in stabilizing folded proteins. We hope that our laboratory's exploration of the principles by which fluorination stabilizes proteins will prove useful for others who may wish to utilize fluorinated proteins for various applications. Although beyond the scope of this review, the excellent NMR properties of fluorine open up many avenues for the study of protein dynamics and their interaction with other biological macromolecules, which we and others have begun to explore [1, 8, 10, 11, 20, 79–81]. Although the examples presented here are primarily focused on α-helical proteins, the stabilizing effects of highly fluorinated amino acids have been demonstrated in β-sheet structures and complex structural motifs found in natural proteins.

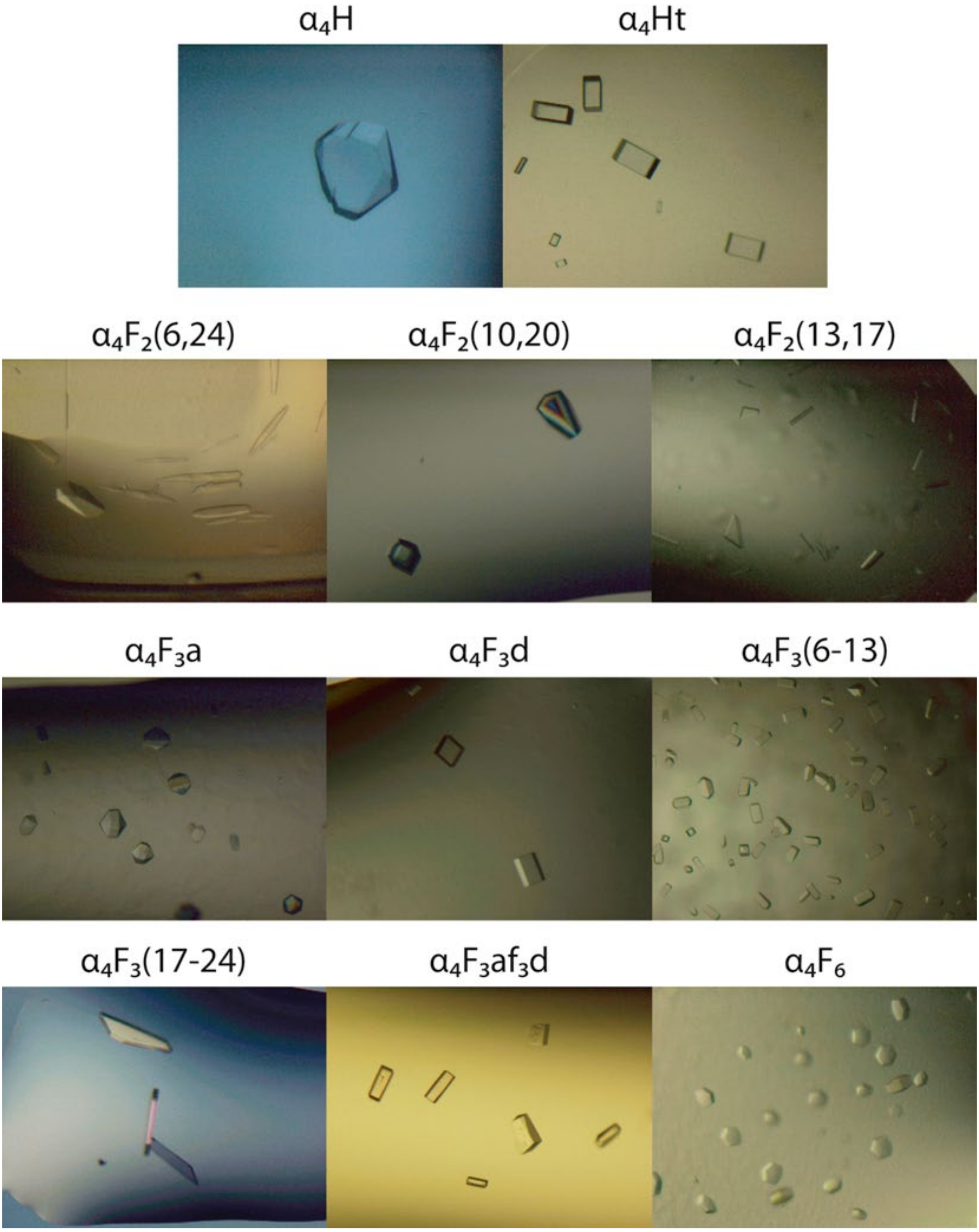


Fig. 5 Representative crystals obtained for 11 variants of α_4

2 Materials

2.1 Fluorinated Amino Acids

Many fluorinated amino acid analogs are commercially available, although at this time hFLeu, which because of its high fluorine content has been extensively used in the design of fluorinated peptides, is not. However, several syntheses of this amino acid, including one from our laboratory, have been reported; a detailed description of its preparation lies outside the scope of this chapter and we refer the reader to the original literature for the synthetic procedures [82–84]. Amino acids most commonly used in the design of fluorinated proteins and peptides are listed in Table 1 together with suppliers if they are commercially available. In some cases, fluorinated amino acids are only available as racemic mixtures, which are obviously not useful for peptide synthesis. In these cases we have found the following protocol for resolving the pure enantiomers useful [85]. We note that the highly fluorinated analogues trifluoroalanine and hexafluorovaline are not included on this list since they are not amenable to incorporation into peptides because they are difficult to couple and easily racemize during peptide synthesis due to the acidity of the α-carbon [86, 87].

2.1.1 Enzymatic Resolution of Racemic tFeG

Some fluorinated amino acids are available commercially only as racemic mixtures and require purification to obtain the needed L-form enantiomer. The enzymatic resolution of tFeG (SynQuest) using porcine kidney acylase I (Sigma) has been successfully performed in our lab.

2.2 Synthesis of Fluorinated Proteins

A typical Boc-protected synthesis is performed on a 0.125 mmol scale with 4-methylbenzhydrylamine (MBHA) resin (AnaSpec Inc.) (this provides an amidated C-terminus; other resins can be chosen if an unmodified C-terminus is desired). Syntheses are performed in a glass sintered reaction vial with vortexing to mix reagents during coupling reactions. Amino acids and coupling reagent are from AnaSpec Inc. and solvents from Fisher. Thorough rinsing with solvent between coupling and deprotection steps is necessary to prevent side reactions.

2.3 Measuring the Thermodynamic Stability of Fluorinated Proteins by Circular Dichroism

2.3.1 Determining $\Delta G°$ from GuHCl-Induced Protein Unfolding

Changes in the ellipticity of a protein resulting from unfolding by GuHCl denaturation are recorded using a spectropolarimeter equipped with an autotitrator; our experiments use an Aviv 62DS spectropolarimeter with a Microlab Series 500 syringe pump. A 1 cm pathlength quartz cuvette (Starna Cells Inc.) is equipped with a stir bar. The automated titration provides a smooth sigmoidal denaturation curve that is then analyzed using MATLAB.

2.3.2 Determining $\Delta H°'$, $\Delta S°'$, and $\Delta C_p°$ from Heat and GuHCl-Induced Protein Unfolding

Heat denaturation experiments are carried out using an Aviv 62DS spectropolarimeter with a temperature controller. A 1 mm pathlength quartz cuvette (Starna Cells Inc.) works best; this pathlength provides a strong protein signal while allowing adequate temperature control.

2.4 Structure Determination via X-Ray Crystallography

Crystallization screens of numerous α_4 analogues yielded single crystals with various precipitants, salts, and buffers. A precipitant condition common to all of the proteins was the presence of high concentrations of small polyethyleneglycol (PEG) molecules between 200 and 600 Da. Armed with this knowledge, a 96-well matrix screen (named BCB-SP) was developed to sample a broad pH range (4.5–9.0) together with varying concentrations of cryoprotecting precipitants including PEG-200, PEG-300, PEG-400, PEG-550MME, and PEG-600, as summarized in Table 2. This screening format can be efficiently set up with either multichannel pipettors or robotic dispensing. The Intelli-Plate 96-3 Crystalliza-tion plate is recommended in conjunction with Art Robbins Instruments robotic dispensing to allow three accurately dispensed protein samples per crystallization condition.

3 Methods

3.1 Synthesis of Fluorinated Proteins

3.1.1 General Protocol for Enzymatic Resolution of Racemic tFeG: Adapted from Ref. 67

1. Dissolve racemic tFeG in 1 M NaOH in an ice bath and then acetylate by slowly adding 2 equiv. acetic anhydride while maintaining pH = ~10. Reaction should be complete after 2 h.
2. Add 6 M HCl to pH = ~2 and extract with eight aliquots of ethyl acetate (or until TLC confirms that no further *N-acetyl-d*,L-tFeG is extracted into organic layer). Remove solvent by rotary evaporation.
3. L-tFeG is enzymatically resolved from *N*-acetyl-D,L-tFeG by porcine kidney acylase I. Dissolve solid *N*-acyl-D,L-tFeG in 1 M LiOH and adjust buffer pH to 7.2 using acetic acid. Add 20 mg porcine kidney acylase I for each gram of amino acid and stir reaction at 37 °C for ~4 h until deacetylation of the L-form is judged complete. This is easily assessed by ^{19}F NMR.
4. The resulting solution is filtered and L-tFeG separated from *N*-acetyl-D-tFeG by chromatography on Dowex 50X8-200 ion-exchange resin (*see* **Note 1**). Condition ~30 mL resin by washing three times with 100 mL 1 M NH_4OH, once with 2 L water, once with 100 mL 1 M HCl, and finally once with 2 L water or until pH = ~7. Apply filtered solution to the column and then reapply flow-through to top of the column. Wash column with 2 L water. Elute the amino acid with 400 mL of 1 M NH_4OH.
5. L-tFeG fractions are identified by staining with ninhydrin and enantiomeric purity confirmed by reaction with Mosher's acid [90] and subsequent ^{19}F NMR.

3.1.2 General In Situ Neutralization Boc-Protected SPPS Procedure

1. Swell resin in *N*,*N*-dimethylformamide (DMF) for 20 min.
2. Deprotect by adding neat TFA to resin and swirl for 1 min. Drain and repeat TFA deprotection for 1 min.

Table 2
BCB-SP screen for crystallizing α_4 proteins

A1	0.1 M	Sodium acetate pH 4.5	50 % PEG 200	E1	0.1 M	Tris pH 8.0	55 % PEG 400
A2	0.1 M	MES pH 5.5	50 % PEG 200	E2	0.1 M	CHES pH 9.0	55 % PEG 400
A3	0.1 M	HEPES pH 7.0	50 % PEG 200	E3	0.1 M	Sodium acetate pH 4.5	60 % PEG 400
A4	0.1 M	Tris pH 8.0	50 % PEG 200	E4	0.1 M	MES pH 5.5	60 % PEG 400
A5	0.1 M	CHES pH 9.0	50 % PEG 200	E5	0.1 M	HEPES pH 7.0	60 % PEG 400
A6	0.1 M	Sodium acetate pH 4.5	55 % PEG 200	E6	0.1 M	Tris pH 8.0	60 % PEG 400
A7	0.1 M	MES pH 5.5	55 % PEG 200	E7	0.1 M	CHES pH 9.0	60 % PEG 400
A8	0.1 M	HEPES pH 7.0	55 % PEG 200	E8	0.1 M	Sodium acetate pH 4.5	40 % PEG 550 MME
A9	0.1 M	Tris pH 8.0	55 % PEG 200	E9	0.1 M	MES pH 5.5	40 % PEG 550 MME
A10	0.1 M	CHES pH 9.0	55 % PEG 200	E10	0.1 M	HEPES pH 7.0	40 % PEG 550 MME
A11	0.1 M	Sodium acetate pH 4.5	60 % PEG 200	E11	0.1 M	Tris pH 8.0	40 % PEG 550 MME
A12	0.1 M	MES pH 5.5	60 % PEG 200	E12	0.1 M	CHES pH 9.0	40 % PEG 550 MME
B1	0.1 M	HEPES pH 7.0	60 % PEG 200	F1	0.1 M	Sodium acetate pH 4.5	45 % PEG 550 MME
B2	0.1 M	Tris pH 8.0	60 % PEG 200	F2	0.1 M	MES pH 5.5	45 % PEG 550 MME
B3	0.1 M	CHES pH 9.0	60 % PEG 200	F3	0.1 M	HEPES pH 7.0	45 % PEG 550 MME
B4	0.1 M	Sodium acetate pH 4.5	45 % PEG 300	F4	0.1 M	Tris pH 8.0	45 % PEG 550 MME
B5	0.1 M	MES pH 5.5	45 % PEG 300	F5	0.1 M	CHES pH 9.0	45 % PEG 550 MME
B6	0.1 M	HEPES pH 7.0	45 % PEG 300	F6	0.1 M	Sodium acetate pH 4.5	50 % PEG 550 MME
B7	0.1 M	Tris pH 8.0	45 % PEG 300	F7	0.1 M	MES pH 5.5	50 % PEG 550 MME
B8	0.1 M	CHES pH 9.0	45 % PEG 300	F8	0.1 M	HEPES pH 7.0	50 % PEG 550 MME

(continued)

Table 2
(continued)

B9	0.1 M	Sodium acetate pH 4.5	50 % PEG 300	F9	0.1 M	Tris pH 8.0	50 % PEG 550 MME
B10	0.1 M	MES pH 5.5	50 % PEG 300	F10	0.1 M	CHES pH 9.0	50 % PEG 550 MME
B11	0.1 M	HEPES pH 7.0	50 % PEG 300	F11	0.1 M	Sodium acetate pH 4.5	55 % PEG 550 MME
B12	0.1 M	Tris pH 8.0	50 % PEG 300	F12	0.1 M	MES pH 5.5	55 % PEG 550 MME
C1	0.1 M	CHES pH 9.0	50 % PEG 300	G1	0.1 M	HEPES pH 7.0	55 % PEG 550 MME
C2	0.1 M	Sodium acetate pH 4.5	55 % PEG 300	G2	0.1 M	Tris pH 8.0	55 % PEG 550 MME
C3	0.1 M	MES pH 5.5	55 % PEG 300	G3	0.1 M	CHES pH 9.0	55 % PEG 550 MME
C4	0.1 M	HEPES pH 7.0	55 % PEG 300	G4	0.1 M	Sodium acetate pH 4.5	40 % PEG 600
C5	0.1 M	Tris pH 8.0	55 % PEG 300	G5	0.1 M	MES pH 5.5	40 % PEG 600
C6	0.1 M	CHES pH 9.0	55 % PEG 300	G6	0.1 M	HEPES pH 7.0	40 % PEG 600
C7	0.1 M	Sodium acetate pH 4.5	60 % PEG 300	G7	0.1 M	Tris pH 8.0	40 % PEG 600
C8	0.1 M	MES pH 5.5	60 % PEG 300	G8	0.1 M	CHES pH 9.0	40 % PEG 600
C9	0.1 M	HEPES pH 7.0	60 % PEG 300	G9	0.1 M	Sodium acetate pH 4.5	45 % PEG 600
C10	0.1 M	Tris pH 8.0	60 % PEG 300	G10	0.1 M	MES pH 5.5	45 % PEG 600
C11	0.1 M	CHES pH 9.0	60 % PEG 300	G11	0.1 M	HEPES pH 7.0	45 % PEG 600
C12	0.1 M	Sodium acetate pH 4.5	45 % PEG 400	G12	0.1 M	Tris pH 8.0	45 % PEG 600
D1	0.1 M	MES pH 5.5	45 % PEG 400	H1	0.1 M	CHES pH 9.0	45 % PEG 600
D2	0.1 M	HEPES pH 7.0	45 % PEG 400	H2	0.1 M	Sodium acetate pH 4.5	50 % PEG 600
D3	0.1 M	Tris pH 8.0	45 % PEG 400	H3	0.1 M	MES pH 5.5	50 % PEG 600
D4	0.1 M	CHES pH 9.0	45 % PEG 400	H4	0.1 M	HEPES pH 7.0	50 % PEG 600

D5	0.1 M	Sodium acetate pH 4.5	50 % PEG 400	H5	0.1 M	Tris pH 8.0	50 % PEG 600
D6	0.1 M	MES pH 5.5	50 % PEG 400	H6	0.1 M	CHES pH 9.0	50 % PEG 600
D7	0.1 M	HEPES pH 7.0	50 % PEG 400	H7	0.1 M	Sodium acetate pH 4.5	55 % PEG 600
D8	0.1 M	Tris pH 8.0	50 % PEG 400	H8	0.1 M	MES pH 5.5	55 % PEG 600
D9	0.1 M	CHES pH 9.0	50 % PEG 400	H9	0.1 M	HEPES pH 7.0	55 % PEG 600
D10	0.1 M	Sodium acetate pH 4.5	55 % PEG 400	H10	0.1 M	Tris pH 8.0	55 % PEG 600
D11	0.1 M	MES pH 5.5	55 % PEG 400	H11	0.1 M	CHES pH 9.0	55 % PEG 600
D12	0.1 M	HEPES pH 7.0	55 % PEG 400	H12	0.1 M	HEPES pH 7.0	45 % PEG 200

3. Dissolve Boc-protected amino acid (4 equiv.) (*see* **Note 2**) and *O*-(6-chlorobenzotriazl-1-yl)-*N*,*N*,*N*′,*N*′-tetramethyl uronium hexafluorophosphate (HCTU, 4 equiv.) in DMF to a concentration of ~0.3 M Boc-amino acid, swirl to dissolve, then add *N*,*N*-diisopropylethylamine (DIEA, 8 equiv.), and swirl manually for ~1 min (*see* **Note 3**). Add activated amino acid solution to resin and vortex for 30 min. Drain solution and wash with DMF (*see* **Note 4**).
4. Perform Kaiser test to determine coupling completion. Repeat **steps 2–3** for each residue.
5. *Kaiser test*: Remove small amount of resin (10–20 particles) and place in 1.5 mL microcentrifuge tube. Wash with DMF (2×) followed by diethyl ether (2×). Add 100 μL Kaiser reagent (Sigma) to the resin. Heat at 110 °C for 2–3 min. Coupling is complete if solution is yellow. Coupling is incomplete if solution is dark green to purple or 1–2 beads are dark purple. Repeat **step 3** if coupling is incomplete.

3.1.3 Cleavage from Resin

1. After coupling of final residue, rinse resin with DMF followed by dichloromethane and then diethyl ether.
2. Purge resin in sintered vial with N_2 to remove residual diethyl ether, place resin in 15 mL conical vial, and dry on high vacuum for 2 h.
3. Cleavage with HF requires specialized equipment due to its toxic and corrosive nature. We have used commercial services, e.g., CS BioCompany Inc. (Menlo Park, CA), to perform the resin cleavage step. The resulting crude peptide powder is subsequently purified by preparative HPLC and identity verified by mass spectrometry (*see* **Note 5**).

3.2 Circular Dichroism Measurements

3.2.1 GuHCl Unfolding

1. Stock protein solutions are prepared as follows: 2.0 mL of peptide (40 μM concentration of monomer) in 10 mM potassium phosphate buffer, pH 7.0 and 10.0 mL of 40 μM peptide in 10 mM potassium phosphate buffer, pH 7.0 with 8.0 M GuHCl (*see* **Note 6**).
2. Equip 1 cm pathlength quartz cuvette with stir bar and add 2.0 mL buffered protein solution.
3. Purge titrator syringes by placing inlet and outlet tubes in water and allow syringes to fill and dispense five times. Repeat syringe purging with protein-GuHCl solution.
4. Leave inlet tube in protein-GuHCl solution and move outlet tube to a waste container. Place titrator cap on the cuvette and wrap with parafilm.

5. Measure ellipticity at 222 nm after stirring for 30 s following each GuHCl addition in 0.1 M increments (*see* **Note 7**).
6. Analyze using MATLAB with code in Subheading 3.2.3.

3.2.2 Heat and GuHCl Unfolding

1. Prepare protein sample of 40 μM peptide (concentration of monomer) in 10 mM potassium phosphate buffer, pH 7.0 in 9–12 different concentrations of GuHCl (*see* **Note 6**).
2. The requisite concentrations of GuHCl needed to observe folded and unfolded baselines are best determined from GuHCl titrations described in Subheading 3.2.1.
3. Place protein sample in a 1 mm pathlength quartz cuvette.
4. For each GuHCl concentration, increase temperature from 4 to 90 °C in increments of 2 °C to obtain a smooth unfolding curve; record ellipticity measurements with a 10-s averaging time to reduce signal noise (*see* **Note 8**).
5. Analyze using MATLAB with code in Subheading 3.2.4 (*see* **Note 9**).

3.2.3 Matlab Code for Determining ΔG° Using GuHCl Unfolding

```
%Change the parameters below
n = 4; %This represents monomer to n-mer folding
P = 40e-6; %Free monomer concentration in Molarity (SI units only!)
T0 = 298.15; %Reference Temperature in K (298.15 is room temperature)

%Make sure the data is saved in two columns where each row is a math of
%(Concentration denaturant, Ellipticity)

%Concentration must be in Molarity
%For reference, %1 kcal = 4184 J exactly

%Do not change anything below this point for standard use
R = 8.3145; %Ideal Gas Constant

data = uiimport; %Import data; it will be saved as 'data'
data = data.data; %MatLab Glitch; this is the workaround
Den = data(:,1);
Theta_Obsd = data(:,2);

K = @(b,Den) exp(-(b(1).*ones(length(Den),1)-b(2).*Den)./(R*T0));
U = @(b,Den) arrayfun(@(k) fzero(@(x) n*x^n+k*x-k*P, [0 P]), K(b,Den));
f = @(b,Den) (b(3).*ones(length(Den),1)+b(5).*Den).*U(b,Den)./P +
(b(4).*ones(length(Den),1)+b(6).*Den).*(P.*ones(length(Den),1)-
U(b,Den))./P;
```

```
%Initial Values Module
beta0 = zeros(6,1);
beta0(3) = max(Theta_Obsd)+1;
beta0(4) = min(Theta_Obsd)-1;
Utest = P.*(Theta_Obsd-
(beta0(4).*ones(length(Theta_Obsd),1)))./(beta0(3)-beta0(4));
Ktest = n.*(Utest).^n./(P.*ones(length(Utest),1)-Utest);
DGtest = -R.*T0.*log(Ktest);

TestMat = ones(length(DGtest),2);
for i = 1:length(DGtest)
    TestMat(i,1) = 1;
    TestMat(i,2) = -Den(i);
end

ParaEst = linsolve(TestMat,DGtest);
beta0(1) = ParaEst(1);
beta0(2) = ParaEst(2);

[beta, r, J, COVB, mse] = nlinfit(Den, Theta_Obsd, f, beta0);

ci = nlparci(beta,r,'covar', COVB);
```

3.2.4 Determining $\Delta H°$, $\Delta S°$, and $\Delta C_p°$ Using Heat and GuHCl Unfolding

```
%Change the parameters below
n = 4; %This represents monomer to n-mer folding
P = 40e-6; %Free monomer concentration in Molarity (SI units only!)
T0 = 298.15; %Reference Temperature in K (298.15 is room temperature)
%Make sure data is saved in three columns:
    %Temp (Celcius), Theta,(Molar)
%For reference, %1 kcal = 4184 J exactly

%Do not change anything below this point for standard use
R = 8.3145; %Ideal Gas Constant in SI units

data = uiimport; %Import data; it will be saved as 'data'
data = data.data; %MatLab Glitch; this is the workaround
Temp = data(:,1) + 273.15; %Converts to Kelvin
Theta_Obsd = data(:,2);
Den = data(:,3);
Cond = [Temp Den];
%Cond is the n x 2 matrix of Temp and Den, and Theta_Obsd is output

%Define additional functions K, U, and f
K =  @(b,Cond)exp(-(b(3)-Cond(:,1)*b(4)+b(5)*(Cond(:,1)-T0 ...
    +Cond(:,1).*log(T0./Cond(:,1)))-Cond(:,2)*b(6))./(R*Cond(:,1)));
U = @(b,Cond) arrayfun(@(k) fzero(@(x) n*x^n+k*x-k*P, [0 P]),
K(b,Cond));
f = @(b,Cond)
1/P*((b(1)*ones(length(Cond(:,1)),1)+b(7).*Cond(:,1)+b(8).*Cond(:,2)).*
U(b,Cond)...

+(b(2)*ones(length(Cond(:,1)),1)+b(9).*Cond(:,1)+b(10).*(Cond(:,2))).*(
P*ones(length(Cond(:,1)),1)-U(b,Cond)));
```

```
%Initial Values Module
beta0 = zeros(10,1);
beta0(1) = max(Theta_Obsd)+1;
beta0(2) = min(Theta_Obsd)-1;
Uest = P*(Theta_Obsd-beta0(2)*ones(length(Theta_Obsd),1))/(beta0(1)-
beta0(2));
Kest = n*Uest.^n./(P*ones(length(Temp),1)-Uest);
DGest = -R*Temp.*log(Kest);

TestMat = ones(length(Temp),4);
for i = 1:length(Temp)
    TestMat(i,1) = 1;
    TestMat(i,2) = -Temp(i);
    TestMat(i,3) = Temp(i)-T0+Temp(i)*log(T0/Temp(i));
    TestMat(i,4) = -Den(i);
end

ParaEst = linsolve(TestMat,DGest);

beta0 = [max(Theta_Obsd);min(Theta_Obsd); ParaEst(1);
ParaEst(2);ParaEst(3);ParaEst(4);0;0;0;0];
%Actual Curve Fitting Portion
[beta, r, J, COVB, mse] = nlinfit(Cond, Theta_Obsd, f, beta0);
ci = nlparci(beta,r,'covar', COVB);
```

3.3 Growing α_4 Protein Crystals

1. Prepare BCB-SP screen (Table 2) in a 96-deepwell block with 2 mL of each condition (*see* **Note 10**).
2. Prepare 75 μL of protein at 1, 3, and 6 mM (concentration of monomer) (*see* **Note 11**).
3. Dispense 70 μL of BCB-SP screen into each of the 96 reservoir wells.
4. Dispense 0.5 μL reservoir solution into each sample well.
5. Dispense 0.5 μL protein solution into each sample well.
6. Seal 96-well plate and store at 4 °C.
7. Check for crystal growth using a light microscope after ~96 h (*see* **Note 12**).

4 Notes

1. Leftover *N*-acyl-D-tFeG can be collected as flow-through from the ion-exchange column.
2. Recommended to use amino acids with acid-stable side-chain-protecting groups for longer proteins, such as Boc-Lys (2-Cl-Z)-OH.
3. Use slightly less coupling reagent (HCTU, HBTU, etc.) than amino acid to prevent formation of a chain terminating tetramethylguanidinium adduct.

4. Rinse resin thoroughly with DCM before and after TFA deprotection following the coupling of Asn or Gln to prevent formation of aspartamide or pyrrolidone carboxylic acid failure sequences.
5. Dissolve crude peptide in HPLC solvent A (95 % H_2O, 4.9 % acetonitrile, and 0.1 % TFA) and pass through 0.22 μm filter prior to purification.
6. To reduce noise, degass all buffers and remove air bubbles by tapping cuvette prior to measurements.
7. To save on protein sample, the titration endpoint GuHCl concentration can be adjusted depending on protein stability.
8. Temperature equilibration time may need to be adjusted depending upon cuvettte holder type and cuvette thickness. Typically 1–2 min is sufficient.
9. Combine data from heat denaturation experiments of a single protein into a three-column spreadsheet containing temperature, ellipticity, and GuHCl molarity.
10. Phosphate buffers should be avoided, as they tend to form inorganic crystals.
11. Stock peptide solutions for crystallography should be prepared in water or a minimal concentration of buffer to allow pH control from the screening conditions.
12. Crystal growth time will vary. Typical crystal growth times we have observed are from 24 h to 1 month.

Acknowledgements

We acknowledge the contributions of the following scientists who have worked on various aspects of the design of fluorinated proteins and peptides in our laboratory: James Anderson, Morris Slutsky, Kyung-Hoon Lee, Hwang-Yeol Lee, Lindsey Gottler, Roberto de la Salud-Bea, Yuta Suzuki, and Benjamin Levin. These projects have been funded, in part, by the following organizations: National Science Foundation (CHE 0640934), the Army Research Office (W911NF-11-1-0251), and the Defense Threat Reduction Agency (HDTRA1-11-1-0019).

References

1. Suzuki Y, Buer BC, Al-Hashimi HM, Marsh ENG (2011) Using fluorine nuclear magnetic resonance to probe changes in the structure and dynamics of membrane-active peptides interacting with lipid bilayers. Biochemistry 50:5979–5987
2. Dalvit C, Vulpetti A (2011) Fluorine–protein interactions and 19F NMR isotropic chemical shifts: an empirical correlation with implications for drug design. ChemMedChem 6:104–114
3. Danielson MA, Falke JJ (1996) Use of F-19 NMR to probe protein structure and confor-

mational changes. Annu Rev Biophys Biomol Struct 25:163–195

4. Gerig JT (1994) Fluorine NMR of proteins. Prog Nucl Magn Reson Spectrosc 26:293–370
5. Luchette PA, Prosser RS, Sanders CR (2002) Oxygen as a paramagnetic probe of membrane protein structure by cysteine mutagenesis and 19F NMR spectroscopy. J Am Chem Soc 124:1778–1781
6. Evanics F, Kitevski JL, Bezsonova I, Forman-Kay J, Prosser RS (2007) 19F NMR studies of solvent exposure and peptide binding to an SH3 domain. Biochim Biophys Acta 1770: 221–230
7. Yu J-X, Kodibagkar VD, Cui W, Mason RP (2005) 19F: a versatile reporter for non-invasive physiology and pharmacology using magnetic resonance. Curr Med Chem 12: 819–848
8. Buer BC, Chugh J, Al-Hashimi HM, Marsh ENG (2010) Using fluorine nuclear magnetic resonance to probe the interaction of membrane-active peptides with the lipid bilayer. Biochemistry 49:5760–5765
9. Khan F, Kuprov I, Craggs TD, Hore PJ, Jackson SE (2006) 19F NMR studies of the native and denatured states of green fluorescent protein. J Am Chem Soc 128:10729–10737
10. Suzuki Y, Brender JR, Hartman K, Ramamoorthy A, Marsh ENG (2012) Alternative pathways of human islet amyloid polypeptide aggregation distinguished by 19F nuclear magnetic resonance-detected kinetics of monomer consumption. Biochemistry 51:8154–8162
11. Suzuki Y, Brender JR, Soper MT, Krishnamoorthy J, Zhou Y, Ruotolo BT, Kotov NA, Ramamoorthy A, Marsh ENG (2013) Resolution of oligomeric species during the aggregation of Aβ1-40 using 19F NMR. Biochemistry 52(11):1903–1912
12. Müller K, Faeh C, Diederich F (2007) Fluorine in pharmaceuticals: looking beyond intuition. Science 317:1881–1886
13. Buer BC, de la Salud-Bea R, Al Hashimi HM, Marsh ENG (2009) Engineering protein stability and specificity using fluorous amino acids: the importance of packing effects. Biochemistry 48:10810–10817
14. Buer BC, Levin BJ, Marsh ENG (2012) Influence of fluorination on the thermodynamics of protein folding. J Am Chem Soc 134:13027–13034
15. Buer BC, Meagher JL, Stuckey JA, Marsh ENG (2012) Structural basis for the enhanced stability of highly fluorinated proteins. Proc Natl Acad Sci U S A 109:4810–4815
16. Buer BC, Meagher JL, Stuckey JA, Marsh ENG (2012) Comparison of the structures and stabilities of coiled-coil proteins containing hexafluoroleucine and t-butylalanine provides insight into the stabilizing effects of highly fluorinated amino acid side-chains. Protein Sci 21:1705–1715
17. Gottler LM, de la Salud-Bea R, Marsh ENG (2008) The fluorous effect in proteins: properties of α4F6, a 4-α-helix bundle protein with a fluorocarbon core. Biochemistry 47:4484–4490
18. Lee H-Y, Lee K-H, Al-Hashimi HM, Marsh ENG (2006) Modulating protein structure with fluorous amino acids: increased stability and native-like structure conferred on a 4-helix bundle protein by hexafluoroleucine. J Am Chem Soc 128:337–343
19. Lee K-H, Lee H-Y, Slutsky MM, Anderson JT, Marsh ENG (2004) Fluorous effect in proteins: de novo design and characterization of a four-α-helix bundle protein containing hexafluoroleucine. Biochemistry 43:16277–16284
20. Buer BC, Levin BJ, Marsh ENG (2013) Perfluoro-tert-butyl homoserine as a sensitive 19F NMR reporter for peptide–membrane interactions in solution. J Pept Sci 19: 308–314
21. Gottler LM, de la Salud Bea R, Shelburne CE, Ramamoorthy A, Marsh ENG (2008) Using fluorous amino acids to probe the effects of changing hydrophobicity on the physical and biological properties of the β-hairpin antimicrobial peptide protegrin-1. Biochemistry 47:9243–9250
22. Gottler LM, Lee H-Y, Shelburne CE, Ramamoorthy A, Marsh ENG (2008) Using fluorous amino acids to modulate the biological activity of an antimicrobial peptide. Chembiochem 9:370–373
23. Bilgiçer B, Fichera A, Kumar K (2001) A coiled coil with a fluorous core. J Am Chem Soc 123:4393–4399
24. Bilgiçer B, Kumar K (2002) Synthesis and thermodynamic characterization of self-sorting coiled coils. Tetrahedron 58:4105–4112
25. Bilgiçer B, Kumar K (2004) De novo design of defined helical bundles in membrane environments. Proc Natl Acad Sci U S A 101: 15324–15329
26. Bilgiçer B, Xing X, Kumar K (2001) Programmed self-sorting of coiled coils with leucine and hexafluoroleucine cores. J Am Chem Soc 123:11815–11816
27. Campos-Olivas R, Aziz R, Helms GL, Evans JNS, Gronenborn AM (2002) Placement of 19F into the center of GB1: effects on structure and stability. FEBS Lett 517:55–60
28. Montclare JK, Son S, Clark GA, Kumar K, Tirrell DA (2009) Biosynthesis and stability of coiled-coil peptides containing (2S,4R)-5,5,5-

trifluoroleucine and (2S,4S)-5,5,5-trifluoroleucine. Chembiochem 10:84–86

29. Son S, Tanrikulu IC, Tirrell DA (2006) Stabilization of bzip peptides through incorporation of fluorinated aliphatic residues. Chembiochem 7:1251–1257

30. Tang Y, Ghirlanda G, Petka WA, Nakajima T, DeGrado WF, Tirrell DA (2001) Fluorinated coiled-coil proteins prepared in vivo display enhanced thermal and chemical stability. Angew Chem Int Ed 40:1494–1496

31. Tang Y, Ghirlanda G, Vaidehi N, Kua J, Mainz DT, Goddard WA, DeGrado WF, Tirrell DA (2001) Stabilization of coiled-coil peptide domains by introduction of trifluoroleucine. Biochemistry 40:2790–2796

32. Tang Y, Tirrell DA (2001) Biosynthesis of a highly stable coiled-coil protein containing hexafluoroleucine in an engineered bacterial host. J Am Chem Soc 123:11089–11090

33. Wang P, Tang Y, Tirrell DA (2003) Incorporation of trifluoroisoleucine into proteins in vivo. J Am Chem Soc 125:6900–6906

34. Woll MG, Hadley EB, Mecozzi S, Gellman SH (2006) Stabilizing and destabilizing effects of phenylalanine → F5-phenylalanine mutations on the folding of a small protein. J Am Chem Soc 128:15932–15933

35. Meng H, Krishnaji ST, Beinborn M, Kumar K (2008) Influence of selective fluorination on the biological activity and proteolytic stability of glucagon-like peptide-1. J Med Chem 51:7303–7307

36. Meng H, Kumar K (2007) Antimicrobial activity and protease stability of peptides containing fluorinated amino acids. J Am Chem Soc 129:15615–15622

37. Niemz A, Tirrell DA (2001) Self-association and membrane-binding behavior of melittins containing trifluoroleucine. J Am Chem Soc 123:7407–7413

38. Chiu H-P, Kokona B, Fairman R, Cheng RP (2009) Effect of highly fluorinated amino acids on protein stability at a solvent-exposed position on an internal strand of protein G B1 domain. J Am Chem Soc 131:13192–13193

39. Chiu H-P, Suzuki Y, Gullickson D, Ahmad R, Kokona B, Fairman R, Cheng RP (2006) Helix propensity of highly fluorinated amino acids. J Am Chem Soc 128:15556–15557

40. Clark GA, Baleja JD, Kumar K (2012) Cross-strand interactions of fluorinated amino acids in β-hairpin constructs. J Am Chem Soc 134: 17912–17921

41. Cornilescu G, Hadley EB, Woll MG, Markley JL, Gellman SH, Cornilescu CC (2007) Solution structure of a small protein containing a fluorinated side chain in the core. Protein Sci 16:14–19

42. Horng J-C, Raleigh DP (2003) Φ-Values beyond the ribosomally encoded amino acids: kinetic and thermodynamic consequences of incorporating trifluoromethyl amino acids in a globular protein. J Am Chem Soc 125:9286–9287

43. Mortenson DE, Satyshur KA, Guzei IA, Forest KT, Gellman SH (2012) Quasiracemic crystallization as a tool to assess the accommodation of noncanonical residues in nativelike protein conformations. J Am Chem Soc 134:2473–2476

44. Naarmann N, Bilgiçer B, Meng H, Kumar K, Steinem C (2006) Fluorinated interfaces drive self-association of transmembrane α helices in lipid bilayers. Angew Chem Int Ed 45: 2588–2591

45. Senguen FT, Doran TM, Anderson EA, Nilsson BL (2011) Clarifying the influence of core amino acid hydrophobicity, secondary structure propensity, and molecular volume on amyloid-β 16–22 self-assembly. Mol Biosyst 7:497–510

46. Wang P, Fichera A, Kumar K, Tirrell DA (2004) Alternative translations of a single RNA message: an identity switch of (2S,3R)-4,4,4-trifluorovaline between valine and isoleucine codons. Angew Chem Int Ed 43:3664–3666

47. Jäckel C, Salwiczek M, Koksch B (2006) Fluorine in a native protein environment—how the spatial demand and polarity of fluoroalkyl groups affect protein folding. Angew Chem Int Ed 45:4198–4203

48. Jäckel C, Seufert W, Thust S, Koksch B (2004) Evaluation of the molecular interactions of fluorinated amino acids with native polypeptides. Chembiochem 5:717–720

49. Pace CJ, Zheng H, Mylvaganam R, Kim D, Gao J (2012) Stacked fluoroaromatics as supramolecular synthons for programming protein dimerization specificity. Angew Chem Int Ed 51:103–107

50. Pendley SS, Yu YB, Cheatham TE (2009) Molecular dynamics guided study of salt bridge length dependence in both fluorinated and non-fluorinated parallel dimeric coiled-coils. Proteins 74:612–629

51. Salwiczek M, Koksch B (2009) Effects of fluorination on the folding kinetics of a heterodimeric coiled coil. Chembiochem 10:2867–2870

52. Zheng H, Comeforo K, Gao J (2008) Expanding the fluorous arsenal: tetrafluorinated phenylalanines for protein design. J Am Chem Soc 131:18–19

53. Zheng H, Gao J (2010) Highly specific heterodimerization mediated by quadrupole

interactions. Angew Chem Int Ed 49: 8635–8639

54. Kwon O-H, Yoo TH, Othon CM, Van Deventer JA, Tirrell DA, Zewail AH (2010) Hydration dynamics at fluorinated protein surfaces. Proc Natl Acad Sci U S A 107: 17101–17106
55. Yoder NC, Yüksel D, Dafik L, Kumar K (2006) Bioorthogonal noncovalent chemistry: fluorous phases in chemical biology. Curr Opin Chem Biol 10:576–583
56. Marsh ENG, Buer BC, Ramamoorthy A (2009) Fluorine—a new element in the design of membrane-active peptides. Mol Biosyst 5: 1143–1147
57. Buer BC, Marsh ENG (2012) Fluorine: a new element in protein design. Protein Sci 21: 453–462
58. Yoder NC, Kumar K (2002) Fluorinated amino acids in protein design and engineering. Chem Soc Rev 31:335–341
59. Salwiczek M, Nyakatura EK, Gerling UI, Ye S, Koksch B (2012) Fluorinated amino acids: compatibility with native protein structures and effects on protein–protein interactions. Chem Soc Rev 41:2135–2171
60. Jäckel C, Koksch B (2005) Fluorine in peptide design and protein engineering. Eur J Org Chem 2005:4483–4503
61. Chothia C (1974) Hydrophobic bonding and accessible surface area in proteins. Nature 248:338–339
62. Eriksson A, Baase WA, Zhang X, Heinz D, Blaber M, Baldwin EP, Matthews B (1992) Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. Science 255:178–183
63. Richards FM (1977) Areas, volumes, packing, and protein structure. Annu Rev Biophys Bioeng 6:151–176
64. Baldwin RL (2013) Properties of hydrophobic free energy found by gas-liquid transfer. Proc Natl Acad Sci U S A 110:1670–1673
65. Marsh ENG (2000) Towards the nonstick egg: designing fluorous proteins. Chem Biol 7: R153–R157
66. Biffinger JC, Kim HW, DiMagno SG (2004) The polar hydrophobicity of fluorinated compounds. Chembiochem 5:622–627
67. Luo ZY, Zhang QS, Oderaotoshi Y, Curran DP (2001) Fluorous mixture synthesis: a fluorous-tagging strategy for the synthesis and separation of mixtures of organic compounds. Science 291:1766–1769
68. Harbury PB, Zhang T, Kim PS, Alber T (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science 262:1401
69. Nicholls A, Sharp KA, Honig B (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11:281–296
70. Wang L, Brock A, Herberich B, Schultz PG (2001) Expanding the genetic code of *Escherichia coli*. Science 292:498–500
71. Wang L, Xie J, Schultz PG (2006) Expanding the genetic code. Annu Rev Biophys Biomol Struct 35:225–249
72. Muir TW (2003) Semisynthesis of proteins by expressed protein ligation. Annu Rev Biochem 72:249–289
73. Muir TW, Sondhi D, Cole PA (1998) Expressed protein ligation: a general method for protein engineering. Proc Natl Acad Sci U S A 95:6705–6710
74. Schnölzer M, Alewood P, Jones A, Alewood D, Kent SBH (1992) In situ neutralization in Boc-chemistry solid phase peptide synthesis. Int J Pept Protein Res 40:180–193
75. Kelly SM, Price NC (1997) The application of circular dichroism to studies of protein folding and unfolding. Biochim Biophys Acta 1338:161–185
76. Jackson SE, Fersht AR (1991) Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Biochemistry 30:10428–10435
77. Kuhlman B, Raleigh DP (1998) Global analysis of the thermal and chemical denaturation of the N-terminal domain of the ribosomal protein L9 in H2O and D2O. Determination of the thermodynamic parameters, ΔH° ΔS° and ΔC° p, and evaluation of solvent isotope effects. Protein Sci 7:2405–2412
78. Yi Q, Scalley ML, Simons KT, Gladwin ST, Baker D (1997) Characterization of the free energy spectrum of peptostreptococcal protein L. Fold Des 2:271–280
79. Liu JJ, Horst R, Katritch V, Stevens RC, Wuthrich K (2012) Biased signaling pathways in {beta}2-adrenergic receptor characterized by 19F-NMR. Science 335:1106–1110
80. Li F, Shi P, Li J, Yang F, Wang T, Zhang W, Gao F, Ding W, Li D, Li J, Xiong Y, Sun J, Gong W, Tian C, Wang J (2013) A genetically encoded 19F NMR probe for tyrosine phosphorylation. Angew Chem Int Ed 52: 3958–3962
81. Pomerantz WC, Wang N, Lipinski AK, Wang R, Cierpicki T, Mapp AK (2012) Profiling the dynamic interfaces of fluorinated transcription complexes for ligand discovery and characterization. ACS Chem Biol 7:1345–1350

82. Anderson JT, Toogood PL, Marsh ENG (2002) A short and efficient synthesis of l-5, 5, 5, 5', 5', 5'-hexafluoroleucine from N-Cbz-l-serine. Org Lett 4:4281–4283

83. Chiu H-P, Cheng RP (2007) Chemoenzymatic synthesis of (S)-hexafluoroleucine and (S)-tetrafluoroleucine. Org Lett 9:5517–5520

84. Xing X, Fichera A, Kumar K (2001) A novel synthesis of enantiomerically pure 5, 5, 5, 5', 5', 5'-hexafluoroleucine. Org Lett 3: 1285–1286

85. Tsushima T, Kawada K, Ishihara S, Uchida N, Shiratori O, Higaki J, Hirata M (1988) Fluorine containing amino acids and their derivatives. 7. Synthesis and antitumor activity of α-and γ-substituted methotrexate analogs. Tetrahedron 44:5375–5387

86. Vine WH, Hsieh K-H, Marshall GR (1981) Synthesis of fluorine-containing peptides. Analogs of angiotensin II containing hexafluorovaline. J Med Chem 24:1043–1047

87. Sani M, Bruché L, Chiva G, Fustero S, Piera J, Volonterio A, Zanda M (2003) Highly stereoselective tandem aza-Michael addition–enolate protonation to form partially modified retropeptide mimetics incorporating a trifluoroalanine surrogate. Angew Chem Int Ed 42:2060–2063

88. Xing X, Fichera A, Kumar K (2002) A simple and efficient method for the resolution of all four diastereomers of 4, 4, 4-trifluorovaline and 5, 5, 5-trifluoroleucine. J Org Chem 67: 1722–1725

89. Erdbrink H, Peuser I, Gerling UI, Lentz D, Koksch B, Czekelius C (2012) Conjugate hydrotrifluoromethylation of α, β-unsaturated acyl-oxazolidinones: synthesis of chiral fluorinated amino acids. Org Biomol Chem 10: 8583–8586

90. Dale JA, Mosher HS (1973) Nuclear magnetic resonance enantiomer regents. Configurational correlations via nuclear magnetic resonance chemical shifts of diastereomeric mandelate, O-methylmandelate, and.alpha.-methoxy-.alpha.-trifluoromethylphenylacetate (MTPA) esters. J Am Chem Soc 95:512–519

Chapter 6

High-Quality Combinatorial Protein Libraries Using the Binary Patterning Approach

Luke H. Bradley

Abstract

Protein combinatorial libraries have become a platform technology for exploring protein sequence space for novel molecules for use in research, synthetic biology, biotechnology, and medicine. To expedite the isolation of proteins with novel/desired functions using screens and selections, high-quality approaches that generate protein libraries rich in folded and soluble structures are desirable for this goal. The binary patterning approach is a protein library design method that incorporates elements of both rational design and combinatorial diversity to specify the arrangement of polar and nonpolar amino acid residues in the context of a desired, folded tertiary structure template. An overview of the considerations necessary to design and construct binary patterned libraries of de novo and natural proteins is presented.

Key words Protein engineering, Protein design, Combinatorial library design, Synthetic biology, Peptide library, Gene library

1 Introduction

Protein combinatorial libraries have become a central strategy for exploring sequence space for novel sequences with desired "phenotypic" functions, including protein stability, enzymatic activity, binding affinity, and specificity. However, while powerful library screening and selection strategies are able to isolate individual sequences with desired properties from large collections of inactive candidates, the presence of truncated gene products and/or encoded misfolded/aggregating full-length proteins lowers the effective diversity of a library, thereby overburdening screens and selections with nonproductive candidate sequences. As the demand (and stringency) for engineered proteins increases for basic research, biotechnology, and the biopharmaceutical industries, the use of combinatorial approaches that increase the likelihood for success is desirable.

Combinatorial libraries are collections of gene sequences, which encode for amino acid diversity at defined positions in a

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_6, © Springer Science+Business Media New York 2014

protein template. The ultimate success of screens and selections is dependent on both the "genotypic" and "phenotypic" quality of the combinatorial library being screened. Genotypic quality refers to the quality of the DNA sequences (i.e., encoding full-length proteins, minimized frameshifts/deletions, diversity encoded in desired positions) in a combinatorial library. While straightforward, the assembly of diverse gene segments without frameshifts and deletions is often technically challenging.

Phenotypic quality refers to the translated gene products of a library that form the designed (typically, a folded and soluble) structure. Because of the enormous combinatorial diversity available for gene-encoded proteins (20^n, twenty naturally encoded amino acids possible at n amino acid residue positions), random protein libraries with greater than eight amino acids (i.e., $20^8 \sim 2.5 \times 10^{10}$) will quickly approach the practical limits of sequences able to be completely screened/selected in the laboratory. Thus for larger random combinatorial libraries, only a representative sampling of the theoretical diversity will be subjected to screens and selections in the laboratory. However, while the theoretical diversity is maximized in a random library, the quality of those sequences is likely to be low in folded and soluble structures, making those having desired properties to be exceedingly rare [1–5]. Thus by utilizing strategies that constrain random sequence space into regions that are likely to yield folded and soluble proteins, the likelihood of isolating functional proteins will be increased.

The binary patterning approach to combinatorial protein library design focuses combinatorial diversity into regions of sequence space rich in folded and soluble structures. By maintaining/defining the positions of polar (P) and nonpolar (N) residues in accordance with a three-dimensional folded structure template, but not their identity, large collections of folded and soluble proteins are able to be obtained. Pioneered with α-helical four-helix bundle and β-sheet de novo scaffolds in the Hecht laboratory, this approach when combined with high-quality gene assembly strategies led to the generation of synthetic combinatorial protein libraries that matched the designed templates, and with subsequent novel activities easily identified [6–11]. Furthermore, this approach has also been successfully translated to natural protein scaffolds [12–15], demonstrating its broad application. This chapter is an update of previous methods chapters [16, 17] outlining the use of binary patterning to design libraries of de novo and native proteins.

2 Materials

All enzymes, reagents, and materials were obtained from commercial sources, with careful consideration of preparation, purification, and quality. All single-stranded oligonucleotides were synthesized by commercial vendors and should include polyacrylamide gel electrophoresis (PAGE) purification prior to use. For overlap extensions during gene library assembly, thermostable polymerases which leave blunt ends (such as Pfu polymerase; Agilent Technologies, Santa Clara, CA) are required.

3 Methods

3.1 High-Quality Phenotypic Library: Design of a Structural Template

The success of binary patterning, for generating libraries of folded and soluble proteins, primarily depends on how well the structural template is designed. While binary patterning has been applied extensively to generate de novo/synthetic protein libraries by utilizing the periodicities found in protein secondary structure [6, 7, 18, 19], the binary code strategy can also be applied to local areas of existing (i.e., natural, evolved) protein scaffolds, such as individual residue positions or segments of the active site, part of the core, an interface, or linker region [12–14]. Several considerations important for library design are described below.

3.1.1 De Novo Binary Patterned Regions

α-Helical Designs

With a repeating periodicity of 3.6 residues per turn, an amphipathic α-helical segment of secondary structure is encoded (P = polar; N = nonpolar) by the pattern of P-N-P-P-N-N-P [16]. The Hecht laboratory has applied this α-helical binary patterning to three different generations (*see* **Note 1**) of the de novo four-helix bundle template, in which the nonpolar face of each amphipathic α-helix forms a central hydrophobic central core and the polar residue positions are oriented toward the aqueous solvent [6, 19, 20]. Extensive biophysical characterization of numerous representative proteins from each library has shown this strategy to yield combinatorial libraries rich in folded and soluble α-helical proteins, with properties of the desired four-helix bundle template [19, 21–23]. This design was further validated by NMR structural determination of second-generation library proteins S-824 and S-836, which show both proteins adopting a well-folded, four-helix bundle structure, as specified by the binary code design (Fig. 1a) [24, 25].

β-Sheet Design

With an alternating (P-N-P-N-P-N-P) periodicity, combinatorial synthetic gene libraries of amphiphilic β-sheets can be created in which opposite faces consist of polar and nonpolar residues, respectively [16, 17]. The Hecht laboratory has applied this binary pattern to a de novo 6 β-strand library template, generating proteins rich in β-sheet secondary structure, but prone to the formation of oligomeric (amyloid-like) structures in aqueous/polar solution [7, 26, 27]. By placing the alternating-patterned proteins in heterogeneous-interface environments, non-fibril assemblies of this template were achieved [28, 29]. Monomeric β-sheet protein libraries were obtained by substituting a lysine into the alternating patterning to disrupt hydrophobic packing, on the edge strand of the β-sheet template [26].

3.1.2 Fixed Regions

1. Fixed (constant) region sequences can be utilized for subcloning (i.e., incorporation of restriction sites) and/or assembly of full-length genes (*see* Subheading 3.3). Other constant amino acid residues/regions/sequences can also be incorporated to improve protein expression, assist in purification and concentration determinations, and increase in vivo protein stability [7, 19, 30–36].
2. In the design of the de novo four-helix bundles and β-sheet protein libraries, turn regions play important roles in forming secondary structure breaks (caps) in the template [16, 17]. Fixed sequences selected for turn regions, in both de novo library designs, were chosen based on their positional frequency and propensity observed in natural proteins [37, 38]. For example, in the first two generations of de novo four-helix bundle libraries, glycine residues were fixed at the helix cap positions and turn regions, due to their frequency at these positions in natural proteins [6, 19, 37]. NMR structure determination of two proteins from the second-generation library shows that these proteins adopt the monomeric four-helix bundle structure as designed (Fig. 1a) [24, 25]. In the third-generation four-helix bundle library, the turn regions were diversified (using the VRS codon which encodes for amino acids Gln, Glu, Asn, Asp, His, Lys, Arg, Ser, and Gly) [11, 16, 20]. While biophysical characterization of proteins from this library were all found to have properties consistent with the previous two generations of four-helix bundle libraries (including α-helical far-UV circular dichroism spectra, cooperative thermal denaturations), a solved crystal structure of a functional individual protein, WA20, showed that the protein formed a dimeric four-helix bundle, with each monomer encoding two long α-helices (Fig. 1b) [39].
3. Fixed regions may serve as self-priming sites for single-stranded synthetic oligonucleotides to anneal and initiate complementary strand enzymatic synthesis during gene assembly (*see* Subheading 3.3) [16]. This enables the combinatorial synthesis of libraries utilizing smaller fragments, which minimizes the incorporation of errors into the library (*see* below) [16]. The lengths of constant regions must be sufficient to promote sequence-specific annealing, with overlaps greater than 12 nucleotides yielding the best results (*see* **Note 2**).

3.1.3 Use of Binary Patterning with Natural Templates

By maintaining the observed patterning of polar and nonpolar residues of a naturally encoded template, sequence diversity can be introduced into a folded/evolved protein scaffold. The binary patterning approach has been applied to create libraries of the α-helical bundle domain of chorismate mutase, with some selected library members having comparable biophysical characteristics as the

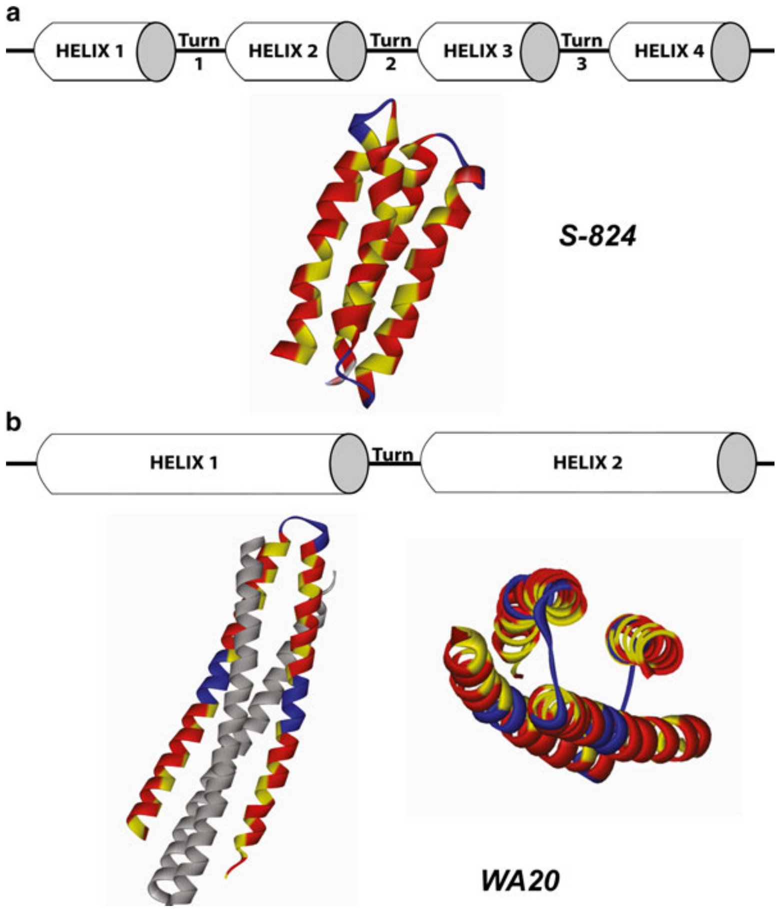


Fig. 1 (**a**) The design template for the elongated second- and third-generation de novo four-helix bundle libraries. Four individual, amphipathic α-helix libraries are encoded by the patterning of polar (*red*) and nonpolar (*yellow*) amino acids in accordance with a repeating periodicity of 3.6 residues per turn. In this template, the hydrophobic face of each helix would be oriented towards the central core of the bundle, while the hydrophilic faces of the helices are exposed to aqueous solvent. NMR structural determination of proteins S-824 (shown; PDB: 1P68) [24] and S-836 [25] from the second-generation library demonstrated that the library protein adopted the tertiary structure as designed. The turn region sequences (*blue*) in the second-generation library were fixed. (**b**) For the third-generation library, the three turn regions were diversified (encoded by the VRS degenerate codon). The X-ray crystallography structure of the functional library protein WA20 (PDB: 3VJF) [39] shows the protein adopting a dimeric four-helix bundle structure (one monomer shaded). WA20 diversified turns 1 and 3 did not form a strong break in the α-helical pattern, as observed with S-824 and S-836

native template and being able to restore cell growth in a chorismate mutase-deficient *Escherichia coli* cell line [12]. Recently, we applied the binary patterning approach to the central linker region of the highly conserved calcium signaling protein calmodulin

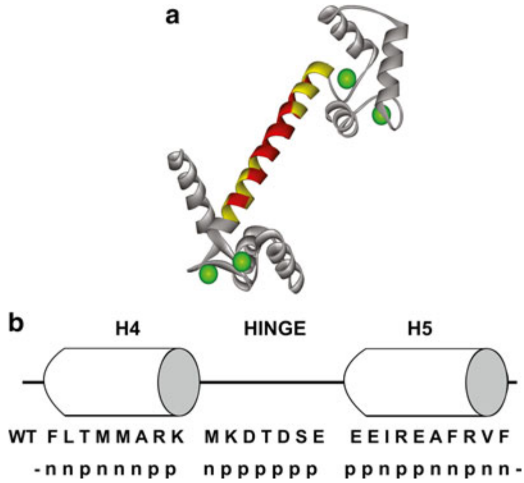


Fig. 2 (**a**) The binary patterning approach was applied to the central linker region of the conserved calcium signaling protein, calmodulin (CaM). Upon the binding of calcium (*green*) in the N- and C-terminal lobes of the protein (*gray*), the CaM central linker (residues: *red*, polar; *yellow*, nonpolar) adopts an open/extended conformation (PDB: 3CLN) [46, 47]. (**b**) The 25 amino acid central linker consists of two α-helices (H4, Helix 4; H5, Helix 5) separated by a seven-residue hinge domain. All library designs are based on the patterning of polar (*p*) and nonpolar (*n*) residues of the mammalian CaM central linker sequence. With six amino acids possible at each polar-encoded position and five amino acids possible at each nonpolar-encoded position, the theoretical diversity of these individual helix and hinge binary patterned libraries is between 2.3×10^5 and 2.4×10^7 sequences. When the individual helix libraries are assembled combinatorially, a Helix 4-Helix 5 library with a theoretical diversity of 1.6×10^{13} sequences is produced. Reprinted from *Protein Expression and Purification*, Vol. 75, Bradley et al., Expression, purification, and characterization of proteins from high-quality combinatorial libraries of the mammalian calmodulin central linker, pp 186–191, Copyright (2011), with permission from Elsevier

(CaM, Fig. 2) [13]. Several different binary patterned libraries were designed by maintaining the observed wild-type patterning of polar and nonpolar amino acid residues in this domain. For example, a combinatorial Helix 4-Helix 5 library with a theoretical diversity of 1.6×10^{13} sequences was constructed by using the seven amino acid (21 nucleotide) central hinge domain of the CaM central linker region as a constant site for annealing two individual binary patterned α-helical libraries (Helix 4 and Helix 5) by spliced overlap extension (Fig. 3) [13]. Characterization of unselected sequences from these libraries demonstrated that this approach yielded high-quality genotypic and phenotypic libraries. All gene sequences examined lacked internal stop codons and the diversity was incorporated as designed, without any duplications (Fig. 3). In addition, all protein sequences examined were able to

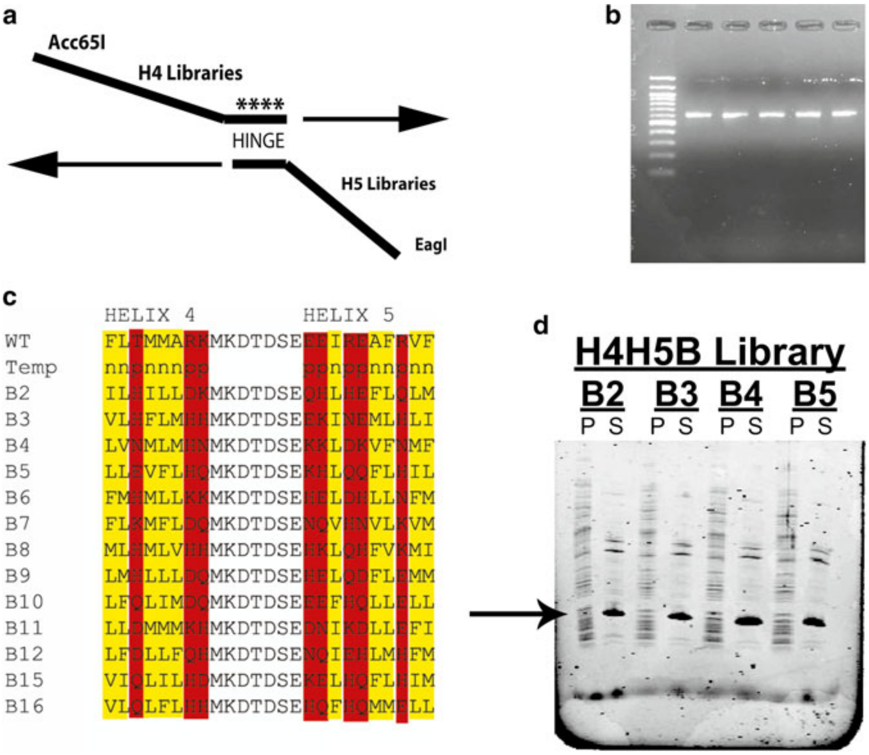

a
Acc65I
H4 Libraries
****
HINGE
H5 Libraries
EagI
b
c
HELIX 4
HELIX 5
WT FLTMMARKMKDTDSEEEIREAFRVF
Temp nnpnnnpp ppnppnnpnn
B2 ILHILLDKMKDTDSEQHLHEFLQLM
B3 VLHFLMHHMKDTDSEEKINEMLHLI
B4 LVNMLMHNMKDTDSEKKLDKVFNMF
B5 LLEVFLHQMKDTDSEKHLQQFLHIL
B6 FMHMLLKKMKDTDSEHELDHLLNFM
B7 FLKMFLDQMKDTDSENQVHNVLKVM
B8 MLHMLVHHMKDTDSEHKLQHFVKMI
B9 LMHLLLDQMKDTDSEHELQDFLEMM
B10 LFQLIMDQMKDTDSEEEFHQLLELL
B11 LLDMMMKHMKDTDSEDNIKDLLEFI
B12 LFDLLFQHMKDTDSENQIEHLMHFM
B15 VIQLILHDMKDTDSEKELHQFLHIM
B16 VLQLFLHHMKDTDSEHQFHQMMELL
d
H4H5B Library
B2 B3 B4 B5
P S P S P S P S

Fig. 3 Assembly of the high-quality Helix 4-Helix 5 (H4H5B) combinatorial gene library. (**a**) A double-stranded H4H5B library gene segment was assembled by spliced overlap extension from two single-stranded oligonucleotides, complementary at the CaM hinge (HINGE, ****) region, encoding helix 4 (coding) and helix 5 (noncoding) libraries, respectively. The Acc65I and EagI restriction sites, for subcloning the assembled library into the pETCaM1c expression vector, are unique and not present in the designed library. (**b**) The assembled, double-stranded H4H5B gene segment ran at the expected size (605 bp) on a 2 % agarose gel run with 100 bp DNA ladder. Bands were excised, restriction digested, and subcloned into pETCaM1c. (**c**) Sequence analysis of randomly selected, transformed *E. coli* found that all sequences were unique, with diversity incorporated as designed. This gene assembly and results were consistent with all other central linker libraries assembled. (**d**) SDS-PAGE analysis of randomly selected members from all assembled binary libraries showed that each library sequence over-expresses in *E. coli* (*arrow*) and was present in the soluble (S) fraction after cell lysis (cell pellet, P). These expression results are representative of all examined (70 total) individual sequences. Reprinted from *Protein Expression and Purification*, Vol. 75, Bradley et al., Expression, purification, and characterization of proteins from high-quality combinatorial libraries of the mammalian calmodulin central linker, pp 186–191, Copyright (2011), with permission from Elsevier

overexpress in *E. coli*, be affinity-purified in a calcium-dependent manner, maintain a strong α-helical CD spectrum, exhibit conformational changes upon the binding of calcium, and undergo posttranslational modification when co-expressed with a novel CaM methyltransferase (Fig. 3) [13, 15]. Collectively, these selected findings suggest that the binary patterning approach is applicable for generating high-quality genotypic and phenotypic libraries, to many different protein scaffolds and domains.

3.2 Codon Usage

As discussed previously, the organization of the genetic code allows for polar and nonpolar amino acid library residue positions to be encoded by defining the middle (second) position of a codon [9, 16, 17]. The degenerate codon N*T*N encodes five nonpolar amino acids (Val, Met, Iso, Leu, Phe), while the degenerate codon V*A*N encodes six polar amino acids (Glu, Asp, Lys, Asn, Gln, His). Standard nucleotide base mixtures are expressed using the International Union of Biochemistry (IUB) degenerate base symbols [40].

3.2.1 The Polar Amino Acid Codon

1. The first position of the VAN codon is encoded by an equimolar mixture of nucleotides G, C, and A [16]. By excluding T at this first codon position, the incorporation of two stop codons (as well as two tyrosine codons) in the gene library is avoided [16].
2. The third position of the VAN codon is occupied by an equimolar mixture of G, C, A, and T, resulting in an equal likelihood of His, Gln, Asp, Lys, Asp, and Glu being incorporated [16].
3. As with any degenerate codon, the compositions of base mixtures can be defined to minimize (or alternatively prefer) the incorporation of certain amino acids into the library. For example in the de novo α-helical libraries, T was omitted from the third codon position (i.e., VAV codon), thereby favoring polar amino acids (Glu, Lys, Gln) with higher propensities to form α-helices [16, 17, 41, 42].

3.2.2 The Nonpolar Amino Acid Codon

1. Both the first and third positions encode for equimolar mixtures of all four bases (A, C, G, T). The NTN codon does not encode for an equal representation of amino acids. Using these equimolar ratios, Leu would, for example, be represented six times more than Met [16].
2. Adjusting the molar ratios at the first and third codon positions minimizes biases or incorporation of unfavorable amino acids into the library. For example, in α-helix four-helix bundle encoded binary patterned libraries [6, 19, 20], the (A:C:G:T) molar ratio of 3:3:3:1 was defined at the first position and the third position was defined with an equimolar of C and G (IUB mixture: S), thereby making Leu less represented (three times more prevalent than Met) and Val (with a low α-helical propensity) limited to ten percent of the library-encoded hydrophobic positions [16].

3.2.3 Consideration of the Host Expression System

1. To facilitate expression and purification, as well as library screens and selections, the library DNA sequences in both the constant and the combinatorial regions of the library should be biased for codons favored in the host expression system. For example, in *E. coli*, the presence of C and G in the third position of a degenerate codon is generally preferred [16, 17, 43].

2. Conversely, codons that are used rarely in the host expression system should be minimized/avoided. Codons in *E. coli* such as CGA, AGA, and AGG (Arg); CTA (Leu); and CCC and ATA (Iso) are rare and may express poorly [16, 17, 43, 44].
3. The above strategies to maintain high protein expression were incorporated in the binary-patterned libraries of the CaM central linker region [13]. Evaluation of expression of randomly selected library members found that all sequences evaluated over-expressed in *E. coli* [13].
4. If it is not possible to avoid the use of certain rare codons, the use of other optimized expression *E. coli* cell lines (such as Rosetta^{TM} (DE3); EMD Millipore) or other host recombinant protein expression systems, with different codon preferences, should be utilized [16].

3.3 High-Quality Assembly of Full-Length Genes

Gene libraries are typically assembled from smaller gene segments. While various methods have been used to assemble full-length genes, this section focuses on considerations for spliced overlap extension (PCR based) gene assembly.

1. For a high-quality genotypic library, it is practical to assemble full-length gene libraries from smaller fragments in order to minimize the inherent errors (mostly deletions and frameshifts) associated with the synthesis of long, degenerate oligonucleotides [16, 17]. All oligonucleotides should include PAGE purification to reduce undesired sequences from being incorporated into the library during gene assembly [16].
2. The presence of interrupted sequences complicates screening strategies by increasing the burden on the selection system while failing to provide additional valid candidate genes for evaluation. It is therefore important to remove incorrect sequences from these libraries prior to screening for function. One strategy to remove incorrect sequences prior to gene assembly is to preselect smaller gene library fragments for open reading frames (*see* **Note 3**). This may be necessary for even relatively high-quality segments. For example, if the sequences of four gene segments are 85 % correct (as designed with no frameshift/deletions), upon combinatorial assembly approximately half of the sequences would encode the desired full-length product (0.85 × 0.85 × 0.85 × 0.85 ~50 %). For the third-generation de novo four-helix bundle library, the pPPV preselection system was developed and used to preselect four gene segment libraries, independent of the polypeptide fragment solubility and structure, before assembly into the full-length library of 102-amino acid sequences [20]. Using this strategy led to a large high-quality genotypic library of sequences, in which virtually all full-length sequences were free of internal stop codons as a result of frameshifts [20].

Furthermore, this high-quality library facilitated the identification of functional synthetic protein sequences that restored cell growth in auxotrophic *E. coli*, under selective conditions [11].

3. The semi-random oligonucleotides may be synthesized as either coding (sense) or noncoding (antisense) strands. Typically, each oligonucleotide encodes an individual library segment. Fixed regions, as described above, serve as sites of self-priming for complementary strand synthesis by DNA polymerase.
4. The use of DNA polymerases that leave blunt ends is required. Other PCR-based DNA polymerases that leave 3′-adenylation will result in the incorporation of frameshifts into the library.
5. Other sequence modifications might need to be considered for facilitating gene assembly and subcloning. Due to the encoded library diversity, it is likely certain that unintended and undesired annealing or restriction sites may be present at a frequency that may lower the genotypic quality of the library. The use of in silico analyses (*see* ref. 16), prior to the final design, is recommended to avoid these potential pitfalls.

4 Notes

1. The first-generation four-helix bundle library constructed in the Hecht laboratory was based on a 74-residue template in which characterized library proteins predominately formed dynamic, molten globular structures [6, 21–23]. Later four-helix bundle library designs included the addition of six library residues (maintaining the binary patterning) to each of the four α-helices [19, 20]. Characterization of proteins from these second- and third-generation 102-residue libraries found a significant increase in well-ordered proteins [19, 24, 25, 39].
2. To further enhance annealing, while still maintaining amino acid sequence diversity, individual nucleotides in the codons immediately preceding and following the fixed regions may be held constant [16, 17]. For example, by defining the third position of the VAN codon as G (VAG codon), two additional nucleotides are held constant for annealing while maintaining amino acid diversity at that residue position by encoding for the polar amino acids Glu, Gln, or Lys.
3. An alternate strategy would be the use of trinucleotide phosphoramidites, which represent entire amino acid codons, thereby eliminating the possibility of single-nucleotide frameshifts during solid-phase oligonucleotide synthesis [45].
4. The availability of large, diverse, and error-free libraries of binary patterned sequences, encoding native-like folded and soluble structures, sets the stage for experiments aimed at the identification of novel proteins with functions [20].

Acknowledgements

Support was obtained from the following sources: a pilot project grant from the NIH National Center for Research Resources (NCRR) Grant P20 RR020171, the Kentucky Science and Engineering Foundation (Grant Agreement # KSEF-148-502-207-201 with the Kentucky Science and Technology Corporation), the University of Kentucky Office of the Vice President of Research, and the University of Kentucky College of Medicine Startup funds.

References

1. Mandecki W (1990) A method for construction of long randomized open reading frames and polypeptides. Protein Eng 3:221–226
2. Davidson AR, Lumb KJ, Sauer RT (1995) Cooperatively folded proteins in random sequence libraries. Nat Struct Biol 2:856–864
3. Prijambada ID, Yomo T, Tanaka F, Kawama T, Yamamoto K, Hasegawa A et al (1996) Solubility of artificial proteins with random sequences. FEBS Lett 382:21–25
4. Yamauchi A, Yomo T, Tanaka F, Prijambada ID, Ohhashi S, Yamamoto K et al (1998) Characterization of soluble artificial proteins with random sequences. FEBS Lett 421: 147–151
5. Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410:715–718
6. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262:1680–1685
7. West MW, Wang W, Patterson J, Mancias JD, Beasley JR, Hecht MH (1999) De novo amyloid proteins from designed combinatorial libraries. Proc Natl Acad Sci U S A 96: 11211–11216
8. Moffet DA, Hecht MH (2001) De novo proteins from combinatorial libraries. Chem Rev 101:3191–3203
9. Hecht MH, Das A, Go A, Bradley LH, Wei Y (2004) De novo proteins from designed combinatorial libraries. Protein Sci 13:1711–1723
10. Cherny I, Korolev M, Koehler AN, Hecht MH (2012) Proteins from an unevolved library of de novo designed sequences bind a range of small molecules. ACS Synth Biol 1:130–138
11. Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH (2011) De novo designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth. PLoS One 6:e15364
12. Taylor SV, Walter KU, Kast P, Hilvert D (2001) Searching sequence space for protein catalysts. Proc Natl Acad Sci U S A 98: 10596–10601
13. Bradley LH, Bricken ML, Randle C (2011) Expression, purification, and characterization of proteins from high-quality combinatorial libraries of the mammalian calmodulin central linker. Protein Expr Purif 75:186–191
14. Sexton T, Hitchcook LJ, Rodgers DW, Bradley LH, Hersh LB (2012) Active site mutations change the cleavage specificity of neprilysin. PLoS One 7:10
15. Magnani R, Chaffin B, Dick E, Bricken ML, Houtz RL, Bradley LH (2012) Utilization of a calmodulin lysine methyltransferase co-expression system for the generation of a combinatorial library of post-translationally modified proteins. Protein Expr Purif 86:83–88
16. Bradley LH, Thumfort PP, Hecht MH (2006) De novo proteins from binary-patterned combinatorial libraries. In: Guerois R, López de la Paz M (eds) Protein design: methods and applications, vol 340, Methods in molecular biology. Humana Press, Totowa, NJ, pp 53–69
17. Bradley LH, Wei Y, Thumfort P, Wurth C, Hecht MH (2007) Protein design by binary patterning of polar and nonpolar amino acids. In: Arndt K, Mueller KM (eds) Protein engineering protocols, vol 352, Methods in molecular biology. Humana Press, Totowa, NJ, pp 155–166
18. Matsuura T, Ernst A, Pluckthun A (2002) Construction and characterization of protein libraries composed of secondary structure modules. Protein Sci 11:2631–2643
19. Wei Y, Liu T, Sazinsky SL, Moffet DA, Pelczer I, Hecht MH (2003) Stably folded de novo proteins from a designed combinatorial library. Protein Sci 12:92–102
20. Bradley LH, Kleiner RE, Wang AF, Hecht MH, Wood DW (2005) An intein-based

genetic selection allows the construction of a high-quality library of binary patterned de novo protein sequences. Protein Eng Des Sel 18:201–207

21. Rosenbaum DM, Roy S, Hecht MH (1999) Screening combinatorial libraries of de novo proteins by hydrogen-deuterium exchange and electrospray mass spectrometry. J Am Chem Soc 121:9509–9513
22. Roy S, Ratnaswamy G, Boice JA, Fairman R, McLendon G, Hecht MH (1997) A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties. J Am Chem Soc 119:5302–5306
23. Roy S, Hecht MH (2000) Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids. Biochemistry 39:4603–4607
24. Wei Y, Kim S, Fela D, Baum J, Hecht MH (2003) Solution structure of a de novo protein from a designed combinatorial library. Proc Natl Acad Sci U S A 100:13270–13273
25. Go A, Kim S, Baum J, Hecht MH (2008) Structure and dynamics of *de novo* proteins from a designed superfamily of 4-helix bundles. Protein Sci 17:821–832
26. Wang W, Hecht MH (2002) Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. Proc Natl Acad Sci U S A 99:2760–2765
27. Xiong HY, Buckwalter BL, Shieh HM, Hecht MH (1995) Periodicity of polar and nonpolar amino-acids is the major determinant of secondary structure in self-assembling oligomeric peptides. Proc Natl Acad Sci U S A 92:6349–6353
28. Brown CL, Aksay IA, Saville DA, Hecht MH (2002) Template-directed assembly of a de novo designed protein. J Am Chem Soc 124:6846–6848
29. Xu G, Wang W, Groves JT, Hecht MH (2001) Self-assembled monolayers from a designed combinatorial library of de novo beta-sheet proteins. Proc Natl Acad Sci U S A 98: 3652–3657
30. Hirel PH, Schmitter JM, Dessen P, Fayat G, Blanquet S (1989) Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino-acid. Proc Natl Acad Sci U S A 86:8247–8251
31. Dalboge H, Bayne S, Pedersen J (1990) In vivo processing of N-terminal methionine in *Escherichia coli*. FEBS Lett 266:1–3
32. Tsunasawa S, Stewart JW, Sherman F (1985) Amino-terminal processing of mutant forms of yeast Iso-1-Cytochrome-C: the specificities of methionine aminopeptidase and acetyltransferase. J Biol Chem 260:5382–5391
33. Huang S, Elliott RC, Liu PS, Koduri RK, Weickmann JL, Lee JH et al (1987) Specificity of cotranslational amino-terminal processing of proteins in yeast. Biochemistry 26: 8242–8246
34. Bowie JU, Sauer RT (1989) Identification of C-terminal extensions that protect proteins from intracellular proteolysis. J Biol Chem 264:7596–7602
35. Parsell DA, Silber KR, Sauer RT (1990) Carboxy-terminal determinants of intracellular protein-degradation. Genes Dev 4:277–286
36. Shoemaker KR, Kim PS, York EJ, Stewart JM, Baldwin RL (1987) Tests of the helix dipole model for stabilization of alpha-helices. Nature 326:563–567
37. Richardson JS, Richardson DC (1988) Amino-acid preferences for specific locations at the ends of alpha-helices. Science 240:1648–1652
38. Hutchinson EG, Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. Protein Sci 3:2207–2216
39. Arai R, Kobayashi N, Kimura A, Sato T, Matsuo K, Wang AF et al (2012) Domain-swapped dimeric structure of a stable and functional de novo four-helix bundle protein WA20. J Phys Chem B 116:6789–6797
40. Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic-acid sequences—recommendations 1984. Nucleic Acids Res 13:3021–3030
41. Chou PY, Fasman GD (1978) Empirical predictions of protein conformation. Annu Rev Biochem 47:251–276
42. Pace CN, Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 75:422–427
43. Gouy M, Gautier C (1982) Codon usage in bacteria—correlation with gene expressivity. Nucleic Acids Res 10:7055–7074
44. Kane JF (1995) Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Curr Opin Biotechnol 6:494–500
45. Virnekas B, Ge L, Pluckthun A, Schneider KC, Wellnhofer G, Moroney SE (1994) Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. Nucleic Acids Res 22: 5600–5607
46. Babu YS, Sack JS, Greenhough TJ, Bugg CE, Means AR, Cook WJ (1985) Three-dimensional structure of calmodulin. Nature 315:37–40
47. Babu YS, Bugg CE, Cook WJ (1988) Structure of calmodulin refined at 2.2 A resolution. J Mol Biol 204:191–204

Chapter 7

Methods for Library-Scale Computational Protein Design

Lucas B. Johnson, Thaddaus R. Huber, and Christopher D. Snow

Abstract

Faced with a protein engineering challenge, a contemporary researcher can choose from myriad design strategies. Library-scale computational protein design (LCPD) is a hybrid method suitable for the engineering of improved protein variants with diverse sequences. This chapter discusses the background and merits of several practical LCPD techniques. First, LCPD methods suitable for delocalized protein design are presented in the context of example design calculations for cellobiohydrolase II. Second, localized design methods are discussed in the context of an example design calculation intended to shift the substrate specificity of a ketol-acid reductoisomerase Rossmann domain from NADPH to NADH.

Key words Protein library design, Codon selection, Protein engineering, Computational protein design, Consensus analysis, Recombination, SCHEMA

1 Introduction

Library-scale design includes many divergent methods, ranging from random mutagenesis (e.g., error-prone PCR) to computational protein design. Library-scale design methods strive to achieve three goals: produce many diverse solutions, maintain folding and functionality in the majority of variants, and maximize ease of interpretation.

In practice, directed evolution (DE) is a remarkably effective library design method; many protein engineering challenges are readily solved via the stepwise accumulation of random mutations [1, 2]. Moreover, whereas a structure is typically a prerequisite for computational protein design (CPD), no special insight into the structure and function of the protein is required for DE methods. However, the search space of all random mutations is enormous; even a high-throughput assay can only sample a tiny fraction of the sequences within a few mutations from a parent sequence. Interpreting randomly accumulated mutations can also be difficult. Only in rare instances can favorable mutations be rationalized from available structural models.

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_7, © Springer Science+Business Media New York 2014

Compared to DE, CPD methods can consider an astronomical number of candidate sequences, including sequences that vary significantly from the initial sequence. CPD can result in impressive changes to the stability [3], aggregation-resistance [4], specificity [5], or enzymatic activity [6, 7], to name a few examples. Despite these successes, the foundation of CPD relies upon approximate models of protein structure and stability. Deficiencies in the scoring function or in the sampling of potential conformations can result in unfolded or inactive design variants [8]. Experimental testing of CPD sequences provides a referendum on the underlying CPD methodology; however, in practice it is difficult to learn from the success or failure of a single design attempt. An unfolded design variant indicates a model deficiency, but usually does not reveal an unambiguous remedy.

The philosophies behind these different methods are divergent: a pure DE scheme can be effective in the absence of protein structure and function information, while an accurate model of protein structure and function is the foundation and goal of CPD. Despite these philosophical differences, the gap between DE and CPD can be quite small in practice. For instance, a design cycle might start by using CPD to identify a low-energy sequence and progress to DE methods [9–11]. Combining the rational methods of CPD with DE screening methods balances search size with diversity. Rather than a search based on a large number of blind guesses (random mutations), one can formulate a search over a discrete set of hypotheses. Library-scale computational protein design (LCPD) methods combine rational and random methods to create a discrete set of hypothesized variants. Ideally, LCPD results in interpretable libraries that (1) are enriched for improved variants and (2) provide useful information for predicting sequence-structure-function relationships.

The appropriate choice of method will depend on the design goal at hand. Our first example discusses LCPD strategies and tools suitable for altering delocalized protein properties. Delocalized properties, such as stability or solubility, are the result of numerous amino acid interactions across a protein. Our second example focuses on protein properties that are localized to a distinct region. Significant variation within localized regions, such as binding pockets or protein interfaces, can create libraries with varying substrate specificity or enzymatic activity.

2 Materials

All computational scripts mentioned in this example are available at www.sharp-n.org. SHARPEN is an open-source C++/Python software library intended to facilitate the development of new algorithms for protein modeling and design.

3 Example I: Delocalized Design Libraries

Diverse libraries sample a broad range of sequence space, farther afield from an initial sequence, and are therefore more likely to contain significant variants of interest. However, when constructing a library of protein sequences, a trade-off is established between sequence diversity and library stability. Library stability is reflected in the properties of the individual sequences in two ways. First, a stable library will have few unfolded sequences. Second, the individual folded sequences within the library will be stable and functional. While allowing a wide range of mutations within a library greatly increases diversity, many mutations will decrease library stability [12]. When available, structural models can guide the selection of stable sequences by providing insight into which mutations are likely to be destabilizing [13].

Current library design methods use sequence and structure information to predict potentially stabilizing mutations. Hecht and co-workers have demonstrated the ability to design de novo proteins with binary patterning of polar and nonpolar amino acids [14–16]. Alternatively, the palette can be designed to ensure that mutations are compatible with the neighboring amino acids, considering multiple amino acid properties, such as volume, charge, and hydrophobicity [17]. Furthermore, information from multiple sequence alignments can be used to identify tolerated or favored mutations at each site [18–21]. Structural models can still be useful in conjunction with sequence-based design methods. For example, a structure can be used to refine ambiguous portions of the alignment (i.e., insertion/deletion sites) and to determine if certain residues (e.g., Pro, Trp) are likely to be incompatible with the protein backbone or the neighboring amino acids.

If detailed structural models are available, combinatorial CPD methods can be used. These methods provide each amino acid with multiple side chain conformations (rotamers) and provide each design site with multiple candidate amino acid identities [22, 23]. The design calculation is thus reduced to the combinatorial optimization problem of finding a rotamer combination of low energy. This problem can be solved using stochastic methods such as simulated annealing. The identification of the global minimum energy combination can also be achieved using methods such as dead-end elimination [24].

To obtain a library of designed sequences, a simple expedient is to repeatedly execute a stochastic design method or to design combinatorial mutation libraries to capture the sequence variation found within the pool of design solutions [25–27]. Such a library, however, will vary largely at sites that the CPD methods are found to be of marginal importance. Furthermore, if the CPD method confidently selects an unfavorable mutation (a systematic error), the poor choice could be present in all members of a library. Such an error could cause the entire library to be unfolded or nonfunctional. For example, a CPD algorithm with insufficient weight for van der Waal interactions might "overpack" the protein core, resulting in a molten globule sequence.

In contrast, an interpretable library of CPD variants could be designed to explicitly uncover and overcome systematic errors. A typical CPD scoring function assesses amino acid interactions as a series of contributions from hydrogen bonding, hydrophobic packing, van der Waals interactions, salt-bridge interactions, and other terms. A favorable design library would serve as a training set suitable for "learning" the weights associated with these different types of interactions. Whereas a CPD method might predict a stabilizing surface salt-bridge, a good library design would test this hypothesis. For example, if the library contains variants with the wild-type interaction, variants with the proposed salt-bridge, and variants with only one partner substituted, there is the possibility of determining the effective contribution of the salt-bridge. The concept is similar to the idea of a double-mutant cycle, although in this case the interaction is assessed in the presence of potentially confounding background sequence variation. If the library contains many such examples, the energy function could be trained to better predict the importance of salt-bridge interactions.

Recombination can be used to generate libraries that reduce the trade-off between library diversity and library stability; sequences generated through recombination are much more likely to retain stability than comparably diverse sequences generated through mutagenesis [28]. Similar to DNA shuffling [29, 30], site-directed recombination diversifies a library by substituting sequence blocks that contain multiple mutations. Recombination of homologous wild-type sequences has proven to be an effective library design method for a variety of protein folds including beta-lactamase [31], cytochrome P450 [32], arginase [33], and several cellulase families [34–37].

Recombination need not be limited to natural sequences; homologous parent sequences identified from directed evolution or CPD methods could also be recombined to create a diverse library. In the example below, we recombine one wild-type parent with two computationally designed sequences. Incorporating computational designs into a recombination library allows the designer to specifically target a library property of interest (e.g., stability at low pH). Energy scoring functions attempt to incorporate many global stability factors, including hydrogen bonds, hydrophobic interactions, packing efficiency, and conformational strain. By searching through a large sequence space, computational designs may identify improved variants that have never occurred in nature. As discussed above, CPD variants are likely to include design errors. Recombining blocks from CPD variants with a wild-type sequence will allow the dissection of stabilizing and destabilizing sequence motifs. Constructing a chimera sequence that incorporates successfully designed blocks from the CPD sequence and leaves out blocks corresponding to CPD failures is likely to result in chimeras that meet the design goals.

In the example below, we demonstrate how LCPD methods might be applied with a model target, cellobiohydrolase II (CBHII) from *Humicola insolens* (PDB entry 1OCN). In the first two sections, parent sequences are designed using CPD. We then recombine the parents to form a chimera library. The final section discusses how information from library screening could be used to enhance subsequent designs. We will not discuss the experimental construction of chimera libraries because protocols for site-directed chimeragenesis are thoroughly described in earlier reports [38, 39].

3.1 Phase I: Create a Design Palette

To begin a protein design problem we define a design palette: the set of candidate amino acids for each design position. Ideally, the design palette should be limited in size so that the resulting sequence space can be computationally searched in a reasonable time frame. Early zinc finger protein design work by Dahiyat and Mayo demonstrated the value of specifying a carefully selected design palette [40]. In this case the palette was restricted to alanine and hydrophilic residues (Ala, Ser, Thr, His, Asp, Asn, Glu, Gln, Lys, and Arg) at surface sites, hydrophobic residues (Ala, Val, Leu, Ile, Phe, Tyr, and Trp) at buried sites, and hydrophilic or hydrophobic residues at boundary sites. Furthermore, two sites with φ angles greater than 0° were restricted to Gly only. Even with this reduced design palette, the small 30-residue protein had a search space of 1.9×10^{27} possible sequences, or 1.1×10^{62} unique conformational variations. Modern computers and search algorithms can effectively search combinatorial solution spaces of this astounding size [24, 41], but such a diverse palette would not be feasible for proteins with hundreds of residues. One reason to use a design palette is to avoid buried hydrophilic amino acids and exposed hydrophobic amino acids. However, Kuhlman and co-workers recently reported a method for avoiding hydrophobic surface patches without eliminating them from the design palette altogether [42].

At the outset of a design challenge it can be difficult to calibrate the design palette. A conservative design palette would consist of relatively few design sites, and would avoid any mutations that are a priori likely to be disruptive. While a non-conservative palette may facilitate the design of a superior sequence, it will also allow more mutations, lead to a diverse library, and may result in a largely unfolded library. The balance between diversity and stability motivates an iterative approach; if the desired library “phenotype” and library stability are not achieved in the first round of library design, the palette can be adapted in subsequent iterations to be more or less conservative.

In this example, we design a very conservative palette intended to engender a largely folded library. First, prevalent mutations are identified from homologous multiple sequence alignments. While mutations commonly seen in consensus alignments are not guaranteed to be stabilizing, the selective pressure of evolution strongly suggests that these mutations are not destabilizing. Second, folding free energy changes are estimated for every point mutation. In principle, excluding mutations predicted to have unfavorable folding free energy changes will result in a smaller, conservative palette that is less likely to harbor destabilizing mutations.

1. Identify sequences with a high sequence identity to the query sequence.

   We used the Basic Local Alignment Search Tool (blast.ncbi.nlm.nih.gov) to identify similar sequences and retained alignments with sequence identity greater than 35 %. For the CBHII consensus analysis, 175 sequences met this cutoff criterion.

2. Identify common amino acids at each site and save in a consensus design palette.

   BLAST results were parsed using *run_alignment.py*. A cumulative approach was used that retained the most common amino acid, second most common amino acid, etc. at each site until 90 % of the sequences had been included.

3. Calculate predicted folding free energy changes ($\Delta\Delta G$) for each point mutation.

   Preparatory steps and FoldX calculations were executed using *run_foldx_multi.py*. All 20 amino acids were considered at each site. *See* **Note 1** for more information.

4. Combine all favorable mutations ($\Delta\Delta \leq 0$) in a secondary palette.

   FoldX outputs were parsed using *run_foldx_analysis.py*.

5. Repeat **steps 3** and **4** with alternate backbone scaffolds to account for slight variations in structure.

   Potential backbone scaffolds 1BVW, 2BVW, 1GZ1, and 1OC5 were identified by BLAST searching against the Protein Data Bank (PDB). Each chain from within a structural model was considered a unique backbone scaffold. *See* **Note 2**.

6. For a conservative design, reduce the design palette to the intersection between the consensus analysis and the multiple structural modeling palettes (Table 1).

   The script *run_consensus_foldx.py* was used to identify the intersection between multiple design palettes. We chose to include mutations allowed in the consensus palette and by the majority of the folding free energy palettes (arbitrarily defined as 2/3).

Table 1
Potential CBHII mutations at selected sites

WT A.A.	Consensus palette	FoldX palette	Intersection palette	Chosen A.A.	Rationale
N103	ANPSK	NP	NP	P	Allows h-bond between Y100 and E154 (Fig. 2a)
R123	AIKNRV	IKLMQRTV	IKRV	I	Computational model predicts favorable energy interactions (Fig. 2b)
Q361	GKLQSV	ILMQV	LQV	Q	Mutating Q361 loses side chain backbone h-bond (Fig. 2c)
K366	AEIKLNQST	FKLY	KL	K	K366L mutation introduces an unfavorable nonpolar surface residue (Fig. 2d)

3.2 Phase II: Select Parent Sequences

There are a few basic principles to consider when selecting parent proteins for recombination. First, parent proteins must have similar structure in order to remain folded upon recombination. If structural models are unavailable, sequence identity can be used to estimate structural similarity. Parent sequences with high sequence identity (60–80 % identity) generally have similar structure [43] and recombination will result in a high fraction of folded chimeras. In contrast, parent sequences with low sequence identity (<40 % identity) are much more likely to engender unfolded chimeras [32, 44, 45]. Second, critical residues should be conserved in each parent. Catalytic active site residues may be considered critical, since variants that do not conserve these amino acids are very unlikely to retain enzymatic activity. Other sites that may be considered critical include disulfide residues, sites for posttranslational modification (e.g., glycosylation sites), and sites that could affect the protein folding mechanism such as *cis*-prolines.

Combinatorial optimization software, such as SHARPEN (www.sharp-n.org), can be used to search for low-energy sequences that meet these criteria [46, 47]. Alternate side chain conformations (rotamers) are included from the backbone-dependent Dunbrack rotamer library [48]. Numerous algorithms exist for finding low-energy sequences and conformations. SHARPEN allows users to easily try a variety of stochastic algorithms (e.g., FasterPacker, SimulatedAnnealingPacker). Because these algorithms may yield different results for each repetition, repeated trials are useful for identifying mutations that are strongly or weakly preferred by the scoring function.

7. Given a design palette, search for low-energy sequences.

   We used the FasterPacker search algorithm in SHARPEN to identify low-energy candidates according to an all-atom Rosetta energy function [49]. This combinatorial optimization routine mimics the single-residue perturbation/relaxation method within the original description of the FASTER algorithm [41]. Briefly, this method systematically attempts to surmount local minima during optimization by temporarily fixing a single side chain in a particular conformation, and then assessing the effect of that perturbation combined with the relaxation/optimization of the neighboring side chains. Design calculations were performed using *run_conservative_design.py*. Separate calculations were run for the conservative and consensus design palettes. A total of 100 repetitions were performed for each design.

8. Sort candidate designs to identify the lowest energy design (Fig. 1).

   The list of designed protein models generated by SHARPEN was sorted using *run_sort_by_energy.py*. For the conservative design, the lowest energy sequence was 38 Rosetta energy

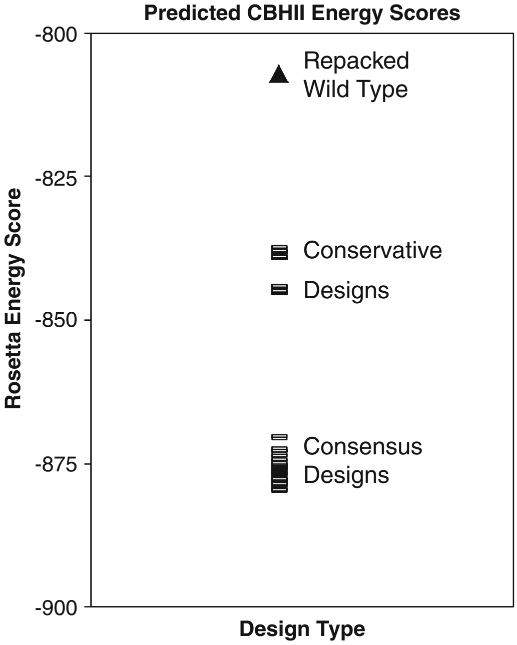


Fig. 1 Using a stochastic search algorithm in a design problem yields variants of differing energies. The distribution of potential low-energy candidates was sampled by performing 100 repetitions for each design palette. The starting energy score of 1OCN.pdb was −501 Rosetta energy units (REU). After repacking to optimize side chain conformations, the energy score was reduced to −807 REU (*triangle*). Searches based on the conservative design palette (intersection of consensus and FoldX methods) achieved an energy reduction of 38 REU (*rectangles*), while the larger consensus palette allowed an energy reduction of 72 REU (*rectangles*)

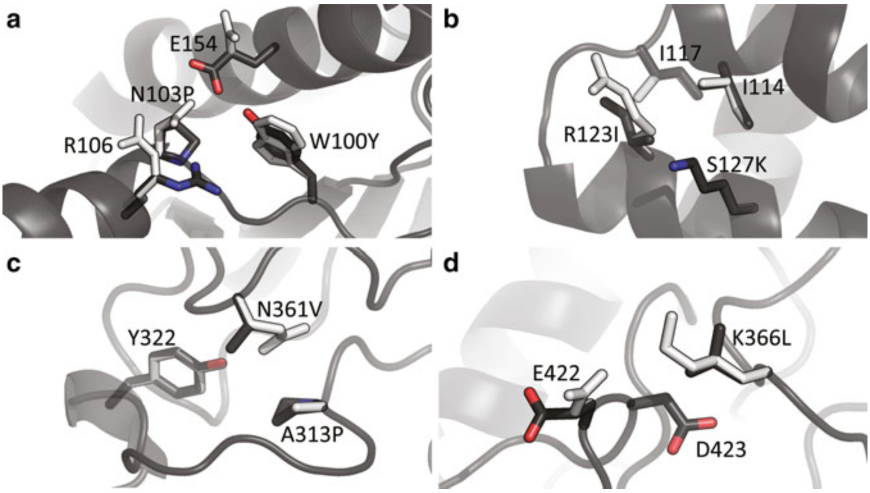


Fig. 2 Visual inspection of potential mutations. Example mutations include (**a**) W100Y and N103P, (**b**) R123I and S127K, (**c**) A313P and N361V, and (**d**) K366L. *See* Table 1 for discussion regarding which mutations were kept or reverted back to wild type

units lower than the wild-type sequence. The larger consensus design palette allowed a slightly more favorable energy change of 73 energy units.

9. Inspect designs to identify common mutations and stabilizing features (Fig. 2a–d).

   Given the limitations of contemporary sampling and scoring in CPD methods, visual inspection of the prospective mutations can provide an additional opportunity to ensure a reasonable design. The script *master.py* incorporates many analysis functions, including multiple sequence alignments (*run_multiple_sequence_alignment.py*), global energy comparisons (*run_compare_pdbs.py*), and amino acid polarity comparisons (*run_polarity_of_mutations.py*). *See* **Note 3** for more information.

The final conservative design contained a total of 58 mutations (84 % sequence identity to wild-type sequence), whereas the consensus design contained 120 mutations (66 % sequence identity to wild-type sequence).

3.3 Phase III: Recombine Parent Sequences to Form a Library

After parent sequences have been finalized, one must select the number of blocks to recombine. Block size influences library interpretability and library size. We define library interpretability as the extent to which it is possible to (1) rationalize the functionality of the library members in terms of structural detail and (2) deploy the experimental data to construct an improved, more predictive model for future designs. Small blocks can greatly improve library interpretability. Namely, small blocks have fewer mutations per block, allowing interesting changes in the protein fitness to be tracked to the responsible mutations. For example, Heinzelman et al. were able to isolate an individual stabilizing mutation C404S from recombined CBH II parents because the cognate block contained only ten other mutations [50]. However, dividing a parent sequence into small blocks can greatly increase library size. Library size can be determined from the number of blocks and the number of parent sequences; for three parent sequences divided into four blocks each, the resulting library size will be 3^4 or 81 chimeras. If the aim is experimental screening of all library members, the library should be sized according to the screening capacity. Recombining more blocks, of smaller size, will increase the library size exponentially.

Given a range of desired block sizes, various structure-guided methods can be used to determine ideal recombination sites. Methods such as SCHEMA [51], SIRCH [52], and OPTCOMB [17] aim to minimize the number of disruptive amino acid contacts in recombined chimeras. Using a slightly different method, FamClash combines clash detection with protein family sequence data to maximize chimera functionality [53]. The protocol in this chapter is built around the recombination as a shortest path problem (RASPP) method [54]. That said, the presented protocol could readily be adapted to incorporate an alternative method.

SCHEMA aims to maximize the number of folded library members by minimizing the number of novel amino acid interactions (not found in parent proteins) [31, 55]. Interactions are defined as heavy atom pairs (excluding backbone O and N and all H) within 4.5 Å in a parent protein. A predictive SCHEMA energy score "E" is assigned to represent the number of novel contacts within each chimera. A diversity parameter "m" specifies the number of mutations between each chimera and the closest parent. The average SCHEMA energy and mutation level of all chimeras within a library are denoted <E> and <m>, respectively. While considering <m> does provide a means of favoring diverse libraries, it does not lend itself to the design of interpretable libraries. We therefore propose a third metric H_{sbs}^{max}, which is the maximum Hamming number for a single block substitution. If a candidate library is dominated by one or a few large blocks H_{sbs}^{max} will be large and the library will be less interpretable because the effect of changing the large blocks will include the aggregate effect of many mutations. A small H_{sbs}^{max} indicates that any block substitution that is found to be important is less likely to have an obscure origin. Multi-scale enzymology, tracing an important block effect to the role of individual mutations, will be more feasible for such a library.

The library containing the minimum number of nonnative amino acid interactions can be determined by formulating the library optimization as a dynamic programming problem [54]. By weighting edges of a graph according to a SCHEMA penalty, a shortest path can be chosen that contains optimum block crossover sites. Further restrictions can be placed on the search space, such as limiting crossover sites to locations where three or four nucleotides

Table 2
RASPP settings used for CBHII recombination

Parameter	Value	Description
pdbfile	"1ocn.A.pdb"	Name of pdb file used to identify native contacts
Cutoff	4.5	Distance cutoff used to identify native contacts (Å)
Skipatoms	['N', 'O', 'H']	Atoms to be skipped when identifying native contacts (skips N and O in backbone only)
Numxo	3	Number of crossover sites (three crossover sites generate four blocks)
Overhang	3	Number of conserved nucleotides required at crossover sites (can be 0 if overhangs are unnecessary for library construction)
Min_lengths	Range (5,7)	Range of minimum block lengths

are preserved in all parent sequences. Conserving nucleotides at crossover sites allows blocks to be recombined using type II restriction enzymes [39].

10. Create a sequence alignment file based on the parent sequences. A number of programs are available for generating sequence alignment files; we used ClustalOmega (www.ebi.ac.uk/Tools/msa/clustalo/) and converted the format using *run_convert_msa_format.py*.

11. Specify the library design parameters (Table 2).
Block size is the primary parameter in recombination problems. However, the provided code is engineered for flexibility. The user can specify how amino acid contacts are defined (minimum cutoff distances, and the heavy atoms considered), and which sites are feasible crossover locations (e.g., the number of nucleotides in overlap regions). These parameters can also be modified in the settings file *raspp_config.py*.

12. Identify potential crossover sites. For each candidate set of crossover sites, calculate <E>, <m>, and H_{sbs}^{max}.
Running *run_raspp_curve.py* generated a list of optimum crossover sites. <E>, <m>, and H_{sbs}^{max} were saved in an output file pareto.csv for each set of sites.

13. Select a candidate library corresponding to a set of crossover sites.
Ideally, the selected library will have low <E>, high <m>, and low H_{sbs}^{max}. Four potential CBHII libraries were identified from a plateau region on the <E>/<m> Pareto front (Fig. 3a). The <E>/ H_{sbs}^{max} Pareto front (Fig. 3b) allowed us to discriminate between these four candidate libraries and select a design that optimized diversity and interpretability. The final design features recombination sites 167, 244, and 345.

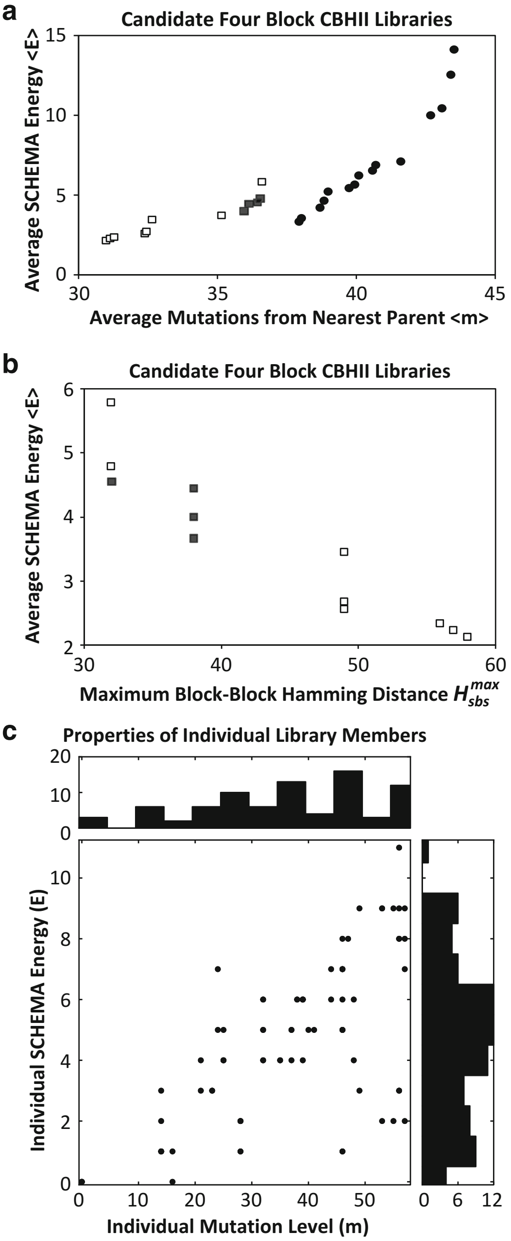


Fig. 3 Multiple design parameters can be considered when selecting a candidate library. (**a**) A Pareto front for four-block CBHII recombination with one wild-type and two conservatively designed parents (*squares*) has a similar average mutation level <m> and average SCHEMA energy <E> as four-block recombination with three wild-type parent sequences *Humicola insolens*, *Chaetomium thermophilum*, and *Hypocrea jecorina* (*filled circles*). Promising candidate libraries have low <E> and high <m> (*filled squares*). (**b**) Maximum block-block hamming distance H_{sbs}^{max} quantifies the interpretability of each library. Four similar candidate libraries from the <m> Pareto front (*filled squares*) are easily distinguished by the H_{sbs}^{max} Pareto curve. (**c**) Within the library featuring crossover sites 167, 244, and 345, chimeras have a distribution of mutation level "m" and SCHEMA energy "E"

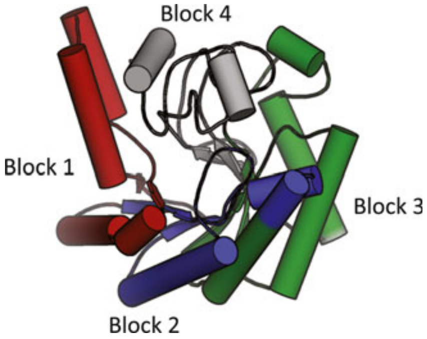


Fig. 4 Structural blocks identified using RASPP methods. Blocks are defined as follows: Block 1—residues 91–166 (*red*), Block 2—residues 167–243 (*blue*), Block 3—residues 244–344 (*green*), Block 4—residues 345–450 (*grey*)

14. Inspect the prospective design.
    (a) Generate histograms of chimera properties (Fig. 3c).
    Do outliers skew the library average properties? Do the distributions show that most library members have acceptable diversity and predicted disruption? Distributions can range from normal to multimodal, depending on the parent proteins. Our selected library showed an approximately normal <E> distribution and a slightly skewed <m> distribution.
    (b) Inspect the crossover sites and structural features of each block (Fig. 4).
    A candidate library design can be inspected using PyMOL (www.pymol.org). First load the parent pdb, and then run the corresponding showcontacts.pml script by typing *@showcontacts.pml.*[*recombination sites*] into the PyMOL command line. Blocks are colored based on selected crossover sites.
    (c) Verify that the library is constructible.
    Are the block sizes compatible with construction? Small DNA fragments could be difficult to purify using gel purification. If using a restriction enzyme-based construction protocol, ensure that the design produces the correct overhangs. If the candidate splice sites are not orthogonal, can alternate codons be used? If necessary, select a new candidate library from the Pareto front.

Another approach for selecting recombination crossover sites is to ignore the protein sequence and select sites solely on the basis of one protein structure. This alternate approach could be useful for preliminary library designs where CPD parent sequences have not yet been determined. In principle, structure-based crossover sites could be selected using a variety of approaches similar to domain detection algorithms [56]. However, to build a recombination library experimentally, the blocks should consist of contiguous residues.

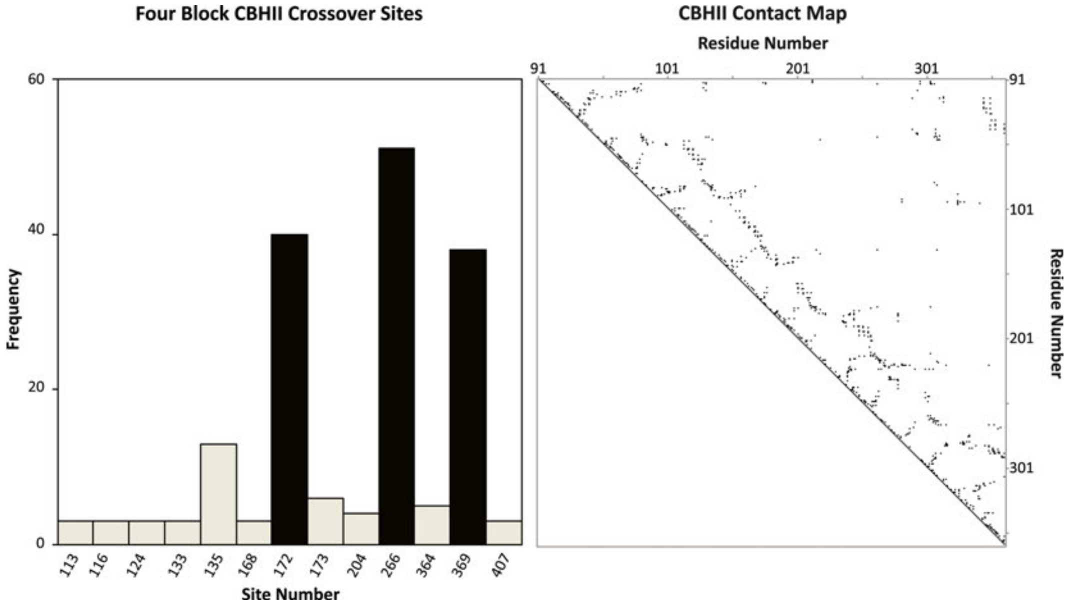


Fig. 5 (**a**) Scanning over a range of minimum block sizes from 5 to 90 creates a range of optimum block recombination sites. To identify preferred sites, all sites occurring in more than two unique libraries are considered. Recombination sites 172, 266, and 369 are preferred for a four-block CBHII library. (**b**) The contact map for 1OCN.pdb shows contacts characteristic of alpha helices and beta sheets. Ideal recombination blocks maximize intra-block contacts and minimize inter-block contacts

Therefore, a simple expedient is to reuse the dynamic programming approach of RASPP, but to replace the SCHEMA penalty matrix with a simple binary contact map. The resulting crossover sites will be those that minimize the number of inter-block contacts (and therefore maximize the number of intra-block contacts). We demonstrate this alternative method using *run_pick_modules.py*.

15. Identify crossover sites for a range of minimum block lengths. We used *run_pick_modules.py* to create a histogram of potential block crossover sites.

16. Select a set of preferred crossover sites and inspect structural blocks.

    In our example, sites 172, 266, and 369 were frequently chosen as crossover sites (Fig. 5a). *Pick_modules.py* strongly favors certain crossover sites that minimize the number of inter-block contacts. Notably, these sites are not obvious from inspection of the protein structure or the contact map (Fig. 5b).

3.4 Phase IV: Evaluate the Library

In any design cycle, the final step involves constructing and experimentally verifying the designs. Selected chimeras can be synthesized using traditional molecular biology techniques [39] or via gene synthesis and assayed to determine the extent of folding or

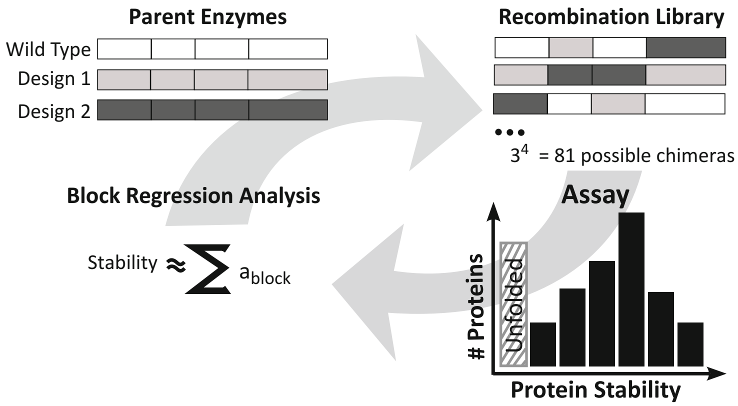


Fig. 6 Library design is an iterative cycle that consists of parent selection, block recombination, experimental testing, and validation of biophysical models

retained activity. In the CBHII example, an activity assay such as the Nelson-Somogyi reducing sugar assay could be performed at a variety of temperatures to test chimera function and stability [34].

Experimental verification of large libraries can be costly and time consuming. One approach is to experimentally screen a small percentage of the library and attempt to use the initial screening data to derive a predictive stability model applicable to the remainder of the library. Simple regression methods that model the stability of each chimera as the sum of contributions from each block have been found to be predictive [57]. The surprising additivity of block contributions to stability can be attributed to sequence conservation among the parents and the partitioning of epistatic interactions into structural modules [45].

To complete the iterative library design cycle, knowledge gained from experimental testing can be incorporated into subsequent designs (Fig. 6). In addition to predicting the fitness of library members, a trained regression model can also guide the refinement of the CPD methodology. For example, if a particular sequence block from one of the CPD design variants was found to be highly destabilizing, the deficiency in the CPD model can be investigated by reexamining the mutations that comprise that block.

4 Example 2: Localized Protein Design Libraries

4.1 Introduction

Many properties of a protein depend critically on a subset of the amino acids. Protein-protein binding, cofactor binding, enzyme specificity, and catalysis are all properties for which structural models can enable hypothesis-driven engineering of specific residues. The applications for focused protein library design are nearly limitless. Below, we briefly survey a selection of such applications before presenting an example protocol.

4.1.1 Protein-Protein Interface Design

Protein-protein interactions (PPIs) are fundamental to many of the biomolecular recognition events that drive biological processes. However, understanding PPIs is difficult because they typically involve many weak noncovalent bonds over large surfaces. The biophysical principles (e.g., extent of buried nonpolar surface area) underpinning protein-protein interfaces vary [58], and not all participating amino acids will contribute equally to the binding affinity [59].

Just as understanding PPIs plays a key role in molecular biology, the ability to control PPIs is key for engineering new therapeutic biomolecules. Baker and co-workers demonstrated an effective protocol de novo protein inhibitor design with a protein that binds an influenza virus stalk site [60, 61]. Engineering new PPIs as a CPD problem extends the methods deployed for monomeric CPD [10, 62]. Combinatorial optimization routines are applied to the interfacial amino acids to optimize a scoring function that includes van der Waals, hydrogen bonding, and electrostatic interactions with the partner protein. Notably, alternate approaches to engineer interactions can circumvent the need to engineer large complementary surfaces. Examples include the addition of shared metal-binding sites [63] or disulfide bonds [64].

An improved understanding of PPIs could also be useful for downstream problems in biotherapeutic development. For example, prevention of aggregation is important to extending the shelf life of therapeutic proteins. Unwanted PPIs could be destabilized through site-specific mutations of existing complementary interfaces or electrostatic repulsion via supercharging [4, 62, 65].

4.1.2 Binding Small Molecules

Binding of metals and small organic molecules is necessary for many proteins to function. Mutations of amino acids in the hydrophobic protein core can result in new cavities for small molecules to bind. For small nonpolar molecules, it is desirable to create a hydrophobic local environment around the cavity. Hecht and co-workers demonstrated that the simple truncation of Phe to Ala in the de novo protein S-824 resulted in the ability to bind small aromatic compounds [66]. Binding polar molecules and metals is more challenging because it requires the installation of complementary electrostatic interactions and hydrogen bonds. Notably, Matthews and co-workers have created cavities in T4 lysozyme that can bind the polar ligands pyridine, phenol, and aniline [67].

4.1.3 Changing Cofactor Specificity

Engineering organisms to produce higher yields of products via knockouts of competing metabolic pathways can create cofactor imbalances. Shifting cofactor specificity may resolve this problem by substituting the limiting cofactor with one that is in excess. For example, in attempts to anaerobically produce isobutanol in *Escherichia coli* via the Ehrlich pathway, NADPH-dependent

Table 3
Translation of degenerate codon base to nucleotides

Degenerate base	Actual base
N	A or C or G or T
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
K	G or T
M	A or C
R	A or G
S	C or G
W	A or T
Y	C or T

enzymes were engineered to shift the specificity preference to NADH. The best variant of the final library exhibited a specificity of 185:1 for NADH to NADPH, a 54,000-fold change from the original variant. By completely removing the dependence on NADPH, isobutanol titres at 100 % theoretical yield were achieved [68].

4.1.4 Degenerate Codons

A widely used approach for introducing amino acid diversity at a particular site is through the use of degenerate codons [69]. Degenerate codons are sets of oligonucleotides that code for multiple amino acids. The standard degenerate codon naming convention used in this text is presented in Table 3. Routine site saturation mutagenesis protocols often employ the degenerate codon NNK, which codes for all 20 amino acids [70]. While site saturation mutagenesis is simple and efficient, combinatorial explosion limits the number of sites that may be targeted. As the number of saturation sites increases, it rapidly becomes infeasible to transform, isolate, and thoroughly screen the resulting library. Even with a very-high-throughput screen, allowing for screening of 10^8–10^{11} targets [71], site saturation mutagenesis can only be performed on eight residues. Furthermore, NNK encodes the amino acids unevenly (Fig. 7). The resulting bias against rare amino acid combinations increases exponentially with the number of sites.

The limitations of site saturation mutagenesis motivate the development of more efficient methods that eschew brute force search. Table 4 illustrates various useful degenerate codon alternatives to NNK. These sets allow for introduction of diversity at a

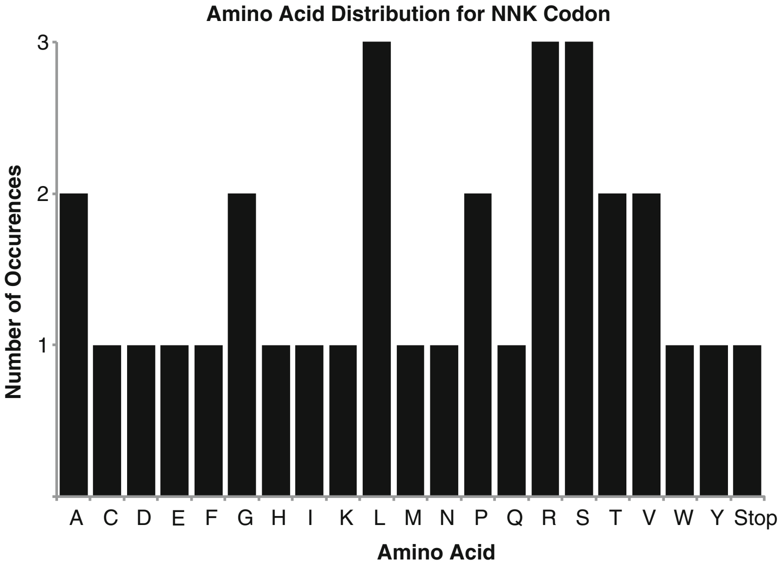


Fig. 7 Selection of the codon NNK unevenly encodes the amino acids. NNK also encodes for a stop codon, which will result in a nonfunctional variant

Table 4
Examples of degenerate codon to amino acid subset

Codon	Type	Amino acids	Stop codons	Unique codons
NNK	All 20 A.A.	All 20	TAG	32
DVT	Hydrophilic	A,C,D,G,N,S,T,Y	None	9
NVT	Charged, hydrophilic	C,D,G,H,N,P,R,S,T,Y	None	12
VVC	Hydrophilic	A,D,G,H,N,P,R,S,T	None	9
NTT	Hydrophobic	F,I,L,V	None	4
TDK	Hydrophobic	C,F,L,W,Y	TAG	6
TTN	Hydrophobic	F,L	None	4
(DSC/DST/DSY)	Small	A,C,G,S,T	None	5
GMT	Single-mutation alanine scanning	A,D	None	2
GMA		A,E	None	2
GST		A,G	None	2
SCA		A,P	None	2
KCC		A,S	None	2
RCT		A,T	None	2
GYT		A,V	None	2

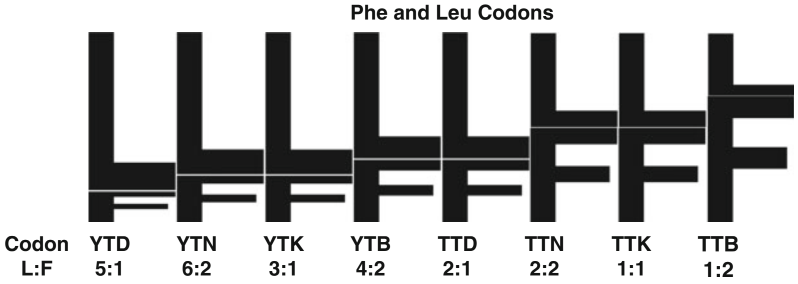


Fig. 8 Visualization of amino acid bias for codons exclusively encoding for L and F. Codon optimization must be performed to discriminate between codons encoding for similar amino acid ratios (i.e., 2:4 and 1:2)

site, but limit the mutations to a set of hypotheses. The foremost factor when considering a degenerate codon is the resulting set of amino acids. A secondary factor to consider is bias. For example, Fig. 8 illustrates how the set of amino acids containing exclusively Phe and Leu can be encoded by eight degenerate codons, with varying bias. Only the degenerate codons TTK and TTN encode Phe and Leu in equal proportions.

4.2 Approach

Due to the problems associated with site saturation mutagenesis, semi-rational methods have been developed for “intelligent” picking of codons to optimize either library size or amino acid ratios [72, 73]. We present a method below that uses an interactive python script (*codons.py*) for selecting site-specific degenerate codons. *Codons.py* allows a user to consider how alternate degenerate codons will drive the distribution of amino acids at a particular site, and to consider the library size and screening requirements that result from degenerate codons at multiple sites. To focus the discussion we will consider an illustrative example consisting of the cofactor switch of NADPH to NADH in ketol-acid reductoisomerase (KARI) [68, 74].

4.3 Phase I: Identification of Site Mutations

We will assume the availability of a structural model. The identification of specific target residues (e.g., active site, cofactor-binding site) can be accomplished via visual inspection in PyMOL [75] or via computational algorithms [76, 77]. Once the positions are selected, continued visual inspection will inform the decision of whether to use site saturation mutagenesis or a more limited subset of amino acids. Just as with site selection, the amino acid design palette can be based on calculations [78] or through biophysical intuition alone. The PyMOL mutation tool is an excellent prospective modeling technique for inspecting candidate mutations, since it allows rotamer sampling and indicates steric clashes. Thus visual inspection of candidate mutations may elucidate amino acids that are too large for the site or cannot avoid a detrimental interaction with existing amino acids/cofactors. Alternately, the scan of potential mutations and the conformations thereof can be automated.

Regardless, a list of favorable and unfavorable amino acids should be tabulated prior to use of *codons.py*. Under most circumstances, it is highly recommended that the wild-type amino acid be included in the design palette. One benefit is practical: including the wild-type amino acid will increase the fraction of the library that retains structure and function. Another benefit is philosophical: if the wild-type amino acid is an option for each design position, then the wild-type sequence should be a member of the library. With sufficient screening, such a library should yield this positive control.

In our cofactor switch example, the structure of the IlvC *E. coli* (without cofactor) was aligned to that of KARI spinach bound with NADPH. Five residues were identified for mutagenesis through proximity to the homologous NADPH position: R68, A71, R76, S78, and Q110. Residues R68, A71, R76, and S78 were selected due to their interactions with the 2′ phosphate group of the bound NADPH. Q110 was selected for its potential to orient the cofactor through interaction with the adenine moiety.

A strategy for shifting specificity from NADPH to NADH is to disrupt the salt bridge between positively charged residues interacting with the NADPH 2′ phosphate by mutations to negatively charged aspartic or glutamic acids [79]. Figure 9 illustrates how unfavorable interactions with NADPH could be formed via mutagenesis of each of the identified residues to Asp. Introducing negatively charged side chains can lower NADPH affinity and is sometimes sufficient to switch the cofactor specificity in favor of NADH [80]. However, improved specificity for NADH is often accompanied by loss of activity. As seen in Fig. 9, mutating S78 not only disrupts the salt bridge of NADPH 2′ phosphate, but it also might create a favorable hydrogen bond with the NADH hydroxyl group. We will use sites R68, R76, and S78 as examples in the degenerate codon design protocol (Table 5).

4.4 Phase II: Codons. py

Codons.py is a user-friendly tool for interactively selecting degenerate codons. The primary function of *codons.py* is to rank all prospective codons according to user-provided design goals. A set of required, taboo, preferred ("good"), and penalized ("bad") amino acids are provided either as arguments or interactive inputs to the main function. Subsequently, simple scoring methods rank the candidate codons that best fulfill the design objectives. Predefined amino acid sets can easily be specified and incorporated into the code. Aliphatic, hydrophobic, hydrophilic, acidic, and basic amino acid sets are predefined and can be input in place of individual amino acids. As outlined in detail below, scoring function options include the number of preferred amino acids encoded by a codon, the number of unique preferred amino acids encoded by a codon, and percentage of preferred amino acids out of amino acids in distribution encoded by a codon. Despite the simplicity of the scoring functions, the results nonetheless facilitate the sifting of many codon possibilities.

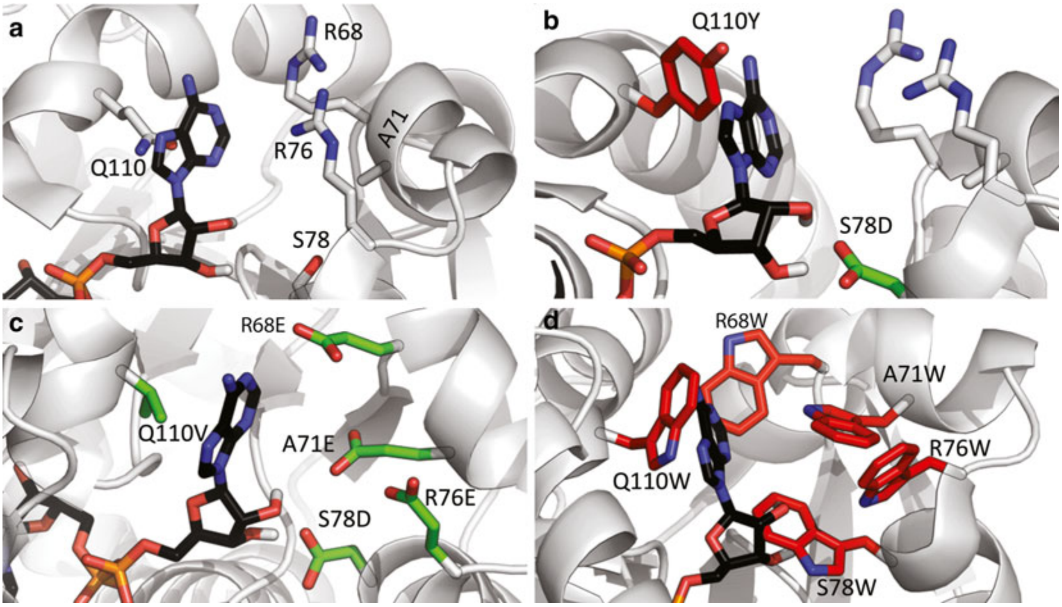


Fig. 9 Visualization of structural alignment between *E. coli* IlvC and Spinach KARI with NADPH bound in the active site. (**a**) Identification of R68, A71, R76, S78, and Q110 as potential residues for mutation in IlvC. (**b**) Depiction of favorable interaction created by S78D mutation and steric clashes created by Q110Y mutation. (**c**) Depiction of potential favorable mutations. (**d**) Depiction of unfavorable interactions from mutations to large residues

The scoring functions could be easily adapted if a more sophisticated scoring scheme is desired. The method for running *codons.py* is as follows:

1. Run *codons.py* interactively by entering the following into the command line: python codons.py (if manual usage is desired, enter python –i codons.py manual; *see* **Note 4** for examples of manual input).
2. The program will interactively ask for arguments (the help screen can be accessed at anytime by entering “?”):
   (a) Enter required amino acids.
   Set of amino acids that *must* be encoded. The wild-type amino acid is highly recommended for this set.
   (b) Enter good amino acids.
   Set of amino acids that give a positive score if encoded.
   (c) Enter bad amino acids.
   Set of amino acids that give a negative score if encoded.
   (d) Enter taboo amino acids.
   Set of amino acids that are not allowed to be encoded. For example, stop codons (denoted by an underscore) are typically designated as taboo.

Table 5
Hypotheses for NADPH cofactor switch example

Target Residue	Required	Preferred	Rationale
R68	R	E,D	Unfavorable interaction with 2′ phosphate group NADPH, potential hydrogen bonding with NADH
A71	A	E,D	Unfavorable interaction with 2′ phosphate group NADPH, potential hydrogen bonding with NADH
R76	R	E,D	Unfavorable interaction with 2′ phosphate group NADPH, potential hydrogen bonding with NADH
S78	S	E,D	Unfavorable interaction with 2′ phosphate group NADPH, potential hydrogen bonding with NADH
Q110	Q	–	No clear preference. Q110 mainly provides steric interaction
Target residue	**Taboo**	**Disfavored**	**Rationale**
R68	Stop	H,K F,W,Y	Favorable interaction with 2′ phosphate on NADPH Size
A71	Stop	P,G,S,T F,W,Y	Disfavored in alpha helix Size
R76	Stop	H,K F,W,Y	Favorable interaction with 2′ phosphate on NADPH Size
S78	Stop	H,K F,W,Y	Favorable interaction with 2′ phosphate on NADPH Size
Q110	Stop	P,G,S,T F,W,Y	Disfavored in alpha helix Size

(e) Enter desired scoring function.

Specify how to score the codons. The default scoring scheme is "distribution."

- *Set*: In this mode, candidate degenerate codons will be assessed using the unique set of encoded amino acids. Scoring is performed by adding 1 if the amino acid is in the preferred set and subtracting 1 if the amino acid is in the bad set. A penalty of −1,000 is included if the amino acid is taboo or if a required amino acid is not encoded by the codon.
- *Distribution*: In this mode, candidate degenerate codons will be assessed using the distribution of amino acids encoded by each codon rather than just the set of unique amino acids. Each amino acid in the codon outcome distribution is scored. Scoring is performed by adding 1 if the amino acid is in the preferred set and subtracting 1 if the amino acid is in the bad set. If the codon includes a taboo amino acid or lacks a required amino acid, the score decreases by 1,000.

- *Percent*: In this mode, candidate degenerate codons will be assessed by scoring the percentage of the outcome amino acids (including the distribution bias) that appear in the "good" set. If required amino acids are not included in the distribution or if taboo amino acids are included a penalty of −1,000 is added.

(f) Specify output cutoff (integer).
Option that only prints codons that score above a value.

(g) Specify the output file name.
Option that prints output to specified file name.

3. Following user input, a table of the ten highest scoring codons will be displayed. If more results are desired, answer "y" to the prompt and type in the desired number of results.
4. After analysis of the table, the user is prompted to select a degenerate codon. Guidelines for selecting degenerate codons are presented in Phase III below.
5. Once a codon is selected for the site, the program asks if another site is desired. If selection of a degenerate codon for another site is desired, answer "y" and **steps 1–4** will be repeated.
6. As the user selects degenerate codons for multiple sites, a multi-site library is defined. Key parameters for a multi-site library include the number of unique variants and the bias in the amino acid distributions at the design positions. The screening (number of random clones) necessary to experimentally observe most of the library (e.g., 95 %) can be estimated using random sampling with the function library_sampling defined within *codons.py*.

Codons.py was run for each mutation site identified in Phase I using hypotheses discussed in Table 5. Sample output from running codons.py for site A71 is represented in Table 6.

4.5 Phase III: Codon Selection

Although the script *codons.py* is interactive, the final selection of a particular codon is manual. On the first attempt at selecting a codon, the ranked candidates should be inspected to determine the frequency of preferred amino acids to non-preferred amino acids (*see* **Note 5**). If the preferred amino acids do not appear frequently enough in the codons, consider rerunning *codons.py* with the preferred amino acid in the required list. The opposite is true as well; if a "bad" amino acid is appearing at too high of a frequency, consider moving that amino acid to the taboo list. Another key aspect of the interactive codon selection is the process of refining the design criteria in the light of the candidate codons. Typically, the candidate list will include codons that result in larger and smaller sets of amino acids, leading naturally to questions of screening capacity. Also, by considering the list of candidates, other tradeoffs are likely to surface. Potentially, one might be selecting between a panel of amino acids that includes all of the desired

Table 6
Sample *codons.py* output for NADPH cofactor switch example

Score	Amino acid distribution	Codons
4	AAAADDEEVVVV	GHN
4	AAAADDEE	GMN
3	AAADEE	GMD/GMV
3	AAADDEVVV	GHB/GHH
3	AAADEEVVV	GHD/GHV
3	AAADDE	GMB/GMH
2	AAEEVV	GHR
2	AAAADDEEKKNNTTTT	RMN
2	AADD	GMY
2	AAADDEKNNTTT	RMB/RMH
2	AADDVV	GHY
2	AADDIINNTTVV	RHY
2	AADEVV	GHK/GHM/GHS/GHW
2	AAADDEIIIKNNTTTVVV	RHH
2	AADDNNTT	RMY
2	AAAADDEEIIIKKMNNTTTTVVVV	RHN
2	AAEE	GMR
2	AADE	GMK/GMM/GMS/GMW
2	AAADDEIIKMNNTTTVVV	RHB
1	ADNT	RMC/RMT

amino acids but also includes an amino acid that is likely to be incompatible with the protein conformation. The user must decide if that codon is preferable to an alternative that avoids the destabilizing option but covers fewer of the favored amino acids. At this stage, it is worth reconsidering how the amino acids that appear in favored codons, but were neither assigned as "good" or "bad," are likely to perform. We suggest evaluating amino acids interactively in PyMOL using the Mutagenesis wizard with the following checklist in mind:

1) Does the mutation clash with the protein backbone?
2) Does the mutation clash with existing side chains?
   (a) If there is a clash with a neighboring side chain, can the neighbor move?

3) Does the mutation clash with an existing water molecule?
   (a) Can the water molecule be displaced without the loss of favorable interactions?
4) Does the mutation clash with a bound substrate?
5) If favorable interaction with a bound substrate is a design criterion, can a candidate mutation make favorable interaction(s) considering size and hydrogen bonding geometry?
6) If an unfavorable interaction with a bound substrate is a design criteria, can a candidate mutation avoid making the unfavorable interactions?

After running *codons.py* for each mutation site, a list of optimum codons was identified (Table 7). Given the availability of a high-throughput screen to determine NADPH/NADH binding (fluorescence of NADPH/NADH) [68], codons encoding high diversity at sites R68 and R76 were allowed. While A71 can accommodate many mutations, the palette was restricted to favor diversity at the neighboring design positions. As a result, codons encoding only the hypothesized residues and the WT were selected. Considering the strong preference for the S78D mutation, D was included in the required set for *codons.py*. The "percent" scoring function was used to identify codons that provided the highest percent coverage of D and E in the resulting amino acid distributions. Finally, due to lack of clear hypotheses for site Q110, only

Table 7
Favorable codons for NADPH cofactor switch example

Site	Candidate codons	Amino acid distribution
R68/R76	VDN	DDEEGGGGHHIIIKKLLLLMNNQQRRRRRRSSVVVV
	RRN	DDEEGGGGKKNNRRSS
	RNN	AAAADDEEGGGGIIIKKMNNRRSSTTTTVVVV
A71	GMN	AAAADDEE
	GMK/GMM/GMS/GMW	AADE
	GMD/GMV	AAADEE
S76	RNN	AAAADDEEGGGGIIIKKMNNRRSSTTTTVVVV
	RRK,RRM,RRS,RRW	DEGGKNRS
	RRC,RRT	DGNS
Q110	VWN	DDEEHHIIIKKLLLLMNNQQVVVV
	VWH	DDEHHIIIKLLLNNQVVV
	VWR	EEIKKLLMQQVV

Table 8
Final codon selection for NADPH cofactor switch example

Site	Final codon	Distribution	Rationale
R68/R76	RNN	AAAADDEEGGGGIIIK KMNNRRSSTTTTVVVV	(1) Introduction of diversity to these sites with good representation of preferred mutations (11 % frequency). (2) No large amino acids included in set. (3) Small frequency of bad amino acids
A71	GMK,GMM, GMS,GMW	AADE	(1) Lowest A:D:E ratio that encodes exclusively for A,D,E. (2) Limited diversity at this site is not unfavorable due to high diversity at other sites. (3) WT contributes to 50 % of encoded distribution—potentially helpful due to high diversity at other sites which might require A71 to avoid steric clashes
S78	RRK,RRM, RRS,RRW	DEGGKNRS	Favorable interaction with hydroxyl group appears at a high frequency (25 %)
Q110	VWN	DDEEHHIIIKKLLLLMN NQQVVVV	(1) Good diversity of smaller amino acids. (2) No large amino acids included in set

disfavored amino acids were specified. Codons at Q110 were thus ranked highly if they encoded high diversity and excluded large residues.

From the selection of the top candidates, final codons were selected as presented in Table 8. Depending on the degenerate codon candidates, codon optimization for the selected expression system (e.g., avoiding rare codons) could help discriminate between candidates that result in different distributions of the same amino acids (e.g., AAAALL and AAL) [81].

4.6 Phase IV: Experimental Synthesis

Commercial oligonucleotide providers (e.g., Integrated DNA Technologies, IDT) can synthesize primers with a mixture of wild-type and non-wild-type nucleotides. If only a single-mutation site or multiple-mutation sites in close proximity are desired, introduction of a degenerate codon can be accomplished with a single PCR [82]. However, if the desired sites are distant from one another, more extensive protocols must be used [70]. If the desired distribution of amino acids is not possible by a degenerate codon, mixing oligonucleotides is an alternative option [78].

5 Notes

1. Numerous programs exist for estimating folding free energy change, including FoldX, I-Mutant2.0, Eris, and sMMGB [83–86]. We chose the commonly used, semiempirical FoldX

for CBHII calculations. Estimating folding free energy changes for 20 amino acids at 358 sites created a computationally intensive calculation. Computing all FoldX calculations for six different backbones took approximately 3 days on a 2.6 GHz CPU.

2. Traditional CPD relies on fixed-backbone combinatorial optimization of side chain positions and amino acid identity. However, small differences in the backbone position can make a large difference in the ability of amino acids to be favorably placed at a given design position. Using an ensemble of reasonable backbone models provides a more realistic approximation of the protein backbone flexibility. This strategy is a partial substitute for true flexible-backbone design algorithms [24, 87].

3. In inspecting the designs, we checked for the loss of hydrogen bonds or the addition of questionable nonpolar surface mutations. Detailed pairwise energy comparisons (*run_evaluate_mutations.py*), combined with visual inspection, constituted the additional analysis of each proposed mutation. If we could not identify the rationale for a mutation chosen by SHARPEN, we performed a secondary search with additional rotamers near the questionable residue (*run_questionable_mutations.py*). All rotamers with chi angles within two standard deviations of the default Dunbrack rotamer library angles were included. Mutations that were still considered favorable in this secondary search were included in the final design. Otherwise, we used *run_mutate_to_wt.py* to revert mutations back to the wild-type amino acid variant.

4. Manual input of arguments in codons.py can be accessed by typing the following into the command line: *python –i codons.py manual*. Manual input allows quick and easy iterations for experienced users. Example inputs are given below for the identification of small replacements for leucine:

   pickcodons(good='AGVLIST', bad='WYFHRKED', taboo='_', required='L', scoring='percent', outfile=codons.txt')

   librarysize=compute_library_size('stringofcodons')

5. As a general rule, it is best to start with soft constraints on required and taboo mutations (i.e., only include WT in required and "_" in taboo). After evaluation of results, if a suitable distribution is not located, or visual inspection merits further discrimination for or against certain amino acids, then mutations may be moved into the required or taboo categories.

6. If a desired amino acid distribution can be encoded by multiple codons, codon optimization can be performed to discriminate between codons that encode for similar ratios.

References

1. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10:866–876
2. Tracewell CA, Arnold FH (2009) Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr Opin Chem Biol 13:3–9
3. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364
4. Miklos AE, Kluwe C, Der BS, Pai S, Sircar A, Hughes RA et al (2012) Structure-based design of supercharged, highly thermoresistant antibodies. Chem Biol 19:449–455
5. Grigoryan G, Reinke AW, Keating AE (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature 458: 859–864
6. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J et al (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190–195
7. Privett HK, Kiss G, Lee TM, Blomberg R, Chica RA, Thomas LM et al (2012) Iterative approach to computational enzyme design. Proc Natl Acad Sci U S A 109:3790–3795
8. Dantas G, Kuhlman B, Callender D, Wong M, Baker D (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol 332:449–460
9. Chica RA, Doucet N, Pelletier JN (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr Opin Biotechnol 16:378–384
10. Karanicolas J, Com JE, Chen I, Joachmiak LA, Dym O, Peck SH et al (2011) A de novo protein binding pair by computational design and directed evolution. Mol Cell 42:250–260
11. Khersonsky O, Kiss G, Röthlisberger D, Dym O, Albeck S, Houk KN et al (2012) Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. Proc Natl Acad Sci U S A 109:10358–10363
12. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. Proc Natl Acad Sci U S A 103:5869–5874
13. Gromiha MM (2007) Prediction of protein stability upon point mutations. Biochem Soc Trans 35:1569–1573
14. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262:1680
15. Bradley LH, Thumfort PP, Hecht MH (2006) De novo proteins from binary-patterned combinatorial libraries. Methods Mol Biol 340: 53–69
16. Bradley LH, Wei Y, Thumfort P, Wurth C, Hecht MH (2007) Protein design by binary patterning of polar and nonpolar amino acids. Methods Mol Biol 352:155–166
17. Pantazes RJ, Saraf MC, Maranas CD (2007) Optimal protein library design using recombination or point mutations based on sequence-based scoring functions. Protein Eng Des Sel 20:361–373
18. Steipe B, Schiller B, Plückthun A, Steinbacher S (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. J Mol Biol 240:188–192
19. Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L et al (2000) From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. Protein Eng 13:49–57
20. Amin N, Liu A, Ramer S, Aehle W, Meijer D, Metin M et al (2004) Construction of stabilized proteins by combinatorial consensus mutagenesis. Protein Eng Des Sel 17:787
21. Kono H, Wang W, Saven JG (2007) Combinatorial protein design strategies using computational methods. Methods Mol Biol 352:3–22
22. Dunbrack RL Jr (2002) Rotamer libraries in the 21st century. Curr Opin Struct Biol 12: 431–440
23. Shetty RP, De Bakker PIW, DePristo MA, Blundell TL (2003) Advantages of fine-grained side chain conformer libraries. Protein Eng 16:963–969
24. Hallen MA, Keedy DA, Donald BR (2012) Dead-end elimination with perturbations (DEEPer): a provable protein design algorithm with continuous sidechain and backbone flexibility. Proteins 81:18–39
25. Mena MA, Treynor TP, Mayo SL, Daugherty PS (2006) Blue fluorescent proteins with enhanced brightness and photostability from a structurally targeted library. Nat Biotechnol 24:1569–1571
26. Allen BD, Nisthal A, Mayo SL (2010) Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. Proc Natl Acad Sci U S A 107:19838–19843

27. Chica RA, Moore MM, Allen BD, Mayo SL (2010) Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. Proc Natl Acad Sci U S A 107:20257–20262

28. Drummond DA, Silberg JJ, Meyer MM, Wilke CO, Arnold FH (2005) On the conservative nature of intragenic recombination. Proc Natl Acad Sci U S A 102:5380

29. Stemmer WPC (1994) Rapid evolution of a protein in vitro by DNA shuffling. Nature 370:389–391

30. Harayama S (1998) Artificial evolution by DNA shuffling. Trends Biotechnol 16:76–82

31. Meyer MM, Silberg JJ, Voigt CA, Endelman JB, Mayo SL, Wang ZG et al (2003) Library analysis of SCHEMA-guided protein recombination. Protein Sci 12:1686–1693

32. Otey CR, Landwehr M, Endelman JB, Hiraga K, Bloom JD, Arnold FH (2006) Structure-guided recombination creates an artificial family of cytochromes P450. PLoS Biol 4:e112

33. Romero PA, Stone E, Lamb C, Chantranupong L, Krause A, Miklos AE (2012) SCHEMA designed variants of human arginase I & Ii reveal sequence elements important to stability and catalysis. ACS Synth Biol 1:221–228

34. Heinzelman P, Snow CD, Wu I, Nguyen C, Villalobos A, Govindarajan S et al (2009) A family of thermostable fungal cellulases created by structure-guided recombination. Proc Natl Acad Sci U S A 106:5610–5615

35. Heinzelman P, Komor R, Kanaan A, Romero P, Yu X, Mohler S et al (2010) Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. Protein Eng Des Sel 23:871–880

36. Komor RS, Romero PA, Xie CB, Arnold FH (2012) Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. Protein Eng Des Sel 25:827–833

37. Smith MA, Rentmeister A, Snow CD, Wu T, Farrow MF, Mingardon F et al (2012) A diverse set of family 48 bacterial cellulases created by structure-guided recombination. FEBS J 279:4453–4465

38. Hiraga K, Arnold FH (2003) General method for sequence-independent site-directed chimeragenesis. J Mol Biol 330:287–296

39. Farrow MF, Arnold FH (2010) Combinatorial recombination of gene fragments to construct a library of chimeras. Curr Protoc Protein Sci Chapter 26, Unit 26.2

40. Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. Science 278:82–87

41. Desmet J, Spriet J, Lasters I (2002) Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. Proteins 48:31–43

42. Jacak R, Leaver-Fay A, Kuhlman B (2012) Computational protein design with explicit consideration of surface hydrophobic patches. Proteins 80:825–838

43. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

44. Meyer MM, Hiraga K, Arnold FH (2006) Combinatorial recombination of gene fragments to construct a library of chimeras. Curr Protoc Protein Sci Chapter 26, Unit 26.2

45. Romero PA, Arnold FH (2012) Random field model reveals structure of the protein recombinational landscape. PLoS Comput Biol 8:e1002713

46. Loksha IV, Maiolo JR 3rd, Hong CW, Ng A, Snow CD (2009) SHARPEN-systematic hierarchical algorithms for rotamers and proteins on an extended network. J Comput Chem 30:999–1005

47. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R et al (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545–574

48. Dunbrack RL Jr, Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6:1661–1681

49. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

50. Heinzelman P, Snow CD, Smith MA, Yu X, Kannan A, Boulware K et al (2009) SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. J Biol Chem 284:26229–26233

51. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH (2002) Protein building blocks preserved by recombination. Nat Struct Mol Biol 9:553–558

52. Moore GL, Maranas CD (2003) Identifying residue–residue clashes in protein hybrids by using a second-order mean-field approach. Proc Natl Acad Sci U S A 100:5091

53. Saraf MC, Horswill AR, Benkovic SJ, Maranas CD (2004) FamClash: a method for ranking the activity of engineered enzymes. Proc Natl Acad Sci U S A 101:4142

54. Endelman JB, Silberg JJ, Wang ZG, Arnold FH (2004) Site-directed protein recombination as a shortest-path problem. Protein Eng Des Sel 17:589–594

55. Silberg JJ, Endelman JB, Arnold FH (2004) SCHEMA-guided protein recombination. Methods Enzymol 388:35–42

56. Ingolfsson H, Yona G (2008) Protein domain prediction. Methods Mol Biol 426:117–143

57. Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, Arnold FH (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat Biotechnol 25:1051–1056

58. Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc Natl Acad Sci U S A 93:13–20

59. Grosdidier S, Fernández-Recio J (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics 9:447

60. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM et al (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332:816–821

61. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C et al (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol 30: 543–548

62. Kortemme T, Baker D (2004) Computational design of protein–protein interactions. Curr Opin Chem Biol 8:91–97

63. Salgado EN, Radford RJ, Tezcan FA (2010) Metal-directed protein self-assembly. Acc Chem Res 43:661–672

64. Ballister ER, Lai AH, Zuckermann RN, Cheng Y, Mougous JD (2008) In vitro self-assembly of tailorable nanotubes from a simple protein building block. Proc Natl Acad Sci U S A 105:3733–3738

65. Lawrence MS, Phillips KJ, Liu DR (2007) Supercharging proteins can impart unusual resilience. J Am Chem Soc 129: 10110–10112

66. Das A, Wei Y, Pelczer I, Hecht MH (2011) Binding of small molecules to cavity forming mutants of a de novo designed protein. Protein Sci 20:702–711

67. Liu L, Baase WA, Michael MM, Matthews BW (2009) Use of stabilizing mutations to engineer a charged group within a ligand-binding hydrophobic cavity in T4 lysozyme. Biochemistry 48:8842–8851

68. Bastian S, Liu X, Meyerowitz JT, Snow CD, Chen MM, Arnold FH (2011) Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in *Escherichia coli*. Metab Eng 13:345–352

69. Tang L, Gao H, Zhu X, Wang X, Zhou M, Jiang R (2012) Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. Biotechniques 52:149–158

70. Georgescu R, Bandara G, Sun L (2003) Saturation mutagenesis. Methods Mol Biol 231:75–83

71. Denault M, Pelletier JN (2007) Protein library design and screening: working out the probabilities. Methods Mol Biol 352:127–154

72. Mena MA, Daugherty PS (2005) Automated design of degenerate codon libraries. Protein Eng Des Sel 18:559–561

73. Patrick WM, Firth AE (2005) Strategies and computational tools for improving randomized protein libraries. Biomol Eng 22:105–112

74. Bastian S, Arnold FH (2012) Reversal of NAD(P)H cofactor dependence by protein engineering. Methods Mol Biol 834:17–31

75. Schrödinger L (2010) The PyMOL molecular graphics system, version 1.3r1

76. Fischer JD, Mayer CE, Söding J (2008) Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24:613–620

77. Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjölander K (2010) Active site prediction using evolutionary and structural information. Bioinformatics 26:617–624

78. Chen MMY, Snow CD, Vizcarra CL, Mayo SL, Arnold FH (2012) Comparison of random mutagenesis and semi-rational designed libraries for improved cytochrome P450 BM3-catalyzed hydroxylation of small alkanes. Protein Eng Des Sel 25:171–178

79. Scrutton NS, Berry A, Perham RN (1990) Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. Nature 343: 38–43

80. Rane MJ, Calvo KC (1997) Reversal of the nucleotide specificity of ketol acid reductoisomerase by site-directed mutagenesis identifies the NADPH binding site. Arch Biochem Biophys 338:83–89

81. Fuglsang A (2003) Codon optimizer: a freeware tool for codon optimization. Protein Expr Purif 31:247–249

82. Chiang LW, Kovari I, Howe MM (1993) Mutagenic oligonucleotide-directed PCR amplification (Mod-PCR): an efficient method for generating random base substitution mutations in a DNA sequence element. PCR Methods Appl 2:210–217

83. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320:369–387

84. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33:W306–W310

85. Yin S, Ding F, Dokholyan NV (2010) Computational evaluation of protein stability change upon mutations. Methods Mol Biol 634:189–201

86. Zhang Z, Wang L, Gao Y, Zhang J, Zhenirovskyy M, Alexov E (2012) Predicting folding free energy changes upon single point mutations. Bioinformatics 28: 664–671

87. Mandell DJ, Kortemme T (2009) Backbone flexibility in computational protein design. Curr Opin Biotechnol 20:420–428

Chapter 8

Symmetric Protein Architecture in Protein Design: Top-Down Symmetric Deconstruction

Liam M. Longo and Michael Blaber

Abstract

Top-down symmetric deconstruction (TDSD) is a joint experimental and computational approach to generate a highly stable, functionally benign protein scaffold for intended application in subsequent functional design studies. By focusing on symmetric protein folds, TDSD can leverage the dramatic reduction in sequence space achieved by applying a primary structure symmetric constraint to the design process. Fundamentally, TDSD is an iterative symmetrization process, in which the goal is to maintain or improve properties of thermodynamic stability and folding cooperativity inherent to a starting sequence (the "proxy"). As such, TDSD does not attempt to solve the inverse protein folding problem directly, which is computationally intractable. The present chapter will take the reader through all of the primary steps of TDSD—selecting a proxy, identifying potential mutations, establishing a stability/folding cooperativity screen—relying heavily on a successful TDSD solution for the common β-trefoil fold.

Key words Symmetric protein design, Protein folding, Protein engineering, Phi-value analysis, β-trefoil, Protein evolution

1 Introduction

Protein design holds a vast, unexploited potential to revolutionize a number of fields, comparable to the great scientific and economic successes of synthetic organic chemistry [1]. However, before the full power of protein design can be leveraged, a solution to the "inverse folding problem" (i.e., how to design an amino acid sequence that will fold into a desired architecture [2]) must be developed. As of yet, it is not possible to routinely identify amino acid sequences that will fold into a predefined architecture, although computational approaches are making significant strides [3]. Thus, modern-day protein design efforts must rely on synergy between both computational and experimental approaches and validation. Despite the tremendous body of literature regarding protein folding and stability, all but the simplest protein design goals represent a massive undertaking.

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_8, © Springer Science+Business Media New York 2014

Top-down symmetric deconstruction (TDSD) is an experimental and computational approach to protein design that generates amino acid sequences that fold into predefined architectures and have biophysical properties favorable for subsequent functional design [4, 5]. The TDSD approach is motivated by three key observations:

First, functional engineering studies would be facilitated by the availability of highly stable, well-behaved protein scaffolds to serve as starting molecules [6]. Rather than attempt to engineer functional residues, specific biophysical properties, and fold-specific interactions simultaneously, the protein scaffold approach posits that the design cycle can be split into two independent steps: scaffold preparation and functional engineering. In this view, there is tremendous utility to generating a "toolkit" of functionally benign protein scaffolds that can later be used for functional engineering studies by other researchers in a community-wide effort.

Second, symmetric protein folds, which encompass approximately one-third of all protein architectures, represent a notable evolutionary solution to the protein folding problem, and are therefore uniquely poised to have broad utility as protein scaffolds [7]. Symmetric proteins satisfy a huge diversity of roles within living systems, ranging from catalysis of redox reactions to protease inhibition, thereby demonstrating an inherent functional flexibility. Thus, by focusing on symmetric protein folds as protein scaffolds, broad functional potential appears feasible. Furthermore, limited evidence suggests that the fundamental symmetric folds are capable of profound thermostability [5].

Third, design of symmetric protein folds can be dramatically simplified by application of a primary structure symmetric constraint in which the sequence of each subdomain is made to be identical. To illustrate this point, consider that the number of possible amino acid sequences for a relatively small 90 amino acid protein is 20^{90}, or about 10^{117}, possible sequences—a staggeringly vast sequence space! Upon application of a three-fold symmetric constraint, the number of possible sequences plummets to 20^{30}, or about 10^{39}, possible sequences. That is, for the case of a 90 amino acid protein, application of a threefold symmetric constraint results in a ~10^{78}-fold reduction in number of possible sequences. Such reductions in complexity may help move the protein design problem from computationally intractable to computationally feasible, and highlight the utility of symmetric protein folds to protein design.

Given the above, the principle goal of TDSD can now be stated more formally: *to generate symmetric protein scaffolds characterized by primary structure symmetry and favorable biophysical properties (high stability and solubility, rapid folding) for future functional engineering studies*. Figure 1 illustrates the key elements of TDSD.

Before proceeding, it should be noted that sequence symmetry was once considered detrimental to protein folding [8–11]. This conclusion, however, was largely on the basis of incomplete and ambiguous results; as such, recent studies directly probing the

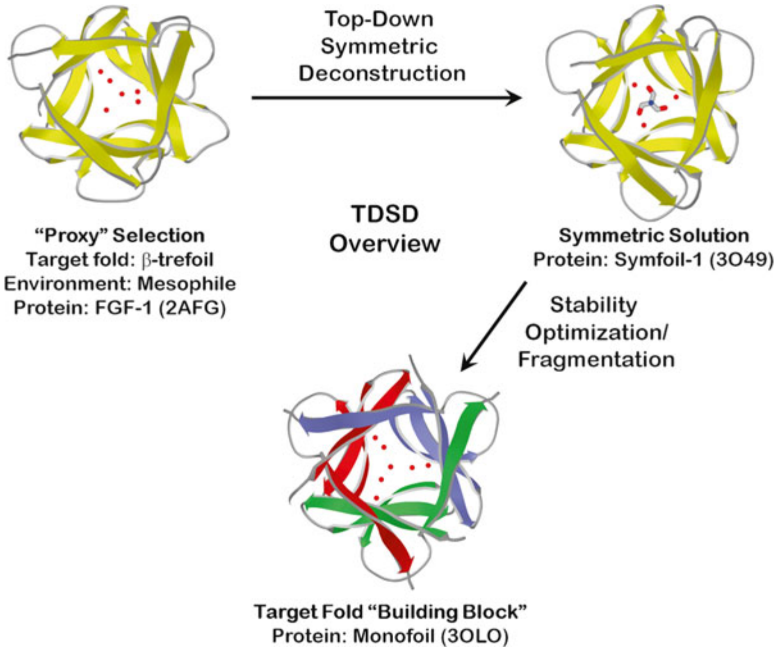


Fig. 1 Overview of TDSD. The principal goals of TDSD are to prepare a fully symmetrized protein scaffold (the "symmetric solution") and to identify a short peptide (~30–50 residues) that can oligomerize to recapitulate a given symmetric architecture (the "peptide building block")

interaction between protein folding and sequence symmetry within the context of a cooperatively folding, single-domain protein (such as a protein scaffold) unambiguously show that sequence symmetry is entirely compatible with—if not supportive of—fast folding, profound thermostability, and high solubility (i.e., ideal properties for a protein scaffold) [5, 12–14].

Materials for TDSD exploit both computational and experimental resources. Furthermore, different model systems or experimental goals may require different materials. Generally speaking, TDSD requires resources associated with the following techniques:

1. Protein expression.
2. Protein purification.
3. DNA mutagenesis and sequencing.
4. Protein stability measurement (circular dichroism polarimeter, fluorescence spectrophotometer, differential scanning calorimeter, refractometer to measure denaturant concentration).
5. High-resolution structural characterization (X-ray crystallography or solution-state NMR).
6. Computational resources (hardware and software) necessary for molecular modeling, data fitting, and prediction of mutant stability effects (i.e., Gibbs energy).

2 Methods

2.1 Top-Down Symmetric Deconstruction in a Nutshell

TDSD begins with the selection of a *proxy*, an efficiently folding and thermostable protein (typically naturally evolved) that adopts the general symmetric architecture of interest (i.e., the "target" architecture of the TDSD) and provides a foldable amino acid sequence to be subjected to iterative symmetrization (Fig. 1). Unlike de novo design approaches, which must directly solve the inverse folding problem, TDSD begins with an amino acid sequence that is known to fold into the correct architecture. Thus, while de novo design approaches must encode fold-specific interactions—such as an efficient folding pathway and a low energy native state, features that largely remain problematic with current computational approaches—TDSD begins with such characteristics intact by coopting them from a naturally evolved protein. Therefore, TDSD dictates that one need only *maintain* features of foldability and stability, thereby significantly reducing the complexity of the design problem and maximizing the probability of success. Finally, whereas failure of a de novo designed sequence to fold may be due to any number of potential problems (and therefore supremely difficult to identify), TDSD circumvents this weakness by employing a *high granularity* of the design cycle; that is, intermediate forms during symmetrization are experimentally characterized at regular intervals such that destabilizing mutations can be readily identified, excluded, and corrected.

Regardless of the architecture of the proxy, TDSD will proceed in the following manner, shown diagrammatically in Fig. 2, and discussed below:

1. Select a proxy for the desired target architecture.
2. Subject the proxy to iterative symmetrization of the core and backbone.
   (a) Identify mutation(s) that will increase the sequence symmetry of the proxy and are predicted to be permissible for stability and folding cooperativity.
   (b) Model the proposed mutation(s) into the available structural data and confirm its (their) compatibility/feasibility with available computational resources.
   (c) Experimentally characterize the stability and folding cooperativity of the proposed mutant protein. If the mutation is benign or stabilizing, and folding cooperativity is maintained, incorporate the mutation into the construct design; otherwise, reject the mutation.
   (d) Return to (2a) until both the protein core and 3° structure backbone are dramatically symmetrized

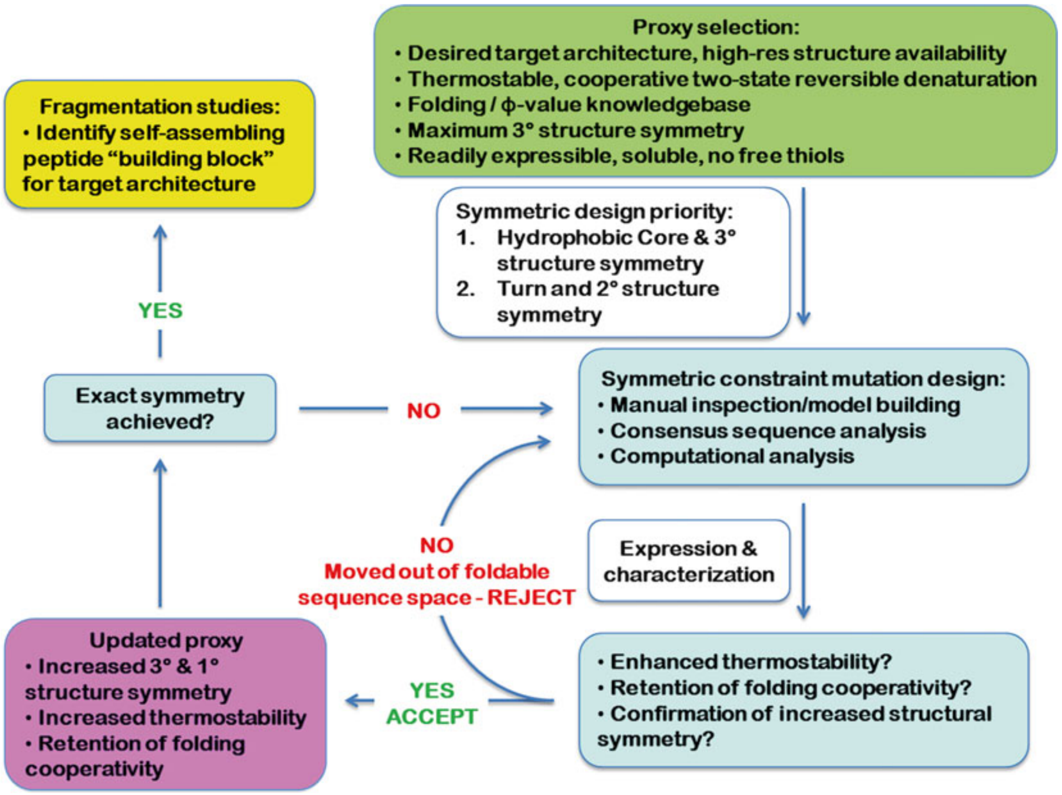
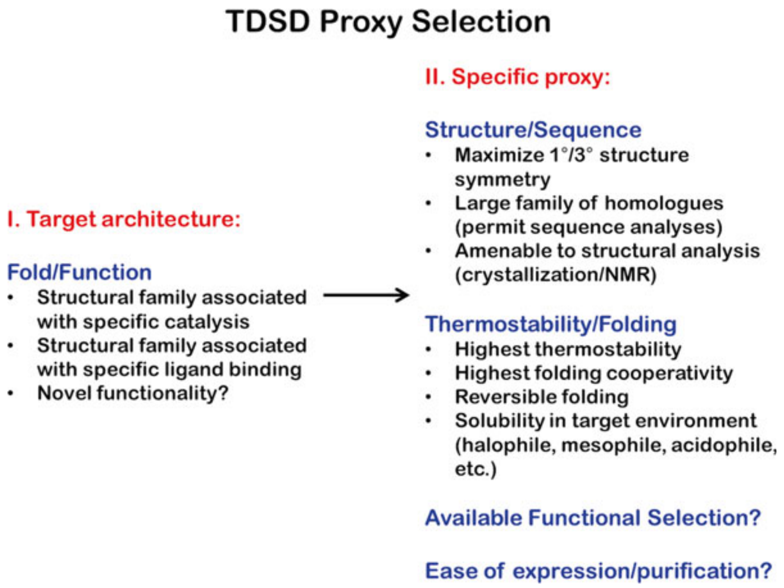

Proxy selection:
• Desired target architecture, high-res structure availability
• Thermostable, cooperative two-state reversible denaturation
• Folding / ϕ-value knowledgebase
• Maximum 3° structure symmetry
• Readily expressible, soluble, no free thiols
Fragmentation studies:
• Identify self-assembling peptide "building block" for target architecture
YES
Symmetric design priority:
1. Hydrophobic Core & 3° structure symmetry
2. Turn and 2° structure symmetry
Exact symmetry achieved?
NO
Symmetric constraint mutation design:
• Manual inspection/model building
• Consensus sequence analysis
• Computational analysis
NO
Moved out of foldable sequence space - REJECT
Expression & characterization
Updated proxy
• Increased 3° & 1° structure symmetry
• Increased thermostability
• Retention of folding cooperativity
YES
ACCEPT
• Enhanced thermostability?
• Retention of folding cooperativity?
• Confirmation of increased structural symmetry?

Fig. 2 Workflow of TDSD. TDSD is an iterative symmetrization approach guided by both computational and experimental validation. See text for a detailed description of each step

3. Once the core and backbone have been significantly symmetrized, either:
   - Proceed as in **step 2** to symmetrize the remaining positions via stepwise mutation until resulting in the *symmetric solution* (that is, full primary structure symmetry).
   - Attempt rapid design of the symmetric solution using a chimera approach.
4. Optimize the stability of the symmetric solution then undertake fragmentation to generate a *peptide building block*—that is, a single peptide typically on the order of 30–50 amino acids long that oligomerizes to recapitulate the target symmetric architecture.

Although many of the steps of TDSD are straightforward in principle, there are several considerations at each level that can greatly mitigate potential problems that may arise during symmetrization. Suggestions for each step of the TDSD approach are included below.



TDSD Proxy Selection

I. Target architecture:

Fold/Function

- Structural family associated with specific catalysis
- Structural family associated with specific ligand binding
- Novel functionality?

II. Specific proxy:

Structure/Sequence

- Maximize 1°/3° structure symmetry
- Large family of homologues (permit sequence analyses)
- Amenable to structural analysis (crystallization/NMR)

Thermostability/Folding

- Highest thermostability
- Highest folding cooperativity
- Reversible folding
- Solubility in target environment (halophile, mesophile, acidophile, etc.)

Available Functional Selection?

Ease of expression/purification?

Fig. 3 Notes on selecting a proxy. Judicious choice of the proxy can speed the TDSD by reducing the required granularity of the study

2.2 Selecting a Proxy

As described above, the *proxy* is the starting molecule, having the desired target architecture, to be symmetrized by TDSD. Although any single-domain globular protein that adopts a structurally symmetric architecture (e.g., β-trefoil, TIM barrel, and β-propeller) can serve as a proxy, some molecules will be more amenable to TSDS than others. In addition, the choice of proxy will influence the properties of the resulting symmetric protein building block. Included below are some general points to consider when selecting a proxy (Fig. 3).

2.2.1 Identify a Protein Architecture Known to Perform the Function of Interest

The principle goal of TDSD is to generate functionally benign protein “canvases” upon which novel functionalities can be “painted.” Thus, when selecting a proxy, *the choice of architecture should be guided primarily by the function of interest*. For illustration, consider the β-trefoil fold. β-trefoils are known to serve as lectins, protease inhibitors, toxins, and cytokines, but are not known to act as enzymes. Therefore, if the function of interest is to catalyze a redox reaction (a role commonly tasked to TIM barrels) selecting a β-trefoil as the proxy may be a poor choice (unless the reader wishes to explore the limits of β-trefoil functionality). Please note, however, that there is no need to consider the *specific function* intrinsic to a given proxy, as this function likely will be entirely lost during symmetrization (and, in fact, is part of the TDSD design goal).

2.2.2 Select a Proxy That Folds Under the Relevant Environmental Conditions

Protein sequences are tuned to operate under specific solvent/environmental conditions: Whereas proteins from mesophiles (organisms that live at moderate temperatures) may have poorly optimized surface electrostatic properties, the proteins of thermophiles (organisms that live at extremes of temperature) make extensive use of optimized surface-exposed salt bridges to gain additional stability [15, 16]. In contrast to both mesophiles and thermophiles, the proteins of halophiles (organisms that live in extremes of salt concentration) have overwhelmingly acidic residues decorating their surface, for improved solubility [17–19]. Thus, the choice of proxy and mutant stability and folding cooperativity screening conditions will influence the target environment under which the resulting symmetric protein building block will be viable. Put simply: If the goal is to prepare an acidophile, select a proxy from an acidophile proteome and screen for stability and folding cooperativity at acidic pH. TDSD is unlikely to identify a single sequence (i.e., peptide building block for the target architecture) for all possible environmental conditions of interest. Thus, TDSD is not a "single solution" experiment as regards each target architecture.

2.2.3 Select a Proxy for Which a High-Resolution Structure Has Been Solved

The first transformation of TDSD is to symmetrize residues that pack to form the hydrophobic core. To do this efficiently, it is absolutely necessary to have high-resolution structural data with which to probe for cavities and model potential mutations. In addition, it will likely be necessary to crystallize the protein or perform high-resolution solution-state NMR structural studies at regular intervals during TDSD to update the protein model; thus, if crystallization or solution-state NMR data collection of a given protein is notoriously difficult, it may be better to choose a proxy for which structural characterization is more tractable.

2.2.4 Preference Proxies for Which There is a Wealth of Folding Data

Functional regions of a protein are least likely to make significant contributions to folding and stability, in accordance with a stability–function [20, 21] and a foldability–function trade-off [22–25]. Thus, the presence of high quality folding data, such as phi-value analysis, can serve as an invaluable guide *to identify those regions of the primary structure that should be preferentially retained during the course of the deconstruction*. Phi-value analysis combines mutational studies of both stability and folding rates—using energy as a probe of structure formation in the transition state—to identify residues that make interactions that are key for protein foldability [26]. Although phi-value data are limited, current results suggest that approximately one-third to one-half of the primary structure of a protein is likely to contribute to the critical folding transition state [22, 27, 28]. Knowledge of the size of the folding nucleus for a given protein architecture can help tune the ultimate design goal by setting the approximate upper limit to the amount of fragmentation that a given architecture can withstand. Such an upper limit is likely a function of the size of a folding nucleus as well as the degree of symmetry (with higher symmetry having a greater entropic penalty for oligomerization). In the case of the β-trefoil, a building block of 42 amino acids was identified that contained the critical folding nucleus and successfully formed a stable homo-trimer oligomer. Although studies of the TIM barrel are ongoing [10, 29, 30], the criteria discussed above suggest that a quarter barrel might be able to spontaneously oligomerize, but that a stable octomer oligomer of a repeating motif seems less likely. Similarly, five-bladed β-propeller studies have identified a two-blade motif as a successful building block able to stably oligomerize to recapitulate the overall β-propeller fold [11, 31].

2.2.5 Proxies Maximizing Existing Structural Symmetry Are Preferred

At the sequence level, greater symmetry directly translates into fewer positions that need to be symmetrized by TDSD (i.e., able to reduce the necessary granularity of the design cycle). Note, however, that successful TDSD has been accomplished starting with a proxy having essentially random primary structure symmetry, as well as tertiary structure asymmetry [5]. Generally, symmetry at some positions is preferable to others: Solvent exposed sites and turns are, on balance, easier to symmetrize than are core positions; thus, when assessing the sequences of potential proxies, *give special consideration to core packing residues.* At the tertiary structure level, two general features should be considered: First, does a potential proxy contain a degenerate structural subdomain? If so, this proxy will almost certainly be harder to symmetrize, as the remaining subdomains will have likely expanded and/or adopted highly asymmetric packing arrangements to compensate for the degenerate subdomain. Second, is there significant backbone symmetry among the core packing groups? Again, the core packing residues are the hardest to symmetrize if repeating subdomains exhibit relative insertions or deletions. In short, *it is unlikely that an optimal symmetric core-packing solution can be identified if the tertiary structure is asymmetric* [13]. Thus, when choosing between potential proxies having differing extents of tertiary or primary structure symmetry, the former is preferred and more likely to enable achievement of the latter. Alternatively, the 3° structure must be initially made symmetric before attempting to enforce 1° structure symmetry.

2.2.6 Select a Proxy That Has High Expression Yields and an Established Purification Protocol

Throughout the course of TDSD, it will be necessary to express and purify intermediate forms for biophysical and structural characterization. Thus, proxies that can be expressed reasonably well in bacteria (that is, proteins that lack glycosylation, disulfide bonds, and do not require chaperons to assist folding) are strongly preferred.

2.2.7 Select a Proxy with Favorable Biophysical Properties, if Possible

Proxies with high stability and high folding cooperativity (discussed in greater detail below) are greatly preferred. Because stability influences a host of biophysical properties, more stable proteins are much easier to work with. In general, increasing protein stability is associated with reduced aggregation/superior solubility [32], which makes biophysical characterization and crystallography more tractable. In addition, proxies that start with greater stability will have an alleviated requirement for stabilization during each round of symmetrization and the goals of stabilization and increasing symmetry can be decoupled. Conversely, when working with mesophilic proteins, it is often necessary to simultaneously stabilize the protein while making symmetry-enhancing mutations. If possible, "two-state" protein folding is preferred, as the interpretation of stability and folding cooperativity for these proteins is greatly simplified [33].

2.2.8 Avoid Proteins with Free Cysteine Residues

Free cysteine residues (that is, cysteine residues that are not participating in a disulfide bond) are extremely disadvantageous to the protein chemist. Thiol-mediated chemistry is a major pathway for protein aggregation and deactivation, resulting in lower solubility, non-two-state folding/irreversible aggregation, and diminished functional half-life [34]. In order to manage the oxidation potential of the buffer, protein chemists routinely include reducing agents (e.g., DTT, βME) to prevent disulfide bond formation. Both of these additives have several disadvantages: DTT is incompatible with differential scanning calorimetry, βME is known to form adducts, and some workers have reported allergies to these compounds. Even worse, mutation of cysteine residues is often nontrivial because of the unique physical and chemical properties of the cysteine side chain. Although the cysteine side chain appears isosteric with serine, sulfur atoms are significantly larger than oxygen atoms; thus, cysteine hydrogen bonding stereochemistry is not necessarily directly compatible with that of serine. For these reasons the isosteric Cys → Ser mutation can be dramatically destabilizing. In practice, all small amino acids (i.e. Ala, Ser, Thr and Val) should be considered.

2.2.9 No Proxy Is Perfect

There is no ideal proxy; instead, individual researchers must weigh the respective importance of each of the above guidelines for themselves. As encouragement, consider the properties of FGF-1, the first protein from which a symmetric protein building block was successfully derived [5, 12]:

- Poor thermostability ($T_m = \sim 45$ °C; $\Delta G = \sim 21$ kJ/mol) [35, 36].
- Non-two-state behavior: folding kinetic frustration, irreversible aggregating upon heating [35].

- Three buried free cysteine residues, known to mediate aggregation and reduce functional half-life [34].
- Negligible sequence conservation between sub-domains (only one position conserved across all three trefoil subdomains); *high degree of backbone tertiary structure asymmetry*.
- Poor solubility (less than 1 mg/mL in low salt).

Symfoil-1, the result of TDSD using FGF-1 as the proxy, is thermostable, soluble, and two-state in its folding properties. Take home message: *even "difficult" proxies can be successfully symmetrized, it will just be a longer, more arduous journey*. Based on the results of Symfoil-1 and others, it is reasonable to project that the result of TDSD will be a more idealized protein than the proxy as regards thermostability and folding behavior.

2.3 Strategies for Identifying Symmetry-Enhancing Mutations

The exact substitutions to be made for symmetrization will vary based on the target architecture, the proxy and, the screening environment. Luckily, there are several straightforward methods to determine what mutations are worth modeling and experimentally characterizing, discussed below:

2.3.1 Propagate the Folding Nucleus

Knowledge of folding nucleus, although not strictly required to perform TDSD, can be greatly beneficial. As illustrated by Fig. 4, the symmetric solution of the TDSD of FGF-1 has significant sequence identity (71 %) to the region identified as the folding nucleus by phi-value analysis [22]. In other words, the application of a stability and folding cooperativity screen protected against

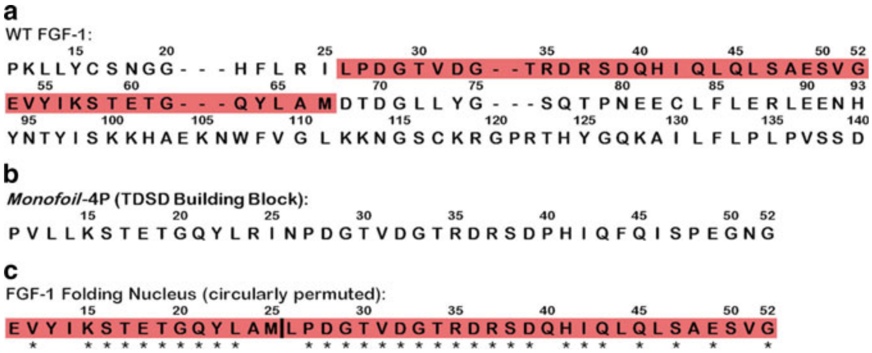

a
WT FGF-1:
15 20 25 30 35 40 45 50 52
P K L L Y C S N G G - - - H F L R I L P D G T V D G - - T R D R S D Q H I Q L Q L S A E S V G
55 60 65 70 75 80 85 90 93
E V Y I K S T E T G - - - Q Y L A M D T D G L L Y G - - - S Q T P N E E C L F L E R L E E N H
95 100 105 110 115 120 125 130 135 140
Y N T Y I S K K H A E K N W F V G L K K N G S C K R G P R T H Y G Q K A I L F L P L P V S S D
b
Monofoil-4P (TDSD Building Block):
15 20 25 30 35 40 45 50 52
P V L L K S T E T G Q Y L R I N P D G T V D G T R D R S D P H I Q F Q I S P E G N G
c
FGF-1 Folding Nucleus (circularly permuted):
15 20 25 30 35 40 45 50 52
E V Y I K S T E T G Q Y L A M | L P D G T V D G T R D R S D Q H I Q L Q L S A E S V G
* *

Fig. 4 TDSD extracts the folding nucleus. Panel **a**: Single letter amino acid code sequence alignment of the three repeating trefoil-fold subdomains in FGF-1 and colored to indicate a major region contributing to formation of the folding transition state as determined by phi-value analysis (*see* Fig. 5, panel **a**) [22]; Panel **b**: Single letter amino acid code of the "building block" 42-mer polypeptide for a stable β-trefoil fold identified by TDSD utilizing FGF-1 as the proxy; Panel **c**: amino acid sequence of the FGF-1 folding nucleus from panel a, represented in a circularly permuted form so as to generate an intact β-trefoil fold; the *asterisks* indicate positions of identity of this folding nucleus with the TDSD solution, revealing that the TDSD method captured a large portion of the critical folding nuclei in FGF-1

mutations that would have a detrimental effect on the foldability of the polypeptide and therefore preserved the folding nucleus. Thus, if knowledge of the region corresponding to the folding nucleus is available, mutations in this region should be limited. Instead, *the sequence of the folding nucleus should be considered a promising template for mutations at symmetry-related positions.*

2.3.2 Target Functional Residues

Functional residues are known to be detrimental to both foldability and the stability, and there is evidence that these residues are segregated from the folding nucleus (Fig. 5, panels a and c) [22–25]. Thus, targeting the substitution of functional residues in many cases is conceptually analogous to the suggestion above, to propagate the folding nucleus. In practice, however, there is a scarcity of detailed folding data for many proxies, and reliance on sequence alignment approaches to identify the folding nucleus is not always motivated, given the poor conservation of amino acids key for folding across homologs [37] and the observation that homologs may or may not fold through similar pathways [38]. Thus, functional residues may serve as a guide to identify the regions of the protein that are most optimized for folding (that is, regions depleted for functional residues). Finally, functional residues are often subject to a "stability–function" trade-off, in which mutation at these positions has a greater chance of improving stability (and obliterating function) [39, 40].

2.3.3 Remove Structural Insertions

In the spirit of the previous two suggestions, structural insertions are often functional and associated with impeded foldability [22–25]. Unlike the previous two suggestions, structural insertions can be readily identified by visual inspection of even a low-resolution structure (manifest as structural "aneurisms" within the framework of a generally symmetric architecture), making them easy targets during deconstruction. In the case of FGF-1, removal of two functional loops had three important consequences (Fig. 5): First, removal of the loops improved the overall backbone symmetry of the protein, which should be the first goal of deconstruction (discussed below). Second, the loops mediated a key function of FGF-1 (heparin binding), which was dramatically attenuated upon their deletion. Finally, removal of the insertions converted the folding of FGF-1 from having two folding phases (that is, non-two state behavior) to have a simple, single folding phase (more ideal two-state behavior, albeit with minor folding-arm rollover).

2.3.4 Consider Consensus Sequences

Consensus sequences are not thought to necessarily retain features of foldability [37, 38]; thus, a purely consensus sequence approach to generate a symmetric solution was unsuccessful [14]. However, in other respects, consensus sequences can provide valuable data: By comparing homolog sequences to that of the proxy, consensus sequence analysis can highlight residues that are key for stability

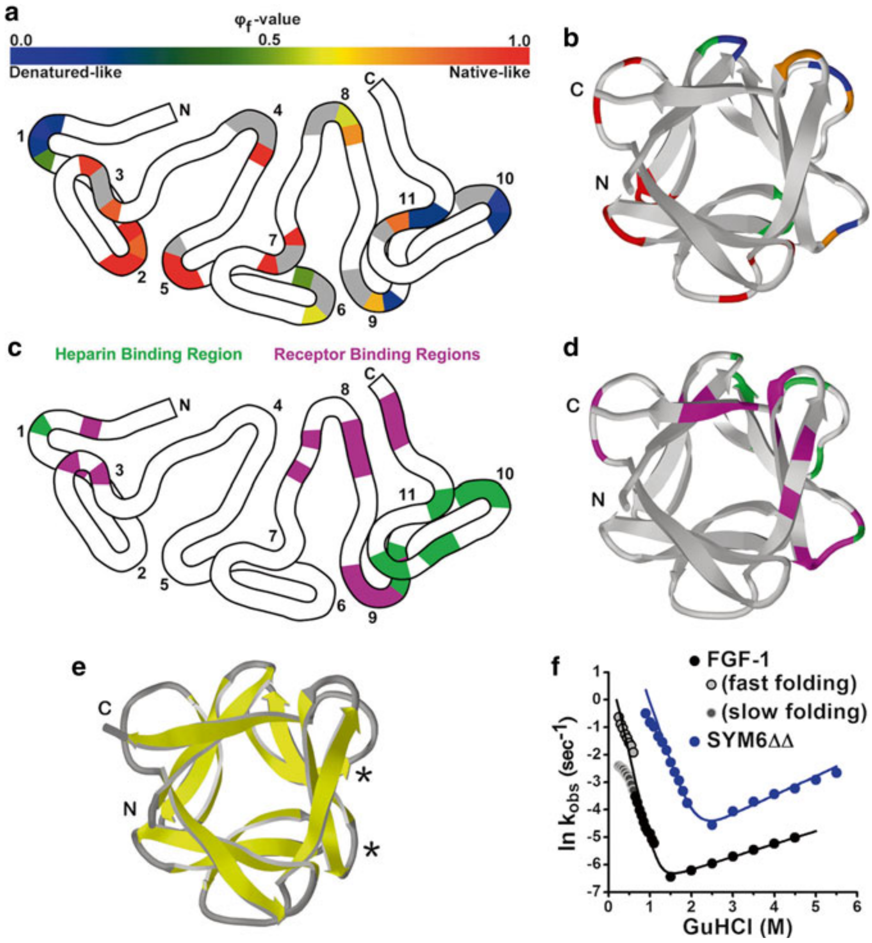


Fig. 5 Foldability–function trade-off in FGF-1. Panel **a**: Flattened ribbon diagram of FGF-1 with a heat map indicating phi-values for all turn regions (*red*: native-like environment in folding transition state; *blue*: denatured-like environment in folding transition state); Panel **b**: Crystal structure of FGF-1 (2AFG) indicating the general structural segregation of regions forming the folding transition state; Panel **c**: Flattened ribbon diagram of FGF-1 colored to indicate functional regions associated with heparin-binding (*green*) and receptor-binding (*magenta*); Panel **d**: Crystal structure of FGF-1 indicating the general structural locations of the heparin-binding and receptor-binding residues; Panel **e**: Crystal structure of a highly symmetric intermediate (Sym6ΔΔ [13]) in the TDSD of FGF-1 and where functional insertions associated with heparin and receptor-binding function have been deleted (improving the 3° structure symmetry; indicated by asterisks); Panel **f**: "Chevron" plot folding kinetic analysis of FGF-1 and Sym6ΔΔ proteins indicating that a kinetically trapped folding intermediate observed in FGF-1 is eliminated in the SYM6ΔΔ symmetric mutant, the folding transition state has been stabilized, and the thermostability has been increased, all at the cost of function deletion [13]

(and will be highly recalcitrant to mutation) as well as residues which may be functionally significant (and be good targets for mutation). Indeed, constructs generated from analyzing the sequences of homologs have been shown to successfully enhance stability for asymmetric protein folds in the absence of a symmetric constraint [41–43]. Consider the consensus sequence for the solvent-exposed Type 1 reverse turn: Asx-Pro-Asx-Gly [44]. Although none of the turns within a given proxy may adopt this exact sequence (as is the case for FGF-1), mutation to the consensus sequence for a Type 1 reverse turn has an improved chance of being stabilizing or neutral, and is thus an excellent way to simultaneously symmetrize the protein while improving thermostability. Turns that are not fully solvent exposed are generally more difficult to optimize, as the packing interactions must also be considered.

2.3.5 Fill Cavities in the Hydrophobic Core

Protein hydrophobic cores were once thought to be well optimized by evolution, suggesting that core repacking would afford modest benefits at best. However, as the availability and quality of structural data improved, it became apparent that many protein cores suffer from packing defects and that core repacking efforts could be a viable strategy to improve protein stability (albeit possibly at the expense of structural dynamics critical for specific functionality) [45–47]. Figure 6 shows an example from the TDSD of FGF-1, in which mutation of Leu44 to Phe efficiently filled an adjacent cavity, increased the symmetry of the primary structure, and stabilized the protein by 2.9 kJ/mol [48].

It is preferable to begin symmetrizing the peptide backbone by (1) removing structural insertions and (2) symmetrizing the hydrophobic core before moving on to optimizing turns and other secondary structure elements. In the case of FGF-1, the hydrophobic core is highly asymmetric and, as a result, symmetrizing mutations at certain sites were highly destabilizing; however, as the symmetry of the core packing groups and backbone improved, mutations that were once not tolerated could be incorporated with an *improvement* of stability (e.g., Met67 → Ile) [13]. This observation underpins the importance of updating the structural model when possible, and teaches that symmetrizing mutations (especially in the core) may not be viable until 3° structure symmetry is established in the course of TDSD.

2.4 The Role of Computation in TDSD

Top-down symmetric deconstruction can greatly benefit from the application of computational approaches and molecular modeling. At present, the “inverse folding problem” remains intractable and cannot be solved computationally and accurate calculations require significant expertise. In other words, TDSD with a granularity of zero (that is, no experimental validation at any intermediate step) is highly unlikely to succeed. Because TDSD starts with a foldable protein, the design challenges, as well as the computational challenges, are greatly alleviated. First, computational approaches are much more capable of differentiating between the stability of two sequences (i.e., $\Delta\Delta G$ quantitation), provided the final folded structure is approximately known, in comparison to determination of absolute $\Delta G_{\text{unfolding}}$ values. Second, core-repacking simulations tend to provide fairly accurate results, in part because core repacking can be viewed as a sort of three-dimensional jigsaw puzzle

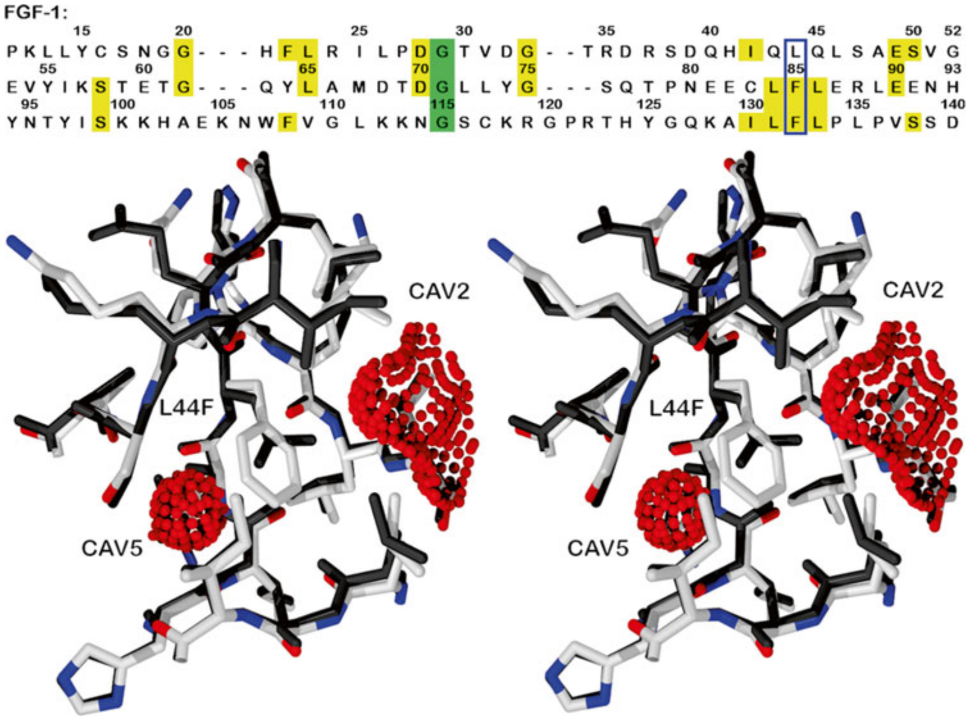

FGF-1:
15 20 25 30 35 40 45 50 52
P K L L Y C S N G G - - - H F L R I L P D G T V D G - - T R D R S D Q H I Q L Q L S A E S V G
55 60 65 70 75 80 85 90 93
E V Y I K S T E T G - - - Q Y L A M D T D G L L Y G - - - S Q T P N E E C L F L E R L E E N H
95 100 105 110 115 120 125 130 135 140
Y N T Y I S K K H A E K N W F V G L K K N G S C K R G P R T H Y G Q K A I L F L P L P V S S D
CAV2
L44F
CAV5
CAV2
L44F
CAV5

Fig. 6 Repacking the hydrophobic core of FGF-1. *Upper panel*: the primary structure of FGF-1 (β-trefoil proxy) aligned to indicate the three repeating trefoil-fold sub-domains. *Yellow shading* indicates symmetry-related positions having two residues in common, and *green shading* indicates symmetry-related positions with all three residues in common. The *blue box* indicates symmetry related positions having two residues in common (Phe, F) that was subjected to a symmetric point mutation (Leu44 → Phe), based upon structural analysis, to enforce a symmetric primary structure constraint at this position. *Lower panel*: Relaxed stereo diagram of wild-type FGF-1 (dark bonds; RCSB accession 2AFG) in the region of position 44 and including local solvent excluded cavities (indicated by *red dots*). Modeling of a F44 symmetric mutation indicated that such packing defects in the core of FGF-1 might readily permit this symmetric mutation. The crystal structure of the resulting mutation (CPK coloring; RCSB accession 1JTC) is overlaid with the wild-type FGF-1 structure and shows that the F44 symmetric mutation was accommodated with minimal structural perturbation. Structural analysis also confirmed an identical rotamer orientation in comparison to the two other symmetry-related Phe residues [48], thus achieving a purely symmetry structural relationship. This large aromatic mutation was the initial step in the TDSD of FGF-1

(i.e., largely limited to solving efficient van der Waals interactions and rotamer selection) [49–51]. Thus, high-level, detailed simulations—which are the purview of the dedicated computational chemist—are not necessary to guide mutational selection of core positions in many cases.

The choice of proxy, as discussed above, will influence the degree of reliance on molecular modeling and computation. Thermophilic proteins have a greater "reservoir" of Gibbs energy (i.e., $\Delta G_{\text{unfolding}}$) to leverage during deconstruction than do their mesophilic homologs. As a result, symmetrizing steps for mesophilic proteins will benefit from a heavier reliance on computational validation (and a greater granularity of the design cycle).

Finally, the limitations of computational design within the context of TDSD should be noted. The design of partially exposed sites tends to be more difficult using computational approaches alone, and experimental validation of these sites is critically important. In addition, the interaction energies employed by most computations are often based upon mesophile, not extremophile, environments; thus, attempts to design into niche environments will likely be met with greater computational inaccuracy. Programs for protein design calculations useful in TDSD include ORBIT [52], MCREM and BMCREM [53], Rosetta Design [54], EGAD [55], Dezymer [56], and SPRUCE [57].

2.5 Notes on Establishing a Screen for Stability and Folding Cooperativity

A key aspect of TDSD is to experimentally screen for stability and folding cooperativity at regular intervals during the design cycle. In doing so, TDSD explicitly traverses a path through stable, foldable sequence space between the proxy and the symmetric solution (with added potential for evolutionary significance [12]). The frequency that symmetrizing constructs are screened is referred to as *granularity*. A high degree of granularity (that is, screening each symmetrizing construct) requires more work than a low-granularity approach; however, the principle benefit of the high-granularity approach is that specific mutations deleterious for stability, folding cooperativity, and solubility (a solubility selection is intrinsic to the expression of mutant proteins) are identified unambiguously, and negative changes can be abandoned before additional mutations are incorporated. In general, proteins that have low to moderate stabilities (roughly in the range of 15–30 kJ/mol) will require a higher a granularity of the design cycle than will more stable proteins. As increasingly accurate computational approaches become available, the need to experimentally screen intermediate forms during TDSD will diminish. It is important to note, however, that some screening will be necessary, even with accurate computational support: features such as folding kinetics, folding cooperativity, the presence of intermediate protein forms (either at equilibrium or during folding), and melting temperature, cannot yet be routinely predicted computationally, especially by nonexperts. In addition, many computations implicitly assume a given set of buffer conditions that would be hard for nonexperts to modify; thus, design strategies focusing on certain niche environments (such as hyper-acidophiles or halophiles) will likely be less computationally tractable.

The specific methodologies used to assess stability should require only small amounts of protein and yield stability data quickly. Take note, however, that within the context of TDSD, "stability" has a formal definition: The Gibbs energy difference between the native state and the denatured state (Fig. 7, panel a). Although stability is related to various other properties (such as protease susceptibility, solubility, and melting temperature) it is

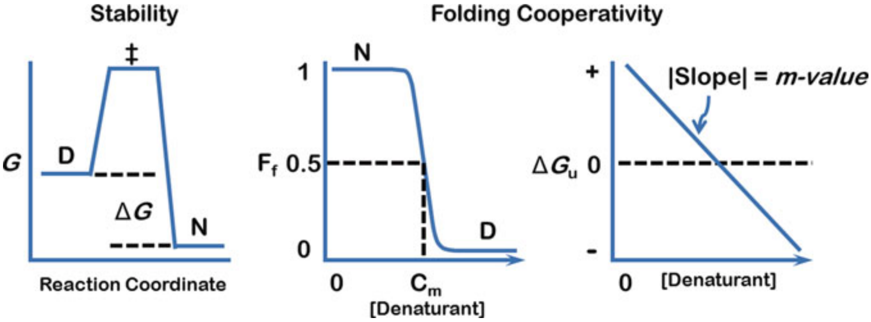


Fig. 7 Two-state model of thermodynamic stability and folding cooperativity. Panel **a**: Gibbs energy diagram of a two-state protein in which the denatured state ensemble (D) and the native state ensemble (N) are separated by a single, high-energy transition state (*double dagger*). Protein stability (ΔG) is taken to be the difference in Gibbs energy between the N and D states. Panel **b**: The fraction of folded protein molecules (F_f) undergoes a sigmoidal transition with respect to denaturant concentration, indicative of folding cooperativity. The steepness of the unfolding transition is dependent upon the *m*-value of the unfolding transition. Panel **c**: Denaturant *m*-values represent the (assumed) linear relationship between ΔG and denaturant concentration

best to measure stability directly, rather than rely on some of the alternatives mentioned above, which can yield misleading results. As such, functional screens/selections, which have been suggested as being useful metrics of protein folding and stability (unfolded proteins are not functional, after all), are generally qualitative, while precise quantitative stability/folding data are essential. Recall, the goal of TDSD is to create a functionless scaffold, which is violated if functional residues are constrained by a functional screen. In this spirit, the use of isothermal equilibrium denaturation (IED) is suggested as an efficient screen of stability and folding cooperativity.

IED experiments spectroscopically (e.g., by circular dichroism or intrinsic fluorescence) quantify protein unfolding upon incubation with chaotropic agents (e.g., guanidinium hydrochloride and urea). Details on how to perform an IED experiment are well established in the literature [58]. A key benefit to using IED over other methodologies lies in the generation of a parameter that reports on the folding cooperativity of the unfolding reaction, the *m*-value (Fig. 7, panels b and c). Within the context of TDSD, the *m*-value corresponds to the length of the protein folding reaction coordinate as regards solvent exposure and relates to the sharpness of the sigmoidal unfolding transition [59]. Mathematically, the *m*-value is defined as the (assumed linear) sensitivity of ΔG to chemical denaturant concentration. Folding cooperativity values scale with overall protein size, and mutations that induce significant denatured-state structure (or native state "fraying") will exhibit reduced *m*-values (see the work of Sánchez and Kiefhaber [60, 61] for an excellent discussion of both equilibrium and kinetic *m*-values).

Denaturant-induced cooperative unfolding may seem an unnecessary goal for the protein engineer, given that the target protein will not likely operate in the presence of guanidine hydrochloride. However, the Gibbs energy profile as a function of denaturant at a fixed temperature is orthogonal to the same energy profile as a function of temperature at fixed denaturant concentration (including 0 M denaturant). High folding cooperativity is a hallmark of evolved proteins, and is exceptionally difficult to design; unlike stability, strategies to enhance folding cooperativity are largely unknown [62]. Thus, while thermostability of proxy proteins can almost always be increased, folding cooperativity appears largely optimized and is, at best, maintained during TDSD. Consider results from the TDSD of FGF-1 (Fig. 8), which suggest that there may be an inverse relationship between thermostability and folding cooperativity (*m*-value) [5]. During the course of TDSD small systematic losses in folding cooperativity may be unavoidable; however, substantial additive effects can lead to a non-cooperatively folding protein.

Although IED studies provide experimental access to stability and folding cooperativity, the conditions required for IED may be incompatible with specific niche environments. For example, it is not generally possible to use IED to measure the stability of proteins in high salt concentrations (2.0–4.0 M NaCl) due to the cosolubility limits of common denaturants and salts. In these instances, orthogonal methods can be used to measure stability, such as differential scanning calorimetry (DSC) or thermal spectroscopy. For a practical reference of biomolecular DSC see Chowdhry and Cole [63].

2.6 The Chimera Approach

As overall structural symmetry improves, the chances of successfully employing a chimera approach (that is, the symmetric solution is taken directly from the sequence of a symmetrized TDSD intermediate) dramatically increases. Because a chimera-based approach is much faster than stepwise symmetrization, it is often worth exploring this option. Before attempting a chimera approach, consider the following:

1. The stability of the intermediate proxy should be quite high (~40+ kJ/mol).
2. The core packing groups, which come together in the chimeric form, must be highly symmetrized. Likewise, modeling and computation should not identify significant predicted clashes or cavities.
3. If possible, the region(s) corresponding to the folding nucleus of the proxy is known and should comprise a significant fraction of, and be contained within, the region(s) comprising the chimera design.

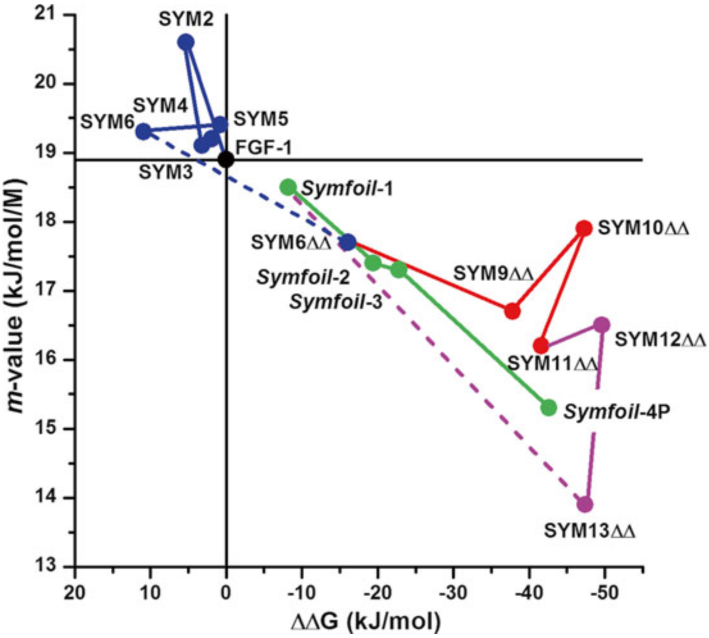


Fig. 8 Inverse linear relationship between protein stability and folding cooperativity. A plot of folding cooperativity (*m*-value) vs. change in protein stability (ΔΔ*G*) for a high-granularity TDSD of the β-trefoil fold (utilizing FGF-1 as the proxy) [5]. Larger *m*-values indicate greater folding cooperativity. A stabilizing mutant is characterized by a negative value for ΔΔ*G*. The TDSD process began with the wild-type FGF-1 proxy (black) and the lines between mutants traces the pathway through thermodynamic stability and folding cooperativity space. The initial transform involved symmetric point mutations within the core region (*blue*). The dashed *blue line* indicates the point at which deletion mutations were introduced in order to increase the 3° structure symmetry (and were essential for achieving a thermostable symmetric core design). The second transform involved symmetric turn mutations (*red*). The third transform involved symmetric β-strand mutations (*magenta*). The *dashed magenta line* indicates the point at which a chimeric mutant design approach yielded a purely symmetric 1° structure (the Symfoil-1 mutant). The fourth transform involved stability enhancement mutations to the symmetric solution (*green*), yielding the hyperthermophile Symfoil-4P symmetric mutant. For the set of TDSD mutants there is an apparent inverse linear correlation between folding cooperativity and thermostability; such that for each kJ/mol of increased thermostability there is an approximately 0.07 kJ/mol/M loss of folding cooperativity. The basis of this property may be an increase in residual structure in the unfolded state, although additional study is needed

4. When considering a chimeric approach, it is not necessary to extract the sequence from within a single structural subdomain; however, it is strongly recommended that the wild-type termini definition be retained. In other words, it is likely permissible (possibly essential) in a chimera solution to circularly permute the 1° structure, but not the 3° structure.

During the TDSD of FGF-1, a chimera approach was successfully employed in the latter stages (at a point of approximately 30 % 1° symmetry), and greatly reduced the amount of time needed to arrive at the symmetric solution. Sequence analysis of the regions used for the chimera and the crystal structure of the resulting protein (Symfoil-1) reveal two key features: First, the regions utilized in construction of the chimera have significant sequence identity to the region identified as the folding nucleus by phi-value analysis (Fig. 4). Thus, information necessary to encode foldability was retained (and perhaps triplicated). Second, in comparison to using the sequence of the folding nucleus directly, systematic improvements in core packing were achieved by the initial transform in the TDSD. This observation is best illustrated by the lack of an aromatic residue at the key symmetry-related core positions 44, 85, and 132 in a folding nucleus-based construct; however, the chimeric construct obtained from a TDSD intermediate has a conserved aromatic residue at these symmetry-related positions, which is critical for stability (see Fig. 4).

2.7 Generating a Peptide Building Block

Once arrived at the symmetric solution, identifying a peptide building block is much less labor intensive, and depending on the goals of the study, potentially optional. To proceed and identify a peptide building block for a given symmetric architecture, one must continue to stabilize the sequence of the symmetric solution (by optimization of secondary structure propensities, core packing, etc.) until fragmentation can be tolerated. For the case of FGF-1, Symfoil-4P ($\Delta G_{\text{unfolding}} = 64$ kJ/mol) was sufficiently stable to withstand fragmentation and retain foldability as a homo-trimer 42-mer polypeptide (see Fig. 1) [5, 12]. In general, it is recommended that fragments not be circularly permutated forms of the proxy 3° structure subdomain definition (although there has been some success with this approach [11]) and that the site of fragmentation be limited to solvent exposed turns.

References

1. Longo LMB, Blaber M (2012) Protein design—a vast unexploited resource. J Protein Proteonomics 3:78–83
2. Yue K, Dill KA (1992) Inverse protein folding problem: designing polymer sequences. Proc Natl Acad Sci U S A 89:4163–4167
3. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT et al (2012) Principles for designing ideal protein structures. Nature 491:222–227
4. Blaber M, Lee J (2012) Designing proteins from simple motifs: opportunities in top-down symmetric deconstruction. Curr Opin Struct Biol 22:442–450
5. Lee J, Blaber SI, Dubey VK, Blaber M (2011) A polypeptide "building block" for the ß-trefoil fold identified by "top-down symmetric deconstruction". J Mol Biol 407:744–763
6. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M et al (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332:816–821
7. Jung J, Lee B (2001) Circularly permuted proteins in the protein structure database. Protein Sci 10:1881–1886
8. Levy Y, Cho SS, Shen T, Onuchic JN, Wolynes PG (2005) Symmetry and frustration in

protein energy landscapes: a near degeneracy resolves the Rop dimer-folding mystery. Proc Natl Acad Sci U S A 102:2373–2378

9. Seitz T, Bocola M, Claren J, Sterner R (2007) Stabilization of a (beta-alpha)8-barrel protein designed from identical half barrels. J Mol Biol 372:114–129
10. Fortenberry C, Bowman EA, Proffitt W, Dorr B, Combs S, Harp J et al (2011) Exploring symmetry as an avenue to the computational design of large protein domains. J Am Chem Soc 133:18026–18029
11. Yadid I, Tawfik DS (2011) Functional β-propeller lectins by tandem duplications of repetitive units. Protein Eng Des Sel 24:185–195
12. Lee J, Blaber M (2011) Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. Proc Natl Acad Sci U S A 108:126–130
13. Brych SR, Dubey VK, Bienkieicz E, Lee J, Logan TM, Blaber M (2004) Symmetric primary and tertiary structure mutations within a symmetric superfold: a solution, not a constraint, to achieve a foldable polypeptide. J Mol Biol 344:769–780
14. Broom A, Doxey AC, Lobsanov YD, Berthin LG, Rose DR, Howell PL et al (2012) Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric globular protein. Structure 20:161–171
15. Fukuchi S, Nishikawa K (2001) Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol 309:835–843
16. Kumar S, Tsai CJ, Nussinov R (2000) Factors enhancing protein thermostability. Protein Eng 13:179–191
17. Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K (2003) Unique amino acid composition of proteins in halophilic bacteria. J Mol Biol 327:347–357
18. Oren A, Larimer F, Richardson P, Lapidus A, Csonka LN (2005) How to be moderately halophilic with broad salt tolerance: clues from the genome of *Chromohalobacter salexigens*. Extremophiles 9:275–279
19. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S (2001) Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res 11:1641–1650
20. Schreiber G, Buckle AM, Fersht AR (1994) Stability and function: two constraints in the evolution of barstar and other proteins. Structure 2:945–951
21. Shoichet BK, Baase WA, Kuroki R, Matthews BW (1995) A relationship between protein stability and protein function. Proc Natl Acad Sci U S A 92:452–456
22. Longo L, Lee J, Blaber M (2012) Experimental support for the foldability-function tradeoff hypothesis: segregation of the folding nucleus and functional regions in FGF-1. Protein Sci 21:1911–1920
23. Capraro DT, Gosavi S, Roy M, Onuchic JN, Jennings PA (2012) Folding circular permutants of IL-1beta: route selection driven by functional frustration. PLoS One 7:e38512
24. Gosavi S, Chavez LL, Jennings PA, Onuchic JN (2006) Topological frustration and the folding of interleukin-1β. J Mol Biol 357: 986–996
25. Gosavi S, Whitford PC, Jennings PA, Onuchic JN (2008) Extracting function from a beta-trefoil folding motif. Proc Natl Acad Sci U S A 105:10384–10389
26. Serrano L, Matouschek A, Fersht AR (1992) The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. J Mol Biol 224:805–818
27. Lowe AR, Itzhaki LS (2007) Rational redesign of the folding pathway of a modular protein. Proc Natl Acad Sci U S A 104(8):2679–2684
28. Liu C, Gaspar JA, Wong HJ, Meiering EM (2002) Conserved and nonconserved features of the folding pathway of hisactophilin, a β-trefoil protein. Protein Sci 11:669–679
29. Richter M, Bosnali M, Carstensen L, Seitz T, Durchschlag H, Blanquart S et al (2010) Computational and experimental evidence for the evolution of a $(\beta\alpha)_8$-barrel protein from an ancestral quarter-barrel stabilized by disulfide bonds. J Mol Biol 398:763–773
30. Carstensen L, Sperl JM, Bocola M, List F, Schmid FX, Sterner R (2012) Conservation of the folding mechanism between designed primordial $(\beta\alpha)_8$-barrel proteins and their modern descendant. J Am Chem Soc 134:12786–12791
31. Yadid I, Tawfik DS (2007) Reconstruction of functional β-propeller lectins via homo-oligomeric assembly of shorter fragments. J Mol Biol 365:10–17
32. Pace CN, Trevino S, Prabhakaran E, Scholtz JM (2004) Protein structure, stability and solubility in water and other solvents. Philos Trans R Soc Lond B Biol Sci 359:1225–1234, discussion 1234–1225
33. Barrick D (2009) What have we learned from the studies of two-state folders, and what are the unanswered questions about two-state protein folding? Phys Biol 6:015001

34. Lee J, Blaber M (2009) The interaction between thermostability and buried free cysteines in regulating the functional half-life of fibroblast growth factor-1. J Mol Biol 393: 113–127

35. Blaber SI, Culajay JF, Khurana A, Blaber M (1999) Reversible thermal denaturation of human FGF-1 induced by low concentrations of guanidine hydrochloride. Biophys J 77:470–477

36. Copeland RA, Halfpenny AJ, Williams RW, Thompson KC, Herber WK et al (1991) The structure of human acidic fibroblast growth factor and its interaction with heparin. Arch Biochem Biophys 289:53–61

37. Larson SM, Ruczinski I, Davidson AR, Baker D, Plaxco KW (2002) Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. J Mol Biol 316:225–233

38. Nickson AA, Clarke J (2010) What lessons can be learned from studying the folding of homologous proteins? Methods 52:38–50

39. Beadle BM, Shoichet BK (2002) Structural basis of stability–function tradeoffs in enzymes. J Mol Biol 321:285–296

40. Rubini M, Lepthie S, Golbik R, Budisa N (2006) Aminotryptophan-containing barstar: structure-function tradeoff in protein design and engineering with an expanded genetic code. Biochim Biophys Acta 1764:1147–1158

41. Steipe B, Schiller B, Pluckthun A, Steinbacher S (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. J Mol Biol 240:188–192

42. Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L et al (2000) From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. Protein Eng 13:49–57

43. Sullivan BJ, Nguyen T, Durani V, Mathur D, Rojas S, Thomas M et al (2012) Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. J Mol Biol 420:384–399

44. Lee J, Dubey VK, Longo LM, Blaber M (2008) A logical OR redundancy with the Asx-Pro-Asx-Gly type I β-turn motif. J Mol Biol 377:1251–1264

45. Karpusas M, Baase WA, Matsumura M, Matthews BW (1989) Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. Proc Natl Acad Sci U S A 86:8237–8241

46. Lassalle MW, Yamada H, Morii H, Ogata K, Sarai A, Akasaka K (2001) Filling a cavity dramatically increases pressure stability of the c-Myb R2 subdomain. Proteins 45:96–101

47. Bernett MJ, Somasundaram T, Blaber M (2004) An atomic resolution structure for human fibroblast growth factor 1. Proteins 57:626–634

48. Brych SR, Blaber SI, Logan TM, Blaber M (2001) Structure and stability effects of mutations designed to increase the primary sequence symmetry within the core region of a β-trefoil. Protein Sci 10:2587–2599

49. Ponder JW, Richards FM (1987) Tertiary templates for proteins—use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 193: 775–791

50. Lesk AM, Branden CL, Chothia C (1989) Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet. Proteins 5:139–148

51. Sandberg WS, Terwilliger TC (1991) Energetics of repacking a protein interior. Proc Natl Acad Sci U S A 88:1706–1710

52. Ross SA, Sarisky CA, Su A, Mayo SL (2001) Designed protein g core variants fold to native-like structures: sequence selection by orbit tolerates variation in backbone specification. Protein Sci 10:450–454

53. Zou J, Saven JG (2000) Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. J Mol Biol 296:281–294

54. Dantas G, Corrent C, Reichow SL, Havranek JJ, Eletr ZM, Isern NG et al (2007) High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. J Mol Biol 366:1209–1221

55. Pokala N, Handel TM (2004) Energy functions for protein design i: efficient and accurate continuum electrostatics and solvation. Protein Sci 13:925–936

56. Wisz MS, Hellinga HW (2003) An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins 51:360–377

57. Jain T, Cerutti DS, McCammon JA (2006) Configurational-bias sampling technique for predicting side-chain conformations in proteins. Protein Sci 15:2029–2039

58. Shaw KL, Scholtz JM, Pace CN, Grimsley GR (2009) Determining the conformational stability of a protein using urea denaturation curves. Methods Mol Biol 490:41–55

59. Myers JK, Pace CN, Scholtz JM (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci 4: 2138–2148

60. Sanchez IE, Kiefhaber T (2003) Evidence of sequential barriers and obligatory intermediates in apparent two-state protein folding. J Mol Biol 325:367–376

61. Sanchez IE, Kiefhaber T (2003) Hammond behavior versus ground state effects in protein folding: evidence for narrow free energy barriers and residual structure in unfolded states. J Mol Biol 327:867–884

62. Aksel T, Majumdar A, Barrick D (2011) The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. Structure 19:349–360

63. Chowdhry BZ, Cole SC (1989) Differential scanning calorimetry: applications in biotechnology. Trends Biotechnol 7:11–18

Chapter 9

Identification of Protein Scaffolds for Enzyme Design Using Scaffold Selection

André C. Stiel, Kaspar Feldmeier, and Birte Höcker

Abstract

The identification of suitable protein structures that can serve as scaffolds for the introduction of catalytic residues is crucial for the design of new enzymes. Here we describe how the automated and rapid scaffold search program ScaffoldSelection can be used to find the best starting points, namely protein structures that are most likely to tolerate the introduction and promote the proper formation of a specific catalytic motif.

Key words Protein design, Active site recapitulation, Computational biology, Structural bioinformatics, Motif search

1 Introduction

Enzymes are molecular machines that carry out many important tasks in the cell. This comparison has become even more applicable since modern protein design techniques allow adjustments, improvements, or even creation of catalytic activities in existing proteins. Most of these methods rely on computational approaches to rationalize the complexity of the protein. One of the most important applications is the *de novo* introduction of a catalytic activity to an unrelated protein. Here, a set of protein structures is searched and proteins are identified that can serve as the "scaffold" for the amino acid arrangement necessary for the catalytic activity, the "motif." Motif perpetuation has to be combined with minimal scaffold perturbation. Consequently, the selection of a good scaffold for a given motif is highly important.

Major achievements employing scaffold search methods have been the creation of an iron superoxide dismutase [1], the design of a retro-aldolase activity [2], and the introduction of a Kemp elimination reaction [3]. However, essential to judge the steadily increasing number of scaffold search algorithms are unified benchmarking sets allowing the proper evaluation of each technique [4, 5].

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_9, © Springer Science+Business Media New York 2014

A number of scaffold search algorithms have been introduced in the last two decades. The SiteSearch algorithm of the program Dezymer [6] for example uses an approach of subsequently checking all backbone positions for the possibility of a successful grafting of a motif amino acid, which is quite time consuming. Later GRAFTER [7] and FITSIDE [8] were presented that construct backbone side chain attachment distance matrices to identify possible motif-grafting sites. Other approaches such as RosettaMatch use inverse rotamer trees or geometric hashing techniques [4]. More recent programs are PRODA_MATCH [9], which uses a rotamer-library-free approach, thus enabling unusual motif geometries, and AUTOMATCH [10], which provides backbone flexibility for the scaffolds by introducing a back-rub-like approach [11]. In this chapter we focus on ScaffoldSelection [5], which uses a very fast approach, by first identifying potential grafting sites based on simple geometrical criteria before the whole motif is introduced and evaluated. Compared to the aforementioned approaches this allows the direct screening of very large sets of PDB files without computational preprocessing (e.g., preparation of hash tables).

In the following, we describe a ScaffoldSelection experiment consisting of the following five elements (1 and 2 can be found in Subheading 2, 3–5 in Subheading 3):

1. The motif – a coordinate set describing the catalytic geometry.
2. The potential scaffolds – a set of PDB structures searched for grafting positions.
3. The algorithm itself.
4. Ranking and weighting of output data.
5. Evaluation of results and post-processing of the data (relaxation and design).

The basic steps apply for all scaffold search algorithms while details are given on ScaffoldSelection only. Towards the end of each paragraph practical information is provided on how to use ScaffoldSelection (file names that directly apply to the program are set in *italic* and program parameters as `Terminal`). An overview of the program and its settings is provided in the flow chart in Fig. 1.

2 Materials

2.1 The Motif

The motif for a scaffold search should encompass all available data on the catalytic residues that can be molded into geometrical information. On the other hand it should be limited to include only the very essential residues since the number of motif side chains to consider determines the computational effort of the approach.

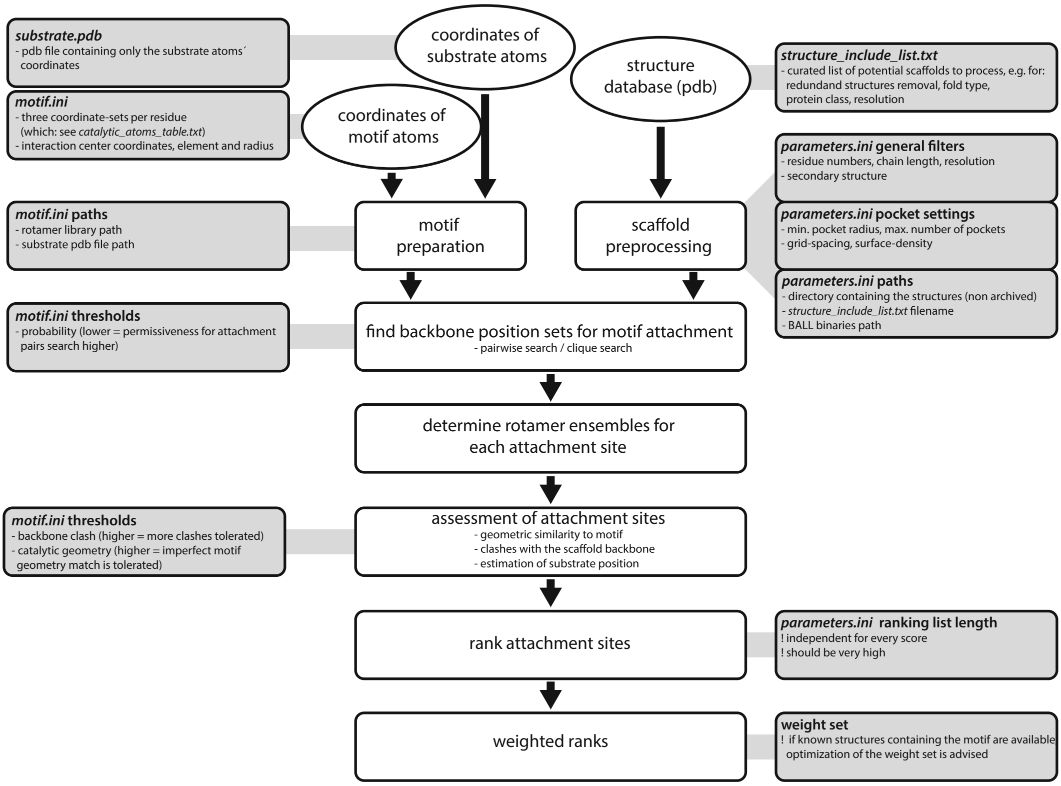

substrate.pdb
- pdb file containing only the substrate atoms´ coordinates
coordinates of substrate atoms
structure database (pdb)
structure_include_list.txt
- curated list of potential scaffolds to process, e.g. for: redundand structures removal, fold type, protein class, resolution
motif.ini
- three coordinate-sets per residue (which: see catalytic_atoms_table.txt)
- interaction center coordinates, element and radius
coordinates of motif atoms
parameters.ini general filters
- residue numbers, chain length, resolution
- secondary structure
motif.ini paths
- rotamer library path
- substrate pdb file path
motif preparation
scaffold preprocessing
parameters.ini pocket settings
- min. pocket radius, max. number of pockets
- grid-spacing, surface-density
parameters.ini paths
- directory containing the structures (non archived)
- structure_include_list.txt filename
- BALL binaries path
motif.ini thresholds
- probability (lower = permissiveness for attachment pairs search higher)
find backbone position sets for motif attachment
- pairwise search / clique search
determine rotamer ensembles for each attachment site
motif.ini thresholds
- backbone clash (higher = more clashes tolerated)
- catalytic geometry (higher = imperfect motif geometry match is tolerated)
assessment of attachment sites
- geometric similarity to motif
- clashes with the scaffold backbone
- estimation of substrate position
rank attachment sites
parameters.ini ranking list length
! independent for every score
! should be very high
weighted ranks
weight set
! if known structures containing the motif are available optimization of the weight set is advised

Fig. 1 Flow chart depicting the core components of ScaffoldSelection (*white boxes*) along with the settings that can be adjusted by the user (*grey boxes*). User-setting statements are positioned along with the program components they influence. However, all settings have to be made in the respective files (set in *italic*) before starting ScaffoldSelection

On average checking one structure with a motif of two to three amino acids takes 82 s (2.5 GHz DualCore, 4 GB RAM). In most cases choosing more than four catalytic amino acids for screening in a large scaffold set is not recommended (*see* **Note 1**).

Structural data on the motif can be gathered in different ways:

1. The most facile approach can be taken if structural data of proteins showing the desired reaction is available. If the catalytic motif in these structures shows only low variation, an average structure can be used to define the motif. If the motif is present in structures that belong to different protein folds and shows distinct differences, separate approaches for the different motives should be considered. Furthermore, such an array of reference structures often provides information on which amino acids are especially conserved and thus relevant for the reaction.

2. If experimental structural information is not available a catalytic geometry can be generated theoretically based on the knowledge of the chemical mechanism (theozyme) [12].

3. Sometimes, both approaches can be combined: structural information from existing motifs can be complemented by knowledge about the importance of specific aspects of the reaction mechanism. For example a strong interaction can be represented in tight constraints on these residues.

Furthermore, geometrical information on a substrate or transition state that interacts with the motif amino acids can be included in the search. It is used by ScaffoldSelection to check for clashes with backbone atoms of the scaffold. Thus, one can ensure that there will be enough space for the ligand. The ligand does not have to be a naturally occurring compound but can also be a combination of multiple states, e.g., educt and product.

A good example illustrating several of the aspects mentioned above is the catalytic triad found in many proteases. Its motif combines three relevant amino acids with a substrate and even includes a distinct interaction with backbone atoms called the oxyanion hole (Fig. 2). To define the motif first an overview of the structural homogeneity of the motif in nature can be obtained. Assuming that more closely related structures have more similar motifs, we can determine the distribution of the motif in different folds. From each fold known to carry this motif we then try to pick at least one crystal structure. In the case of the serine triad, we observe mainly two geometries: one common to subtilisin-like and one typical for

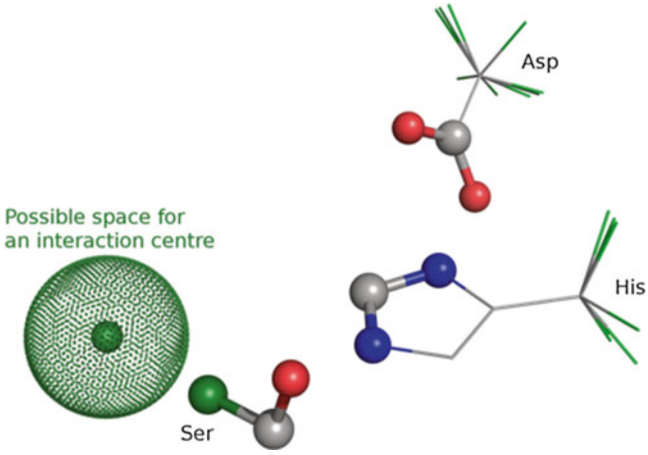


Fig. 2 Illustration of the motif components using serine proteases as an example. For every catalytically important amino acid rotamers are shown that meet the motif definition. Three atoms per residue that are required by ScaffoldSelection to unambiguously define the residue geometry are depicted as *balls* (coloring: oxygen *red*, nitrogen *blue*, carbon *grey*). Cα atoms that provide a grafting position for the scaffold are depicted in *green*. The interaction center is shown as a *dotted sphere* with its center as a *green ball*. The sphere radius defines acceptable positions as set by the user

the trypsin-like proteases. αβ-hydrolases that also carry the triad show both geometries. Thus, for this reaction two different motif geometries, one for the subtilisin geometry and one for the trypsin geometry, can be defined (*see* **Note 2**).

We now create the *motif.ini* file that describes the geometry for ScaffoldSelection, involving all atoms required for catalysis. First we provide a name for the motif and state the number of amino acids it contains:

```
[motif]
name=triad
residue_number=3
```

ScaffoldSelection uses three coordinate sets per residue that unambiguously describe the positions of the relevant side chain atoms of the motif amino acids to each other (Fig. 2). The file *catalytic_atoms_table.txt* specifies which three atoms are needed for every amino acid. For serine it shows the following entry:

```
Amino_acid     atom_1     atom_2     atom_3
Ser            CA         CB         OG
```

The exact positions of the amino acids involved in the catalytic motif can be taken from a PDB most common for this geometry (e.g., an average structure after clustering). For every amino acid in the motif an entry in *motif.ini* has to be made with the three coordinate sets for the corresponding atoms, e.g., for serine (*see* **Note 3**):

```
[residue1]
type=SER
ID=221
atom1=17.876    3.000    26.298
atom2=18.348    3.044    24.839
atom3=17.413    2.478    23.945
```

Since we also want to look for an oxyanion hole, represented by a nitrogen in a certain area, we add an interaction center to the search motif at the corresponding position. An interaction center defines a region (restricted by `cutoff_radius`) in which an atom has to be found in the scaffold. This allows looking for the presence of a distinct atom in a defined area that can be used for catalysis. The parameters for a nitrogen interaction center are described in the following: In the `[scores]` section of *parameters.ini* the use of interaction centers has to be enabled by adding:

```
[scores]
calculate_interaction_center=true
```

Moreover, the details for the interaction center have to be added to the *motif.ini*. The first statement sets the number of centers, in this case one. The second statement passes the center position and indicates that only grafting positions in scaffolds with a nitrogen within 3 Å radius of this position will be regarded:

```
interaction_center_number=1
[interaction_center1]
position=14.773  29.919      13.617
cutoff_radius=3
expression=element(N)
```

Further, we have to provide information on the substrate we want to use in our search. In this case we use a small peptide fragment containing only the atoms involved in the reaction (*see* **Note 4**). The position of the substrate fragment has to be defined relative to the catalytic motif. Its coordinates will be stored in a separate file (e.g., *substrate.pdb*). This PDB file will later be used by ScaffoldSelection to search for clashes of the substrate with backbone atoms of the protein. The path for this file has to be defined in the *motif.ini* by:

```
[paths]
substrate_pdb_file=substrate.pdb
```

Another important path to be specified here is the path to the rotamer library used for the motif side chains in the search (*see* **Note 5**):

```
rotamer_library_file =/resources/bbind02.May.library
```

2.2 The Scaffolds

All existing protein structures as deposited in the protein data bank (PDB; http://www.rcsb.org) can be considered as potential scaffolds and thus could be searched. Nonetheless, in most cases it is reasonable to make a preselection since the calculation time depends on the number of scaffold candidates to be screened. A selection can be made in different ways:

1. The PDB can be curated for structures of a certain resolution, multimeric state, method of structure determination, etc. Additionally, redundant structures, defined by a sequence similarity threshold, can be excluded (*see* **Note 6**).
2. Furthermore, potential scaffolds can be preselected for attributes such as size, fold type, or parent organism (*see* **Note 7**).
3. Finally, only a handpicked selection of structures can be screened to analyze a promising subset in a more detailed way (e.g., motif with more amino acids or interaction centers).

For an unobstructed scaffold search it is recommended to "normalize" all scaffold structures; that is, structures containing atoms other than H, C, N, O, and S should be excluded or modified (e.g., selenomethionines should be changed to methionines). Furthermore, continuous residue numbering should be ensured and the atom naming should meet PDB standards. In general, everything that does not look like a perfectly normal amino acid should be excluded. The outcome is used as the input list of scaffolds to be screened (*see* **Note 8**).

In the following example we pass the set of structures to be searched to ScaffoldSelection and introduce several parameters, which restrict the search space within the algorithm. The file *structure_include_list.txt* tells ScaffoldSelection which PDB files will be searched. It has the following format:

```
1EMA.pdb
2LYZ.pdb
...
```

The path to the *structure_include_list.txt* and to the directory containing the respective structures in .pdb format has to be defined in the path section of the *parameter.ini* file (*see* **Note 9**).

```
[paths]
structure_directory =/Database/
include_list_file=structure_include_list.txt
```

Furthermore, the file *parameters.ini* encompasses several thresholds that restrict the number of considered scaffolds as well as grafting sites within ScaffoldSelection. The parameters `minimal_residue_number` and `maximal_residue_number` restrict the search space to structures with the corresponding number of amino acids. `minimal_chain_length` on the other hand restricts the search space to proteins with the corresponding chain length. This can be used for example to exclude small peptide fragments from the search. The parameter `minimal_resolution` filters based on the resolution of a crystal structure. `helix_boundary_length` and `strand_boundary_length` are restrictions on amino acid placement with respect to secondary structure elements (determined by DSSP [13]). If for example `helix_boundary_length` is set to 5, only helical positions that are not more than 5 amino acids away from the ends of the helix will be taken into account.

Finally, the parameter `ranking_list_length` determines the number of possible insertion sites that will be saved (all positions in all scaffolds together). If additional positions are found, ScaffoldSelection will delete the lowest scoring ones (*see* **Note 10**).

A typical section in *parameters.ini* amounts to the following:

```
[thresholds]
minimal_residue_number=100
maximal_residue_number=1000
minimal_chain_length=20
minimal_resolution=2.5
helix_boundary_length=5
strand_boundary_length=5
ranking_list_length=500000
```

If, as in most cases, the motif should be a catalytic site that is shielded from solvent, e.g., in an accessible cavity, ScaffoldSelection, and in a similar way most other programs, can reduce the search space further by identifying suitable pockets in the protein. ScaffoldSelection identifies pockets on the protein surface using LIGSITEcs [14] and restricts the search to attachment positions near these pockets (as defined by `pocket_radius` and `number_of_pocket_clusters`). These settings have to be passed within the `ligsitecs` section in the *parameters.ini*, e.g.:

```
[ligsitecs]
compute_surface_pockets=true
pocket_grid_space=1.0
number_of_pockets=3
SSSthreshold=5
surface_density=0.5
pocket_radius=8.0
```

3 Methods

3.1 The Algorithm

Using the information of the motif and the scaffold set, the program ScaffoldSelection employs a two-step approach to search structures for positions where the motif amino acids can be grafted onto the backbone. In a first step a pair-wise search is carried out for Cα-Cβ vectors of the scaffolds side chains that can support two given motif side chains in a predefined geometry. To this end the distance as well as the angles between two motif components are measured and compared to the motif (`probability_threshold` controls permissiveness) resulting in a list of potential attachment pairs. This step is followed by sampling potential rotamers at these positions, which expands the pair list for several potential rotamer combinations. Pairs that have common positions are then stepwise combined to complete motif attachment sites. These attachment sites are then checked for motif integrity (maximal_catalytic_geometry_penalty controls permissiveness) and, furthermore, backbone clashes as well as substrate clashes are controlled (if at one position no clash-free rotamer is found, the whole group is discarded; maximal_backbone_clash_penalty controls permissiveness for the backbone and maximal_substrate_clash_penalty for the substrate).

These penalty parameters, along with further basic settings are described in the following example. In *parameters.ini* we add the path to the BALL library [15] to the [paths] section; thus the complete paths entry for the *parameters.ini* should now look like this:

```
[paths]
structure_directory =/Database/
include_list_file=structure_include_list.txt
ball_directory =/BALL/
```

Then the scores that should be calculated have to be enabled (e.g., if you are not searching with a ligand/substrate calculate_substrate_clashes can be set to false). Including the already set interaction center switch, this section amounts to:

```
[scores]
calculate_interaction_center=true
calculate_backbone_clashes=true
calculate_substrate_clashes=true
calculate_rotamer_probabilities=true
calculate_interaction_center=true
```

Furthermore, settings have to be added to the file *motif.ini* defining the search permissiveness. Lower probability_threshold results in taking more positions into account, while higher maximal_backbone_clash_penalty results in more backbone clashes being tolerated for a hit. Finally a higher maximal_catalytic_geometry_penalty allows for grafting sites that do not properly meet the geometry (*see* **Note 11**). A characteristic setting for the threshold section in *motif.ini* would be the following:

```
[thresholds]
probability_threshold=0.01
maximal_backbone_clash_penalty=1.0
maximal_catalytic_geometry_penalty=1.5
```

Now ScaffoldSelection can be run with the files *motif.ini*, *parameter.ini*, *substrate.pdb*, *structure_include_list.txt* and a directory containing the PDBs as defined in *structure_include_list.txt* (*see* **Note 12**).

3.2 Output Data: Ranking and Weighting

Finally, the ScaffoldSelection run will lead to the identification and scoring of a number of potential attachment sites. Several scores are calculated for each found site (hit) that reflect the different aspects of the search. Key criteria for a good hit are the following:

1. The geometry between the newly found attachment site and that of the defined motif should be as similar as possible (`catalytic_geometry_penalty_score`).
2. The used rotamers should be frequently observed in natural structures, which indicates that they are energetically favorable (`rotamer_probability_score`).
3. The perturbation of the scaffold protein should be as low as possible. That is, no steric clashes between the grafted side chains of the motif and the backbone of the scaffold protein should occur (`backbone_clash_penality_score`).
4. In case of a motif including a substrate or ligand, its correct fit in the geometric assembly of the grafted motif has to be considered (`substrate_clash_penality_score`).
5. If an interaction center was used, it is relevant, how precisely (distance to center) the interaction center criteria have been met (`interaction_center_score`; *see* **Note 13**).
6. Not directly involved in the scaffold search but highly relevant for proper motif accommodation is that the side chains surrounding the grafted motif will need as little changes as possible to make the motif fit properly.

All these scores can be combined and weighted by the user (e.g., by simply multiplying with weight factors) to obtain a total score indicative of the quality of the motif grafting positions (*see* **Note 14**).

Our exemplary ScaffoldSelection run would result in different score output files (*example_backbone_clash_score.txt*, *example_cgp_score.txt*, *example_rotamer_freq_score.txt*, *example_rotamer_indices.txt*, *example_ia_center_score.txt* and *example_substrate_clash_score.txt*). Each of these files contains a list, which is already sorted by the corresponding score and contains only the maximum number of hits as defined by `ranking_list_length` (in this respect a very high setting for this parameter is advisable, see above). However, this can result in not necessarily all lists containing the same attachment scaffolds/positions. Consequently, it is required to collect the different ranks for a given position. It is a good idea to start with the positions performing best in the `catalytic_geometry_penalty_score` and search the other lists for those hits. In the end one should have a list with the most promising position and the ranks of all scores. These ranks are then multiplied with weights and summed up to a final score. Overall, it is recommended to put most emphasis on the `catalytic_geometry_penalty_score`.

If possible we optimize these weights and parameters employing a training-set with structures known to carry the functional catalytic geometry. ScaffoldSelection should rediscover the respective motif and score its position. Good parameters that give a relatively high number of well scoring hits in the training-set should consequently lead to an enrichment of credible positions in the complete search. Thus, favorable weights and parameters can be deduced by comparison. Interesting scaffolds/positions from ScaffoldSelection are then extracted for further analysis.

3.3 Post-Scaffold Search Approaches

ScaffoldSelection provides the user with the additional ability to actually build a hit (scaffold with grafted motif) as a PDB structure. Besides being able to visually inspect promising hits in more detail, the built structures can be used to apply further computational (design) steps. As a first measure the built structure can be relaxed, which is necessary since ScaffoldSelection just grafts the motif amino acids but does not reflect the conformational change of the scaffold protein caused by the change. For example RosettaRelax [16, 17] can be used as a fast and efficient approach, but its changes of the backbone conformation are limited. Alternatively, molecular dynamics might be applied as a more physically realistic, but time-consuming approach. Relaxing the hit structure will very likely display severe steric clashes between the grafted motif amino acids and the scaffold side chains (the `backbone_clash_penality_score` only covers backbone clashes). At this point, computational design methods such as RosettaDesign [17] or PocketOptimizer [18] can be used to change the motif surrounding amino acids, thus perfecting its accommodation. The design should be combined with relaxation steps to adapt to global changes of the structure caused by changed amino acids (*see* **Note 15**).

In the following an example is given on how to use the MotifConstruction module of ScaffoldSelection. The motif builder requires the input file *mc_parameters.ini*. It provides information on where the motif amino acids have to be grafted, the path to the scaffold structure to use and an output path. A typical *mc_parameters.ini* file would look like this:

```
[attachment_position1]
index=65
chain=A
…
[paths]
scaffold_file=1EMA.pdb
output_file=1EMA_grafted_motif_out.pdb
```

Given the scaffold PDB file, the *mc_parameters.ini* and the *motif.ini* from the respective ScaffoldSelection run are present, MotifConstruction is started by:

```
motifConstruction mc_parameters.ini outputname
```

The resulting structure is the scaffold PDB with the motif amino acids grafted onto the backbone.

4 Notes

1. Binary packages of ScaffoldSelection including all third party software can be obtained from (https://webdav.tue.mpg.de/u/birtehoecker). Packages are available as installer scripts for Linux, OSX, and Windows operating systems.
2. If the motif consists of more than four catalytic amino acids one could do an initial search on a large number of potential scaffolds with a restricted motif followed by a thorough search with the complete motif on the top hits of the previous search.
3. Structures considered for motifs should have at least an $R_{\text{free}} \leq 0.3$ and a resolution ≤ 2.5 Å.
4. The `ID=` statement in *motif.ini* is voluntary, e.g., representing the amino acid numbering in the original protein for reference reasons (`ID=221`).
5. Often structural information on a substrate present in the catalytic center is only available through structures with bound inhibitors. From these the coordinates for the substrate can be deduced by superposition of the desired substrate with the corresponding part of the inhibitor.
6. The Dunbrack rotamer library (edit 2002) can be obtained from http://svn.cgl.ucsf.edu/svn/chimera/trunk/libs/Rotamers/Dunbrack/bbind02.May.lib.
7. Lists of curated pdb-IDs can be obtained from the Dunbrack Lab web server ("http://dunbrack.fccc.edu/Guoli/PISCES_OptionPage.php").
8. If available for the proteins in question, variants of thermophilic organisms are favorable, since their proteins are often more robust due to their thermostability.
9. As scaffolds one might consider to regard only soluble, monomeric proteins below a certain size that do not contain highly specialized cofactors or disulfide bridges, since their versatility and handling (mutagenesis, expression, etc.) is easier. In later stages of a scaffold search (when looking only at top hits) it

might also be advisable to take additional information into account, such as: parent organism, information about function of the designated motif grafting positions (e.g., is this region important for folding?).

10. In order to run ScaffoldSelection on a large number of structures those have to be split into manageable sets. In our approach lists with a length of 100 scaffolds worked well. For every *structure_include_list.txt* a separate run of ScaffoldSelection is started in a separate folder. After completion of all runs, the identified positions and their corresponding scores have to be collated into single files.
11. The `ranking_list_length` should always be set very kigh (e.g., 1,000,000) to ensure listing of all possible grafting solutions.
    Note: `ranking_list_length` as Terminal
12. For an initial search the parameters *`probability_threshold, maximal_backbone_clash_penalty`*, and *`maximal_catalytic_geometry_penalty`* should be set very permissive. according to the number of relevant hits ScaffoldSelection identifies they can be made more restrictive in subsequent runs, starting with adjustments to the `maximal_catalytic_geometry_penalty` setting.
13. Commenting in all ScaffoldSelection files can be done with "`;`" to add user info like:

    ```
    ; section only set for experiment…
    ```
14. Since reaction centers are very versatile tools, their score can be interpreted in different ways. For a singular prosthetic atom, e.g., a magnesium ion, the score might be handled relatively tightly, while a diffuse interaction center like the oxyanion hole of proteases requires a more permissive usage of the score. These notions can be taken into account by appropriate weight sets.
15. It is recommended to train the weight and parameter sets, if possible, using known structures able to catalyze the reaction of interest. The scores of these "known" hits yield proper settings for the various starting parameters of the search algorithm.
16. The number of motif amino acids should always be reduced to the essential set, since every changed scaffold amino acid reduces the likelihood of yielding the natural stabilized fold.

References

1. Pinto AL, Hellinga HW, Caradonna JP (1997) Construction of a catalytically active iron superoxide dismutase by rational protein design. Proc Natl Acad Sci U S A 94:5562–5567
2. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A et al (2008) De novo computational design of retro-aldol enzymes. Science 319:1387–1391
3. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J et al (2008) Kemp elimination catalysts by computational enzyme design. Nature 453: 190–195
4. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA et al (2006) New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 15:2785–2794
5. Malisi C, Kohlbacher O, Höcker B (2009) Automated scaffold selection for enzyme design. Proteins 77:74–83
6. Hellinga HW, Richards FM (1991) Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. J Mol Biol 222:763–785
7. Hearst DP, Cohen FE (1994) GRAFTER: a computational aid for the design of novel proteins. Protein Eng 7:1411–1421
8. Hornischer K, Blocker H (1996) Grafting of discontinuous sites: a protein modeling strategy. Protein Eng 9:931–939
9. Lei Y, Luo W, Zhu Y (2011) A matching algorithm for catalytic residue site selection in computational enzyme design. Protein Sci 20:1566–1575
10. Zhang C, Lai L (2012) AutoMatch: target-binding protein design and enzyme design by automatic pinpointing potential active sites in available protein scaffolds. Proteins 80: 1078–1094
11. Friedland GD, Linares AJ, Smith CA, Kortemme T (2008) A simple model of backbone flexibility improves modeling of side-chain conformational variability. J Mol Biol 380:757–774
12. Tantillo DJ, Chen J, Houk KN (1998) Theozymes and compuzymes: theoretical models for biological catalysis. Curr Opin Chem Biol 2:743–750
13. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
14. Huang B, Schroeder M (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 6:19
15. Kohlbacher O, Lenhof HP (2000) BALL—rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. Bioinformatics 16:815–824
16. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93
17. Leaver-Fay A, Tyka M, Lewis SM et al (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545–574
18. Malisi C, Schumann M, Toussaint NC et al (2012) Binding pocket optimization by computational protein design. PLoS One 7(12): e52505

Chapter 10

Computational Design of Novel Enzymes Without Cofactors

Matthew D. Smith, Alexandre Zanghellini,
and Daniela Grabs-Röthlisberger

Abstract

In this review we present a recently developed computational method to design de novo enzymes. Starting from the three-dimensional arrangement of the transition state structure and the catalytic side chains around it (theozyme), RosettaMatch identifies successful placements of the theozyme into protein scaffolds. Subsequently, RosettaEnzDes (for EnzymeDesign) redesigns the active site around the theozyme for binding and stabilization of the transition state and the catalytic residues. The resulting computationally designed enzymes are expressed and experimentally tested for catalytic activity.

Key words Computational enzyme design, Novel enzyme, Rosetta, RosettaMatch, Computational protein design, Enzyme engineering

1 Introduction

Computational design of novel protein catalysts is a truly interdisciplinary endeavor that brings together the fields of biochemistry, chemistry, biophysics, and the power of computational methods. The design of enzymes for arbitrary chemical reactions has the potential to greatly impact many fields and industries by allowing the creation of valuable chemicals and breakdown of pollutants or toxins. Furthermore, missing links in biochemical pathways could be filled in, bringing us closer to the dream of the cell as a customizable, miniature chemical factory. We review here a recent method for developing novel protein catalysts. This method has been successfully applied to the design of novel catalysts for the retro-aldol cleavage reaction [1], the Kemp elimination reaction [2], ester hydrolysis [3], and the example we review here, the Diels-Alder 2+4 cycloaddition reaction [4].

Before starting the computations for a novel enzyme, there are several points to consider. First, the ease of enzyme design can vary greatly for different chemical reactions, but it is often difficult to determine this in advance as the current methods (such as QM/

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_10, © Springer Science+Business Media New York 2014

MM) to assess chemical reactivity *in silico* have shown moderate accuracy and are extremely computationally intensive. The enzyme design strategy presented here relies primarily on the preferential binding and stabilization of the transition state(s) of the reaction by providing hydrogen bond donors or acceptors, stabilizing charges, and/or pre-organizing the orientation of the substrate(s). The enzyme thereby reduces the energy barrier and catalyzes the reaction. While we see a great range of uncatalyzed rates for the substrates of natural enzymes, we work under the assumption that the higher the activation energy of a reaction, the more difficult it will be to create an enzyme for it. Additionally, it has been found empirically that design is more successful when applied to substrates that contain both hydrophobic and hydrogen-bonding groups.

Second, the readout of the enzymatic reaction and the sensitivity of the assay must be considered. In the successful application of de novo enzyme design we present here, a variety of readouts—colorimetry, fluorimetry, as well as mass spectrometry—have been used. These methods vary in their ease and speed, which affects the throughput and extent of downstream characterization. It is expected that initial computationally designed lead enzymes will exhibit low activities. In order to detect these low activities, assays that can measure low turnover numbers or product formation in the low μM range are preferred. Figure 1 depicts the substrates, transition state, and product for the Diels-Alder reaction, the example reaction we chose to illustrate computational enzyme design. In this case, there was no colorimetric or fluorimetric readout, so a liquid chromatography assay coupled to a mass spectrometer (for greater sensitivity) was used to measure both the formation and the stereochemistry of the reaction products.

Third, a critical step in enzyme design is to determine the transition states of the reaction and the core catalytic machinery that

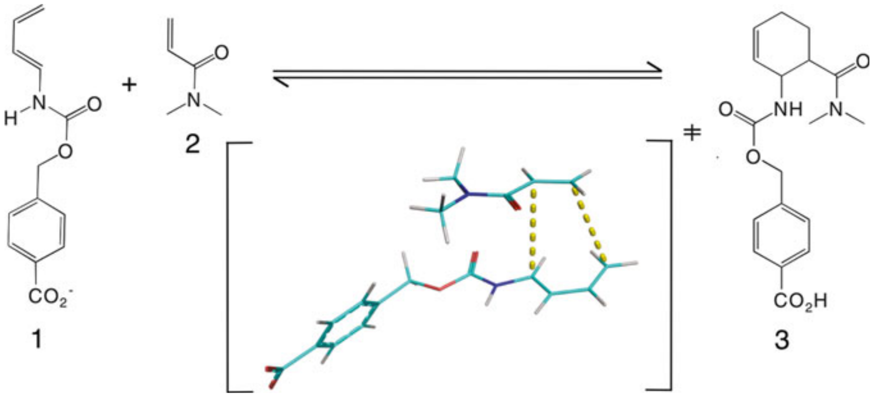


Fig. 1 Substrates (diene 1 and dienophile 2) and product (3) for Diels-Alder enzyme design project. The transition-state structure of the reaction is depicted in *cyan*

can catalyze the reaction. In the case of reactions with multiple transition states, such as the retro-aldol reaction [1], one typically picks the transition state with the highest activation-energy barrier, although building "consensus" transition state models is a possibility [1]. One way to elucidate and model transition states is through careful study of the literature on the reaction of interest and/or obtaining transition state structures from computational chemistry experiments. Once a reasonable transition state model has been made, the next challenge is to determine the arrangement of amino acid side chains around the transition state to best catalyze the reaction of interest (this arrangement of amino acid side chains and transition state is called a theozyme) [5]. We have had success working with quantum chemists, studying transition state models and enzyme mechanisms, and applying general quantum chemistry principles (stabilization/destabilization of orbitals of interest as they evolve along the reaction path and stabilization of charges as they develop in the transition state) as well as general principles from the vast literature on enzyme mechanisms in nature. In cases where small-molecule catalysts have been found for the reaction of interest, their structures may be used to guide theozyme design. In more than one example from our work designing enzymes, we have seen it necessary to try multiple theozymes, as no activity was seen for the first models [1, 3].

For the Diels-Alder reaction in general and the particular example we review here (Fig. 1), much was already known about the mechanism of catalysis [6, 7]. There are many small-molecule catalysts for the Diels-Alder reaction [8], and extensive mechanistic studies have elucidated relatively confident models of the orbital dynamics of the reaction. In addition, catalytic antibodies had been selected which carried out this Diels-Alder reaction [9]. In this case, we looked to frontier molecular orbital theory to guide the active site design. Transition-state stabilization can be achieved by raising the energy of the HOMO (highest occupied molecular orbital) and lowering the energy of the LUMO (lowest unoccupied molecular orbital), thereby narrowing the energy gap between the two orbitals and reducing the free energy of activation. The energy of the HOMO is raised by positioning a hydrogen bond acceptor (such as a carbonyl group of glutamine or asparagine) to stabilize the developing positive charge on the carbamate NH of the diene, and the energy of the LUMO is lowered by positioning a hydrogen bond donor (such as serine, threonine, or tyrosine) to stabilize the developing negative charge on the carbonyl of the dienophile. Quantum mechanical (QM) calculations were carried out to determine the geometry of the lowest free energy barrier transition state between substrates and product in the presence of these hydrogen-bonding groups [4] which resulted in a three-dimensional representation of the theozyme, depicted in Fig. 2.

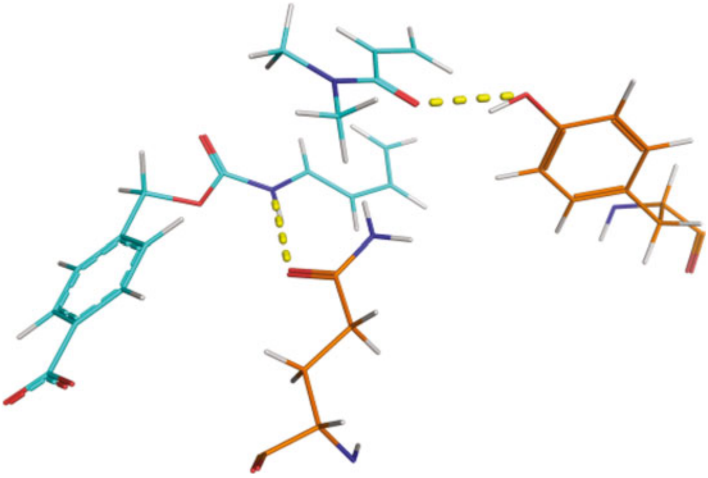


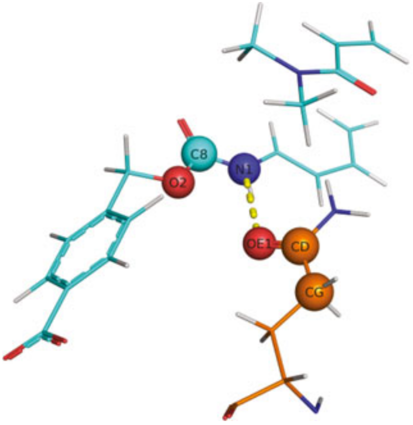
Fig. 2 Theozyme structure for the Diels-Alder reaction used in computational enzyme design [4]. A glutamine and a tyrosine serve as catalytic residues (in *orange*) to stabilize the transition state (in *cyan*)

Once the initial development of a sensitive assay and the calculation of an optimal theozyme have been completed, one can start the Rosetta enzyme design protocol which is divided into four stages:

1. Preparation of enzyme specification and scaffold set.
2. Matching of the theozyme into the scaffold set.
3. Design of the active site.
4. Post-design filtering.

2 Preparation of Enzyme Specification and Scaffold Set

To describe the theozyme we require a specification of the geometric relationships between the desired active site residues and the transition state. This specification is called a constraint file. From the theozyme structure of the Diels-Alder reaction (Fig. 2), it is easy and convenient to extract the geometry of interactions between all catalytic side chains (glutamine and tyrosine in this case) and the transition state. An important point to consider is the conformational flexibility in the transition-state structure. Rotatable bonds should be "sampled" in the transition-state structure to allow more variability in compatible active sites (*see* Supplementary material in ref. 4, and Supplementary material in ref. 1 for examples of the ensemble used for matching). For Rosetta enzyme design, the convention is to name the ensemble of transition state structures the *downstream partner* and each catalytic side chain in



```
Atom names:          N1-C8-O2 -- CG-CD-OE1
Generalized names: D1-D2-D3 -- U3-U2-U1
Object names:        ----A---    ----B---
```

Degree of freedom:	Definition:
Distance	Distance D1-U1
Angle A	Angle D2-D1-U1
Angle B	Angle D1-U1-U2
Torsion A	Torsion angle D3-D2-D1-U1
Torsion AB	Torsion angle D2-D1-U1-U2
Torsion B	Torsion angle D1-U1-U2-U3

Fig. 3 Example of a theozyme interaction from a glutamine (in *orange*) to the diene (in *cyan*) as well as the definition of the degrees of freedom used in the constraint file

the theozyme as the *upstream partner*. Their interaction geometries can be uniquely specified by the six degrees of freedom between three atoms of either partner (Fig. 3).

A constraint file contains one block (starting with CST::BEGIN and ending with CST::END) for each catalytic side chain. For the Diels-Alder reaction we would need two such blocks (one for glutamine and one for tyrosine) to make up the full theozyme definition (note that the text in curly brackets throughout the chapter indicates comments and should not be included in the actual files):

```
CST::BEGIN
TEMPLATE:: ATOM_MAP: 1 atom_name: N1 C8 O2
   {downstream partner}
TEMPLATE:: ATOM_MAP: 1 residue3: TRS
TEMPLATE:: ATOM_MAP: 2 atom_name: OE1 CD CG
   {upstream partner}
TEMPLATE:: ATOM_MAP: 2 residue3: GLN
CONSTRAINT:: distanceAB: 2.80   0.20   100.  0  2
CONSTRAINT::    angle_A: 113.5  10.0   10  360. 2
CONSTRAINT::    angle_B: 120.   10.0   10  360.1
CONSTRAINT::  torsion_A: 0.  10.0  10     360.  2
CONSTRAINT:: torsion_AB:  0.  30.0  0.0    90.   1
CONSTRAINT::  torsion_B: 180.  20.0  10  360.  3
CST::END
```

Each degree of freedom is defined by five numbers:

```
CONSTRAINT:: torsion_AB: 0. 30.0 0.0 90. 1
```

The first number is the optimal value of the degree of freedom: either a distance, angle, or torsion angle. Here, torsion_AB (the torsion angle of D2-D1-U1-U2) is defined to be 0°. The next number, 30.0, defines the allowed deviation, for a final allowable range of $0° \pm 30°$. The third number, 0.0, defines the "weight" on this constraint—the magnitude of the energetic penalty for deviating from the allowed range, which is used during minimization and design but not during matching [10]. In our example of the catalytic geometry between the glutamine and the transition state, the weight is zero as this degree of freedom is free to rotate and should not be penalized during design. The fourth number, 90, defines the periodicity of this value. A periodicity of 90 means that this degree of freedom is allowed every 90° starting from the initial value. In this example, the angles $0° \pm 30°$, $90° \pm 30°$, $180° \pm 30°$, and $270° \pm 30°$ are allowed catalytic geometries. The final number, 1, specifies the fineness of sampling to perform in the matching stage. The matching algorithm will sample this degree of freedom at $2*n+1$ points between the allowed deviation from the specified value (with this much sampling at each point of periodicity). This gives a total of 12 samples, 1 at every 30°, for even sampling of this degree of freedom.

Next, we need to prepare a set of scaffolds of protein structures into which we will attempt to build our enzymes. In general, we choose protein structures with a reasonable resolution (3 Å or better) that contain small molecules (to ensure the presence of a cleft or pocket) and for which the protein is expressed in *E. coli* (for ease of expression and purification). There are a number of other criteria (like multimerization state, number of residues, protein origin, or protein fold) that can be applied as well. Proteins from thermophilic organisms are good candidates for the scaffold set and have been preferred in de novo enzyme design due to the fact that they are usually more tolerant to destabilizing mutations arising in the design of novel catalytic activity [11]. Once a set of protein structures has been obtained for use as scaffolds for enzyme design, we must prepare them for use with Rosetta. For each protein structure we generate two files: a coordinate file (encoded in the Protein Data Bank—PDB—format [12]) with just the amino acids of the protein present and a position file which contains a list of residue numbers (or amino acid positions) which will be considered as catalytic residue placements during matching. While theozyme placement in the core of the protein will most probably unfold the protein, theozyme placement on the surface of the protein will be difficult as only a limited number of residues are available for interactions with the transition state. Therefore, residues lining a pocket or a cleft have the highest chance to successfully build an active design and should be included in the positions file. A straightforward way to identify these residues is to select residues within 5–7 Å of the natural ligand in the crystal structure.

3 Matching

The goal of the matching stage is to efficiently find scaffolds where we can place the theozyme (including both amino-acid catalytic side chains and transition-state structure) in a geometric orientation consistent with catalysis. In this section, we go through the matching protocol to give a broad overview of how it works and to provide an understanding of the options and parameters which can be set or changed (for a detailed technical discussion *see* refs. 10 and 13). To keep track of the location of the transition state, the matching algorithm first divides the active site pocket of the scaffold into geometrical boxes, or "bins." The bin size can be set with the –euclid_bin_size option. It then places each discrete conformation of the catalytic residue (termed rotamer) independently at each allowed position in the scaffold (defined in position file). Then the transition state is placed according to the defined catalytic geometries (defined in constraint file) for this catalytic residue. If the atoms of the catalytic residue or the transition-state structure do not overlap with the rest of the protein main chain, the bin in which the transition-state structure resides, the orientation of the transition-state structure, the rotamer, and the position of the catalytic residue are recorded in a "hash table." A hash table is an efficient way to store and resolve possible theozyme placements. At the end of matching, the entries in the hash table are compared to identify successful theozyme placements. A successful placement is defined as one where all catalytic residues, originating from different positions, place the transition state into the same spatial bin in roughly the same orientation. How exact the orientation has to overlap is defined by the combination of –euler_bin_size and –euclid_bin_size options, and are key parameters for a matching run. For each such placement or "match," a PDB file is written out which is then used as input for the design stage.

The match executable located in your Rosetta directory (*see* **Note 1**) carries out the matching using the following command line options (refer to the documentation and manual for a complete set of command line options):

```
-database      your_directory/rosetta/rosetta_
   database/{pointing to Rosetta database}
-extra_res_fa TRS.params    {params file for
   transition state}
-in:file:s scaffold.pdb {scaffold file used for
   matching}
-match:scaffold_active_site_residues scaffold.
   pos {position file for scaffold}
```

```
-match:lig_name TRS {name of transition state as defined in params file}
-match:orientation_restype TRS {use transition state to calculate orientation during matching}
-match:orientation_atoms N1 C8 O2 {use these three atoms to calculate orientation}
-match:geometric_constraint_file TRS.cst {describes catalytic geometry of theozyme}
-match:euclid_bin_size 1.50 {bin size in Ångstrom}
-match:euler_bin_size 40.0 {angle deviation of orientation}
-match:bump_tolerance 0.6 {allowed overlap between transition state and catalytic side chain/backbone}
-packing:ex1 {add extra rotamers at ±1 standard deviation for chi1)
-packing:ex2 {add extra rotamers at ±1 standard deviation for chi2)
-packing:ex3 {add extra rotamers at ±1 standard deviation for chi3)
-packing:ex4 {add extra rotamers at ±1 standard deviation for chi4)
-nstruct 1 {repeat this protocol once}
```

The more hits recorded in the hash table and compared at the end the more memory is used and the more files are written as successful matches (which takes time and disc space). There is thus a fine balance between finding enough matches (in the hundreds or thousands) and running out of memory or disc space. There are two categories of parameters one can adjust: those that control the precision of the match (euler and euclid bin) and those that sample the different degrees of freedom (conformations of the transition state, allowed catalytic geometries, extra rotamers for catalytic side chains, etc.) (*see* **Note 2**). The pragmatic approach is to take two or three scaffolds with different folds, perform a test run, and look at the results. If you get too few matches or none at all, you may want to loosen the precision of the match parameters, increase the sampling, or both. If you get too many matches you do the opposite. When you are happy with the result you can then launch a production run (matching on all scaffolds) and carry the output forward to the design stage.

4 Design of Active Site

Once the theozyme has been matched into scaffolds, the next step is to redesign the rest of the active site pocket away from its previous function and towards activity for the reaction of interest. Quite often the introduced transition-state structure and catalytic side chains do not fit well into the pocket or cleft provided by the wild-type sequence. There can be steric overlap ("clashes") between the placed theozyme and the rest of the protein side chains of the original scaffold, or cavities where the protein side chains provide no interaction with the transition-state structure. The design protocol relieves these clashes by moving or mutating the impinging residues and introduces interactions that further bind and stabilize the transition state, which acts both to create affinity for the substrate as well as accelerate the rate of the reaction. Furthermore, the design approach works towards native-like "backing up" of the catalytic residues, trying to achieve good interactions to stabilize these residues in the desired, catalytically productive orientation. All of this is accomplished simultaneously with the Metropolis criterion Monte Carlo sampling approach used in RosettaDesign [14, 15] to identify the most energetically favorable enzyme active site.

While it is in theory feasible to redesign the entire scaffold on which the theozyme has been placed, one typically restricts redesign to the active site pocket. This is done to make the problem more computationally tractable and to minimize mutations that may destabilize the protein scaffold. Therefore, we set cutoffs for amino acid positions that are to be mutated (6–8 Å from the placed transition-state structure, with catalytic residues excluded from design) and for those residues that are to be repacked, meaning to have their rotamer but not identity changed (10–12 Å from the transition state). These cutoffs can be set with the options -enzdes:cut1 (all residues closer to the transition-state structure and then cut1 can be mutated), -enzdes:cut2 (all residues between cut1 and cut2 and pointing towards the transition-state structure can be mutated, the others repacked), -enzdes:cut3 (all residues between cut2 and cut3 are repacked), and -enzdes:cut4 (all residues between cut3 and cut4 and pointing towards the transition-state structure are repacked, all others stay in their crystal structure conformation). Since during the matching stage we allowed for some deviation from the ideal catalytic geometry, the enzyme design protocol first optimizes all constraints (-cst_opt option) by minimizing the transition-state structure position (done automatically), the chi angles of the catalytic side chains (-chi-min option), and the backbone conformation (-bb_min option). It then cycles between designing the active site (-enzdes:design option) and minimizing it (-enzdes:cst_min option) for a given number of times (-enzdes:design_min_cycles <value> option). The core of the design process is Monte Carlo sampling of amino acid rotamers. For those positions to be mutated, all rotamers for all allowed amino acids are available during sampling. For those that are to be repacked, only rotamers for the current amino acid are accessible during sampling. If different conformations for the transition-state structure are defined, then those are sampled at this stage as well (*see* **Note 3**).

Monte Carlo sampling requires an evaluation function. For design, Rosetta uses a full-atom scoring function, modified for use with enzymes. For information about the energetic terms in this scoring function (which include electrostatics, the Lennard-Jones potential, hydrogen bonding, and other terms), refer to ref. 16. In addition, energetic penalties are computed to assess when the theozyme's degrees of freedom are out of the range specified in the constraint file. The balance between these two contributors to total score (which determines selection of amino acid identity as well as their rotameric state) can be changed by increasing or decreasing the weight for each degree of freedom within the constraint file.

The executable EnzdesFixBB performs the active site design. Typical command line options for enzyme design are:

```
-database/your_directory/rosetta/rosetta_
   database/{pointing to Rosetta database}
-extra_res_fa TRS.params {params file for
   transition state}
-enzdes:cst_opt {optimize all constraints}
-enzdes:cst_design {design active site}
-enzdes:cst_min {minimize active site}
-enzdes:chi_min {minimize chi angles of side
   chains}
-enzdes:bb_min {minimize backbone during cst_
   opt and cst_min}
-enzdes:bb_min_allowed_dev 0.1 {allowed
   deviation of Cα before penalty applies}
-enzdes:cstfile TRS.cst {describes catalytic
   geometry}
-enzdes:detect_design_interface {define
   active site residues around transition state}
-enzdes:cut1 6.0 {all residues within 6A of the
   transition state}
-enzdes:cut2 8.0 {all residues within 8A of the
   transition state}
-enzdes:cut3 10.0{all residues within 10A of
   the transition state}
```

```
-enzdes:cut4 12.0{all residues within 12A of
   the transition state}
-enzdes:lig_packer_weight 1.8 {increase
   ligand weights during scoring by 1.8}
-enzdes:design_min_cycles 3 {cycle between
   design and minization three times before out-
   put a structure}
-packing:ex1     {add extra rotamers at ±1 stan-
   dard deviation for chi1)
-packing:ex2     {add extra rotamers at ±1 stan-
   dard deviation for chi2)
-packing:ex1aro  {add extra rotamers for chi1
   for aromatic residues)
-packing:extrachi_cutoff 1 {add extra rotam-
   ers for residues with a neighbor count of 1
   or more}
-packing:use_input_sc {include rotamer from
   original pdb structure}
-packing:soft_rep_design {decrease overlap
   penalty during design (and not minimization)
   stage}
-nblist_autoupdate {update neighborlist
   during minimization}
-linmem_ig 10    {use linear memory interaction
   graph}
-nstruct 100     {perform this entire protocol
   100 times and output 100 structures}
```

EnzdesFixBB produces two types of output: design structures and score data. The design structures are PDB files containing the coordinates of the design, showing the protein, redesigned active site, and transition state. The score data contains per-residue (contained in the corresponding PDB structure file) and total (contained in a separate file, the score file) score data (with these scores determined by the score terms and weighting described above) for the output design.

5 Post-design Filtering

After collecting all the output from the design stage one is left with the nontrivial task of selecting active designs and filtering out the inactive ones. Despite significant progress in the field of enzyme design this is still essential, especially if the number of designs that can be experimentally tested is limited. A wide variety of filtering criteria can be used to rank or filter the designs. While there are some general rules (low total energy of the protein indicates likely expression and solubility, low transition-state energy indicates good binding, good theozyme geometry (as constraint score) is necessary for catalytic activity), each theozyme will require its unique set of filters. In order to filter designs, one first has to define a list of features or characteristics that are important for the reaction to happen (a specific distance, angle, hydrogen bond interaction, etc.). After defining a metric for each of these features one can look at a handful of randomly chosen designs and determine a threshold for acceptance/rejection of a design. Applying these filters to the entire data set should result in the "best" designs. Depending on the assay throughput and cost to synthesize genes, one may want to be more or less stringent in their filtering.

As an illustrative example, we present here the concrete cases of filtering on total protein score and catalytic geometry. When filtering total protein score it is important to consider each scaffold independently as the Rosetta score for each scaffold can vary significantly. The total protein score can be found in the first column of the enzdes.scores file, labeled "total_score." Sorting this column, one typically keeps the best 10–50 % of the designs. A similar approach is taken for the catalytic geometry. The total score for the catalytic geometry can be found in the fourth column of the enzdes.scores file, labeled "all_cst." In this case, the aim is to keep all designs that have a catalytic geometry as close as possible to the geometry specified by the theozyme. The appropriate threshold for this value can be defined by looking at a handful of designs spanning the entire score range. Designs with a cst score higher than the threshold are discarded.

Finally, the sequence of the selected genes is synthesized and cloned into an expression vector with appropriate tags for affinity purification (*see* **Note 4**). Standard protocols for expression and purification of recombinant proteins can be used to obtain pure protein, and this protein can then be tested for the enzyme activity of interest (Fig. 4).

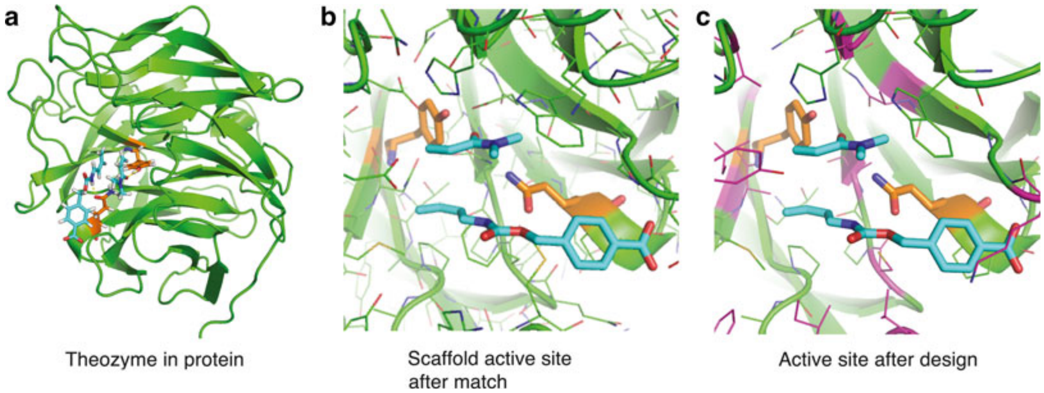


Fig. 4 Progress of design from theozyme (**a**) to match (**b**), to final design (**c**). During the design process, wild-type residues (in *green lines*) are mutated (*pink lines*) if more favorable interactions can be made

6 Conclusion

The de novo computational design method presented in this review emphasizes the importance of quantum effects in enzymes (through the concept of the theozyme) and of transition-state stabilization (as implemented through the design process). While this method of enzyme design has proved successful in a number of cases, we believe that other considerations may be necessary to continue improving the method. These include treatment of quantum mechanics within the design state, evaluating the effect of protein motions, or the effect of long-range electrostatic interactions. Enzyme design is a fast-growing field and has the potential to create commercially valuable enzymes as well as useful tools to probe biology. Finally, this method leads us closer to the dream of the cell as a customizable chemical factory, allowing economical and green chemical processing.

7 Notes

1. Rosetta is available for licensing and download at http://www.rosettacommons.org

   It is free for academic and nonprofit users and is available at a competitive licensing rate for commercial users. Included with the distributed source code are instructions and manuals for downloading, building, and installing the software. The RosettaCommons website is also a good resource, with manuals and support on how to run the Rosetta software.

2. For catalytic side chain to transition-state interactions that have multiple degrees of freedom (and therefore require extensive sampling), a complementary matching approach called "secondary matching" is available in the RosettaMatch application. A detailed description and discussion of the advantages and disadvantages of this approach can be found in [13].

3. Designing a large number of matches can be quite computationally intensive. If computing power is limited, a two-stage approach to design may be used. In a first design round minimize high-intensive sampling (like extra rotamers, backbone minimization, or extra design cycles) and filter the output. In a second design round, use only the best designs as input instead of the original matches and follow the described protocol.

4. It is good practice to check the original scaffold for missing density or compare the final design sequence with the sequence of the original scaffold from UniProt or GenBank before ordering or synthesizing designed sequence. Rosetta will omit residues that are missing in the crystal structure without a warning

or comment in the output PDB. A protein with a missing surface loop may be output as a design in this way, and such a protein might not express in soluble form, precluding experimental testing of the computationally designed novel enzyme.

References

1. Althoff EA, Wang L, Jiang L, Giger L, Lassila JK, Wang Z et al (2012) Robust design and optimization of retroaldol enzymes. Protein Sci 21:717–726
2. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J et al (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190–195
3. Richter F, Blomberg R, Khare SD, Kiss G, Kuzin AP, Smith AJT et al (2012) Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. J Am Chem Soc 134:16197–16206
4. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, StClair JL et al (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329:309–313
5. Tantillo DJ, Chen J, Houk KN (1998) Theozymes and compuzymes: theoretical models for biological catalysis. Curr Opin Chem Biol 2:743–750
6. Kim SP, Leach AG, Houk KN (2002) The origins of noncovalent catalysis of intermolecular Diels-Alder reactions by cyclodextrins, self-assembling capsules, antibodies, and RNAses. J Org Chem 67:4250–4260
7. Blake JF, Lim D, Jorgensen WL (1994) Enhanced hydrogen bonding of water to Diels-Alder transition states. Ab initio evidence. J Org Chem 59:803–805
8. Breslow R, Dong SD (1998) Biomimetic reactions catalyzed by cyclodextrins and their derivatives. Chem Rev 98:1997–2012
9. Yli-Kauhaluoma JT, Ashley JA, Lo C-H, Tucker L, Wolfe MM, Janda KD (1995) Anti-metallocene antibodies: a new approach to enantioselective catalysis of the Diels-Alder reaction. J Am Chem Soc 117:7041–7047
10. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA et al (2006) New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 15:2785–2794
11. Besenmatter W, Kast P, Hilvert D (2007) Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. Proteins 66:500–506
12. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K et al (2002) The Protein Data Bank. Acta Crystallogr Sect D: Biol Crystallogr 58:899–907
13. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D (2011) De novo enzyme design using Rosetta3. PLoS One 6:e19230
14. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368
15. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97: 10383–10388
16. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R et al (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545–574

Chapter 11

De Novo Design of Peptide Scaffolds as Novel Preorganized Ligands for Metal-Ion Coordination

Aimee J. Gamble and Anna F.A. Peacock

Abstract

This chapter describes how de novo designed peptides can be used as novel preorganized ligands for metal ion coordination. The focus is on the design of peptides which are programmed to spontaneously self-assemble into α-helical coiled coils in aqueous solution, and how metal ion binding sites can be engineered onto and into these structures. In addition to describing the various design principles, some key examples are covered illustrating the success of this approach, including a more detailed example in the case study.

Key words De novo peptide design, Coiled coil, α-helix, Coordination chemistry, Metal ions, Templated ligands

1 Introduction

Metallo-proteins represent nearly a third of all proteins and are capable of coordination chemistry typically inaccessible with simple (small molecule) coordination complexes in aqueous solution. For example, coordinatively unsaturated metal ions can be achieved which do not result in dinuclear degradation, due to the steric bulk of the protein ligand. The metallo-protein exists in water, yet the protein environment is capable of controlling the dielectric. Challenging asymmetric coordination environments are common, and furthermore, second coordination sphere residues (for example with acidic or basic groups) are extremely important in the resulting metal ion chemistry.

The de novo design of peptides with predictable secondary, tertiary and quaternary structure, for metal ion coordination, offers the synthetic advantages of small molecules, whilst retaining the core elements of natural proteins. These peptides can be readily synthesized using the same amino acids and ligands as proteins, but also offer the opportunity for inclusion of non-coded amino acids or functionalities. Ultimately they offer a simplified scaffold with which one can more readily establish important structure–function relationships.

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_11, © Springer Science+Business Media New York 2014

The first step in the design of a metallo-peptide is the selection of a suitable scaffold. Though peptide scaffolds based on β-sheet structures [1, 2], β-hairpin peptides [3], γ-turns [4] and $\beta_2\alpha$ motifs similar to zinc-finger domains [5, 6], have been explored, the majority of work has focused on the design of miniature α-helical coiled coil scaffolds [7]. The coiled coil is a structural motif generated by two or more α-helices which wrap around each other in a left-handed superhelical fashion; some examples of which are shown in Fig. 1. These miniature protein folds are the most commonly used de novo designed peptide scaffolds as they contain a single secondary structure unit, the α-helix, yet possesses both tertiary and quaternary structure. An attractive feature of the coiled coil is its predictable and programmable structure based on the primary amino acid sequence, which can in some cases be modified without significantly affecting the peptide folding. Furthermore its hydrophobic core is akin to the buried hydrophobic core of a protein and represents an attractive location for metal ion coordination, as described later in the Chapter.

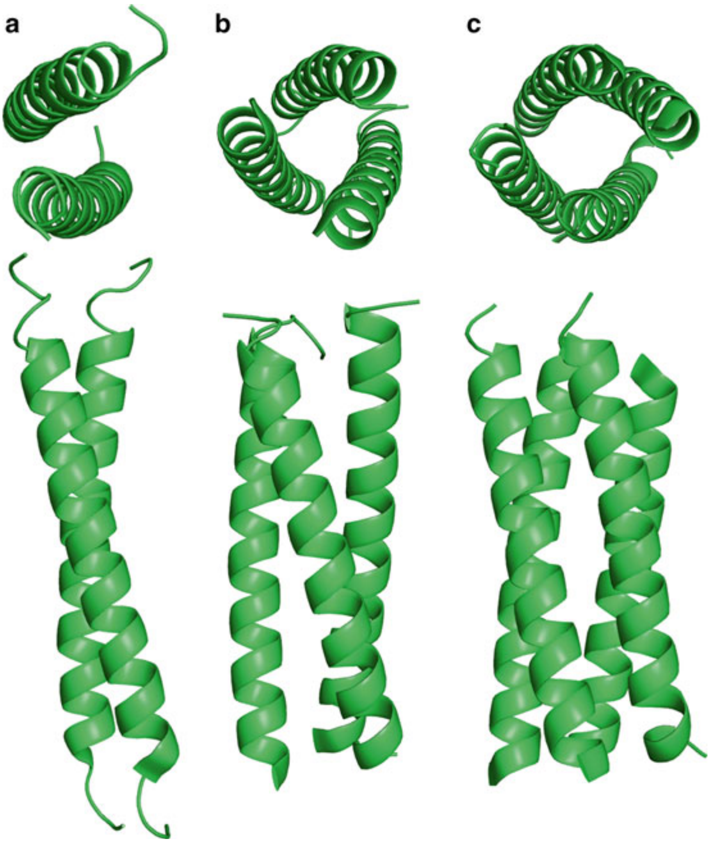


Fig. 1 PyMOL representations of parallel homo—(**a**) dimer (pdb 1C94 [67]), (**b**) trimer (pdb 1GCM [68]), and (**c**) tetramer (pdb 3R4H [16]) coiled coil scaffolds, with top-down (*top*) and side-on (*bottom*) views. Shown are main chain atoms as ribbons. Side chains and solvent molecules have been omitted for clarity

This chapter will first describe key guidelines in the rational design of a coiled coil scaffold. This will be followed by a brief discussion of the ligands available for metal ion coordination, followed by the various strategies by which these can be introduced into the coiled coil scaffold. Special attention will be paid to the incorporation of a metal binding site within the hydrophobic core, and a select number of examples will be discussed in more detail. Finally, a short case study, which illustrates some important design opportunities, will be discussed.

2 The Design of Coiled Coils

The most common approach in the design of a metallo-peptide, involves the selection of an appropriate peptide scaffold, for example a specific coiled coil, and subsequent modification of the primary amino acid sequence, so as to incorporate the desired metal ion binding site into the scaffold. Here we will cover some important guidelines for the rational design of coiled coil structures, providing the basis by which a suitable scaffold may be designed for subsequent metal ion coordination.

2.1 Features of a Coiled Coil

2.1.1 The Hydrophobic Core

As for natural proteins, one of the main driving forces for coiled coil folding is the hydrophobic effect. The structure of a folded α-helix consists of 3.6 residues per turn; therefore placement of hydrophobic groups at the first and fourth position forms an amphipathic α-helix with one hydrophobic face. This is the basis of the heptad repeat approach, which defines approximately two turns of the α-helix and is described as $(abcdefg)_n$; *a*–*g* represent amino acids in the first to the seventh position of the heptad, and n is the number of heptad repeats (generally $n \geq 3$ for a well folded coiled coil) [8]. The hydrophobic core residues occupy the *a* (first) and *d* (fourth) sites in this notation, which is applicable to dimers, trimers, and tetramers (Fig. 1), as well as higher order oligomeric structures. Helical wheel diagrams illustrating the location of residues in the resulting coiled coil are shown in Fig. 2. As the heptad does not exactly describe two turns of the α-helix, the hydrophobic face migrates around the α-helix in a coiled fashion in the opposite direction to the helix backbone. So as to minimize contact of the hydrophobic core with the solvent, the bundle of α-helices does not form a geometrically parallel structure, but that of a left-handed super-coil.

The *a* and *d* positions of coiled coils are normally populated by aliphatic hydrophobic resides, in particular Ile and Leu (and to a lesser extent Ala and Val), as opposed to the aromatic side chains of Phe and Trp [8], although coiled coils based around the latter have been prepared [9, 10]. The choice of residues to incorporate into these sites and the effect they have on the oligomerization of the coiled coil assembly is discussed in Subheading 2.2.

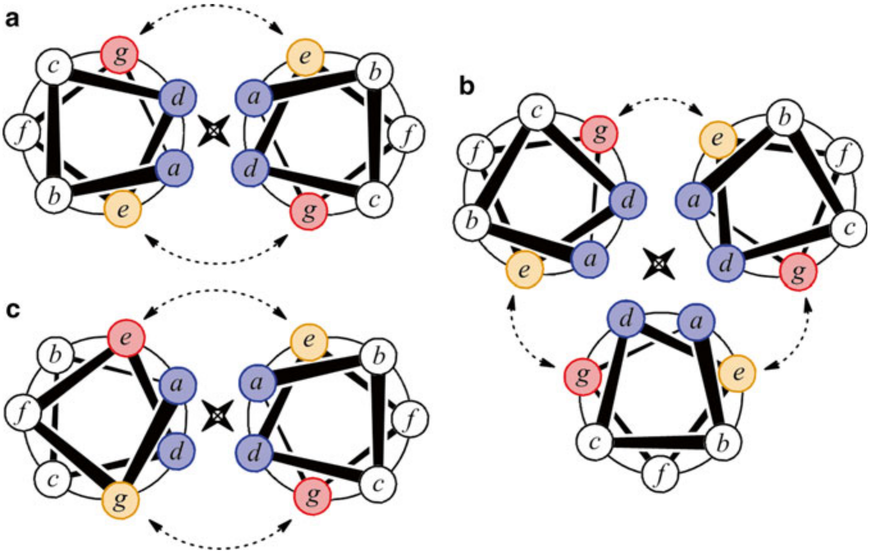


Fig. 2 Helical wheel diagrams of (**a**) a parallel dimer, (**b**) a parallel trimer, and (**c**) an antiparallel dimer coiled coil. *Blue*: hydrophobic core residues, *red/orange* oppositely charged salt-bridging residues, *white*: exterior residues, *solid arrows*: hydrophobic interactions, *dashed arrows*: electrostatic interactions

2.1.2 Salt Bridges and Aggregate Stability

The residues directly adjacent to the hydrophobic core in the *e* and *g* positions (*see* Fig. 2) are important for both helix specificity and the overall stability of the coiled coil. The effects the *e* and *g* sites have on the oligomerization number are briefly discussed in Subheading 2.2. The *e* residue of one α-helix in a coiled coil is located in close proximity to the *g* residue of an adjacent strand and is therefore able to participate in a $g_n{:}e_{n+1}'$ interaction (*see* Fig. 3). Introduction of oppositely charged residues in these positions (e.g., Lys:Glu) stabilizes the coiled coil structure through the formation of favorable salt bridges. It has been reported that each salt bridge contributes to the stability of the coiled coil by approximately 1.5 kJ mol^{-1} [11]. Extension of the coiled coil with additional heptads (i.e., greater n) yields scaffolds with increased stability, due to both the greater number of favorable inter-helical salt bridges and hydrophobic layers within the core [12, 13].

2.1.3 Exterior Residues

The remainder of the heptad residues (*b*, *c*, and *f*) form the exterior of the coiled coil (*see* Fig. 2). Two types of residues are often incorporated into these sites; water solubilizing groups and α-helix promoters. Scholtz and Pace assessed the helix propensity of solvent-exposed residues in α-helices (defined as the frequency of occurrence), of nineteen natural amino acids. Pro was excluded from the study as it is known to distort α-helices and is more commonly found in turn motifs [14]. Ala has the highest propensity, with Leu, Met, Lys, Gln, Glu, and Ile also strongly favoring a

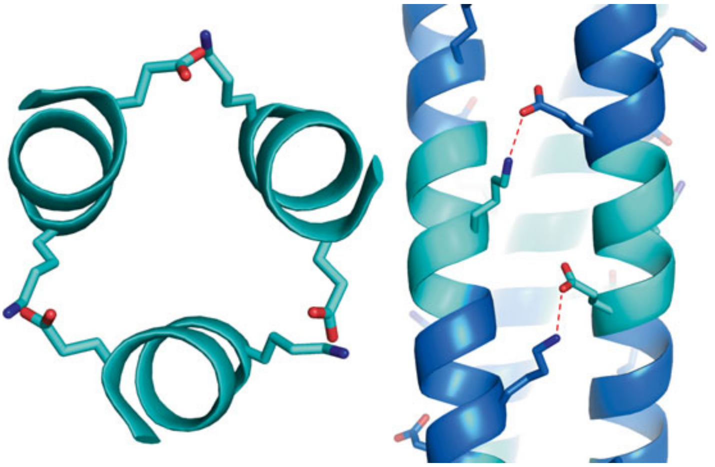


Fig. 3 PyMOL representations of a portion of the homotrimer CC-pII (pdb 4DZL [16]), with *e* and *g* side chains forming the favorable *e*:g_{n+1}′ electrostatic interactions shown from both a top-down (*left*) and side-on (*right*) view. Alternate heptads (*a*–*g*) are colored *light* and *dark grey*

helical structure, with Gly the lowest [14, 15]. This is reflected in many de novo designed coiled coils, with Ala, Lys, Gln, and Glu often populating the *b*, *c*, and *f* sites [8].

Finally, residues providing functionality may be incorporated into the coiled coil structure using the exterior *f* site, as this site is not directly involved in any coiled coil forming interactions. Groups can be introduced which aid in the characterization of the coiled coil, for example using Trp (UV chromophore) and 4-iodo-L-phenylalanine (X-ray structure determination) [16], or by locating reactive sites for the coupling of non-native functionality. The latter can include azides for Cu(I) catalyzed 3 + 2 cycloaddition reactions and alkynes for thiol-ene click reactions [17].

2.2 Controlling the Number of α-Helices

The dominating factor in influencing the number of α-helices in a coiled coil, is the steric packing of side chains within the hydrophobic core. Therefore, to design a coiled coil with a specific oligomerization number, one can take advantage of a set of guidelines based on which residues are located in the *a* and *d* sites, and may consist of both positive (stabilizing a specific oligomer) or negative design (destabilizing other oligomers). The most commonly used guidelines are based on the highly branched Ile and Leu residues, where the combination *a* = Ile, *d* = Leu tends to form a dimer; *a* = *d* = Ile or Leu a trimer; and *a* = Leu, *d* = Ile a tetramer [8]. This is due to the different orientation of side chains when located in an *a* or *d* site, and hence the reversal of *a* and *d* residues can induce an alternative oligomer [8]. A recent review by Klok and coworkers provides an excellent summary of the various oligomers and the residues most commonly observed in the corresponding *a* and *d* sites [7].

These design principles should be viewed only as guidelines, as other residues or sites in the sequence are also capable of influencing the coiled coil oligomerization number. For example, placing polar residues (e.g., Lys) in an *a* or *d* site can destabilize higher oligomerization numbers, as the ability of the polar side chain to protrude from the core and interact with solvent at the α-helical interface is least hindered in dimers [7]. Similarly, higher oligomers (≥5) can be achieved by the incorporation of hydrophobic residues in the *b*, *c*, *e* and *g* sites, so as to minimize contact area with the solvent [7], or by construction of the coiled coil around the bulkier Phe or Trp residues [9, 10]. Woolfson and coworkers have reported an analysis of coiled coil structures in the Protein Data Bank and the occurrence of residues in the *a* and *d* sites. Notably, they found that the presence of Asn in an *a* site favors the formation of a dimer, whereas Asn in a *d* site favors a trimeric structure [16].

2.3 Helix Selectivity and Orientation

The examples described so far correspond to parallel homo-coiled coils, which are coiled coil structures with identical α-helices all orientated in the same direction (*see* Fig. 1). These scaffolds are therefore only capable of providing a symmetric metal ion coordination environment. In contrast non-symmetrical coordination environments can be achieved using antiparallel coiled coils, hetero-coiled coils, or α-helical bundles.

Both α-helix selectivity and orientation are largely dependent on the ionic and polar interactions between the *e* and *g* positions. Placement of same-charged residues in all *e* and g sites of a heptad destabilizes a homo-coiled coil, by preventing formation of interhelical salt bridges and through electrostatic repulsion of the peptide strands. Only on introduction of the complementary (oppositely charged) strand will the favorable g_n:e_{n+1}' interactions form on assembly of a hetero-coiled coil (*see* Fig. 4a) [18]. However, as this approach relies upon oppositely charged α-helices, only even numbered structures (e.g., dimers and tetramers) are accessible. Hetero-trimer designs are still accessible by *e* and *g* site recognition by engineering α-helices with complementary electrostatic patterns throughout the coil (*see* Fig. 4b) [19]. This principle may be extended to the design of antiparallel coiled coils by introduction of favorable electrostatic *e*:*e*′ and *g*:*g*′ interactions between α-helices [7, 8].

The orientation and selectivity of α-helices can also be modified by disrupting the “knobs into holes” packing of the hydrophobic core. The creation of a cavity within the hydrophobic core, for example through introduction of a single Ala residue in GCN4-p1, has been reported to sufficiently destabilize the parallel dimer, resulting in the formation of an antiparallel trimer so as to minimize cavity formation (Fig. 4c) [20]. In contrast, the parallel trimer was obtained in the presence of benzene, which docks into the cavity [21]. The “knobs into holes” approach can be expanded

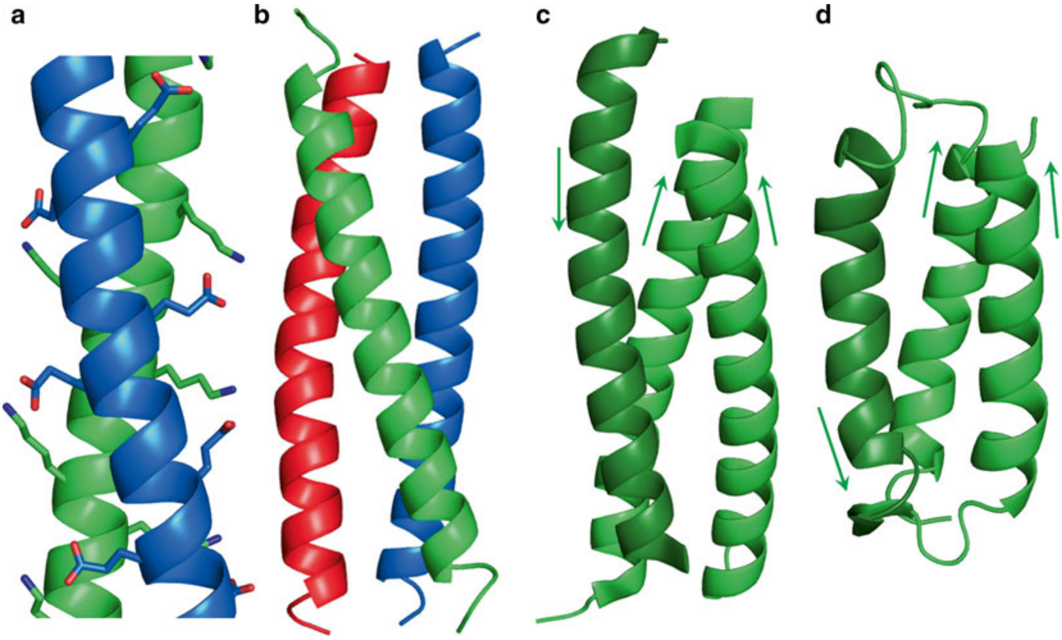


Fig. 4 PyMOL representations of examples of helix selectivity and antiparallel orientation: (**a**) portion of an acid–base (*blue* and *green*, respectively) parallel heterodimer with same-charged residues (shown) in the *e* and *g* sites of the helices forming a complementary pair (pdb 1KD8 [18]), (**b**) an A–B–C parallel heterotrimer, designed by the specific matching of *e* and *g* residues in each helix (pdb 1BB1 [19]), (**c**) an antiparallel homotrimer, resultant from the creation of a hydrophobic core cavity using Ala in a parallel homodimer, where an antiparallel trimer configuration reduces the cavity size (pdb 1RB5 [20]) and (**d**) solution NMR structure of a three-stranded helical bundle, α_3D (pdb 2A3D [23]). Shown are main chain atoms as ribbons or unfolded loops. Side chains (apart from those in **a**) and solvent molecules have been omitted for clarity. *Arrows* (towards N terminus) indicate helix direction in the antiparallel structures

to engineer complementary α-helices for the design of hetero-coiled coils. Tanaka and co-workers demonstrated this approach by modification of a trimeric coiled coil to incorporate an Ala layer (IZ-2A); the matching of this α-helix with a complimentary Trp substituted peptide (IZ-2W) resulted in preferential packing as an AAB trimer with an Ala-Ala-Trp contact [22].

Hetero-parallel and antiparallel coiled coils offer a key advantage over homo-parallel coiled coils—the ability to introduce an asymmetrical metal ion coordination environment. In order to circumvent the issue of the self-assembly of monomer units, an alternative approach to generate an asymmetric ligand environment is to use α-helical bundles, such as the de novo three helix bundle, α_3D (Fig. 4d), prepared by DeGrado and coworkers [23]. These are single peptide sequences which cooperatively fold to generate a helical bundle, based on helix-loop-helix motifs, in a clockwise or anticlockwise arrangement of successive α-helices [24]. Pecoraro and coworkers incorporated a heavy metal binding site into the hydrophobic core of the anticlockwise α_3D scaffold (Fig. 4d). A trigonal Cys site was introduced towards the *C* terminal end as this is known to tolerate residue mutations. The peptide, α_3DIV was demonstrated to bind Cd^{2+}, Hg^{2+}, and Pb^{2+} with high affinity [25]. Future studies focused on this peptide are likely to include the development of α-helical bundles with asymmetric binding sites.

3 Metallo-coiled Coils

A peptide scaffold is selected depending on the symmetry required. For example, a metal ion coordination site with threefold symmetry, would require a three stranded parallel homo-coiled coil. A square planar metal ion coordination geometry with a mixed nitrogen/sulfur coordination sphere is likely to require a four stranded parallel hetero-coiled coil or a four stranded antiparallel coiled coil. Once a scaffold has been selected, the sequence can be altered to introduce ligands capable of coordinating the metal ion. Here, we will discuss the methods by which a metal-binding site can be engineered into the primary sequence, including the location of the site within the scaffold, paying particular attention to those generated in the hydrophobic core of coiled coils. Examples will be used to highlight the advantages of coiled coils as ligands for metal ions, where specific coordination geometries are achieved which may not be accessible using small molecule ligands in aqueous solution.

3.1 Ligand Selection

3.1.1 Natural Amino Acids

Of the twenty naturally occurring amino acids, a number are capable of metal ion coordination via their side chains. Those that are most often found in the primary coordination sphere of metals in biological systems are the imidazole of His, the carboxylate of Asp and Glu and the thiolate of Cys. Other residues which participate in metal ion coordination include Asn, Thr, Ser, Tyr, Met, and Gln, but to a lesser extent.

A study of the Protein Data Bank for the most commonly occurring metals in biological systems (with the exception of Na and K) quantified the occurrence of each of the amino acids in the primary coordination sphere of these metal ions, and is summarized in Table 1 [26]. This provides a useful guide for the selection of amino acids for a specific metal ion. For example, the soft thiolate group of Cys is often found coordinated to soft metals (e.g., Cd^{2+}) and the hard carboxylate of Asp/Glu binds to hard metals (e.g., Mn^{2+} and Mg^{2+}). The imidazole of His is ubiquitous, in that it is observed coordinated to both hard and soft metal ions.

It is therefore unsurprising that these residues have been used to coordinate metal ions to de novo designed peptide scaffolds. For example, Cu^{2+}, Ni^{2+}, and Zn^{2+} have been coordinated to His residues within the hydrophobic core of trimeric coiled coils [27, 28]. The $Zn(His)_3$ site will be discussed in more detail later [28].

Table 1
Ranking of the occurrence of amino acids in the primary coordination sphere of the most commonly found biological metal ions (except Na and K) in the PDB (cutoff limit at 5 %) [26]

M^{n+}	Coordinating amino acid
Cu^{2+}	His >> Cys > Met
Fe^{2+}	His >> Glu ≈ Cys ≈ Asp ≈ Met
Ni^{2+}	His >> Cys ≈ Glu ≈ Asp
Zn^{2+}	Cys > His >> Asp ≈ Glu
Cd^{2+}	Cys > His ≈ Glu ≈ Asp
Ca^{2+}	Asp >> Glu
Co^{2+}	His > Asp ≈ Glu > Cys
Mn^{2+}	Asp > His > Glu
Mg^{2+}	Asp > Glu ≈ His > Asn

Hard metals, such as Ca^{2+} and Sr^{2+} have been shown to coordinate to Glu and Gln residues located in the core of coiled coils (*see* also Fig. 7c) [29], and Cys residues have been used to coordinate a range of soft metals, including $[AuPEt_3]^+$ [30], Pb^{2+} [31], Cd^{2+} [32, 33], and Hg^{2+} [34], some of which will be discussed in more detail in Subheadings 3.3 and 4.

3.1.2 Non-natural Amino Acids

The variety of binding sites that can be engineered into de novo designed structural motifs is not limited to those accessible to nature. Solid-phase peptide synthesis allows non-natural amino acids to be readily included in the design, and these can possess side chains which closely resemble ligands used in traditional coordination chemistry. For example the 2-amino-3-(2,2′-bipyridyl)propanoic acids (Fig. 5a) provides a 2,2′-bipyridine side chain capable of binding a range of metal ions [35]. Incorporation of this residue in a β-hairpin structure resulted in a preorganized tetrahedral $(bipy)_2$ cage capable of binding Zn^{2+} ions. Although any ligand of choice can theoretically be incorporated via the side chain of an amino acid, the steric restrictions associated with the group needs to be carefully considered on its introduction into the peptide scaffold.

Non-natural amino acids can more closely resemble natural amino acids. For example, penicillamine (Pen, *see* Fig. 5b) can be viewed as the bulkier analogue of Cys, with methyl groups in place of the β-methylene protons, or alternatively as the thiol containing analogue of Val [36]. The use of Pen in designing a metal binding site is discussed in the case study in Subheading 4. To accommodate the high coordination numbers preferred by lanthanide ions,

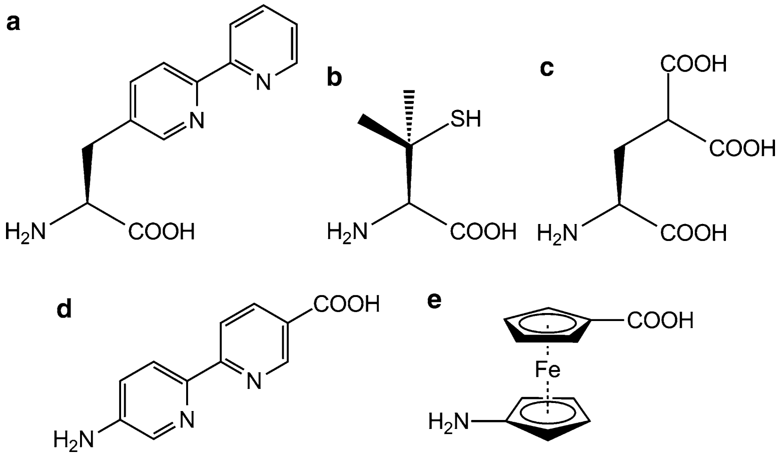


Fig. 5 Some examples of non-natural amino acids used for metal ion binding in de novo peptide design. This can be achieved through the non-natural amino acid side chain: (**a**) 2-amino-3-(2,2′-bipyrid-5-yl)propanoic acid, (**b**) L-PENICILLAMINE (Pen), and (**c**) γ-carboxyglutamic acid (Gla); or by directly incorporating the ligand into the peptide backbone: (**d**) 5′-amino-2,2′-bipyridine-5-carboxylic acid, and (**e**) 1′-aminoferrocene-1-carboxylic acid

γ-carboxyglutamic acid (Gla, Fig. 5c), a derivative of Glu has been employed for the binding of Ln^{3+} ions to coiled coils [37, 38]. Gla has also been used to replace a hydrophobic core with a Ca^{2+} based zipper, by introducing it in all the *a* and *d* sites in the peptide sequence [39].

An alternative approach involves incorporating the ligand directly into the peptide backbone, for example using 5′-amino-2,2′-bipyridine-5-carboxylic acid (Fig. 5d) [40]. Two or more of these residues in a sequence can enforce a peptide conformational change on metal ion coordination [41]. Alternatively, a metal complex can itself be integrated into the peptide backbone, for example 1′-aminoferrocene-1-carboxylic acid (Fig. 5e) [42]. However, the introduction of residues with geometries which are incompatible with those found in the α-helix restricts their use to other peptide scaffolds.

3.2 Metal Ion Site Location

3.2.1 The N-Terminus

A coiled coil can accommodate a metal ion in a variety of different locations along its structure. Solid-phase peptide synthesis provides a convenient opportunity to include a ligand group via a peptide bond at the N-terminus. Considerable work has focused on 5-carboxylic acid-2,2′-bipyridine as a capping group providing a chelating ligand for metal ion coordination [43, 44]. When in a parallel homotrimer, a preorganized octahedral metal ion binding site is generated which is capable of binding metal ions such as Fe^{2+}, Co^{2+}, Ni^{2+}, and Ru^{2+} as $[M(\text{bipy-peptide})_3]^{n+}$ [45, 46]. Artificial homo-dimerization of α-helical bundles was achieved by coupling of a single 5,5′-dicarboxylic acid-2,2′-bipyridine to two peptides at the N-terminus, to generate bipy(peptide)$_2$. This ligand has subsequently been coordinated to metals, such as Ru^{2+} to generate [Ru(bipy)$_2${bipy(peptide)$_2$}]$^{2+}$ [47]. A metal coordinating group at the N-terminus can be incorporated by a range of different reactions. For example a chelating hydroxy-phenyl oxazoline group capable of binding lanthanides, was introduced at the N-terminus of a trimeric coiled coil, through the Cu(I) catalyzed 3 + 2 cycloaddition reaction [48].

3.2.2 The f Site and Inter-helical Interface

As described earlier, the *f* site can readily be altered, often without affecting the coiled coil structure, and therefore ligands suitable for metal ion coordination can be introduced at this position. However, metal ion coordination to these sites is still capable of influencing the coiled coil structure. Ogawa and coworkers prepared a coiled coil with 4-pyridylalanine in the *f* site of two consecutive heptads, which in the absence of metal adopted a dimeric structure [49]. On addition of a [Pt(en)]$^{2+}$ moiety (where en = ethylenediamine), intra-strand f_n:f_{n+1}' metallo-bridges are formed inducing a tetrameric structure, which achieves a favorable metal complex geometry whilst minimizing the f_n:f_{n+1}' distance.

Metal ions have also been coordinated to residues located at the *e* and *g* sites. Hodges and co-workers reported a disulfide bridged two-stranded peptide where the electrostatic repulsions of Gla or Glu in the *e* and *g* positions destabilized the folding of the peptide [38, 50]. However, folding was induced on coordination of La^{3+} or Yb^{3+} to the Gla/Glu (located in the *e* and *g* sites) and neutralization of the electrostatic repulsions.

3.2.3 Designing a Metal Ion Binding Site Within the Hydrophobic Core

Among the first examples of metal complexation within a de novo α-helical structure was the introduction of heme centers. Placement of His residues in the interior of four helix bundles, consisting of a pair of two helix bundles (helix-turn-helix) dimerized by a disulfide bridge, allowed the preparation of mono-heme and multi-heme maquettes [51, 52], and further optimization of this design resulted in functional heme bundles capable of binding O$_2$ [53].

This design strategy has since been extended to the introduction of a metal ion binding site within the hydrophobic core of a coiled coil. Replacement of a single hydrophobic amino acid located in either an *a* or *d* site with a residue capable of coordinating metal ions, should result in the formation of a preorganized metal ion binding site in a parallel homo-coiled coil. The degree of oligomerization can be exploited to influence the number of ligating residues located in the binding site, and therefore the coordination number available to the metal ion. A dimer, trimer, and tetramer coiled coil can easily offer linear, trigonal-planar, and square-planar ligand geometries, respectively. The examples discussed below focus primarily on trimers, which can be used to achieve trigonal and tetrahedral coordination geometries.

The metal-coordinating ligand can be introduced into either an *a* or *d* site, but will be orientated differently in each of these. In a parallel homotrimer the side chain of the *d* residue is directed towards the center of the coiled coil and may at first appear ideal for ligand placement [8]. However, the naturally occurring amino acids used most often as metal ligands in coiled coils are functionalized on the γ-carbon (e.g., Cys, Asp, and His) and therefore the coordinating group can be directed away from the center of the super-helix, and appears to require preorganization prior to metal ion binding. In contrast the *a* substituted site appears ideal for metal ion coordination; the γ-carbon is directed away from the coiled coil center, in return prepositioning the coordinating group for optimal metal ion coordination. This is reflected in the crystal structures of the trimeric peptides $(CSL9C)_3$ and $(CSL19C)_3$, where a Cys residue is introduced into an *a* and *d* site, respectively (*see* Table 2 and Fig. 6) [54].

The location of the metal ion binding site within the coiled coil (i.e., which heptad), can also influence the coordination sphere of the metal. Proximity to the N- or C-terminus provides an opportunity for improved solvent access to the metal, as fewer hydrophobic residues between the metal and terminus result in fraying of the coiled coil. Alternatively, locating the metal in the center of the coiled coil minimizes solvent accessibility and reduces the flexibility of the metal cavity. The case study (Subheading 4) describes the preparation of a metal ion site in which an exogenous solvent molecule contributes to the primary coordination sphere.

Table 2
Parallel homotrimer coiled coil peptide sequences discussed in this chapter

Peptide	Sequence *a b c d e f g*
CoilSer	Ac-E WEALEKK LAALESK LQALEKK LEALEHG-NH_2
CSL9C	Ac-E WEALEKK CAALESK LQALEKK LEALEHG-NH_2
CSL19C	Ac-E WEALEKK LAALESK LQACEKK LEALEHG-NH_2
CSL9PenL23H	Ac-E WEALEEK PenAALESK LQALEKK HEALEHG-NH_2
TRI	Ac-G LKALEEK LKALEEK LKALEEK LKALEEK G-NH_2
TRIL9C	Ac-G LKALEEK CKALEEK LKALEEK LKALEEK G-NH_2
TRIL12C	Ac-G LKALEEK LKACEEK LKALEEK LKALEEK G-NH_2
TRIL16C	Ac-G LKALEEK LKALEEK CKALEEK LKALEEK G-NH_2
TRIL9CL23H	Ac-G LKALEEK CKALEEK LKALEEK HKALEEK G-NH_2
TRIL16Pen	Ac-G LKALEEK LKALEEK PenKALEEK LKALEEK G-NH_2
TRIL12AL16C	Ac-G LKALEEK LKAAEEK CKALEEK LKALEEK G-NH_2
TRIL12L_DL16C	Ac-G LKALEEK LKAL$_D$EEK CKALEEK LKALEEK G-NH_2

Pen = L-Penicillamine, L_D = D-Leucine

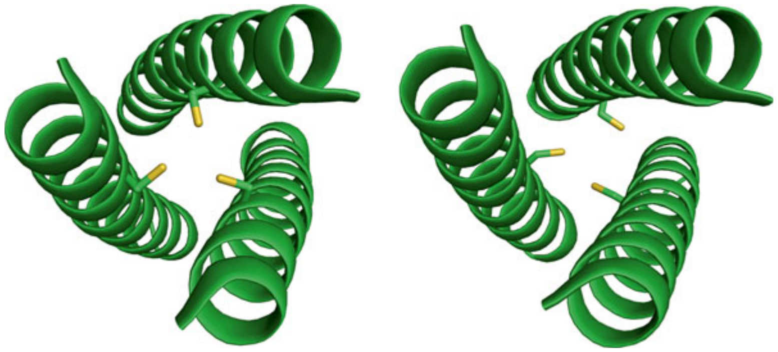


Fig. 6 PyMOL representations of parallel homotrimers with Cys introduced in an *a* (CSL9C, *left*, major form of pdb 3LJM) or a *d* site (CSL19C, *right*, pdb 2X6P) [54]. Placement of Cys in the *a* site appears more preorganized for metal ion binding compared to the *d* site, where the ligand groups are directed away from the center of the coiled coil. Shown are main chain atoms as ribbons and the Cys side chain in stick form (thiol in *orange*). All remaining side chains and solvent molecules have been omitted for clarity

3.3 Select Examples of Hydrophobic Core Binding Sites

3.3.1 TRI and Mercury: Access to Unusual Coordination Geometries

The TRI peptide family (Table 2) exhibit pH dependant aggregation, with dimers dominating at pH <5 and trimers at pH values >6 [55]. This peptide was modified so as to introduce a soft metal ion, including Hg^{2+}, binding site. This was achieved by substituting a Leu residue located in either an *a* or *d* site with Cys (e.g., TRIL12C and TRIL16C, Table 2), to create a preorganized trigonal thiol binding site within the hydrophobic core [56]. The coordination of Hg^{2+} to the *a* and *d* sites was reported to be slightly different, and is discussed in more detail in the literature [56].

Hg^{2+} binding to TRILXC was reported to enhance the stability of the resulting coiled coil. Intriguingly the oligomeric state of the TRILXC peptide was dependent on the peptide–Hg^{2+} ratio [56]. A dimer was exclusively obtained at a ratio of 2:1 at both low and high pH values, resulting in linear HgS_2, thus illustrating the preference of Hg^{2+} to form linear complexes with thiols. Only on addition of further peptide, giving a ratio of 3:1, was a pH-dependent trimeric structure obtained. This yielded a fully three-coordinate, trigonal-planar Hg^{2+} thiolate complex at high pH (>8.5), as characterized by ^{199}Hg NMR, UV-visible, and ^{199m}Hg Perturbed Angular Correlation (PAC) spectroscopy [57]. This represented the first water-stable trigonal planar HgS_3 spectroscopic model for the binding of Hg^{2+} to MerR, a Hg^{2+} metalloregulatory protein [34].

3.3.2 CoilSer and Arsenic: A Potential ArsR Structural Model

The trigonal thiolate binding site within the hydrophobic core of TRI peptides (Table 2), has also been investigated for the coordination of other soft metals, for the purpose of creating structural models for heavy metal ion binding (based on $M(SR)_3$) in proteins. As^{3+} coordination to TRIL9C (*see* Table 2) results exclusively in the formation of As(TRIL9C)$_3$, which was proposed to be a good model for As^{3+} bound to the repressor protein ArsR [58]. However, it proved challenging to fully characterize the As^{3+} coordination geometry. Therefore a Leu in position 9 (*a* site) of CoilSer (*see* **Note 4**) was replaced with Cys to generate CSL9C, and was reacted with As^{3+} to form As(CSL9C)$_3$ [59]. Crystals of sufficient quality were obtained and the resulting X-ray crystal structure showed that the As^{3+} ion is coordinated to the Cys side chains in an *endo* position. The metal is situated *below* the thiolate groups and level with the methylene side chain of Cys, and the As^{3+} lone pair is directed towards the C-terminus (Fig. 7a). This was in stark contrast to small molecule models of As(SR)$_3$, which had previously suggested that the As^{3+} adopts an *exo* configuration, where the ion is located above the thiolate groups of a trigonal S$_3$ binding site. This example therefore demonstrates that studying the binding of metal ions within coiled coils, may provide an alternative, and possibly even more suitable model for metal ion coordination in larger protein structures.

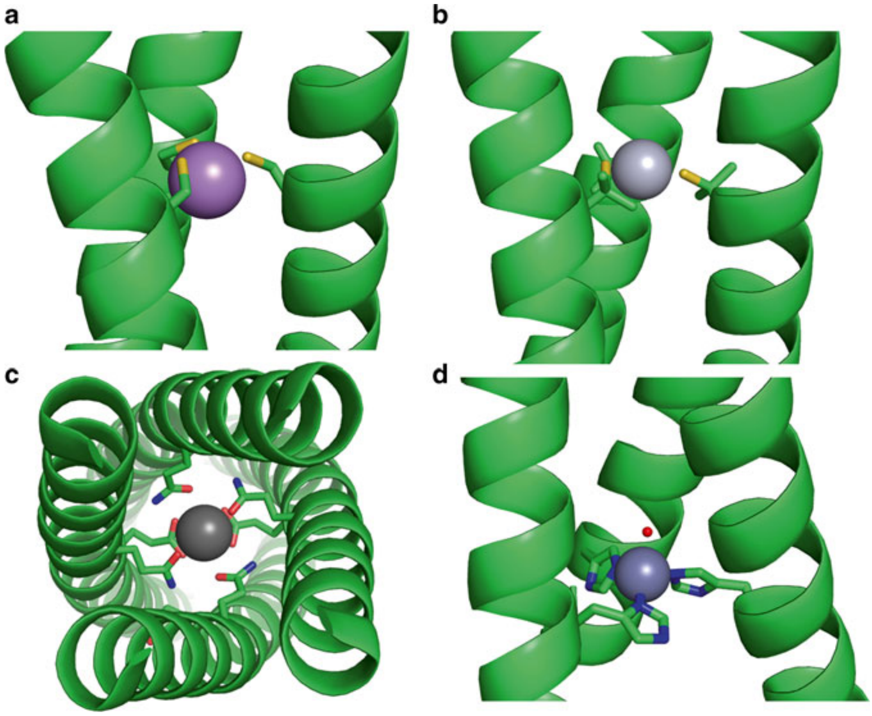


Fig. 7 PyMOL representations of metal-ion binding sites designed in the hydrophobic core of homo-parallel coiled coils. (**a**) As^{3+} (grey) binding in As(CSL9C)$_3$ (pdb 2JGO [59]), (**b**) and (**d**) Hg^{2+} (**b**, *light grey*) and Zn^{2+} (**d**, *grey*) binding in (CSL9PenL23H)$_3$, respectively (pdb 3PBJ [28]) and (**c**) octahedral Ca^{2+} (*dark grey*) coordination geometry (pdb 2O1K [29]). Shown are main chain atoms as ribbons. Side chains (except metal-binding residues which are shown in stick form) and solvent molecules have been omitted for clarity

3.3.3 A Functional Zinc Carbonic Anhydrase Mimic

A logical evolution to the use of coiled coils as structural models is the development of functional models, for example by incorporation of a catalytic site within the scaffold. The Zn(His)$_3$O catalytic site of carbonic anhydrase was selected by Pecoraro and coworkers for its high catalytic activity [28]. The three-fold symmetry meant that a parallel homotrimer, TRI, was selected as the scaffold. His residues were introduced into the hydrophobic core by replacing the Leu at position 23. This represents an *a* site located in close proximity to the C-terminus. The opportunity to expand at the C-terminus was thought to be necessary in order to accommodate the bulkier Zn(His)$_3$ site. The introduction of His was anticipated to be destabilizing and so a Hg(Cys)$_3$ site was introduced at position 9, located toward the N-terminus, in an effort to confer additional stability to the resulting scaffold (TRIL9CL23H, Table 2). The binding of Zn^{2+} and Hg^{2+} to the peptide was confirmed by X-ray crystallography with the related CoilSer derivative (CSL9PenL23H, Table 2), *see* Fig. 7b, d [28]. Most importantly, the Zn(His)$_3$ site is structurally extremely similar to the Zn^{2+} active site in carbonic anhydrase, whereas small molecule models of Zn(His)$_3$O often suffer from dinuclear degradation in aqueous solution. The peptide model is also able to mimic the catalytic activity (activity of *p*-nitrophenyl acetate hydrolysis and CO$_2$ hydration were reduced by only 100- and 500-fold respectively compared to carbonic anhydrase II). The [HgZn(TRIL9CL23H)$_3$]$^{n+}$ complex, along with the CoilSer analogue, are the first reported de novo designed hydrolytic metalloenzymes.

4 Case Study: Design of CdS$_3$ and CdS$_3$O Sites

The TRI peptides described in Subheading 3.3, fold to generate parallel homotrimer scaffolds. A Cys layer can be introduced within the hydrophobic core, resulting in the formation a preorganized trigonal thiolate site, capable of coordinating heavy metal ions such as As^{3+} and Hg^{2+}. Similar binding studies were subsequently performed with Cd^{2+} and TRIL16C, where TRIL16C has the Leu at position 16 (*a* site) replaced with Cys, to form Cd(TRIL16C)$_3^-$ (as confirmed by UV-visible titrations). Efforts were then directed towards fully characterizing the resulting complex and specifically the Cd^{2+} coordination site. The UV-visible spectrum displayed the characteristic Cd-S ligand-to-metal charge transfer (LMCT) at 232 nm, with an extinction coefficient of 22 600 M^{-1} cm^{-1} [60]. A key advantage of cadmium is that it is a spectroscopically rich metal ion and experiments performed with isotopically enriched ^{113}Cd(NO$_3$)$_2$ allowed ^{113}Cd(TRIL16C)$_3^-$ to be probed by ^{113}Cd NMR. This resulted in a single resonance with a chemical shift of 625 ppm [60], which falls well within the range reported for CdS$_3$ (CdS$_3^-$ 570–660 ppm) [32]. Similarly EXAFS data was best fit to three sulphurs bound to the Cd^{2+} with a bond distance of 2.49 Å [60]. Again this value is close to those previously reported which range from 2.42 to 2.48 Å [32]. The data obtained so far could therefore be consistent with the intended CdS_3; however, it could also apply to a CdS_4 or CdS_3X (where X=O or N) coordination sphere. ^{111m}Cd perturbed angular correlation (PAC) spectroscopy was subsequently employed to characterize the Cd^{2+} coordination environment. Intriguingly these results could be fit to a mixture of two species; 40 % CdS_3 (as intended) and 60 % CdS_3X, where X=an exogenous water (solvent) molecule. A single resonance is observed for these two species in the ^{113}Cd NMR spectrum as the water exchange occurs rapidly on the NMR timescale (ms) but the two species can be observed separately on the faster PAC timescale (ns).

Cd^{2+} binding to a trigonal Cys site in $(TRIL16C)_3$ had been confirmed; however, this was as a mixture of two species. It therefore was an attractive challenge to use de novo peptide design to prepare exclusive CdS_3O and CdS_3 sites, respectively. It was reasoned that by removing the steric bulk above the Cd^{2+} binding site, by replacing the Leu at position 12 with the sterically less demanding Ala residue, would generate a cavity that could accommodate a water molecule (*see* Fig. 8). Cd^{2+} binding to this peptide was investigated using both ^{113}Cd NMR and ^{111m}Cd PAC. A single resonance at 574 ppm was observed for $^{113}Cd(TRIL12AL16C)_3^-$, and this was confirmed by ^{111m}Cd PAC to be due to a single species consistent with a tetrahedral CdS_3O site [61, 62].

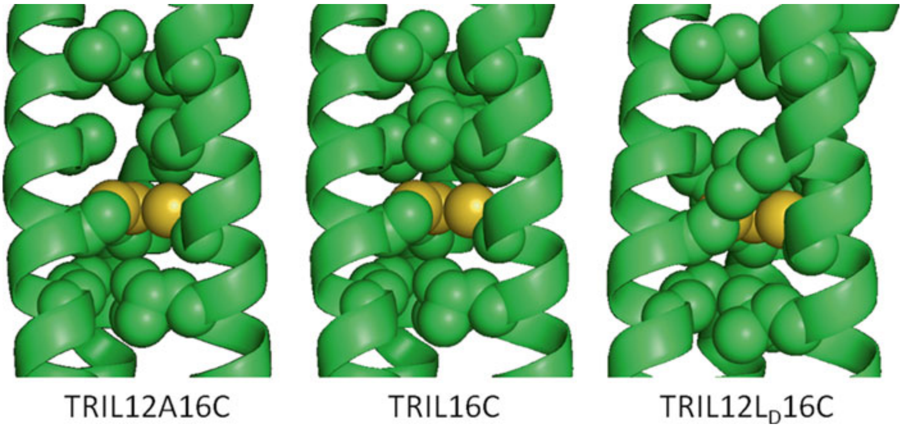


Fig. 8 PyMOL models of parallel homotrimers (based on TRIL16C) containing a single trigonal Cys site within the hydrophobic core. A cavity is formed by replacing a leucine at position 12 with an Ala in TRIL12AL16C, which can increase the hydration state of a coordinated Cd^{2+} ion. In contrast, the L- to D-Leu substitution at position 12 in TRIL12L$_D$L16C, repositions the side chain towards the Cd^{2+} binding site, excluding the coordination of exogenous water molecules. Shown are main chain atoms as ribbons, and the internal residue (9, 12, 16, and 19) side chains as *spheres*, carbons (*green*), and thiols (*yellow*)

Attempts were made to generate a peptide scaffold capable of coordinating Cd^{2+} as exclusively CdS_3, however, this proved to be more challenging. This was accomplished when non-natural residues were introduced into the design. The introduction of L-Penicillamine (Pen), the bulkier analogue of Cys or the thiol containing analogue of Val, resulted in a peptide TRIL16Pen which was still capable of folding (as confirmed by circular dichroism, CD) to generate the required α-helical coiled coil scaffold. The resulting coiled coil $(TRIL16Pen)_3$ was capable of binding Cd^{2+} as exclusively CdS_3, again confirmed by both ^{113}Cd NMR and ^{111m}Cd PAC. A single resonance at 684 ppm was observed for $^{113}Cd(TRIL16Pen)_3^-$, and a single species consistent with a trigonal planar CdS_3 site, was confirmed by ^{111m}Cd PAC [62].

Though this design was successful, an alternative approach was reported which achieved CdS_3 coordination to Cys rather than Pen. This approach again utilized a non-natural residue in the design, however, a second sphere non-coordinating residue was altered. It was reasoned that the steric bulk above the Cd^{2+} plane, a Leu residue, was directed slightly towards the C-terminus of the peptide, however on altering the chirality at the α-C (L- to D-Leu) the side chain would be positioned towards the opposite N-terminus and therefore directly towards the Cd^{2+} binding site, preventing water coordination (*see* Fig. 8). CD studies again confirmed that the inclusion of the non-natural residue had a negligible effect on the structure of the coiled coil. Cd^{2+} binding to the resulting peptide TRIL12L_DL16C (L_D = D-Leu) was again probed by ^{113}Cd NMR (697 ppm) and ^{111m}Cd PAC, and confirmed to be 100 % trigonal planar CdS_3 [63]. This work was further developed when a single parallel homotrimer was designed with two Cd^{2+} binding sites. These both consisted of a Cys layer in an *a* site, differing only in the nature of the residue directly above the Cd^{2+} coordination site. Introduction of an Ala and a D-Leu residue in these locations, resulted in a coiled coil capable of coordinating one Cd^{2+} as tetrahedral four-coordinate CdS_3O and a second as trigonal planar CdS_3 [63].

These examples clearly demonstrate the huge potential of de novo designed peptide scaffolds, and particularly coiled coils, as ligands for metal ions. Not only is the symmetry of the peptide scaffold, the identity and location of the coordinating residue important for determining the metal ion coordination geometry, but the nature of a non-coordinating second sphere residue can be important. This case study demonstrates that these second sphere residues can be exploited in order to generate a four coordinate site within a three stranded coiled coil with an exogenous fourth ligand, CdS_3O (Ala), or alternatively a coordinatively unsaturated CdS_3 (D-Leu) site.

5 Notes

1. On introduction of metal binding residues within a peptide scaffold, attention must be paid to the nature of the primary amino acid sequence of the scaffold, so as to avoid/reduce competition binding sites. For example, many de novo coiled coils (e.g., TRI and CoilSer) contain acidic residues (e.g., Asp or Glu) which are either important for the formation of favorable salt bridges or for solubilizing the scaffold. These sequences may therefore be less appropriate for modification to create an Asp or Glu metal-binding site within the hydrophobic core.
2. The introduction of a metal-binding site within a hydrophobic core requires the disruption of at least one hydrophobic layer. This will destabilize the coiled coil structure (can be monitored by CD), and can be further exacerbated by introducing polar and charged residues in their place [64]. Additional coiled coil stability may be required to facilitate metal binding (e.g., for the preorganization of the binding site), and can be achieved by addition of an extra heptad repeat [13], covalent tethering of the helices with disulfide bridges [50], or through the introduction of a structural metal into the coiled coil [28, 64].
3. CoilSer was one of the first de novo peptides to be characterized by X-ray crystallography, and was solved to yield an antiparallel trimeric structure [65]. This was initially thought to be due to the steric bulk of the *N*-terminal Trp residues, however, the solution chemistry of CoilSer and subsequent crystal structures of Cys and Pen containing derivatives, all suggest that the structure is in fact a parallel trimer [54, 66]. Despite the controversy surrounding CoilSer, it remains an attractive peptide scaffold to adopt as it is so amenable to crystallization studies.
4. Woolfson and co-workers have reported a very useful de novo designed "basis set" of parallel homo-coiled coils, designed to give predictable and stable structures based on a four-heptad repeat peptide with a hydrophobic core of Ile or Leu [16]. The set currently consists of a dimer, trimer, and tetramer, however, more up-to-date information about this can be found at the following Web site: http://coiledcoils.chm.bris.ac.uk/pcomp/pcomps.php.

Acknowledgment

AFAP and AJG thank the School of Chemistry at the University of Birmingham and the EPSRC (EP/J014672-1) for financial support.

References

1. Venkatraman J, Naganagowda GA, Sudha R, Balaram P (2001) De novo design of a five-stranded beta-sheet anchoring a metal-ion binding site. Chem Commun 24:2660–2661
2. Yang W, Jones LM, Isley L, Ye Y, Lee H-W, Wilkins A et al (2003) Rational design of a calcium-binding protein. J Am Chem Soc 125:6165–6171
3. Platt G, Chung CW, Searle MS (2001) Design of histidine-Zn^{2+} binding sites within a beta-hairpin peptide: enhancement of beta-sheet stability through metal complexation. Chem Commun 13:1162–1163
4. Bonomo RP, Casella L, De Gioia L, Molinari H, Impellizzeri G, Jordan T et al (1997) Metal ion and proton stabilisation of turn motif in the synthetic octapeptide histidyltris(glycylhistidyl) glycine. J Chem Soc Dalton Trans 14:2387–2389
5. Krizek BA, Merkle DL, Berg JM (1993) Ligand variation and metal-ion binding specificity in zinc finger peptides. Inorg Chem 32:937–940
6. Ende CWA, Meng HY, Ye M, Ye M, Pandey AK, Zondlo NJ (2010) Design of lanthanide fingers: compact lanthanide-binding metalloproteins. Chembiochem 11:1738–1747
7. Apostolovic B, Danial M, Klok H-A (2010) Coiled coils: attractive protein folding motifs for the fabrication of self-assembled, responsive and bioactive materials. Chem Soc Rev 39:3541–3575
8. Woolfson DN (2005) The design of coiled-coil structures and assemblies. In: Parry D, Squire J (eds) Fibrous proteins: coiled-coils, collagen and elastomers, vol 70, 1st edn. Elsevier and Academic, Boston, MA, pp 79–112, Adv Protein Chem
9. Liu J, Yong W, Deng YQ, Kallenbach NR, Lu M (2004) Atomic structure of a tryptophan-zipper pentamer. Proc Natl Acad Sci U S A 101:16156–16161
10. Liu J, Zheng Q, Deng Y, Kallenbach NR, Lu M (2006) Conformational transition between four and five-stranded phenylalanine zippers determined by a local packing interaction. J Mol Biol 361:168–179
11. Zhou NE, Kay CM, Hodges RS (1994) The role of interhelical ionic interactions in controlling protein-folding and stability—de novo designed synthetic 2-stranded alpha-helical coiled-coils. J Mol Biol 237:500–512
12. De Crescenzo G, Litowski JR, Hodges RS, O'Connor-McCourt MD (2003) Real-time monitoring of the interactions of two-stranded de novo designed coiled-coils: effect of chain length on the kinetic and thermodynamic constants of binding. Biochemistry 42:1754–1763
13. Su JY, Hodges RS, Kay CM (1994) Effect of chain-length on the formation and stability of synthetic alpha-helical coiled coils. Biochemistry 33:15501–15510
14. Pace CN, Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 75:422–427
15. Strehlow KG, Robertson AD, Baldwin RL (1991) Proline for alanine substitutions in the C-peptide helix of ribonuclease-A. Biochemistry 30:5810–5814
16. Fletcher JM, Boyle AL, Bruning M, Bartlett GJ, Vincent TL, Zaccai NR et al (2012) A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology. ACS Synth Biol 1:240–250
17. Mahmoud ZN, Gunnoo SB, Thomson AR, Fletcher JM, Woolfson DN (2011) Bioorthogonal dual functionalization of self-assembling peptide fibers. Biomaterials 32: 3712–3720
18. Keating AE, Malashkevich VN, Tidor B, Kim PS (2001) Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. Proc Natl Acad Sci U S A 98:14825–14830
19. Nautiyal S, Alber T (1999) Crystal structure of a designed, thermostable; heterotrimeric coiled coil. Protein Sci 8:84–90
20. Holton J, Alber T (2004) Automated protein crystal structure determination using ELVES. Proc Natl Acad Sci U S A 101:1537–1542
21. Gonzalez L, Plecs JJ, Alber T (1996) An engineered allosteric switch in leucine-zipper oligomerization. Nat Struct Biol 3:510–515
22. Kashiwada A, Hiroaki H, Kohda D, Nango M, Tanaka T (2000) Design of a heterotrimeric alpha-helical bundle by hydrophobic core engineering. J Am Chem Soc 122:212–215
23. Walsh STR, Cheng H, Bryson JW, Roder H, DeGrado WF (1999) Solution structure and dynamics of a de novo designed three-helix bundle protein. Proc Natl Acad Sci U S A 96: 5486–5491
24. Baltzer L, Nilsson H, Nilsson J (2001) De novo design of proteins—what are the rules? Chem Rev 101:3153–3163
25. Chakraborty S, Kravitz JY, Thulstrup PW, Hemmingsen L, DeGrado WF, Pecoraro VL (2011) Design of a three-helix bundle capable of binding heavy metals in a triscysteine environment. Angew Chem Int Ed 50: 2049–2053

26. Dokmanic I, Sikic M, Tomic S (2008) Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. Acta Crystallogr D Biol Crystallogr 64:257–263

27. Tanaka T, Mizuno T, Fukui S, Hiroaki H, Oku J, Kanaori K et al (2004) Two-metal ion, Ni(II) and Cu(II), binding alpha-helical coiled coil peptide. J Am Chem Soc 126:14023–14028

28. Zastrow ML, Peacock AFA, Stuckey JA, Pecoraro VL (2012) Hydrolytic catalysis and structural stabilization in a designed metalloprotein. Nat Chem 4:118–123

29. Bowman GD, Nodelman IM, Levy O, Li SL, Tian P, Zamb TJ et al (2000) Crystal structure of the oligomerization domain of NSP4 from rotavirus reveals a core metal-binding site. J Mol Biol 304:861–871

30. Peacock AFA, Bullen GA, Gethings LA, Williams JP, Kriel FH, Coates J (2012) Gold-phosphine binding to de novo designed coiled coil peptides. J Inorg Biochem 117:298–305

31. Neupane KP, Pecoraro VL (2010) Probing a homoleptic PbS_3 coordination environment in a designed peptide using Pb-207 NMR spectroscopy: implications for understanding the molecular basis of lead toxicity. Angew Chem Int Ed 49:8177–8180

32. Peacock AFA, Iranzo O, Pecoraro VL (2009) Harnessing natures ability to control metal ion coordination geometry using de novo designed peptides. Dalton Trans 13:2271–2280

33. Peacock AFA, Pecoraro VL (2013) Natural and artificial proteins containing cadmium. In: Sigel A, Sigel H, Sigel RKO (eds) Cadmium: from toxicity to essentiality, vol 11, Metal ions in life sciences. Springer Science + Business Media B.V, Dordrecht, pp 303–307

34. Pecoraro VL, Peacock AFA, Iranzo O, Iranzo O, Luczkowski M (2009) Understanding the biological chemistry of mercury using a de novo protein design strategy. In: Long E, Baldwin M (eds) Advances in inorganic biochemistry: from synthetic models to cellular systems. ACS symposium series, vol 1012, pp 183–197

35. Cheng RP, Fisher SL, Imperiali B (1996) Metallopeptide design: tuning the metal cation affinities with unnatural amino acids and peptide secondary structure. J Am Chem Soc 118:11349–11356

36. Petros AK, Shaner SE, Costello AL, Tierney DL, Gibney BR (2004) Comparison of cysteine and penicillamine ligands in a Co(II) maquette. Inorg Chem 43:4793–4795

37. Kashiwada A, Ishida K, Matsuda K (2007) Lanthanide ion-induced folding of *de novo* designed coiled-coil polypeptides. Bull Chem Soc Jpn 80:2203–2207

38. Kohn WD, Kay CM, Sykes BD, Hodges RS (1998) Metal ion induced folding of a de novo designed coiled-coil peptide. J Am Chem Soc 120:1124–1132

39. Dai Q, Dong M, Liu Z, Castellino FJ (2011) Ca^{2+}-induced self-assembly in designed peptides with optimally spaced gamma-carboxyglutamic acid residues. J Inorg Biochem 105:52–57

40. Torrado A, Imperiali B (1996) New synthetic amino acids for the design and synthesis of peptide-based metal ion sensors. J Org Chem 61(25):8940–8948

41. Rama G, Arda A, Marechal J-D, Gamba I, Ishida H, Jiménez-Barbero J et al (2012) Stereoselective formation of chiral metallopeptides. Chemistry 18:7030–7035

42. Barisic L, Rapic V, Metzler-Nolte N (2006) Incorporation of the unnatural organometallic amino acid 1′-aminoferrocene-1-carboxylic acid (Fca) into oligopeptides by a combination of Fmoc and Boc solid-phase synthetic methods. Eur J Inorg Chem 20:4019–4021

43. Doerr AJ, McLendon GL (2004) Design, folding, and activities of metal-assembled coiled coil proteins. Inorg Chem 43:7916–7925

44. Schneider JP, Kelly JW (1995) Templates that induce alpha-helical, beta-sheet and loop conformations. Chem Rev 95:2169–2187

45. Lieberman M, Sasaki T (1991) Iron(III) organizes a synthetic peptide into 3-helix bundles. J Am Chem Soc 113:1470–1471

46. Ghadiri MR, Soares C, Choi C (1992) A convergent approach to protein design—metal-ion assisted spontaneous self-assembly of a poly peptide into a triple-helix bundle protein. J Am Chem Soc 114:825–831

47. Mihara H, Nishino N, Hasegawa R, Fujimoto T, Usui S, Ishida H et al (1992) Design of a hybrid of 2 alpha helix peptides and ruthenium trisbipyridine complex for photoinduced electron-transfer system in bilayer-membrane. Chem Lett 9:1813–1816

48. Samiappan M, Alasibi S, Cohen-Luria R, Shanzer A, Ashkenasy G (2012) Allosteric effects in coiled-coil proteins folding and lanthanide-ion Binding. Chem Comm 48: 9577–9579

49. Tsurkan MV, Ogawa MY (2007) Metal-mediated peptide assembly: Use of metal coordination to change the oligomerization state of an alpha-helical coiled-coil. Inorg Chem 46:6849–6851

50. Kohn WD, Kay CM, Hodges RS (1998) Effects of lanthanide binding on the stability of

de novo designed alpha-helical coiled-coils. J Pept Res 51:9–18

51. Choma CT, Lear JD, Nelson MJ, Dutton PL, Robertson DE, DeGrado WF (1994) Design of a heme-binding 4-helix bundle. J Am Chem Soc 116:856–865
52. Robertson DE, Farid RS, Moser CC, Urbauer JL, Mulholland SE, Pidikiti R et al (1994) Design and synthesis of multi-heme proteins. Nature 368:425–431
53. Koder RL, Anderson JLR, Solomon LA, Reddy KS, Moser CC, Dutton LP (2009) Design and engineering of an O_2 transport protein. Nature 458:305–309
54. Chakraborty S, Touw D, Peacock AFA, Stuckey J, Pecoraro VL (2010) Structural comparisons of apo- and metalated three-stranded coiled coils clarify metal binding determinants in thiolate containing designed peptides. J Am Chem Soc 132:13240–13250
55. Dieckmann GR, McRorie DK, Tierney DL, Utschig LM, Singer CP, O'Halloran TV et al (1997) De novo design of mercury-binding two- and three-helical bundles. J Am Chem Soc 119:6195–6196
56. Dieckmann GR, McRorie DK, Lear JD, Sharp KA, DeGrado WF, Pecoraro VL (1998) The role of protonation and metal chelation preferences in defining the properties of mercury-binding coiled coils. J Mol Biol 280:897–912
57. Iranzo O, Thulstrup PW, Ryu S-B, Hemmingsen L, Pecoraro VL (2007) The application of Hg-199 NMR and Hg-199 m perturbed angular correlation (PAC) spectroscopy to define the biological chemistry of Hg-II: a case study with designed two- and three-stranded coiled coils. Chemistry 13:9178–9190
58. Farrer BT, McClure CP, Penner-Hahn JE, Pecoraro VL (2000) Arsenic(III)-cysteine interactions stabilize three-helix bundles in aqueous solution. Inorg Chem 39:5422–5423
59. Touw DS, Nordman CE, Stuckey JA, Pecoraro VL (2007) Identifying important structural characteristics of arsenic resistance proteins by using designed three-stranded coiled coils. Proc Natl Acad Sci U S A 104:11969–11974
60. Matzapetakis M, Farrer BT, Weng TC, Hemmingsen L, Penner-Hahn JE, Pecoraro VL (2002) Comparison of the binding of cadmium(II), mercury(II), and arsenic(III) to the de novo designed peptides TRI L12C and TRI L16C. J Am Chem Soc 124:8042–8054
61. Lee KH, Matzapetakis M, Mitra S, Marsh EN, Pecoraro VL (2004) Control of metal coordination number in de novo designed peptides through subtle sequence modifications. J Am Chem Soc 126:9178–9179
62. Lee KH, Cabello C, Hemmingsen L, Marsh EN, Pecoraro VL (2006) Using nonnatural amino acids to control metal-coordination number in three-stranded coiled coils. Angew Chem Int Ed 45:2864–2868
63. Peacock AFA, Hemmingsen L, Pecoraro VL (2008) Using diastereopeptides to control metal ion coordination in proteins. Proc Natl Acad Sci U S A 105:16566–16571
64. Ghosh D, Pecoraro VL (2004) Understanding metalloprotein folding using a de novo design strategy. Inorg Chem 43:7902–7915
65. Lovejoy B, Choe S, Cascio D, McRorie DK, DeGrado WF, Eisenberg D (1993) Crystal-structure of a synthetic triple-stranded alpha-helical bundle. Science 259:1288–1293
66. Peacock AFA, Stuckey JA, Pecoraro VL (2009) Switching the chirality of the metal environment alters the coordination mode in designed peptides. Angew Chem Int Ed 48:7371–7374
67. Mittl PRE, Deillon C, Sargent D, Liu N, Klauser S, Thomas RM et al (2000) The retro-GCN4 leucine zipper sequence forms a stable three-dimensional structure. Proc Natl Acad Sci U S A 97:2562–2566
68. Harbury PB, Kim PS, Alber T (1994) Crystal-structure of an isoleucine-zipper trimer. Nature 371:80–83

Chapter 12

Computational Design of Metalloproteins

Avanish S. Parmar, Douglas Pike, and Vikas Nanda

Abstract

A number of design strategies exist for the development of novel metalloproteins. These strategies often exploit the inherent symmetry of metal coordination and local topology. Computational design of metal binding sites in flexible regions of proteins is challenging as the number of conformational degrees of freedom is significantly increased. Additionally, without pre-organization of the primary shell ligands by the protein fold, metal binding sites can rearrange according to the coordination constraints of the metal center. Examples of metal incorporation into existing folds, full fold design exploiting symmetry, and fold design in asymmetric scaffolds are presented.

Key words Symmetry, Coordination geometry, Active site, Flexibility, Simulation

1 Introduction

Nearly one-third of proteins in our body utilize metals to control folding, stability, or functionality. Designing de novo metalloproteins provides a platform to test our understanding of biochemical structure and function [1]. It also provides a tool for creating protein and peptide ligand metal binding sites for enhancing specificity, folding, and functionality [2]. Metalloproteins play central roles in most natural processes such as, photosynthesis, water oxidation and nitrogen fixation, among many others. Designed metalloproteins may be tailored and are tunable systems for new biomedical, industrial, and material applications of the future. This chapter presents the computational strategies used to design metal driven protein assembly. The challenges of metalloprotein design can be divided into three topics: first, we will describe the computational design of metal binding sites into existing protein templates of known structure; next, the power of symmetry is exploited in the design of de novo metalloproteins; and third we discuss computational design of metal sites into flexible regions of proteins, the difficulties, and possible solutions.

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_12, © Springer Science+Business Media New York 2014

2 Computational Design of Metal Binding Sites into Existing Protein Templates of Known Structure

Generally, metal binding sites are introduced in proteins using a rational approach to achieve a desired protein structure and function. In 1990, Hellinga and Richards developed the computational program DEZYMER [3] where they introduced new ligand binding sites into proteins of known three-dimensional structure. DEZYMER keeps the backbone of the protein fixed, and scans for adjacent positions in the structure where first-shell ligand mutations may be placed. Sites are scored based on how well they satisfy geometric ligand–metal interaction constraints. The DEZYMER program can search for any coordination geometry, number and differing combination of amino acids [3, 4] by searching for rotamer clusters of particular amino acids that can be accommodated by a protein's backbone geometry. In a particularly noteworthy example of the design of a new metal binding site in *E. coli* thioredoxin [5] (Fig. 1) Mercury(II) bound in the intended manner; however copper(II) bound to only two of the designed ligand residues and additionally to two native residues in thioredoxin. It was concluded that in addition to designing specific metal binding geometries, it is also necessary to prevent competing binding geometries of one or more metals and native residues. The platform has been used to design metal binding sites for many other complex ligands [8, 9].

Clarke and Yuan developed in 1995 another computer program, METAL-SEARCH [10], which also aids in designing metal binding sites in proteins of known structure. Like DEZYMER this

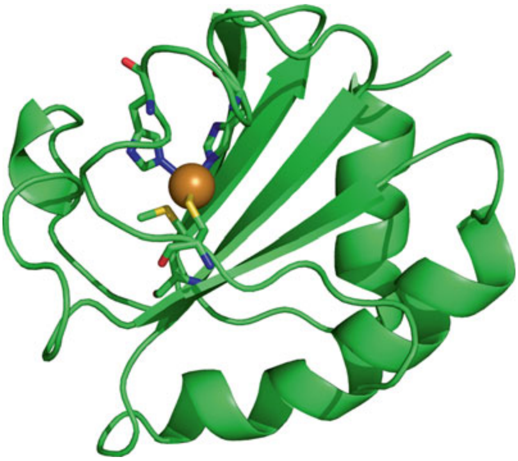


Fig. 1 Design model for the introduction of metal binding sites into *E. coli* thioredoxin (PDB: 2TRX [6]) where a new buried metal site is formed by cysteine, 2 histidines, and methionine with Cu (II) (shown as a *orange color sphere*). The model was built with the protein design software package, protCAD [7] and Pymol

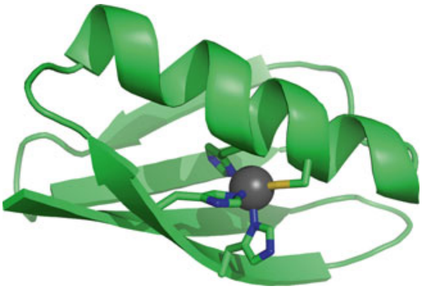


Fig. 2 Model of the design of a tetrahedral binding site in streptococcal protein G PDB:1GB1 [13] between His_3Cys and Zn (II) (shown as a *grey sphere*). Built with the protein design software package, protCAD [7] and Pymol

program keeps the backbone of the protein static and changes the amino acid sequences and the positions of the corresponding side chains. METAL-SEARCH specialized on tetrahedrally coordinated metal centers with cysteine and histidine ligands [11]. This program was used to design the tetrahedral His_3Cys Zn(II) binding sites in a small domain of streptococcal protein G [12]. The designed protein displays a high binding affinity ($K_d \sim 10^{-9}$ M) for Zn (II) without affecting the secondary and tertiary structure of the native protein (Fig. 2). Despite its simplicity, METALSEARCH has been used successfully in the design of multiple metal binding sites in proteins [14–16].

The recently developed RosettaMatch algorithm is based in part on METAL-SEARCH. The Baker lab [17] developed the RosettaMatch approach for designing catalytically active binding sites, in which they employ a hashing technique for searching favorable backbone conformations which can accommodate the catalytically active site geometry, in addition to searching for rotamer clusters that may suit the geometry required for catalysis. RosettaMatch follows a four step protocol. First positions in a set of protein scaffolds that match the desired catalytic sites' geometry are identified. In a second step the active site geometry is optimized and clashes are removed. In the third step RosettaDesign optimizes residue identities near the active site. In the final step the generated models are ranked by the computed binding energy. The Kuhlman lab used RosettaMatch [18] to design a protein monomer which forms a symmetric homodimer in the presence of zinc. By introducing metal binding sites at the interface of the proteins they successfully designed a high-affinity protein–protein interaction.

DEZYMER and METAL-SEARCH have two very different approaches to tackle the complexity in designing metal binding sites. DEZYMER uses "depth first pruning" and METAL-SEARCH uses "on the fly binning" [11]. The key distinction between these two approaches is, that in "depth first pruning" the geometrical search is carried out by identifying a set of rotamers of neighboring residues that can accommodate the metal binding geometry and discarding all related backbone branches that do not meet the requirements, whereas METALSEARCH uses an "on the fly binning" search which proceeds more incrementally, by identifying individual residues that can bind the metal, then searching for spatial overlap between them. RosettaMatch is the most commonly used algorithm nowadays and is an extension of METAL-SEARCH's "on the fly binning" technique, in addition to an inverse rotamer library where the sidechains are fixed in a preferred binding configuration and favorable backbone conformations for each binding residue are explored.

The Tezcan lab designed a metal-template surface by exploiting a minimal number of mutations on a monomeric protein, cytochrome cb_{562} to enable self-association of the protein block, which otherwise is a non-self-associating protein [19]. Subsequently they used ROSETTA to computationally optimize the interface of the self-associated protein blocks to achieve protein–protein interaction in the absence of metal [20]. Directionality and strength of metal coordination interactions have also been utilized on protein interfaces to create homooligomeric protein assemblies [21].

3 Advantages of Symmetry in De Novo Design

A key insight that has emerged from the study of natural metalloproteins is the relationship between elements of symmetry in the metal binding site and the overall protein fold. Rotational symmetry of the metal center and surrounding ligands is reflected in the symmetry of the secondary structural elements comprising the metal binding portion of the protein (Fig. 3). Symmetry can be exploited to simplify the design process by reducing both conformational and sequence degrees of freedom.

Symmetry reduces the size of sequence space to be sampled in patterning the sequence for a target fold. Consider a four α-helix protein, with each helix consisting of twenty amino acids. For a naïve design approach, the total number of possible sequences would be 20^{80} or ~10^{104}. Assuming twofold rotational symmetry would result in 20^{40} or ~10^{52} sequences, where two of the helices have the same sequence as the opposing pair. Appropriate fourfold symmetry would result in ~10^{26} sequences—a reduction in complexity of nearly 80 orders of magnitude relative to the original design task. When two orders of magnitude in computational complexity can mean the difference between obtaining a result after a few days versus 1 year, the advantages of such large reductions in sequence space become obvious.

A similar argument can be made for the advantages of constraining conformational sampling space using known symmetry. In de novo design, the structure needs to be specified in addition

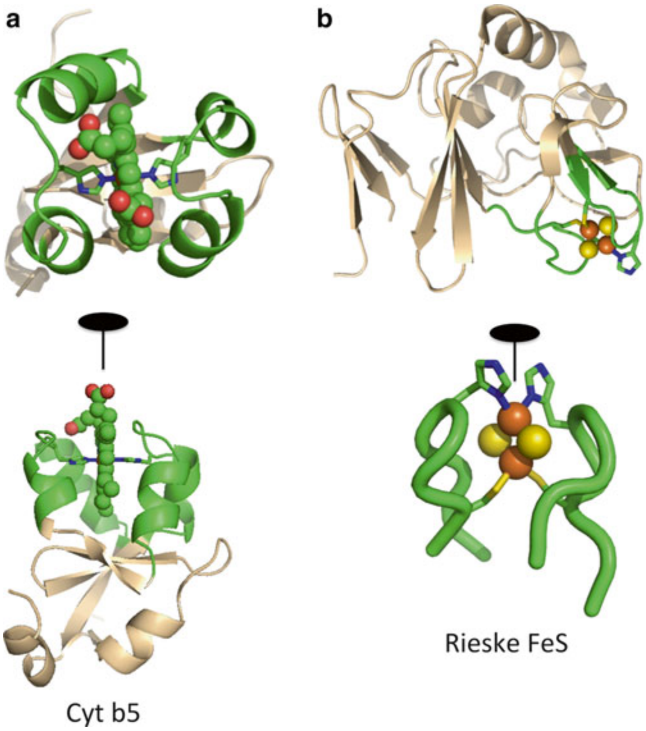


Fig. 3 Symmetry in natural metalloproteins. (**a**) Cytochrome b5 binds a heme cofactor between a pair of helix-loop-helix hairpins. Both the histidine ligand and the helical hairpins are related by a twofold rotational axis. (**b**) Similarly, the metal binding domain in a Rieske Fe_2S_2 protein shows two beta hairpins coordinating the metal by a pair of Cys/His sites related by a twofold rotational axis

to the sequence. Accomplishing this by sampling backbone torsional degrees of freedom is inefficient. In the aforementioned four-helix bundle, this would be approximately 160° of freedom—the backbone ϕ and ψ torsions for each of the 80 amino acids. Assuming idealized α-helices and rigid body sampling of conformations, this is now reduced to 24° of freedom—three rotational and three translational degrees per helix. Symmetry allows one to relate transformations of one helix to another, reducing the space even further. The combined simplifications in sequence and conformational degrees of freedom afforded by symmetry constraints make many design problems tractable [22–25].

The coordination geometry of a metal or cofactor determines the orientation of key ligands in the first-shell, which in turn constrain second-shell residues and in many cases, the protein topology. This relationship can be used in de novo metalloprotein design to simplify the number of potential sequences and the conformational degrees of freedom to be sampled.

3.1 General Steps for Metal-Centric Design

1. Assess high-resolution protein structures containing the metal cofactor of interest and identify elements of symmetry. This can be done manually using molecular visualization tools, or quantitatively using software to generate alignments for multiple structures [26–29]. Structures can be obtained from the Protein Data Bank (www.rcsb.org/pdb).
2. Identify the *keystone residues*—first shell amino acids are typically histidine, cysteine, aspartate, or glutamate. If multiple structures are available, determine geometric constraints between keystone residues and the cofactor such as bond distances and angles, or side chain rotameric configurations.
3. Isolate local secondary structural elements that present the keystone residues. Often, their placement matches the symmetry of the first shell interactions. Multiple alignments may make such symmetries more obvious.
4. Create idealized versions of the secondary structure element(s) containing the keystone residue(s) and generate a set of backbone topologies that match observed ligand–cofactor geometric constraints. This is typically done by rigid-body transformations using a reduced set of degrees of freedom as specified by the design target.
5. Use sequence-patterning tools [7, 30–33] to design sequences onto the specified backbone scaffolds. The identities of equivalent positions in sequence as specified by symmetry may be linked to reduce the total space to be sampled.

The specifics of each step are highly dependent on the design target. A few case studies are presented, describing potential variations.

3.2 Example: Constructing a Redox Active Rubredoxin Mimic

Several metalloproteins such as rubredoxin and ferredoxin serve as electron shuttles within the cell. There is significant interest in developing model electron transfer proteins to study the effect of structure on redox activity. Iron–sulfur proteins such as these are difficult to design as most natural examples place the metal atom or cluster in the loop regions of the fold. Additionally, the proteins must withstand redox cycling where loss of the metal can result in unfolding or loss of ligand availability due to disulfide crosslinking in the binding site.

In the design of a redox active rubredoxin mimic, the structure of a naturally occurring rubredoxin was analyzed and found to have internal twofold symmetry at the active site (Fig. 4). The iron sits between two short beta hairpins, coordinated by four cysteines in the turns of the hairpins. These keystone cysteines are stabilized by networks of backbone hydrogen bonds originating from the turn itself. Once the symmetry and keystone interactions were identified, a scaffold was constructed by replacing one of the

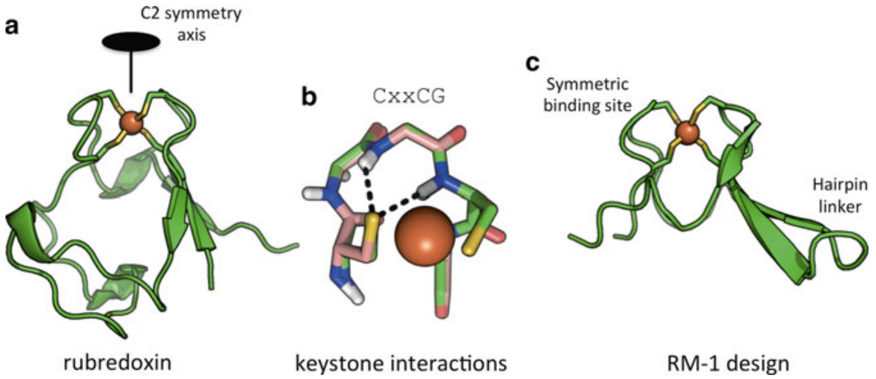


Fig. 4 Symmetry guided design of a rubredoxin mimic. (**a**) The structure of *P. furiosis* rubredoxin shows a twofold rotational symmetry axis for two β-hairpins around the metal site. (**b**) Keystone interactions in each of these hairpins are formed by a CxxCG motif with backbone amide to cysteine sulfur hydrogen bonds. (**c**) The design of Rubredoxin Mimic 1 (RM-1) utilized the twofold symmetry plus a short hairpin linker to create a redox-active metalloprotein [34]

hairpins with a replica of the other, created by a 180° rotation around a central axis passing through the metal site. A short linker was inserted to connect the two hairpins into a single polypeptide chain. All residues were patterned computationally using the SCADS platform [32] except for keystone residues and the linker, which was based on a previous rational design [28]. The final molecule showed correct metal binding stoichiometry and was stable over multiple redox cycles [34].

In this example, the backbone was generated from existing protein structures, rather than idealized secondary structure elements. This type of fragment-based approach is powerful as it often preserves molecular details of local interactions which may be crucial for a successful design. This strategy could possibly be extended to more complex iron–sulfur sites such as the Rieske complex shown in Fig. 3b, which exhibits similar symmetry and secondary structure.

3.3 Example: Binding Non-natural Porphyrins

In the case of the rubredoxin mimic, it was possible to utilize backbone elements obtained from existing natural counterparts. A more challenging target is one that has no direct natural counterpart. In this example, a four-helix bundle was developed to bind a pair of synthetic porphyrins. These cofactors differed from natural porphyrins such as hemes by the structure and placement of the pendant groups (Fig. 5). Previous heme binding bundles had been designed using the metal-centric approach. Natural examples of protein topologies with D2 symmetry—having two orthogonal twofold rotational symmetry axes—were found in cytochrome bc1 [36, 37]. However, these natural backbones were not appropriate for the synthetic cofactor, whose aromatic pendant groups would not be accommodated into the existing scaffold.

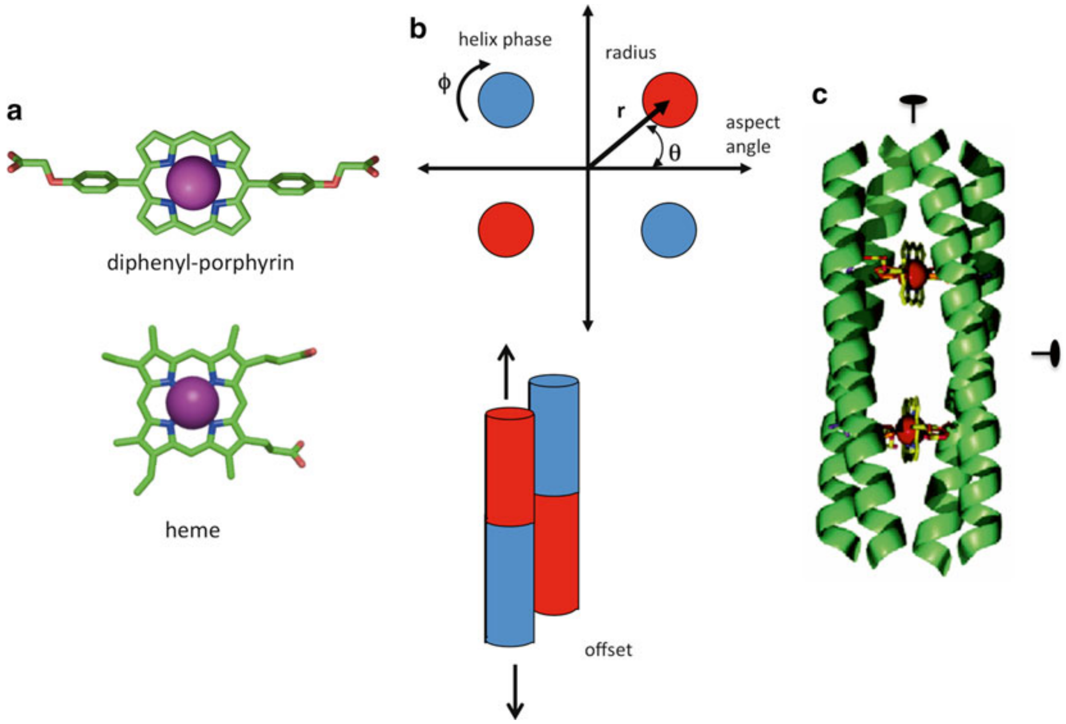


Fig. 5 Design of a non-natural metallocofactor binding protein. (**a**) The target cofactor differs from natural heme in the presentation of pendent groups off the porphyrin ring. The cofactor has several orthogonal pseudo symmetric twofold rotational axes. (**b**) A set of rigid body rotations and translations are used to generate an ensemble of backbones for binding the cofactor. (**c**) Final model topology for the diphenyl-porphyrin binding protein [35]

To address this obstacle, a series of backbones were modeled using rigid body transformations of an idealized helix—creating a series of D2 symmetric coiled-coils. The scaffolds that were best suited to accommodate keystone interactions—a pair of histidine–metal bonds and a second-shell threonine hydrogen bond to the histidine—were subjected to further sequence patterning. The final design was able to bind the target cofactor tightly, but was unable to coordinate natural heme, indicating that both affinity and specificity were achieved [35].

3.4 Example: A Novel Fe_4S_4 Binding Topology

In the first two examples, the symmetry of natural proteins was used to guide the design of synthetic metalloproteins. In cases where natural examples of symmetric topologies do not exist, it may be possible to build directly from the symmetry of the cofactor itself. This was the approach utilized in the design of a Fe_4S_4 binding four-helix bundle. While four-helix bundles are not unique and are commonly utilized in design, no currently known natural examples of such a Fe_4S_4 protein topology exist.

The cubane Fe_4S_4 cluster has tetrahedral symmetry. Within tetrahedral symmetry are twofold and three-fold axes of rotational

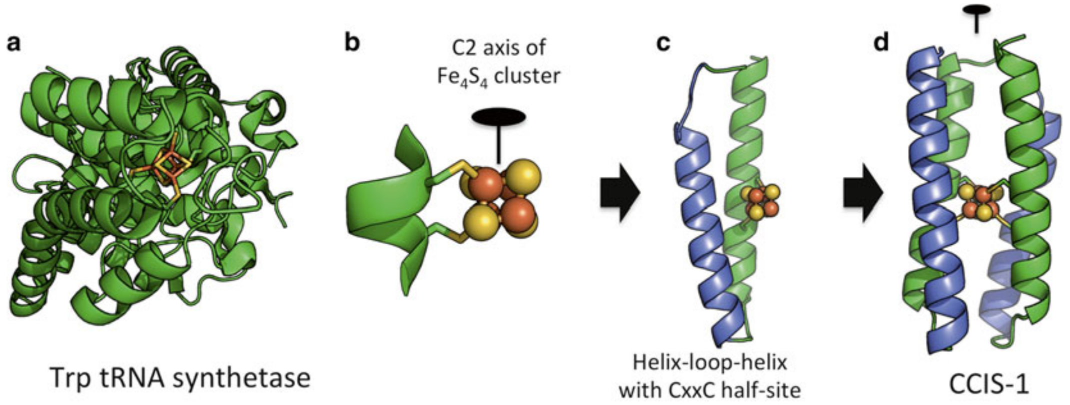


Fig. 6 (**a**) Trp tRNA synthetase contains symmetrical iron–sulfur cluster, but the binding site and local topology does not match any axes of symmetry in the cluster. (**b**) One of the twofold axes of the cluster is extended to dictate the overall symmetry of the protein fold. The twofold symmetry axis is applied to a helix-loop-helix motif (**c**) to generate the final topology (**d**) [38]

symmetry. However, most cysteine ligands are presented by turns or loops that are at the ends of secondary structure elements. After extensive manual inspection of existing folds, one example was found where two of the four cysteine ligands were presented by a single α-helix (Fig. 6). By considering one of the twofold rotation axes of the Fe_4S_4 cluster, it was possible to generate a second helix that was parallel to the original. Based on this, a series of two-helix bundles were created using rigid-body transformations, searching for one with optimal keystone cysteine to cluster interactions. Two additional helices were then docked in an antiparallel orientation to the existing design such that short loops could be inserted to create a single chain construct. This scaffold was then subjected to sequence patterning using a combination of the protCAD and ROSETTA design platforms [7, 33]. The resulting design bound Fe_4S_4 clusters with the correct stoichiometry and binding induced a helical fold [38, 39].

3.5 Applications of Symmetry

The symmetry of designed folds can be exploited to create multi-cofactor containing proteins based on existing single-cofactor designs. Numerous examples exist in nature where single cofactor binding sites are concatenated to create electron transfer pathways through large proteins—essentially acting as protein wires. The ferredoxin fold is a fusion of two Fe_4S_4 domains. Alone, ferredoxin serves as an electron shuttle between other proteins. A series of proximal ferredoxin-like domains are found in hydrogenases, creating an electron transfer pathway. The periodicity of an α-helix allows the facile extension of a single cofactor binding motif to a design with multiple cofactors presented in a linear array. This strategy was utilized in the design of a four-porphyrin binding

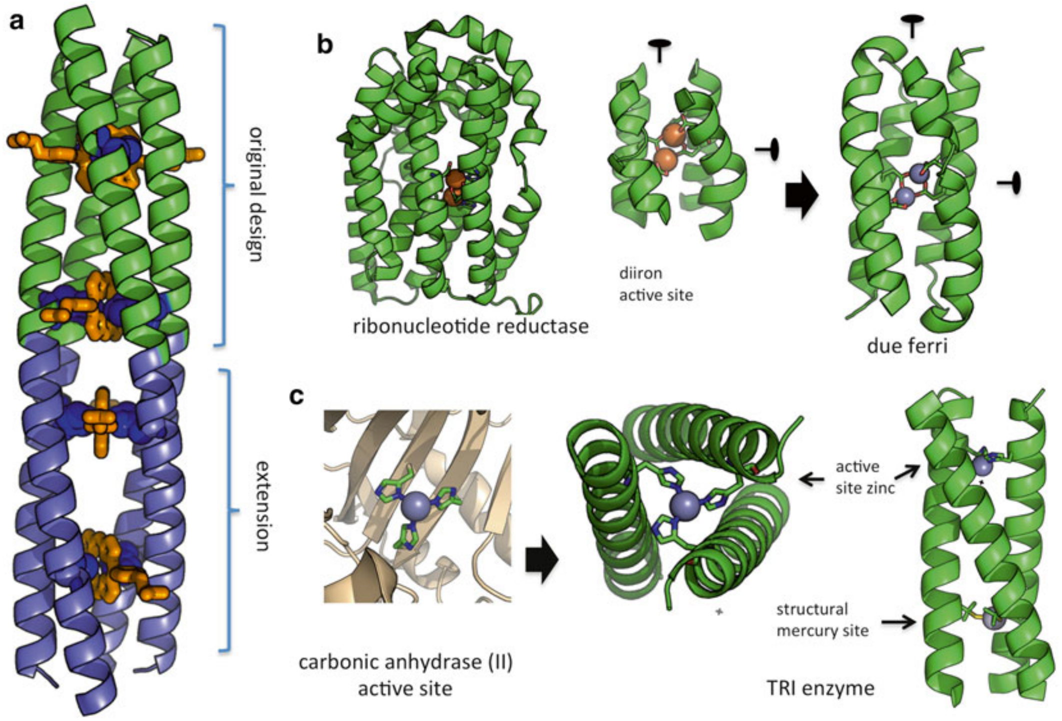


Fig. 7 Applications of symmetry guided metalloprotein design. (**a**) Periodic helical symmetry of the diphenyl porphyrin bundle was used to build a four-cofactor binding protein [40]. (**b**) Core D2 symmetry in di-metal oxidoreductases was exploited to create a minimal four-helix metalloenzyme [28, 29]. (**c**) Threefold rotational symmetry of the active site in CA(II) was used to generate a three-helix bundle TRI capable of binding two metal centers—one structural mercury ion and one active site zinc ion

chain created by extending the scaffold of the original molecule with the same geometric parameters (Fig. 7a) [40]. A similar approach could be used to create helical bundle Fe_4S_4 wires that mimic the structures created by tandem ferredoxin domains.

Symmetry has also been applied to the design of metalloenzymes. A number of oxidoreductases such as methane monooxygenase and ribonucleotide reductase are able to carry out challenging catalytic reactions at a di-metal containing active site. Analysis of the symmetry of a number of related oxidoreductases revealed common D2 symmetry and consistent keystone residues that specify the topology of an antiparallel four-helix bundle (Fig. 7b). This has been used in the development of the Due Ferri (di-iron) series of de novo metalloenzymes, which have served as powerful model systems for studying natural oxidoreductase mechanisms [41, 42] and as starting points for synthetic catalysts [43, 44].

The design of a C3 symmetric metalloenzyme capable of CO_2 hydration was developed based on the active site of carbonic anhydrase II (CAII). In this case, the topology of the natural enzyme did not reflect the threefold axis of the tetrahedral symmetry of the first shell histidines (Fig. 7c). The active site was accommodated in the core of a three-helix bundle. Similar threefold symmetry of trigonal-planar thiol–mercury complexes allowed a combined design including both the structural mercury and an active site Zn. The resulting TR1 design had catalytic activity within 100-fold of the CAII and was better than small molecule catalysts by a similar order of magnitude [45, 46].

4 Computational Design of Metal Binding Sites into Flexible Regions of Proteins

Most successful metalloproteins designs consist of well-defined tertiary structure and a sizeable hydrophobic core. The constraint for designing a high affinity metal binding site is that proteins should consist of rigid ligands where the geometrical arrangement is pre-organized to match the coordination of the metal ions. Certain proteins lack regular tertiary structure with a well-defined hydrophobic core and also lack highly rigid sites for introducing the ideal geometry of creating metal–ligand sites. Examples include metallothioneins, ferredoxins, and collagen.

4.1 Example: Metallothioneins and Ferredoxin

Metallothioneins are cysteine rich (up to 30 % of the total amino acid content), low molecular weight (approx. 6–7KDa) proteins with a complete lack of aromatic amino acids in the primary sequence [47]. The tertiary structures of all metallothioneins are dominated by the formation of metal–thiol clusters which involve terminal and non-terminal bridging of cysteinyl thiolate groups [48] (Fig. 8). Loop-rich proteins such as metallothioneins are challenging to deconstruct using symmetry and topology-based approaches. Despite the lack of extensive secondary structure, the metal imposes significant constraints on the accessible topology in these mini-proteins. This was explored in detail in a case study of the Fe_4S_4 ferredoxin fold.

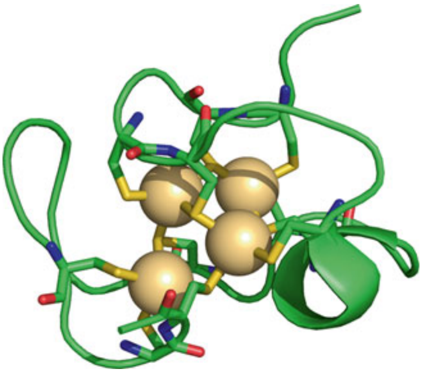


Fig. 8 Structure of human metallothionein-2 bound to four cadmium ions (PDB: 1MHU [49])

Only the right-handed topology found in ferredoxin presented a network of hydrogen bonds that stabilized the cluster-peptide complex [50].

It was noted that the four first-shell cysteine ligands of Fe_4S_4 binding sites could be grouped in most instances to a CxxCxxC stretch, followed by a distant fourth cysteine. Looking down from the fourth cysteine toward the metal cluster, the CxxCxxC circumscribed the cluster in a counterclockwise, or right-handed direction, going from the N-terminal cysteine to the C-terminal one in nearly all cases. Given the lack of clear secondary structure in this region (Fig. 9), it was interesting to assess whether such a conformation was essential for binding, or if it had arisen by chance early in protein evolution.

To assess this, all possible backbone conformations of a short CGGCGGC heptapeptide were enumerated and evaluated for their ability to form a viable Fe_4S_4 binding site based on geometric constraints using the protCAD and AMBER software platforms [50, 51]. The central cysteine was joined to the cluster, creating a composite Cys-Fe_4S_4 residue. Cluster placement was constrained by the two sidechain rotamers of this compound residue. It was found that both left and right-handed conformations could be generated, but only right-handed ones donated a significant network of backbone amide to cluster sulfur hydrogen bonds. Such interactions would stabilize binding and serve to tune the midpoint potential of the cofactor.

Fragment based simulations such as these may prove useful in the design of larger metalloproteins with loop-rich metal coordination sites. Libraries of metal-constrained conformational motifs could be combined with existing fragment-assembly design methods [52] to create novel metal protein folds.

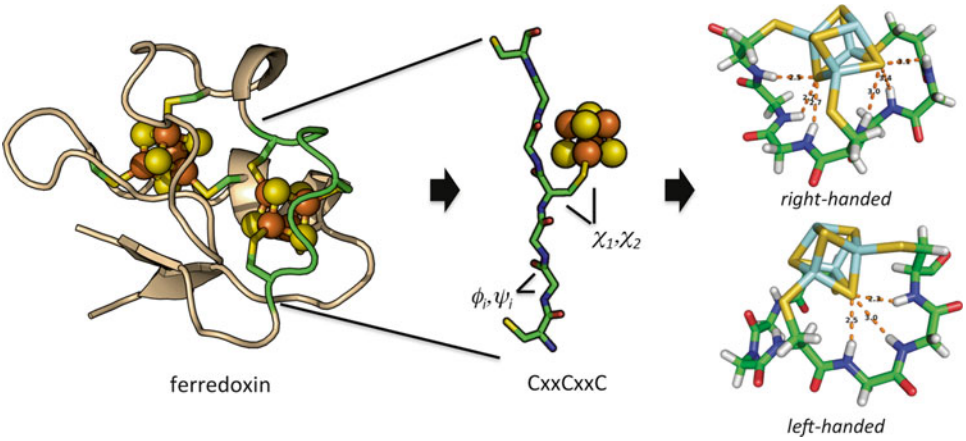


Fig. 9 Deconstruction of metal binding constraints on the fold of bacterial ferredoxins. The common CxxCxxC binding site for Fe4S4 in ferredoxin lacks distinct secondary structure. Conformational enumeration of backbone and sidechain degrees of freedom in a heptapeptide revealed a number of left and right-handed topologies

4.2 Example: Metal Template Folding of Collagen

Metal binding sites engineered at the ends of oligomeric proteins can be used to tune their stability. Based on the coupled thermodynamics of folding and metal binding, it could be shown that the stability of trimer formation around a metal site in large libraries of α-helical monomers could be screened by adjusting the metal concentration [53–55]. Using a similar approach we wanted to design a collagen heterotrimer by introducing metal binding sites at the end of the triple helix. Unlike the helical designs which had three-fold rotational symmetry, collagen has a screw symmetry such that the termini of the chains are not adjacent in the structure. Collagen has an extended structure with a limited number of tertiary contacts (Fig.10a). This makes it challenging to engineer a metal binding site with desired stability and specificity. In natural collagens, fibrillar regions extend over a thousand amino acids, making them difficult to express and characterize. As such, short collagen mimetic peptide (CMPs) systems have been essential tools in exploring the molecular basis for stability, specificity and higher order assembly. The most stable CMPs using biogenic amino acids consist of repeating (Gly-Pro-Hyp)$_n$ triplets (amino acid code for hydroxyproline => Hyp< or >O<, respectively). A structural metal has been rationally designed in CMP heterotrimers. This was accomplished by introducing bidentate 2.2′-bipyridyl ligands at

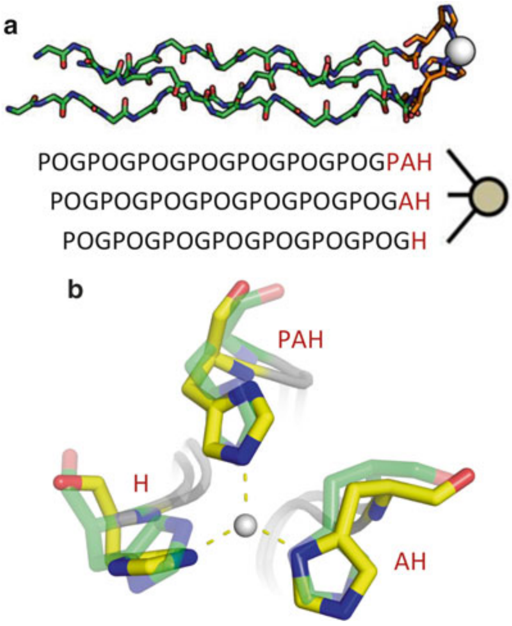


Fig. 10 Model and design of a metal binding heterotrimer: (**a**) Model of peptides containing histidines at the C-terminus coordinating a zinc (II) ion, to stabilize the heterotrimer. (**b**) protCAD (*green*) design and AMBER (*yellow*) minimized structure showing zinc binding to histidine for the formation of a metal binding heterotrimer

the N-terminus for the formation of an octahedral metal binding site [56]. The analysis of the rationally designed models revealed that metal binding on the terminus of the triple-helix requires conformational rearrangement of ligands for formation of a heterotrimer. In order to achieve the specific metal–ligand geometry of the terminus, multiple conformations must be evaluated and computational methods can expedite this analysis.

A possible way to introduce a target metal binding site is to attach a His-metal binding site at the C-terminus of a collagen triple helix (Fig. 10a). In order to promote the geometric vicinity of the three His residues, the sequences of a blunt-ended heterotrimeric triple helix should be designed by attaching Gly-Ala-His, Ala-His, and His to three chains $(POG)_7PAH$, $(POG)_7AH$, $(POG)_7H$, respectively (Fig. 10). These were built with the protein design software package, protCAD [7], and attached in the model to the C-terminus of a high resolution X-ray crystal structure of $(PPG)_{10}$ (PDB ID: 1K6F) [57]. The binding-site conformation was optimized by minimizing interaction scores using a Monte Carlo simulated annealing protocol [58]. The interaction scores included van der Waals energy and a pseudo-energy term measuring the geometry of metal–His interactions. The backbone conformations of residues involved in the binding site were optimized by adjusting their ψ, φ angles according to the sidechain-dependent frequencies observed in crystal structures [59]. Histidine–metal interactions of all the conformers were optimized with protCAD to select the best conformer as an initial structure for further molecular minimization. Should this strategy be successful, it might also be used to promote the folding and stability of some known charged heterotrimers [60, 61].

Computational design of metal binding sites at flexible regions of a protein presents many challenges and advantages for metal-binding specificity. One of the advantages of putting the binding sites at the flexible region of a protein is that the ligands can adjust to bind the metals in their preferred geometry [62], possibly involving hinge-bending motion [63]. The challenge hereby is constraining the flexibility to the formation of a specific heterotrimeric blunt-ended metal binding site. Terminal flexibility comes at the expense of specificity, where the homotrimers (Fig. 11) are formed in addition to the blunt-ended heterotrimer (Fig. 10). To overcome this may require a more refined computational approach that can sample sufficient conformational space.

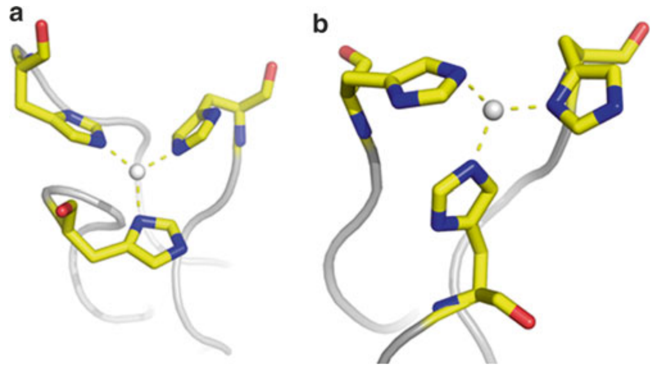


Fig. 11 Model and design of metal binding homotrimers showing zinc binding to histidine for the formation of metal binding homotrimers (**a**) $(POG)_7AH$; (**b**) $(POG)_7PAH$

Acknowledgements

This work was supported by NIH R01 GM-089949.

References

1. Gibney BR, Huang SS, Skalicky JJ, Fuentes EJ, Wand AJ, Dutton PL (2001) Hydrophobic modulation of heme properties in heme protein maquettes. Biochemistry 40:10550–10561
2. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A (1999) De novo design and structural characterization of proteins and metalloproteins. Annu Rev Biochem 68:779–819
3. Hellinga HW, Richards FM (1991) Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. J Mol Biol 222:763–785
4. Hellinga HW (1996) Metalloprotein design. Curr Opin Biotechnol 7:437–441
5. Hellinga HW, Caradonna JP, Richards FM (1991) Construction of new ligand binding sites in proteins of known structure. Ii Grafting of a buried transition metal binding site into escherichia coli thioredoxin. J Mol Biol 222:787–803
6. Katti SK, LeMaster DM, Eklund H (1990) Crystal structure of thioredoxin from escherichia coli at 1.68 a resolution. J Mol Biol 212:167–184
7. Summa CM (2002) Computational methods and their applications for de novo functional protein design and membrane protein solubilization. In: School of Medicine Ph. D. University of Pennsylvania, Philadelphia, PA
8. Pinto AL, Hellinga HW, Caradonna JP (1997) Construction of a catalytically active iron superoxide dismutase by rational protein design. Proc Natl Acad Sci U S A 94: 5562–5567
9. Coldren CD, Hellinga HW, Caradonna JP (1997) The rational design and construction of a cuboidal iron-sulfur protein. Proc Natl Acad Sci U S A 94:6635–6640
10. Clarke ND, Yuan SM (1995) Metal search: a computer program that helps design tetrahedral metal-binding sites. Proteins 23:256–263
11. Desjarlais JR, Clarke ND (1998) Computer search algorithms in protein modification and design. Curr Opin Struct Biol 8:471–475
12. Klemba M, Gardner KH, Marino S, Clarke ND, Regan L (1995) Novel metal-binding proteins by design. Nat Struct Biol 2:368–373
13. Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT et al (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein g. Science 253:657–661
14. Klemba M, Regan L (1995) Characterization of metal binding by a designed protein: single ligand substitutions at a tetrahedral cys2his2 site. Biochemistry 34:10094–10100
15. Regan L (1995) Protein design: novel metal-binding sites. Trends Biochem Sci 20: 280–285

16. Regan L, Clarke ND (1990) A tetrahedral zinc(ii)-binding site introduced into a designed protein. Biochemistry 29:10878–10883
17. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA et al (2006) New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 15: 2785–2794
18. Der BS, Machius M, Miley MJ, Mills JL, Szyperski T, Kuhlman B (2012) Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. J Am Chem Soc 134:375–385
19. Salgado EN, Ambroggio XI, Brodin JD, Lewis RA, Kuhlman B, Tezcan FA (2010) Metal templated design of protein interfaces. Proc Natl Acad Sci U S A 107:1827–1832
20. Salgado EN, Radford RJ, Tezcan FA (2010) Metal-directed protein self-assembly. Acc Chem Res 43:661–672
21. Brodin JD, Ambroggio XI, Tang C, Parent KN, Baker TS, Tezcan FA (2012) Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. Nat Chem 4:375–382
22. Betz SF, DeGrado WF (1996) Controlling topology and native-like behavior of de novo-designed peptides: design and characterization of antiparallel four-stranded coiled coils. Biochemistry 35:6955–6962
23. Plecs JJ, Harbury PB, Kim PS, Alber T (2004) Structural test of the parameterized-backbone method for protein design. J Mol Biol 342:289–297
24. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS (1998) High-resolution protein design with backbone freedom. Science 282:1462–1467
25. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, Andre I et al (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336:1171–1174
26. Sippl MJ, Wiederstein M (2012) Detection of spatial correlations in protein structures and molecular complexes. Structure 20:718–728
27. Thompson KE, Wang Y, Madej T, Bryant SH (2009) Improving protein structure similarity searches using domain boundaries based on conserved sequence information. BMC Struct Biol 9:33
28. Lombardi A, Summa CM, Geremia S, Randaccio L, Pavone V, DeGrado WF (2000) Retrostructural analysis of metalloproteins: application to the design of a minimal model for diiron proteins. Proc Natl Acad Sci U S A 97:6298–6305
29. Summa CM, Lombardi A, Lewis M, DeGrado WF (1999) Tertiary templates for the design of diiron proteins. Curr Opin Struct Biol 9:500–508
30. Chowdry AB, Reynolds KA, Hanes MS, Voorhies M, Pokala N, Handel TM (2007) An object-oriented library for computational protein design. J Comput Chem 28:2378–2388
31. Pokala N, Handel TM (2005) Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J Mol Biol 347:203–227
32. Kono H, Saven JG (2001) Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. J Mol Biol 306:607–628
33. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368
34. Nanda V, Rosenblatt MM, Osyczka A, Kono H, Getahun Z, Dutton PL et al (2005) De novo design of a redox-active minimal rubredoxin mimic. J Am Chem Soc 127:5804–5805
35. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ et al (2005) Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. J Am Chem Soc 127:1346–1347
36. Ghirlanda G, Osyczka A, Liu W, Antolovich M, Smith KM, Dutton PL et al (2004) De novo design of a d2-symmetrical protein that reproduces the diheme four-helix bundle in cytochrome bc1. J Am Chem Soc 126:8141–8147
37. North B, Summa CM, Ghirlanda G, DeGrado WF (2001) D(n)-symmetrical tertiary templates for the design of tubular proteins. J Mol Biol 311:1081–1090
38. Grzyb J, Xu F, Weiner L, Reijerse EJ, Lubitz W, Nanda V et al (2010) De novo design of a non-natural fold for an iron-sulfur protein: alpha-helical coiled-coil with a four-iron four-sulfur cluster binding site in its central core. Biochim Biophys Acta 1797:406–413
39. Grzyb J, Xu F, Nanda V, Luczkowska R, Reijerse E, Lubitz W et al (2012) Empirical and computational design of iron-sulfur cluster proteins. Biochim Biophys Acta 1817:1256–1262
40. McAllister KA, Zou H, Cochran FV, Bender GM, Senes A, Fry HC et al (2008) Using alpha-helical coiled-coils to design nanostructured metalloporphyrin arrays. J Am Chem Soc 130:11921–11927

41. Maglio O, Nastri F, Pavone V, Lombardi A, DeGrado WF (2003) Preorganization of molecular binding sites in designed diiron proteins. Proc Natl Acad Sci U S A 100:3772–3777

42. DeGrado WF, Di Costanzo L, Geremia S, Lombardi A, Pavone V, Randaccio L (2003) Sliding helix and change of coordination geometry in a model di-mnii protein. Angew Chem Int Ed Engl 42:417–420

43. Reig AJ, Pires MM, Snyder RA, Wu Y, Jo H, Kulp DW et al (2012) Alteration of the oxygen-dependent reactivity of de novo due ferri proteins. Nat Chem 4:900–906

44. Kaplan J, DeGrado WF (2004) De novo design of catalytic proteins. Proc Natl Acad Sci U S A 101:11566–11570

45. Zastrow ML, Pecoraro VL (2013) Influence of active site location on catalytic activity in de novo-designed zinc metalloenzymes. J Am Chem Soc 135:5895–5903

46. Zastrow ML, Peacock AF, Stuckey JA, Pecoraro VL (2012) Hydrolytic catalysis and structural stabilization in a designed metalloprotein. Nat Chem 4:118–123

47. Stillman MJ (1995) Metallothioneins. Coord Chem Rev 144:461–511

48. Kagi JHR (1991) Overview of metallothionein. Meth Enzymol 205:613–626

49. Messerle BA, Schaffer A, Vasak M, Kagi JH, Wuthrich K (1990) Three-dimensional structure of human [113cd7]metallothionein-2 in solution determined by nuclear magnetic resonance spectroscopy. J Mol Biol 214:765–779

50. Kim JD, Rodriguez-Granillo A, Case DA, Nanda V, Falkowski PG (2012) Energetic selection of topology in ferredoxins. PLoS Comput Biol 8:e1002463

51. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KM Jr et al (2005) The amber biomolecular simulation programs. J Comput Chem 26:1668–1688

52. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225

53. Roy L, Case MA (2010) Protein core packing by dynamic combinatorial chemistry. J Am Chem Soc 132:8894–8896

54. Case MA, McLendon GL (2004) Metal-assembled modular proteins: toward functional protein design. Acc Chem Res 37:754–762

55. Cooper HJ, Case MA, McLendon GL, Marshall AG (2003) Electrospray ionization fourier transform ion cyclotron resonance mass spectrometric analysis of metal-ion selected dynamic protein libraries. J Am Chem Soc 125:5331–5339

56. Lebruin LT, Banerjee S, O'Rourke BD, Case MA (2011) Metal ion-assembled micro-collagen heterotrimers. Biopolymers 95: 792–800

57. Berisio R, Vitagliano L, Mazzarella L, Zagari A (2002) Crystal structure of the collagen triple helix model [(pro-pro-gly)(10)](3). Protein Sci 11:262–270

58. Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

59. Srinivasan R, Rose GD (1995) Linus: a hierarchic procedure to predict the fold of a protein. Proteins 22:81–99

60. Xu F, Zahid S, Silva T, Nanda V (2011) Computational design of a collagen a:B:C-type heterotrimer. J Am Chem Soc 133: 15260–15263

61. Parmar AS, Zahid S, Belure S, Young R, Hasan N, Nanda V (2014) Design of net-charged abc-type collagen heterotrimers. J Struct Biol 185:163–167

62. Wray JW, Baase WA, Ostheimer GJ, Zhang XJ, Matthews BW (2000) Use of a non-rigid region in t4 lysozyme to design an adaptable metal-binding site. Protein Eng 13:313–321

63. Zhang XJ, Matthews BW (1995) Edpdb—a multifunctional tool for protein-structure analysis. J Appl Crystallogr 28:624–630

Chapter 13

Incorporation of Modified and Artificial Cofactors into Naturally Occurring Protein Scaffolds

Koji Oohora and Takashi Hayashi

Abstract

As a possible modification of cofactor-containing proteins, cofactor-substitution typically leads to drastic changes of protein function. In particular heme, a porphyrin iron complex, is a representative, replaceable cofactor for this methodology and numerous cofactor-modified hemoproteins (reconstituted hemoproteins) have been prepared with the goal of elucidating their operational mechanism and/or engineering the protein function. In a series of hemoproteins, myoglobin, an oxygen storage hemoprotein, is one of the most rewarding scaffolds to generate a modified protein with an artificial cofactor. In this chapter, we describe practical procedures for the preparation of apomyoglobin and incorporation of zinc porphyrin as an artificial cofactor. Furthermore, we discuss the methodology to characterize the obtained cofactor-substituted proteins and the design of several artificial cofactors.

Key words Cofactor design, Heme, Protein engineering, Hemoprotein reconstitution

1 Introduction

In the area of protein engineering, the substitution of a native cofactor with a different one is among the most powerful methods to investigate, improve, or modify the native function. For example, several cofactors such as heme, flavin, quinone, or cobalamin as shown in Fig. 1 are exchangeable [1–4]. Generally, a cofactor is bound to a specific site in the corresponding protein matrix and plays a central role as the reaction center, enabling redox reactions, bond activation, gas binding, and so on. Therefore, the incorporation of a nonnatural cofactor prepared by chemical synthesis into a cofactor binding site will have a significant direct influence on the protein function. The modified protein is often called a "reconstituted protein." Figure 2 depicts a typical scheme for the generation of a reconstituted protein. Over four decades, many groups have reported various reconstituted proteins not only for the investigation of their native function and underlying mechanisms [5] but also for the creation of artificial enzymes [6], biomaterials and biodevices [7].

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216, DOI 10.1007/978-1-4939-1486-9_13, © Springer Science+Business Media New York 2014

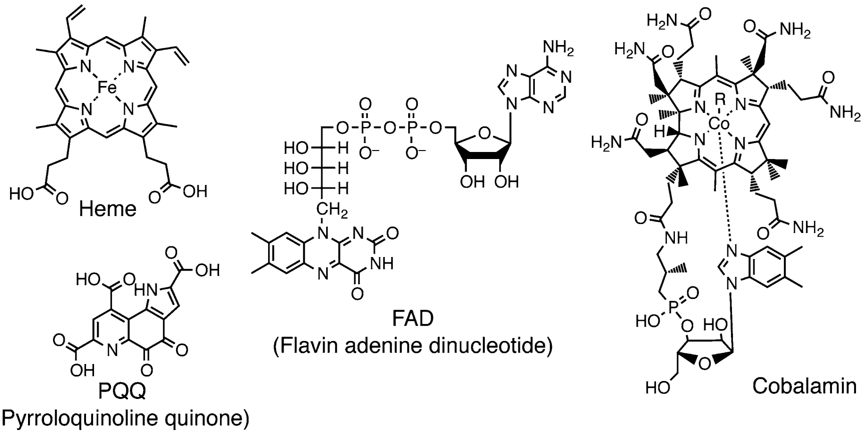


Fig. 1 Molecular structures of representative cofactors for proteins in biological systems

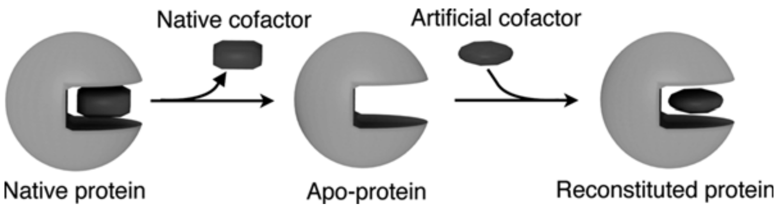


Fig. 2 Schematic representation of cofactor substitution

Hemoproteins being rich and diverse in function are particularly rewarding targets for the application of functional modification by the aforementioned reconstitution method [8]. Myoglobin, an oxygen storage hemoprotein, has one protoheme IX, which is bound by coordinative bonding, hydrogen bonding, and hydrophobic and electrostatic interactions with amino acid residues of the protein scaffold [9]. Figure 3 shows the crystal structure of myoglobin (PDB ID: 2MBW). Under physiological conditions, heme is tightly bound to the protein matrix, but dissociates completely at pH 2–3 in the case of myoglobin. Teale demonstrated that the dissociated heme can be easily removed and subsequently extracted with a suitable organic solvent such as 2-butanone. The following neutralization yields apomyoglobin, a cofactor-free protein [10]. Into the apoprotein, native heme or a well-designed artificial cofactor can be easily (re-)incorporated. A similar strategy can be applied to other hemoproteins such as cytochrome P450$_{\mathrm{cam}}$, horseradish peroxidase (HRP), cytochrome *b*$_{562}$ and others [11–15]. Figure 4 shows representative examples of artificial cofactors for the reconstitution of myoglobin. Incorporation of modified heme **1** where one of the propionate side chains is linked to a branched aromatic amide structure—believed to serve as a substrate binding site—converts myoglobin mutant H64D into an highly active

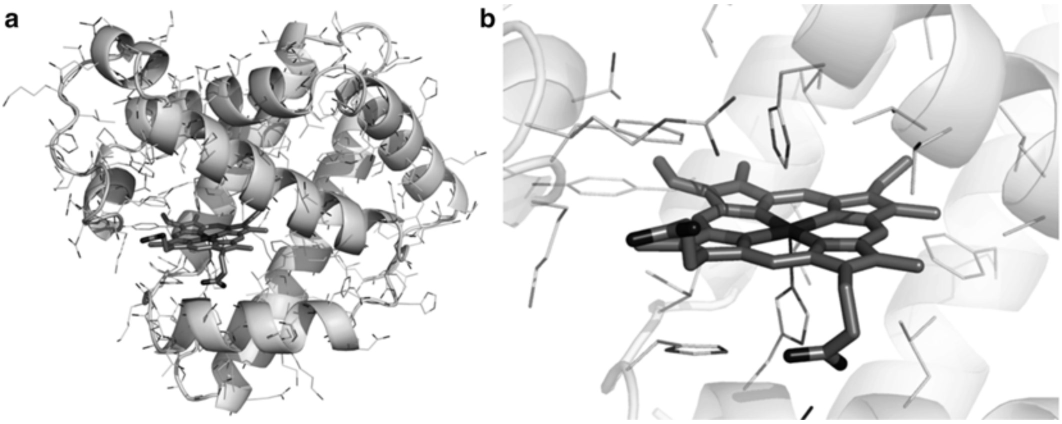


Fig. 3 Crystal structure of myoglobin (PDB ID: 2MBW). (**a**) Overall structure. (**b**) Structure of heme and several interacted amino acid residues

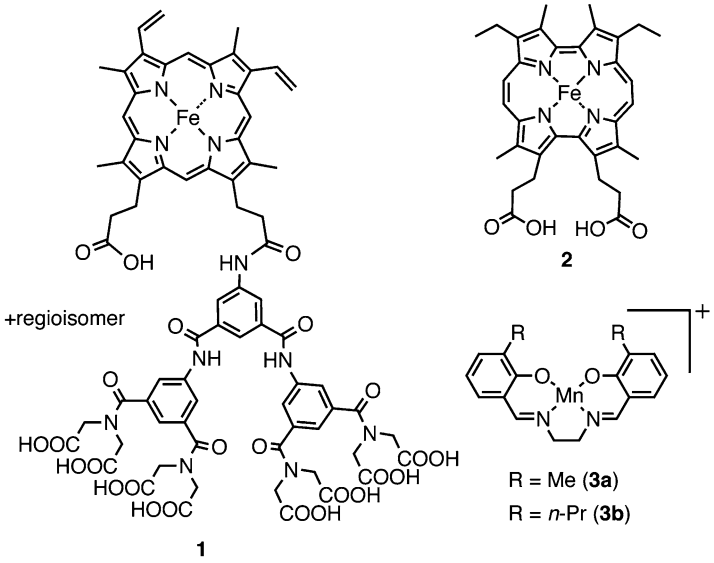


Fig. 4 Molecular structures of representative artificial cofactors for myoglobin

peroxidase for 2-methoxyphenol oxidation [16]. A constitutional isomer of heme, iron porphycene **2**, shows a 2,300-fold higher affinity for dioxygen in the apomyoglobin matrix compared to that of native heme [17]. Manganese salen **3** as a cofactor enables enantioselective sulfoxidation when incorporated into the apo-form of the myoglobin mutant H64D-A71G [18].

The corresponding apoprotein of a range of flavoproteins can be prepared by dialysis under carefully chosen nonnative conditions such as acidification or the addition of a denaturant. Furthermore, chromatographic methods to obtain reconstituted flavoproteins have been developed [3]. For example a His-tagged flavoprotein can be immobilized on a nickel-nitrilotriacetic acid column and an eluent containing a denaturant enables the removal of the native cofactor. On the beads, the reconstitution of the apoprotein with an artificial flavin derivative can be performed and the reconstituted protein is finally eluted with an imidazole gradient. These strategies for cofactor exchange in hemo- and flavoproteins are likely to be extendable to other proteins containing noncovalently bound cofactors [2].

Most recently, the reconstitution methodology has been applied to construct artificial biodevices and biomaterials [7]. In such systems, incorporation of the artificial cofactor into the scaffold establishes an interface to connect proteins with other attractive materials. Due to the specificity and rigidity of the interaction between apoprotein and corresponding cofactor a high degree of directionality can be realized in the inter-system linkage. Willner and coworkers reported unique constructs by linking FAD via carbon nanotubes to a gold electrode and thereby enabled electric contact between the active site of Glucose oxidase and the electrode surface [19]. In a related manner, metalloporphyrin moieties were immobilized via covalent bond linkage onto an electrode towards efficient photocurrent generation [20]. Furthermore, the interaction of heme with a hemoprotein matrix has been employed for specific and tight protein–protein association to create supramolecular nanostructures towards unique biomaterials [21, 22].

In this chapter, we describe the details of practical and typical procedures for the reconstitution of apomyoglobin with an artificial heme derivative. In particular, we detail the preparation and the design strategy of the artificial cofactors and the characterization of the reconstituted protein.

2 Materials

Prepare all aqueous solutions using ultrapure water (deionized water with an electrical resistivity of 18 MΩ cm). Chemicals of the highest available grade should be used as obtained from commercial sources. Typically no further purification is required unless indicated otherwise.

2.1 Synthesis of Zinc Porphyrin

1. Dimethylformamide (DMF).
2. Methanol.
3. Protoporphyrin IX (Frontier Scientific, Inc.).
4. $Zn(OAc)_2$.

2.2 Preparation of Apomyoglobin

1. Horse heart myoglobin (Sigma Aldrich; if needed, the purchased protein can be further purified by cation exchange chromatography (Whatman CM-52) with a linear gradient: 10 mM potassium phosphate buffer, pH 6.0, to 100 mM potassium phosphate buffer, pH 7.0. Collect the fractions whose ratios of absorption at 408 and 280 nm are over 5.)
2. HCl aqueous solution (0.1 M).
3. Cooled 2-butanone at 4 °C.
4. Dialysis membrane (Spectrum Laboratories, molecular weight cut off: 6–8 kDa): Before use, steep the membrane in boiling water for 3 min and rinse it subsequently with cold water to remove the glycerol protection.
5. Potassium phosphate buffer (pH 7.0, 100 mM): Prepared by mixing two solutions of 100 mM K_2HPO_4 and KH_2PO_4 while monitoring the pH value.

2.3 Insertion of the Cofactor and Purification of the Obtained Protein

1. Dimethyl sulfoxide (DMSO).
2. Amicon Ultra-4 (MW: 10,000, Merck Millipore).
3. Econo-Column (Bio-Rad).
4. Sephadex G-25 (GE Healthcare); before packing the column, the resin should be well swollen with ultrapure water.

2.4 Characterization of Reconstituted Protein

1. Ammonium acetate buffer (5 mM): A freshly prepared solution shows a near neutral pH (6.8–7.0).

3 Methods

3.1 Design Considerations for an Artificial Cofactor

The hemoprotein cofactor protoheme IX consists of a porphyrin framework with a coordinated iron atom and three types of peripheral groups, namely two propionate side chains, four methyl groups and two vinyl groups which precisely interact with amino acid residues in the heme pocket. Thus, a designed artificial heme cofactor should generally maintain an appropriate molecular shape and several essential interactions between the cofactor and the protein. Considering the character of the heme–protein interaction, there are at least four methods to construct an artificial cofactor (Fig. 5):

1. Metal substitution such as substitution of iron for zinc, cobalt, manganese and so on.
2. Introduction of an additional peripheral group to the porphyrin framework.
3. Modification of the propionate side chains.
4. Exchange of the porphyrin framework for an artificial porphyrinoid.

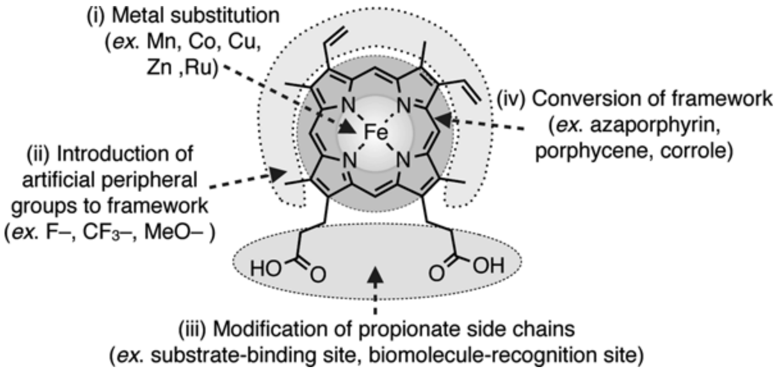

(i) Metal substitution
(ex. Mn, Co, Cu, Zn ,Ru)
(ii) Introduction of artificial peripheral groups to framework
(ex. F–, CF3–, MeO–)
(iii) Modification of propionate side chains
(ex. substrate-binding site, biomolecule-recognition site)
(iv) Conversion of framework
(ex. azaporphyrin, porphycene, corrole)
N
N
Fe
N
N
HO
O
O
OH

Fig. 5 Strategies for the design of an artificial cofactor for a series of hemoproteins

In practice, all of the above types of modifications have been reported: (1) To investigate the native function, an artificial cofactor using metal substitution of heme is attractive, especially cobalt porphyrin is rewarding for spin state studies. In addition, zinc porphyrin is suitable for the introduction of a photochemical probe into the protein. (2) The addition or exchange of a functional group in the porphyrin framework is one of the most effective methods to tune the function. For example, the substitution of a methyl and/or vinyl group for a trifluoromethyl group moderately changes the oxygen binding properties and contributed to the elucidation of the autoxidation mechanism [23]. (3) The propionate side chains being directed to the exterior of the protein allow for wide ranging modifications. Dendritic structures of for example polyanion-, polycation-, peptide-, oligonucleotide-, or carbohydrate-tethered heme have been reported for the construction of rationally functionalized hemoproteins [1, 2, 8]. (4) The exchange of the porphyrin framework leads to the most dramatic functional changes since the physicochemical properties of the metal center in metalloporphyrins is strongly controlled by the porphyrin framework [8, 14, 17]. This strategy is sometimes extremely challenging due to the encompassed synthetic difficulties en route to the envisaged porphyrinoids. In contrast, Watanabe and his coworkers demonstrated that a readily prepared salen metal complex consisting of a substantially different framework to porphyrin acts as an active cofactor for the apo-from of myoglobin mutant H64D-A71G [18]. This example illustrates that the reconstitution method is not limited to porphyrinoid metal complexes. Finally the selection from the above strategies must be appropriate for the envisioned design goal be that the creation of new artificial metalloenzymes, biomaterials or biodevices. The analysis and consideration of the above concepts will hopefully help in the design of artificial cofactors for the apo-forms of other cofactor-dependent proteins.

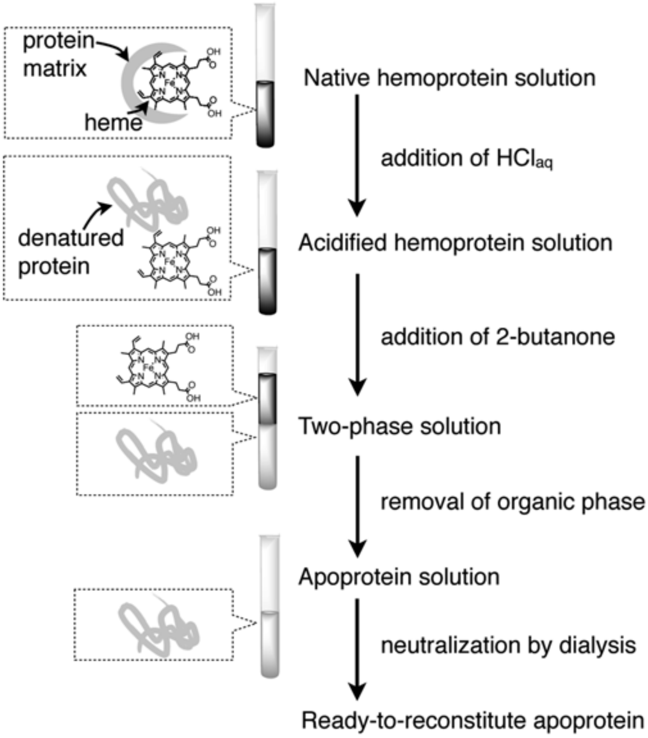
3.2 Synthesis of Zinc Porphyrin

Zinc porphyrin is easily synthesized by the reported procedure [24] with only small modifications. Both the free-base porphyrin and its zinc complex are highly photoactive. Therefore, the synthetic procedure should be performed under the exclusion of light.

1. Dissolve protoporphyrin IX (50 mg, 89 μmol) in DMF (5 mL).
2. Add 0.5 mL of a methanolic solution of $Zn(OAc)_2$ (200 mg, 1.1 mmol) (*see* **Note 1**).
3. Stir the mixture at 60 °C overnight.
4. After cooling in an ice bath, pour cold methanol (30 mL) into the reaction mixture.
5. Collect the precipitants by filtration and wash them well with cold methanol.
6. Dry the purple solid in vacuo and store in the freezer.

3.3 Preparation of Apomyoglobin

A typical procedure for the preparation of apo-hemoprotein is illustrated in scheme 1. Myoglobin is the most convenient scaffold for reconstitution because heme is readily replaceable while the scaffold displays relatively high heme-affinity. A number of examples have been published [1, 2, 8]. However, as some research groups have



Scheme 1 Procedures for the preparation of apo-hemoprotein

pointed out, overall the structures of almost all apoproteins are generally less stable than those of the corresponding holoproteins. Thus, apoproteins should be handled with care, stored at the appropriate temperature and the contamination with organic solvent, leading to irreversible aggregation, should be avoided [25].

1. Dissolve 10 mg of holo-myoglobin in 5 mL of ultrapure water in a glass cuvette (*see* **Note 2**).
2. In an ice bath, acidify the solution to pH 2.2 using 0.1 M HCl aqueous solution while monitoring the pH value (*see* **Notes 3** and **4**).
3. Add 5 mL of cooled 2-butanone to this aqueous solution and gently shake by repeated inverting of the capped cuvette.
4. Leave to settle or centrifuge the mixture at 4 °C and remove the red-colored organic phase by means of a pipette (*see* **Note 5**).
5. Repeat **steps 3** and **4** at least four times (*see* **Note 6**).
6. Confirm that the obtained aqueous phase is colorless. Transfer the solution into a dialysis membrane.
7. Dialyze the solution with 1 L of potassium phosphate buffer (100 mM, pH 7.0) for 2 h at 4 °C. Repeat the dialysis process three times to remove 2-butanone (*see* **Note** 7). The resulting solution should be stored at 4 °C.
8. (Optional step) If you do not need to insert a native or artificial cofactor into the apoprotein at this time, repeat the dialysis with 1 L of water at least three times and lyophilize the solution. The resulting powder can be stored at –80 °C for at least ten months.

3.4 Insertion of the Artificial Cofactor into Apomyoglobin and Purification of the Obtained Reconstituted Protein

An artificial cofactor with efficient affinity and specificity for a heme pocket will be incorporated into a corresponding apoprotein. In the case of myoglobin, the dropwise addition of a concentrated cofactor solution at 4 °C will readily give the corresponding reconstituted protein. If the cofactor is insoluble in aqueous media, a small amount of organic solvent such as DMSO or pyridine should be used to obtain a homogeneous cofactor solution.

1. To 10 μM solution of apomyoglobin in 100 mM potassium phosphate buffer (20 mL), pH 7.0, at 25 °C, add stepwise (2 μL per step) the DMSO solution of zinc porphyrin (3 mM) under the exclusion of light and monitor the incorporation by UV–Vis spectroscopy (*see* **Notes 8** and **9**). In general, within 5 min after each addition step the incorporation is complete.
2. Plot the absorbance at 428 nm against the total amount of zinc porphyrin after each addition step. Confirm the change in the slope of the curve at approximately 1 equivalent zinc porphyrin relative to the initial amount of apomyoglobin. The addition of a small excess (ca. 1.2–1.5 equivalents) of zinc

porphyrin to the protein is usually required to reach saturation (*see* **Note 10**).

3. After mild shaking over 2 h (*see* **Note 11**), concentrate the solution using an ultrafiltration membrane (e.g., Amicon Ultra-4) (*see* **Note 12**). The final volume of the concentrated solution should be smaller than 2 % of the column volume of the following gel-filtration.
4. Prepare the gel-filtration column (Sephadex G-25, column diameter: 1 cm, column length: 50 cm) by the manufacturer supplied procedure and equilibrate it with elution buffer (120 mL of 100 mM potassium phosphate buffer, pH 7.0). Carefully load the protein solution with a minimum amount of the eluent and then collect the colored solution after passing through the column while taking several fractions (*see* **Note 13**). Check the UV–Vis spectra of each fraction and combine the fractions with the highest ratio of absorbance at 428 nm versus 280 nm.
5. Concentrate the obtained protein solution to higher 1 mM but no more than 3 mM by means of ultrafiltration and store the concentrate in the freezer.

3.5 Characterization of Reconstituted Myoglobin

Characterization of reconstituted myoglobin is performed by UV–Vis, circular dichroism (CD), and electrospray ionization mass spectroscopic (ESI-MS) methods (*see* **Notes 14** and **15**).

1. UV–Vis spectral measurement and determination of extinction coefficient.

   Prepare the diluted protein samples with various concentrations (1–50 μM). Check each absorption maxima at each protein concentration. Next, measure the metal concentration for each protein concentration by means of inductively coupled plasma MS (ICP-MS) to determine the concentration of Zn-porphyrin.
2. CD spectrum measurement.

   To reduce noise, use a quartz cell with a short light path (1 mm path length is recommended). Prepare a diluted solution (less than 10 μM) of your reconstituted protein to check the folding in the wavelength region from 190 to 300 nm, where the signals caused by α-helices provide very high intensity. To check the signal assigned to cofactor absorption, use in contrast highly concentrated protein solution (over 100 μM) and measure in the region between 550 and 300 nm to get a conclusive readout.
3. Characterization of reconstituted protein by ESI-MS.

   For ESI-MS measurements, exchange the buffer salts to volatile ones. Common buffer components such as potassium phosphate are not vaporized and lead to clogging of the needle in the

MS during the ionization operation. A useful buffer for the measurement is a 5 mM NH_4OAc solution (*see* **Note 16**). Exchange the buffer by repeating concentration by ultrafiltration and dilution with NH_4OAc buffer at least 10 times (*see* **Note 17**). To check the formation of a cofactor-apoprotein complex, a slightly concentrated protein solution (~10 μM) is appropriate. To detect the mass of the cofactor-bound protein, the acceleration voltage in the detector should be as low as possible. For example on a Bruker micrOTOF mass spectrometer, we typically employ an acceleration voltage of 5.0 V, (60.0 and 55.0 V for the capillary exit and skimmer 1, respectively). To promote ionization, the solution is typically acidified by addition of AcOH. Take care of the pH value to avoid the denaturation of protein or the uncoupling of the supramolecular complex. For the above example a pH >5.5 and <7.5 is recommended. Multiple m/z values caused by species with varying degrees of protonation need to be deconvoluted with the appropriate software to extract the molecular mass of the analyte.

4 Notes

1. $Zn(OAc)_2$ should be applied in the synthesis as a saturated methanolic solution and typically a suspension of $Zn(OAc)_2$ in methanol is used.
2. Glass ware is most suitable due to its durability against organic solvents. However, a polypropylene (not polystyrene) centrifuge tube is also acceptable in this experiment.
3. If necessary, you can add L-histidine to support the dissociation of heme by acidification, for the preparation of the apoprotein under milder condition. For cytochrome $P450_{cam}$ and HRP, L-histidine is often added in concentrations up to 200 mM [12, 13]. For HRP, re-neutralization is required immediately after the extraction of heme. As an alternative to acidification, denaturation by guanidinium hydrochloride has been used to dissociate native heme before extraction by 2-butanone [20].
4. A method using acidified acetone, which is prepared by adding 2.5 mL of 2 M HCl to 1 L of acetone is suitable for the preparation of apohemoglobin [11].
5. Centrifugation is strongly recommended to efficiently separate the two phases.
6. In the final extraction step, the volume of 2-butanone should be reduced by about half compared to that of the initial extraction.

7. Although a threefold repetition of dialysis is not sufficient to remove 2-butanone entirely from the aqueous layer, the levels are typically sufficiently low for the reconstitution of the protein. If high purity apoprotein is required dialysis should be performed for another three cycles.

8. If there is no significant change in the UV–Vis spectra after incorporation of the artificial cofactor, try to monitor the incorporation by CD spectroscopy.

9. The incorporation of a cofactor into a series of apoproteins should be performed in aqueous solution without any organic solvents, if the cofactor is sufficiently soluble in aqueous media. If organic solvent is needed to dissolve the cofactor, its amount should be kept as low as possible (less than 1 %(v/v) in the final protein solution is recommended). The rate for the incorporation strongly depends on the content of organic solvent, temperature, the concentration of apoprotein and the solubility of the cofactor.

10. Excess amount of artificial cofactor sometimes causes problems for the subsequent purification, because often artificial cofactors tend to interact with the protein surface nonspecifically. Additionally, highly concentrated cofactor can undergo undesired aggregation which may suppress the insertion of the cofactor into the heme pocket.

11. Asymmetric heme cofactors such as protoheme or deuteroheme can bind in two modes, the forward and the backward form, in the heme pocket. Generally, it takes 5–10 h to reach at the equilibrium between the two configurations [5].

12. Artificial cofactor added in excess often precipitates during ultrafiltration In such a case, dialyze the solution for 2 h immediately after the addition of the cofactor to the apoprotein solution and then remove the precipitate by centrifugation before ultrafiltration, instead of following the procedure including the equilibration step (*see* **Note 11**).

13. Instead of the gel-filtration column, an ion exchange column is also useful. In the case of myoglobin, a purification on a DEAE (diethylaminoethyl) column can help to remove the excess of the artificial cofactor. To remove an excess of apomyoglobin, a purification on a SP (sulfopropyl) column is suitable.

14. With modern mass spectroscopy techniques, a large hemoprotein such as cytochrome P450$_{\text{cam}}$ (over 40 kDa) can be ionized in such a way that the supramolecular complex with an artificial cofactor can be maintained. It is even possible to characterize a cofactor-mediated supramolecular protein assembly, which reaches 100 kDa, by ESI-MS spectroscopy.

15. If a 3D structure of a native scaffold protein has been revealed by X-ray crystallography, the crystallization of the reconstituted protein should be tried under similar conditions.
16. An aqueous solution of $(NH_4)_2CO_3$ can also be useful for the measurement, although it affects positive ionization due to the higher pH.
17. To exchange the buffer component, gel filtration is also recommended. Utilization of a desalting column such as Hitrap desalting or PD-10 (GE healthcare) allows for efficient buffer exchange.

Acknowledgement

The preparation of this chapter was supported by Grants-in-Aid for Scientific Research from MEXT and JSPS.

References

1. Hayashi T, Hisaeda Y (2002) New functionalization of myoglobin by chemical modification of heme-propionates. Acc Chem Res 35:35–43
2. Fruk L, Kuo C-H, Torres E, Niemeyer CM (2009) Apoenzyme reconstitution as a chemical tool for structural enzymology and biotechnology. Angew Chem Int Ed 48:1550–1574
3. Hefti MH, Vervoort J, van Berkel WJH (2003) Deflavination and reconstitution of flavoproteins. Eur J Biochem 270:4227–4242
4. Zhou K, Oetterli RM, Brandl H, Lyatuu FE, Buckel W, Zelder F (2012) Chemistry and bioactivity of an artificial adenosylpeptide B_{12} cofactor. ChemBioChem 13:2052–2055
5. La Mar GN, Pande U, Hauksson JB, Pandey RK, Smith KM (1989) Proton nuclear magnetic resonance investigation of the mechanism of the reconstitution of myoglobin that leads to metastable heme orientational disorder. J Am Chem Soc 111:485–491
6. Ward TR (2009) Top organomet chem, Bio inspired catalysts. Springer, Berlin
7. Willner B, Katz E, Willner I (2006) Electrical contacting of redox proteins by nanotechnological means. Curr Opin Biotechnol 17: 589–596
8. Hayashi T (2013) Generation of functionalized biomolecules using hemoprotein matrices with small protein cavities for incorporation of cofactors. In: Ueno T, Watanabe Y (eds) Coordination chemistry in protein cages: principles, design, and applications. Wiley, Hoboken, pp 87–110
9. Hargrove MS, Wilkinson AJ, Olson JS (1996) Structural factors governing hemin dissociation from metmyoglobin. Biochemistry 35:11300–11309
10. Teale FW (1959) Cleavage of the haem-protein link by acid methylethylketone. Biochim Biophys Acta 35:543
11. Asoli F, Fanelli MR, Antonini E (1981) Preparation and properties of apohemoglobin and reconstituted hemoglobins. Methods Enzymol 76:72–87
12. Wagner GC, Perez M, Toscano WA Jr, Gunsalus IC (1981) Apoprotein formation and heme reconstitution of cytochrome P-450cam. J Biol Chem 256:6262–6265
13. Ator MA, David SK, Ortiz de Montellano PR (1989) Stabilized isoporphyrin intermediates in the inactivation of horseradish peroxidase by alkylhydrazines. J Biol Chem 264:9250–9257
14. Matsuo T, Murata D, Hisaeda Y, Hori H, Hayashi T (2007) Porphyrinoid chemistry in hemoprotein matrix: detection and reactivities of iron(IV)-oxo species of porphycene incorporated into horseradish peroxidase. J Am Chem Soc 129:12906–12907
15. Itagaki E, Palmer G, Hager LP (1967) Studies on cytochrome b_{562} of Escherichia coli. II. Reconstitution of cytochrome b_{562} from apoprotein and hemin. J Biol Chem 242: 2272–2277
16. Matsuo T, Fukumoto K, Watanabe T, Hayashi T (2011) Precise design of artificial cofactors for enhancing peroxidase activity of myoglobin:

myoglobin mutant H64D reconstituted with a "single-winged cofactor" is equivalent to native horseradish peroxidase in oxidation activity. Chem Asian J 6:2491–2499

17. Hayashi T, Dejima H, Matsuo T, Sato H, Murata D, Hisaeda Y (2002) Blue myoglobin reconstituted with an iron porphycene shows extremely high oxygen affinity. J Am Chem Soc 124:11226–11227
18. Ueno T, Koshiyama T, Ohashi M, Kondo K, Kono M, Suzuki A et al (2005) Coordinated design of cofactor and active site structures in development of new protein catalysts. J Am Chem Soc 127:6556–6562
19. Patolsky F, Weizmann Y, Willner I (2004) Long-range electrical contacting of redox enzymes by SWCNT connectors. Angew Chem Int Ed 43:2113–2117
20. Onoda A, Kakikura Y, Uematsu T, Kuwabata S, Hayashi T (2012) Photocurrent generation from hierarchical zinc-substituted hemoprotein assemblies immobilized on a gold electrode. Angew Chem Int Ed 51: 2628–2631
21. Kitagishi H, Oohora K, Yamaguchi H, Sato H, Matsuo T, Harada A et al (2007) Supramolecular hemoprotein linear assembly by successive interprotein heme–heme pocket interactions. J Am Chem Soc 129:10326–10327
22. Oohora K, Onoda A, Hayashi T (2012) Supramolecular assembling systems formed by heme–heme pocket interactions in hemoproteins. Chem Commun 48:11714–11726
23. Shibata T, Matsumoto D, Nishimura R, Tai H, Matsuoka A, Nagao S et al (2012) Relationship between oxygen affinity and autoxidation of myoglobin. Inorg Chem 51:11955–11960
24. Smith KM (1975) Porphyrins and metalloporphyrins. Elsevier, Amsterdam
25. Fändrich M, Forge V, Buder K, Kittler M, Dobson CM, Diekmann S (2003) Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments. Proc Natl Acad Sci U S A 100:15463–15468

Chapter 14

Computational Redesign of Metalloenzymes for Catalyzing New Reactions

Per Jr. Greisen and Sagar D. Khare

Abstract

The ability to design novel activities in existing metalloenzyme active sites is a stringent test of our understanding of enzyme mechanisms, sheds light on enzyme evolution, and would have many practical applications. Here, we describe a computational method in the context of the macromolecular modeling suite Rosetta to repurpose active sites containing metal ions for reactions of choice. The required inputs for the method are a model of the transition state(s) for the reaction and a set of crystallographic structures of proteins containing metal ions. The coordination geometry associated with the metal ion (Zn^{2+}, for example) is automatically detected and the transition state model is aligned to the open metal coordination site(s) in the protein. Additional interactions to the transition state model are made using RosettaMatch and the surrounding amino acid side chain identities are optimized for transition state stabilization using RosettaDesign. Validation of the design is performed using docking and molecular dynamics simulations, and candidate designs are generated for experimental validation. Computational metalloenzyme repurposing is complementary to directed evolution approaches for enzyme engineering and allows large jumps in sequence space to make concerted sequence and structural changes for introducing novel enzymatic activities and specificities.

Key words Enzyme design, Rosetta software, Enzyme redesign, Metalloenzymes, Zinc ions

1 Introduction

Metal ions are versatile catalysts for carrying out biological and non-biological reactions, affording rates and reaction mechanisms not accessible in conventional acid–base or covalent catalysis. Considerable effort has been made to design artificial metalloproteins [1, 2], however the de novo design of metal-dependent enzyme active sites has been challenging because of stringent design requirements: (a) multiple flexible, polar residues are necessary to bind the metal and these must be held in place by additional second-shell residues, (b) destabilization of alternative conformations that would disrupt the designed conformation is necessary (negative design), and (c) second- and third-shell effects can be critical for modulating the electrostatic environment of the active

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9_14, © Springer Science+Business Media New York 2014

site that is a key determinant of metal reactivity. In view of these requirements, the repurposing of existing metal binding sites for accessing new chemical reactivity offers a relatively simple, yet effective strategy for design.

One of the most common metal ions in biology is the zinc ion (Zn^{2+}). Zn^{2+} is coordinated by proteins in three different coordination geometries, tetrahedral, trigonal bipyramidal, and octahedral. The reactivity of the site is regulated by the coordinating groups of the metal which are usually histidine (His), aspartic acid (Asp), glutamic acid (Glu), or cysteine (Cys) [3]. The coordination sphere of the metal ion is flexible and it is possible to tune the reactivity of the site by means of the ligands. In enzymes, Zn^{2+} can act as a Lewis acid for example in alcohol dehydrogenase, or it activates a water molecule to perform nucleophilic attack on a substrate for example in carbonic anhydrase, or both [4, 5]. Therefore, zinc metalloenzymes provide a viable platform for the introduction of novel activities using computational repurposing.

Reusing some or all of the catalytic elements in existing enzyme active sites for new chemistry is a common theme in natural enzyme evolution, and underlies the functional diversification seen in enzyme superfamilies. In contrast, de novo computational enzyme design aims at placing catalytic elements in otherwise inert scaffolds to introduce new reactivity. We have implemented a computational design strategy that is inspired by natural enzyme evolution in that it reuses existing catalytic elements of enzyme active sites but also uses de novo enzyme design methods to rationally engineer new activities in the framework of the macromolecular modeling suite Rosetta. Application of this method to a set of mononuclear zinc enzymes led the design of organophosphate hydrolysis activity in an adenosine deaminase [6] (Figs. 1 and 2). The method described below illustrates the approach used for the above design project, but has been extended to binuclear metal sites and for a variety of other reactions including *s*-triazine, beta-lactam, and cyanuric acid hydrolysis (unpublished data).

2 Methods

Starting from a transition state model of the reaction under consideration and a set of zinc-containing PDB files as inputs, we generate a design model and evaluate it. The overall workflow involves the following steps:

1. Generation of the TS ensemble.
2. Analysis of metal site in the PDB file(s) and classification.
3. Alignment of TS ensemble to the curated active site set.
4. Minimization of TS ensemble in a polyAla pocket (optional).

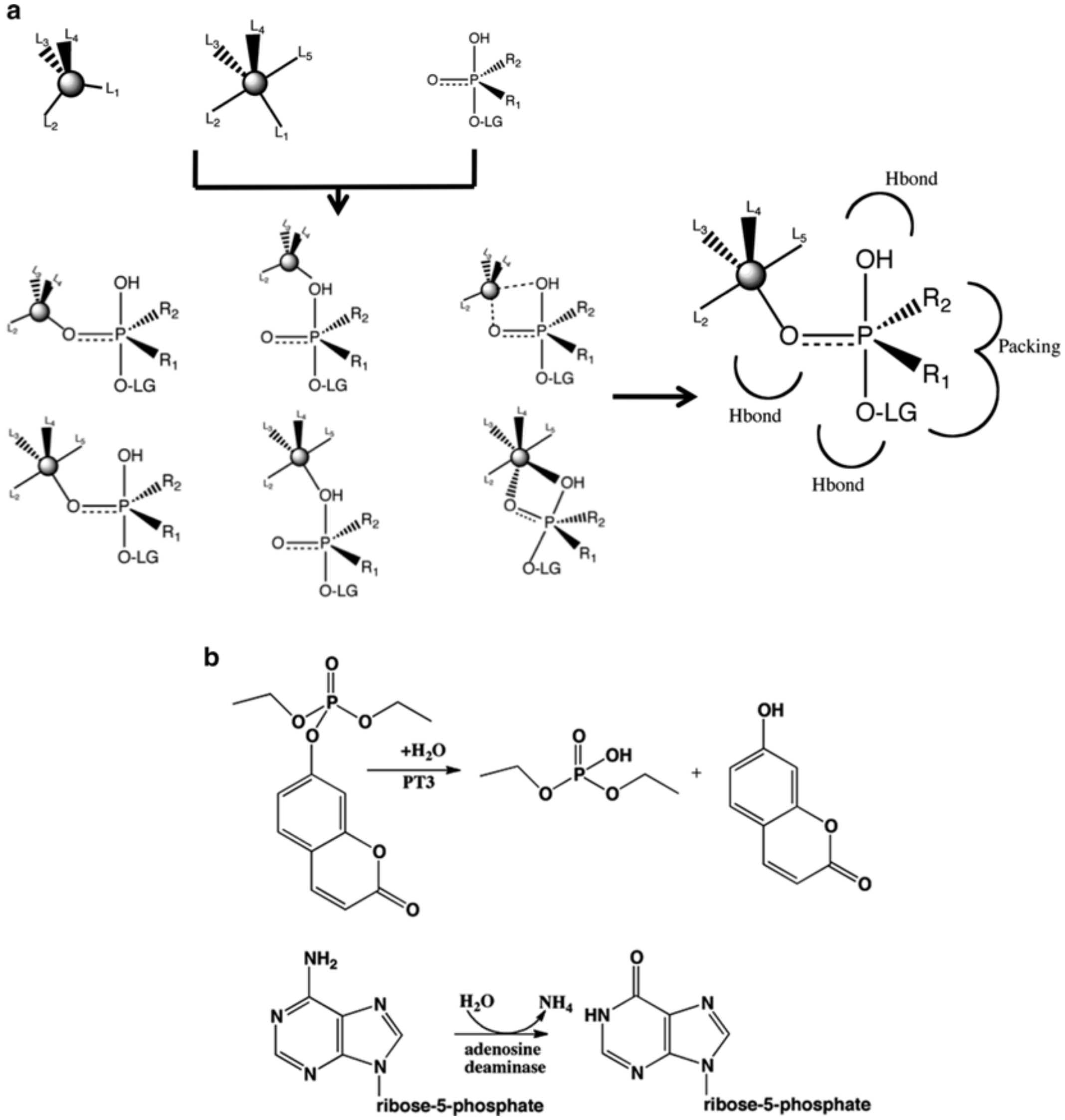


Fig. 1 (**a**) Scheme for the computational repurposing of active sites. Different zinc coordination sites found in crystallographic structures in the Protein Databank are curated (e.g., tetrahedral and trigonal bipyramidal) and a model of the transition state of the reaction under consideration is superimposed on the open coordination site(s) of the metal ion in each PDB file. LG is leaving group, and L_1, L_2 etc. represent zinc ligands. RosettaMatch and RosettaDesign are used to design additional TS stabilizing interactions. (**b**) Using this approach, organophosphate hydrolysis activity (*top*) was designed into an adenosine deaminase (*bottom*)

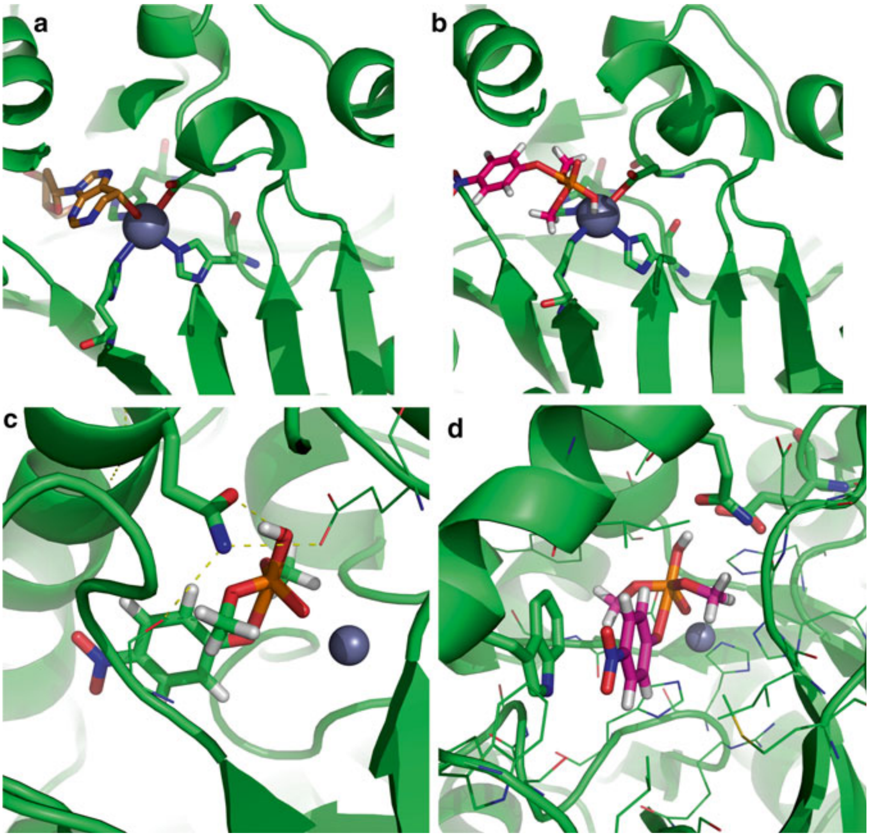


Fig. 2 Example of computational enzyme repurposing. (**a**) The original adenosine deaminase crystal structure bound to an inhibitor (PDB code 1A4L). (**b**) Structure of the active site after the new organophosphate hydrolase TS was superimposed. (**c**) RosettaMatch was used to identify additional hydrogen bonds to the TS. The residue Gln58 was placed to interact with the attacking nucleophile. (**d**) RosettaDesign was used to identify additional TS stabilizing interactions. The residue W65 was found to make pi–pi stacking interactions with the leaving group

5. Introducing additional catalytic residues.
6. Sequence design to maximize TS affinity.
7. Reversion of destabilizing sequence changes to wild type identities.
8. Docking the TS models in the designed active site (validation) using RosettaDock and molecular dynamics simulations (Fig. 3).
9. Protein expression, purification, and experimental characterization (not discussed here).

2.1 Transition State Ensemble

The TS ensemble is generated using a TS analog structure and/or using quantum chemical simulations of the reaction under consideration. For our purpose, we assume that a molecular model of the TS ensemble can be obtained. For constructing the Rosetta model, typically we start from a molfile representation, and

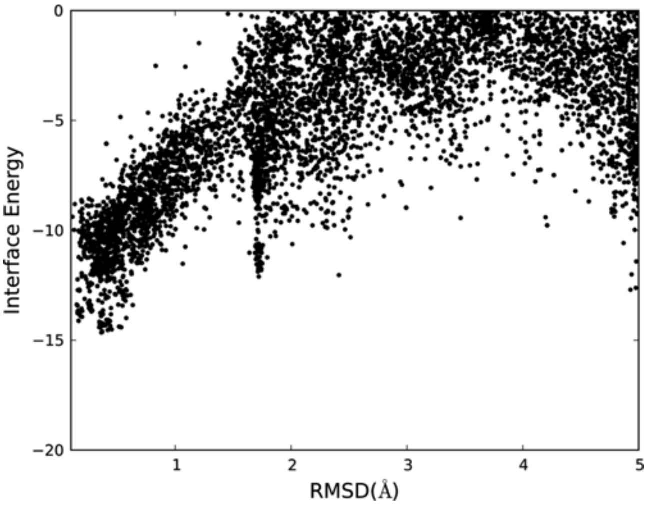


Fig. 3 Design validation using docking. RosettaDock was used to interrogate the energy landscape of the designed protein bound to the TS model. A robust funnel indicated by low interface energies corresponding to the conformations similar to the designed position of the TS (low RMSD) suggests that alternative binding modes are disfavored

convert it to Rosetta parameters using a script provided with the Rosetta software: /path/to/rosetta/rosetta_source/src/python/apps/public/molfile_to_params.py.

2.2 Analysis of Metal Site in the PDB File(s) and Classification by Coordination Geometry

To select protein scaffolds suitable for design, all protein structures from the Protein Data Bank (PDB) are collected. The search is limited to high-resolution structures (<2.9 Å) and a sequence identity cutoff of 90 % is used to rule out structures that differ only slightly from another. The proteins are chosen such that they have been expressed in *E. coli* and contain Zn^{2+} in the structures. Non-catalytic sites as well as surface-bound zinc ions from crystal structures or active sites with multiple Zn^{2+} are excluded: (a) all Zn^{2+} sites with less than 3 coordinating atoms from the protein, (b) sites where there are two or more metal ions such that the metal–metal distance is less than 5 Å, and (c) structural sites defined as being coordinated to 4 Cys/His residues are removed. The PDB files are further modified by removing redundant protein chains or protomers. Alternate side-chain positions and atomic coordinates are discarded keeping one protein chain along with its metal site. In our study, the total number of selected proteins was 105 and included a variety of protein folds and enzymes classes.

2.3 Alignment of the TS Ensemble to the Curated Active Site Set

An algorithm was developed to align the TS model onto the Zn in the native protein scaffold. It uses the classification of the metal site in the protein described above and the coordinating atoms to identify the direction of the enzymatic pocket. To explore as many possibilities for the chemical reaction and to take advantage of the functional diversity of the zinc ion, different alignments were performed in our case: monodentate with the hydroxyl or the phosphoryl formally coordinating the Zn^{2+} and a bidentate alignment with both the hydroxyl and phosphoryl aligning the Zn^{2+} (Fig. 1). Depending on the alignments, the heteroatom coordinating the Zn^{2+} in the PDB file was used to superimpose the TS model.

Coordinate constraints were generated from the crystal structures to keep the Zn^{2+} in the same position during the design process. The ligand-protein interaction was optimized using a TS conformer library such that clashes with the protein backbone and sidechain clashes with zinc-coordinating residues were minimized. At this stage, all amino acids of the protein except those coordinating the zinc ion were converted into alanines, and an energy function dominated by the Lennard-Jones repulsive potential was used for steepest descent minimization [7].

2.4 Finding Additional Interactions to Buttress the TS Using RosettaMatch

As the newly placed TS model can have unsatisfied hydrogen bond donors/acceptors it is important to introduce additional interactions that can further stabilize the TS model and hence enhance catalysis. The secondary matching algorithm [8] implemented in Rosetta is used to introduce additional interactions to the TS model. Briefly, it goes through all the positions on a protein scaffold to see if it can “match” the interactions required (e.g., hydrogen bonds) to any sidechains on the protein. Here, a secondary match was performed for either a base or an oxyanion hole depending on the reaction mechanism under consideration.

2.5 Protein Design of Matches

All hits from the secondary matches contained one modification to their sequences positioning either a base, an oxyanion hole, or the attacking hydroxyl. To stabilize the modifications introduced, the energy function in Rosetta for protein [9] and ligand [10] was used. The interaction between ligand and protein was optimized to decrease the interaction energy between the TS model and the protein as well as to stabilize the protein sequence for folding. All designs having an interface energy <-3 Rosetta units (RU), solvent accessible surface area (SASA) between 0.8 and <1, and a constraint energy less than 3 RU were collected for further analysis.

2.6 Evaluation of Designs

2.6.1 Reversion to Wild Type Amino Acid Identity

To evaluate the mutations introduced by Rosetta, a multiple sequence alignment of the protein scaffold homologs was generated to evaluate how conserved the mutated residues were. This was performed to remove substitutions that might be important for folding. All designed sequences were evaluated using the ConSurf server [11, 12]. Residues that had a high conservation score and were known to not be a part of the native catalytic machinery were reverted back to their native residue type as they were assumed important for folding or solubility. In general, glycine (Gly) and proline (Pro) residues were reverted back to the wild type residues in the designed protein. The designs were further evaluated using the visual interface to Rosetta—Foldit [13].

2.6.2 Evaluation of Alternative Binding Pockets Using RosettaDock

The TS model was docked into the protein to observe if other local energy minima existed. The TS model was docked into the design structure 10,000 times using the docking algorithm implemented in Rosetta. If other local binding minima existed, the binding of the TS model was reoptimized by attempting to disfavor alternative binding modes by making new amino acid substitutions.

2.6.3 Molecular Dynamics Simulation of Designs

To explore the phase space of the substrate in the designed structure a MD simulation was performed using Amber10. The tautomer state for histidine residues was established by visual inspection of the proteins (native and designed protein). The parameters for the substrate were generated using the module Antechamber [14] within AmberTools [15]. The metal site was fixed by applying bond, angle, and dihedral constraints to the site. The weights were varied between 10 and 100 and the charge of the metal ion was set to +2. The structures were minimized for 10,000 steps with 5,000 steps using steepest descent. The complexes were solvated and neutralized by placing ions in the simulation box. The water molecules were parameterized with the TIP3P water model [16]. After minimization, the system was heated up to 300 K using weak restraints on the protein complex for 100 ps with an integration step of 1 fs. The volume was fixed for equilibration of the pressure. Next, the system was equilibrated for 1 ns in the NPT ensemble keeping a pressure of 1 atm and the temperature constant at 300 K using a Langevin thermostat with a collision frequency of 1 ps^{-1} and the integration time step was changed to 2 fs. All hydrogens were fixed using SHAKE [17] and full electrostatics were computed using the Particle Mesh Ewald [18] with periodic boundary conditions. The MD simulation was run for 20 ns using the ff99SB force field [19].

3 Notes

While the ultimate goal of computational design methods is to automate all design steps, in practice most protocols rely upon the chemical intuition and domain knowledge of the user. Our method is no exception and so below we give some suggestions about aspects that need to be considered by the user while evaluating the designs generated by the protocol described above.

1. The Rosetta force field, as other molecular mechanics force fields, does not accurately model interactions of protein functional groups with metal ions and especially not with metal ions in active sites, therefore, it is necessary to treat these interaction with restraints. The weights used in the restraints

will be system dependent but in the final models one should end up with a metal site geometry similar to the one from the starting crystal structure with some small deviation. If the metal site is completely distorted, the weights of the restraints should be increased to keep the geometry fixed.

2. In the generation of the transition state ensemble or during minimization with Rosetta, it is useful to vary some internal angles and distances of the TS model, which will change the TS model geometry either calculated from quantum mechanics or generated from empirical knowledge. During our use of the protocol, the TS model was varied in order to sample different placements of the model in the active sites. Again, the variance of models will be case-dependent and theoretical or experimental knowledge on the reaction mechanism should be included when one decides how much variation to include in the models.
3. Another metric that is currently evaluated by human intuition in our protocol is whether the substrate can enter (and product can leave) the pocket of the active site and that access to the active site has not been blocked by new mutations introduced in the design protocol. Conformational changes upon substrate binding are not modeled and system-dependent knowledge of the dynamics of the closure and opening of the active site should be kept in mind when picking out scaffolds for design and evaluating designs by inspection.
4. Many substitutions can be introduced but as a designer one should also make sure that the initial protein scaffold can accommodate these changes in the absence of any substrate, otherwise the enzyme will either not express or be unfolded.
5. Chemical intuition is almost always required to evaluate the goodness of designs.

References

1. Lu Y, Yeung N, Sieracki N, Marshall NM (2009) Design of functional metalloproteins. Nature 460:855–862
2. Zastrow ML, Pecoraro VL (2013) Designing functional metalloproteins: from structural to catalytic metal sites. Coord Chem Rev 257: 2565–2588
3. Vallee BL, Auld DS (1990) Zinc coordination, function, and structure of zinc enzymes and other proteins. Biochemistry 29:5647–5659
4. Vallee BL, Hoch FL (1955) Zinc, a component of yeast alcohol dehydrogenase. Proc Natl Acad Sci U S A 41:327–338
5. Christianson DW, Cox JD (1999) Catalysis by metal-activated hydroxide in zinc and manganese metalloenzymes. Annu Rev Biochem 68: 33–57
6. Khare SD, Kipnis Y, Greisen P, Takeuchi R, Ashani Y, Goldsmith M et al (2012) Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. Nat Chem Biol 8:294–300
7. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93
8. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D (2011) De novo enzyme design using Rosetta3. PloS One 6:e19230
9. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a

novel globular protein fold with atomic-level accuracy. Science 302:1364–1368

10. Meiler J, Baker D (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins 65:538–548
11. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38(Web Server issue):W529–W533
12. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T et al (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33(Web Server issue):W299–W302
13. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M et al (2010) Predicting protein structures with a multiplayer online game. Nature 466:756–760
14. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model 25:247–260
15. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr et al (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668–1688
16. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
17. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical-integration of Cartesian equations of motion of a system with constraints – molecular-dynamics of N-alkanes. J Comput Phys 23:327–341
18. Darden T, York D, Pedersen L (1993) Particle mesh Ewald – an N.Log(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
19. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins Struct Funct Bioinf 65: 712–725

Index

Valentin Köhler (ed.), *Protein Design: Methods and Applications*, Methods in Molecular Biology, vol. 1216,
DOI 10.1007/978-1-4939-1486-9, © Springer Science+Business Media New York 2014