Protein Engineering and Design

Edited by
Sheldon J. Park
Jennifer R. Cochran

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Petitioner Merck, Ex. 1017, p. Cover

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2010 by Taylor and Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-7658-5 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Protein engineering and design / editors, Sheldon J. Park, Jennifer R. Cochran.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-1-4200-7658-5 (hardcover : alk. paper)
1. Protein engineering. I. Park, Sheldon J. II. Cochran, Jennifer R. III. Title.
[DNLM: 1. Protein Engineering. 2. Models, Molecular. 3. Proteins--chemistry. QU 450 P9668 2010]

TP248.65.P76P738 2010
660.6'3--dc22 2009024409

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface.....vii
Editors.....ix
Contributor List.....xi

Chapter 1 Phage Display Systems for Protein Engineering.....1
Andreas Ernst and Sachdev S. Sidhu

Chapter 2 Cell Surface Display Systems for Protein Engineering.....23
Sarah J. Moore, Mark J. Olsen, Jennifer R. Cochran, and Frank V. Cochran

Chapter 3 Cell-Free Display Systems for Protein Engineering.....51
Pamela A. Barendt and Casim A. Sarkar

Chapter 4 Library Construction for Protein Engineering.....83
Daša Lipovšek, Marco Mena, Shaun M. Lippow, Subhayu Basu, and Brian M. Baynes

Chapter 5 Design and Engineering of Synthetic Binding Proteins Using Nonantibody Scaffolds.....109
Shohei Koide

Chapter 6 Combinatorial Enzyme Engineering.....131
Patrick C. Cirino and Christopher S. Frei

Chapter 7 Engineering of Therapeutic Proteins.....153
Fei Wen, Sheryl B. Rubin-Pitel, and Huimin Zhao

Chapter 8 Protein Engineered Biomaterials.....179
Cheryl Wong Po Foo and Sarah C. Heilshorn

Chapter 9 Protein Engineering Using Noncanonical Amino Acids.....205
Deniz Yüksel, Diren Pamuk, Yulia Ivanova, and Krishna Kumar

Petitioner Merck, Ex. 1017, p. v

Chapter 10 Computer Graphics, Homology Modeling, and Bioinformatics 223

David F. Green

Chapter 11 Knowledge-Based Protein Design 239

Michael A. Fisher, Shona C. Patel, Izhack Cherny, and Michael H. Hecht

Chapter 12 Molecular Force Fields..... 257

Patrice Koehl

Chapter 13 Rotamer Libraries for Molecular Modeling and Design of Proteins..... 281

Hidetoshi Kono

Chapter 14 Search Algorithms..... 293

Julia M. Shifman and Menachem Fromer

Chapter 15 Modulating Protein Structure..... 313

M.S. Hanes, T.M. Handel, and A.B. Chowdry

Chapter 16 Modulation of Intrinsic Properties by Computational Design..... 327

Vikas Nanda, Fei Xu, and Daniel Hsieh

Chapter 17 Modulating Protein Interactions by Rational and Computational Design..... 343

Jonathan S. Marvin and Loren L. Looger

Chapter 18 Future Challenges of Computational Protein Design 367

Eun Jung Choi, Gurkan Guntas, and Brian Kuhlman

Index..... 389

Preface

Proteins possess a broad range of structural and functional properties that are unmatched by any other class of biological molecules. Amazingly, nature has arranged simple atoms and chemical bonds in such a way to facilitate complex biological processes like molecular recognition and catalysis. Nature has also inspired many scientists and engineers to design and create their own customized proteins. These engineered proteins can serve as novel molecular tools for scientific, medical, and industrial applications, thus addressing many needs unmet by naturally occurring proteins.

Protein engineering requires identification of particular amino acid sequences that will result in desired structural and functional properties. Despite recent advances in the field, however, protein engineering remains as much an art as it is a science. Engineering an arbitrary protein structure or function remains a formidable challenge, because the rules defining sequence-structure-function relationships are still not well understood. Even with refined quantitative models, the large degrees of freedom present in a typical protein do not easily allow identification of optimal sequences using currently available computational techniques. Furthermore, the complexity of proteins present engineering challenges whose solutions will most likely require a combination of experimental and computational approaches.

This book discusses two general strategies commonly used to engineer new proteins: diversity-oriented protein engineering and computational protein design. Diversity-oriented protein engineering, or directed evolution, identifies protein variants with desired properties from a large pool of mutants. As such, its success depends on generating sufficient sequence diversity and employing sensitive high-throughput assays. Computational protein design, on the other hand, generates and screens protein sequences *in silico* before synthesizing them in the laboratory. This is still an unfamiliar concept to many, so an important goal of this book is to demystify the subject by describing its development and current implementations. Structure-based protein engineering similarly uses computation to facilitate the discovery of interesting protein sequences. However, computational protein design places emphasis on both engineering new, useful proteins and on testing sequence-structure relationships. In this regard, it shares a deep philosophical root with protein folding, which similarly seeks to understand the relationship between protein sequence and tertiary structure.

The book is organized into two sections. The first half of the book discusses experimental approaches to protein engineering and starts by describing several high-throughput protein engineering platforms (Chapters 1–3). This is followed by a chapter on key techniques used for diversity generation (Chapter 4). The next few chapters present examples of therapeutics, enzymes, biomaterials, and other molecules that were engineered by rational or combinatorial-based approaches (Chapter 5–8). The section finishes with a chapter on the use of unnatural amino acids in protein engineering (Chapter 9).

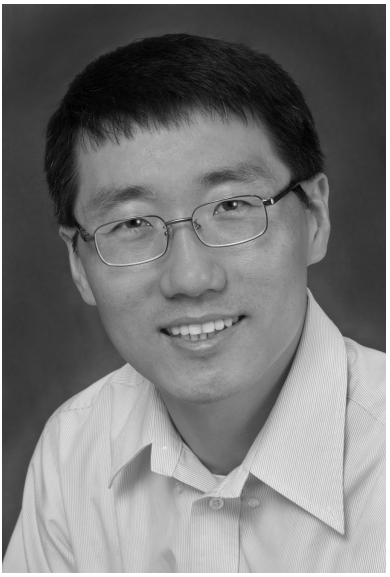
Petitioner Merck, Ex. 1017, p. vii

The second half of the book introduces computational protein design, which designs new sequences by quantitatively modeling sequence-structure relationships. Despite their unique approaches, protein engineering and design are increasingly developing a synergistic relationship. To that end, more and more experimentalists are recognizing computation as an important molecular tool for protein engineering, and vice versa. These days, it is routine for those planning a protein engineering project to first perform sequence analysis and to visualize protein structures in a molecular viewer. It is thus appropriate to start this section with a chapter on the common use of computers and informatics in protein engineering (Chapter 10). Examples of heuristic protein design are described in Chapter 11, before the core components of computational protein design are discussed in detail in Chapters 12–14. Subsequent chapters present examples of computationally designed proteins that played critical roles in advancing the use of computers in protein engineering (Chapters 15–17). The field has not yet fully matured and there are difficulties that remain to be resolved; these challenges are discussed in the last chapter of the book (Chapter 18).

Modern biology has provided a deep understanding of the molecular nature of biological processes. In particular, we now have a variety of tools that can be used to analyze and control key biological processes with molecular precision. Protein engineering and design are attempts to accomplish exactly these goals. As examples throughout the book show, certain categories of problems have attracted attention from scientists and engineers with a diverse range of technical expertise. We hope these studies will help the reader identify potential opportunities to bridge experimental protein engineering and computational protein design and will lead to exciting breakthroughs in biotechnology and medicine.

Sheldon J. Park
Jennifer R. Cochran

Editors



Sheldon Park holds a B.A. in math and physics from the University of California (Berkeley), an M.S. in physics from Massachusetts Institute of Technology, and a Ph.D. in biophysics from Harvard University. He studied protein engineering and design while working as a postdoc for Dr. Jeffery Saven and Dr. Eric Boder at the University of Pennsylvania. Since 2006, he has been a professor of chemical and biological engineering at University at Buffalo. In his research, Dr. Park uses modeling and simulation to analyze protein molecules and uses high-throughput screening to engineer protein molecules of various structure and function. He is particularly interested in developing efficient methods of engineering complex protein molecules with potential biotechnological and biomedical applications.

Jennifer Cochran holds a B.S. in biochemistry from the University of Delaware and a Ph.D. in biological chemistry from Massachusetts Institute of Technology (MIT). She studied and developed combinatorial protein engineering methods while a postdoctoral fellow in the lab of K. Dane Wittrup in the Department of Biological Engineering at MIT. Since 2005, she has been a professor of bioengineering at Stanford University. Dr. Cochran's laboratory uses interdisciplinary approaches in chemistry, engineering, and biophysics to study complex biological systems and to create designer protein therapeutics and diagnostic agents for biomedical applications. She is interested in elucidating molecular details of receptor-mediated cell signaling events and at the same time developing protein and polymer-based tools that will allow manipulation of cell processes on a molecular level.

Petitioner Merck, Ex. 1017, p. ix

10 Computer Graphics, Homology Modeling, and Bioinformatics

David F. Green

CONTENTS

Primary Sequence Analysis 224
Engineering through Consensus Motifs 224
Pairwise Interactions from Sequence Analysis 227
Graphical Analysis of Protein Structure 228
Homology Modeling and Structure Visualization 229
Modulation of Stability and Affinity through Steric Complementarity 231
Modulation of Affinity through Electrostatic Complementarity 232
Fast Methods for Mutational Evaluation 234
Affinity Enhancement through Peripheral Electrostatic Interactions 234
Summary 236
References 236

Computational methods play a range of roles in protein engineering from the simple use of visualization to guide rational design to fully automated *de novo* design algorithms. In the next chapters, many of these approaches will be discussed. Here we will focus on the former, that is, computational methods that complement human insight in rational protein engineering. The approaches can loosely be grouped into three classes: (1) methods based on analysis of primary sequence; (2) the visual analysis of protein structure; and (3) fast estimation of mutational effects. The mechanistic details of performing sequence and structural analysis have been extensively discussed in other texts, and thus the focus here is on the application of these approaches. The approaches discussed here all involve use of software that is either available as a Web service or as a freely available, downloadable program. The Web locations of key tools are summarized in the tables.

Petitioner Merck, Ex. 1017, p. 223

PRIMARY SEQUENCE ANALYSIS

Evolution provides a tremendously useful model for protein design. As seen in previous chapters, several approaches to mimicking evolution in the laboratory have been demonstrated to be powerful methods for the engineering of improved or novel function, but we may also take advantage of the results of natural evolution. Many families of proteins contain hundreds or thousands of members, spread across diverse species. By considering the common features of the sequences of these proteins, it is possible to deduce the key elements that determine protein structure and function—even in absence of any explicit structural information. In order to take this approach, several tools are needed. First, given one (or a few) sequences of a target structure, it is necessary to be able to search through the vast array of known sequences for related proteins. Second, this large family of related proteins must be aligned such that conserved positions are in register with one another. From this point, analysis of the degree of conservation at each position can give important insight applicable to protein engineering. Computational methods for this analysis of primary sequence are well established, and many tools are available through Web-based servers (Table 10.1).

Engineering through Consensus Motifs

One of the most straightforward applications of primary sequence data in protein engineering is the use of multiple-sequence alignments to define consensus motifs for a particular structure or function. These sequence signatures focus on the common features of a class, while not corresponding to any natural sequence. As a result, the resulting sequence may be expected to share the features that all members of the family have in common (such as a particular structure) without the specific features of particular family members (such as affinity for a specific binding partner). One of the first applications of this approach was in the design of a consensus-based zinc finger protein (Krizek et al. 1991). Subsequently, it was demonstrated that frequency of occurrence in a multiple-sequence alignment was a

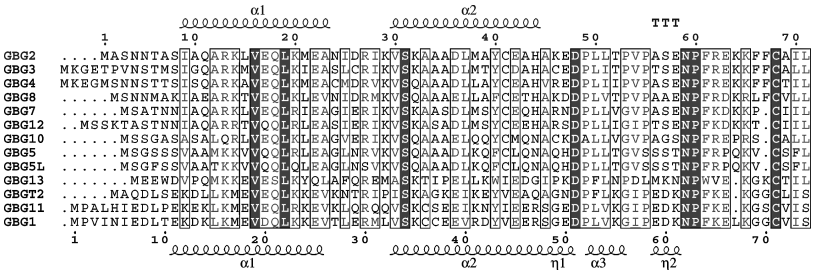
TABLE 10.1
Web Services and Databases for Primary Sequence Analysis

Service	Web Location (URL)	Description
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/	Repository of all publicly available nucleotide sequences.
Swiss-Prot	http://ca.expasy.org/sprot/	Annotated database of protein sequences.
BLAST	http:/www.ncbi.nlm.nih.gov/BLAST/	Online service to search for related sequences.
ClustalW2	http://www.ebi.ac.uk/tools/clustalw2/index.html	Online service for multiple sequence alignment.
ClustalW	http://www.clustal.org/	Downloadable software for multiple sequence alignment.

good predictor of the effects of point mutations on the stability of an immunoglobulin domain (Steipe et al. 1994). These early implementations took a fairly simple approach; relatively small numbers of sequences were used, and global differences in amino-acid frequencies were not considered. An example of the approach is shown in Figure 10.1. More recently, these approaches have been refined and applied to a number of diverse systems.

The fundamental approach is as follows.

1. Select a set of known sequences of the target protein family.
2. Use the profile of these sequences to search global sequence databases for additional family members.
3. Perform multiple sequence alignment on this large set of sequences.
4. Compute statistical enrichment measures for the occurrence of each amino acid at each position.
5. Use this information to bias the selection of sequences in an engineering context.



(A)

```
100%  -------------V--L-----------S----------------D----------NP-------C---

 50%  -----------K-VEQLK-EA---R-KVSKAAADL--YCE--A--DPLL-GVP-S-NPFRE-K--C-IL

 20%  MNSTASIAQARKEVEQLKLEAGLERIKVSKAAADLMAYCEQHAREDPLLTGVPASENPFREKKGFCAIL
```

(B)

FIGURE 10.1 (see color insert following page 178) Multiple sequence alignment and consensus sequence determination. (A) A multiple sequence alignment highlights key similarities between the sequences. Here, a red background indicates 100% identity across all sequences, blue boxes indicate largely conserved residues, with similar residue types in red, and black indicates unconserved residues. The known regions of helical secondary structure for the top and bottom sequences are also shown. Figure generated with ESPRIPT (Gouet et al. 1999). (B) From the multiple sequence alignment, consensus sequences can be derived. Only seven residues are strictly conserved across all sequences (100%), while for 39 positions, there is one amino acid that occurs in the majority of sequences (50%). A complete sequence of 69 residues can be obtained by choosing the most common amino at each position (20% of the chosen amino acids all occur in at least three of the sequences). The sequences displayed are the set of γ-subunits from human heterotrimeric G-proteins.

Regan and coworkers have applied this approach to the design of several repeat proteins, including the tetratricopeptide repeat proteins (Main et al. 2003). The tetratricopeptide-repeat (TPR) family of proteins is a class of proteins consisting of repeated units of a small domain. In natural proteins, a TPR domain consists of 34 amino acids, and a typical TPR protein contains 3 to 16 repeats. Main et al. took the sequences of 1837 domains from 107 naturally occurring proteins and derived site-specific global amino-acid propensities based on this alignment. The most common amino acid at every position was significantly enriched over the average amino-acid frequency across all proteins, with enrichment factors ranging from 2.5 to 11.9.

Often, consensus sequences are defined by selecting positions that are conserved above a certain threshold. For example, for the TPR domain, Trp at position 4, Tyr at position 11, Gly at position 15, Tyr at position 17, Ala at position 20, Tyr at position 24, Ala at position 27, and Pro at position 32 all show enrichment of at least sixfold above the average amino-acid frequency. In terms of understanding the key sequence determinants of the protein fold, this is all the information that is needed. When designing a protein, however, one clearly needs a strategy for constructing a fully defined sequence.

Main et al. defined the sequence corresponding to the amino acid with the highest propensity at each site. Such a sequence would capture the most common features of this domain, but does not correspond to any natural protein. Repeats of one to three units of this engineered domain were constructed, with a few slight modifications: The single position where Cys was the most enriched residue was replaced with Ala (the second choice) in all repeats; a three residue helix-capping motif was added at the N-terminus of the protein; a solvating helix, corresponding to the first helix of the consensus motif with large aromatic residues, was replaced with Lys and Gln.

When they synthesized the model domain, even single domains of this protein were found to be well-structured, although they had relatively low stability. Repeats of this consensus sequence are very stable, with a repeat of three having a melting temperature of 83ºC and a repeat of two melting at 74ºC. In comparison, a naturally occurring three-repeat domain from PP5 has a melting temperature of 47ºC, comparable to that of the single repeat consensus protein (49ºC). These results suggest that the observed minimal repeat length of three (in natural systems) is not a result of stability requirements, but rather of an alternate reason, such as the requirement to bind protein targets.

Similar results have been obtained in numerous systems, including the ankyrin repeat proteins (Kohl et al. 2003), the leucine-rich repeat proteins (Binz et al. 2003), phytases (Lehmann et al. 2000), and antibodies (Knappik et al. 2000). The rationale for why this approach works is quite simple. Proteins have evolved both for stability and for function, and in a family of conserved fold, features defining protein stability will be conserved, while those that define diverse functions will not. In any given natural sequence, some of the residues that provide stability will likely be varied in order to accommodate function, but the particular residues that are varied will differ from protein to protein. Thus, the consensus sequence will define the underlying common feature that all members of the family share—an ability to stably fold into the target structure. As the consensus sequence does not contain those variations that create specific function at the expense of stability, it may form a much more stable structure.

A highly stable protein engineered through this approach may form a good starting point for engineering novel function, either through directed evolution or rational, computed-aided design. When the initial sequence is particularly stable, sequence variations that contribute strongly to a particular function at the expense of stability are more easily accommodated.

Pairwise Interactions from Sequence Analysis

Consensus-based engineering assumes an independence of each position in the primary sequence. That is, it is presumed that combining the most common residues at each position will produce a stable protein. However, the contributions of residues in folded proteins (both to stability and to function) are known to be strongly coupled in many cases. More detailed analysis of sequence conservation from multiple sequence alignments can provide insight into this coupling, which can subsequently be applied in an engineering context.

One such approach is statistical coupling analysis (Lockless and Ranganathan 1999). This method quantifies the difference in amino-acid frequency at one position when sequence subsets containing only a single type of amino acid are considered at a second position. For example, if two positions contain Lys or Glu with roughly equal probabilities, the sequences with Lys at the first position are unlikely to contain Lys at the second position, and vice versa. Such statistical coupling may be due to an easily interpretable structural feature, for example, the presence of a salt-bridge, or to more subtle functional interactions.

In certain cases, these coupled interactions may be essential characteristics that must be considered when engineering a protein through sequence analysis. Ranganathan and coworkers have considered this problem in the context of the WW-domain family of proteins (Socolich et al. 2005). A set of 120 sequences was aligned, and statistical enrichments for each amino acid at each position were computed, as were the coupling parameters between each position and a number of moderately conserved sites. Two sets of novel protein sequences were generated from this data. The first were sequences randomly generated based on the site-independent enrichments; these are not consensus sequences per se, but rather sequences with similar amino-acid distributions to the natural set. The second set of sequences was generated through a computational procedure designed to match both the site-independent distributions and the pairwise coupling values. Forty-three proteins of each type were expressed in *E. coli* and tested whether they fold to a well-defined, native-like structure. Control sets of 42 natural sequences and 19 random sequences with the same overall mean amino-acid frequency as the other sets were also considered. None of the site-independent derived set were natively folded, although 70% were expressed and soluble. In comparison, 28% of the coupled-conservation set were natively folded, with a similar total number of soluble sequences. Of the random set, less than 50% were expressed and soluble (and none were folded), while 84% of the native set were soluble and 67% folded. These results clearly suggest that statistical correlations between sites are essential in sequence-based design.

At first glance, the results from these two studies seem contradictory. In the first case, a site-independent consensus sequence resulted in a highly thermostable

protein, while in the second, designed proteins based on a site-independent analysis did not fold to the native state. One possible explanation is simply that each protein family may act differently. Site-independent information may completely determine the structure for some proteins, for example, zinc-fingers, phytases, antibodies, and the repeat proteins (TPR, ankyrin, and leucine-rich). For others, including the WW domains, the coupling between particular residues may be essential.

However, an alternate explanation is also possible. In the work of Ranganathan and coworkers, the site-independent sequences were not consensus sequences; that is, they did not use the most enriched amino acid at each site. Instead, they designed sequences with an overall amino-acid conservation profile similar to wild-type proteins. Thus, in the context of a given wild-type protein, the coupling between positions may be essential. That pairwise interactions can be important is, of course, well known—a deleterious mutation at one position can be rescued by a compensating mutation at a second site—and it is useful to note that a purely sequence-based analysis can aid in determining which interactions are functionally important. The consensus-based protein, however, is a single sequence with the most enriched amino acids at each position. Thus, some of the coupling information may be implicitly taken into account. Consider, for example, two positions that are strongly coupled—every sequence in the set contains KE or EK, but never KK or EE. If the KE motif is seen in just slightly more sequences, then Lys will be the preferred choice at site one while Glu is the preferred choice at site two. Thus, the consensus sequence will contain the KE motif, which is an acceptable choice. If, however, the sequences are derived randomly but biased by the independent conservation profiles, they will contain KK and EE in a significant number of cases.

Statistical coupling analysis can also be used together with consensus-based design, as was done by Magliery and Regan for the TPR-repeat designs, discussed previously (Magliery and Regan 2004). This analysis identified three strongly coupled networks of residues. While the consensus contains one choice of residues at each of these, other alternatives are possible. Additionally, the consensus-derived TRP sequence is highly negatively charged (–6), compared to natural sequences that predominantly have net charges between –3 and +3. Consideration of the statistical coupling between charged residues revealed subtle effects by which natural sequences contain compensating pairings of positive and negative residues. As these are at relatively weakly conserved sites and the pairings are not unique, this information is lost in the consensus sequence.

GRAPHICAL ANALYSIS OF PROTEIN STRUCTURE

Many protein-engineering applications involve the creation of a small number of mutations to a naturally occurring protein so as to enhance its function in a well-defined manner. In these cases, a structural biologist's intuition is often an important tool in the design of the desired variants, an approach that may be termed *structure-based protein design* to borrow a term from the drug design field. Visualization of the known reference structure is a key component of this. For example, visualization can identify unsatisfied hydrogen bond donors or acceptors that may be mutated

to increase stability or affinity. Similarly, visualizing steric interactions can help engineer interactions to discriminate among several potential binding targets.

Homology Modeling and Structure Visualization

In many cases, protein engineering targets a protein whose structure has not been solved. This does not preclude the use of structure-based methods, as the known structures of related proteins can be used to create model structures though the process of homology modeling. Briefly, homology modeling consists of a number of conceptual steps, which may or may not be performed independently in a given program. The first consists of mapping the backbone of homologous residues from a protein of unknown structure onto a known structure. Next, the side chains of these residues must be packed into the structure defined by the threaded backbone. For highly homologous proteins this may use the reference structure as an aid, while for less-similar proteins an alternate search procedure must be used. In many cases, however, there are regions of nonhomologous sequence, even in highly homologous proteins. Therefore, a new backbone, with appropriately placed side chains, must be constructed for these regions. As these sequences most often occur in loop regions, this procedure is referred to as *loop building*. Finally, the initial model structure is refined using a minimization protocol based on an empirical force field. Numerous homology modeling programs are available for the purpose, including several Web-based servers (Table 10.2). Each method produces slightly different models, but generally the closer the sequence, the more accurate the homology model will be. When a model of relatively low sequence similarity is desired, the use of multiple reference structures can improve the accuracy. Figure 10.2 displays a homology model of a designed dimeric protein complex, based on the known structure of a monomeric protein from which the complex was engineered.

Numerous tools are available for visualizing protein structure, many of which are freely available (Table 10.3). While every atom in a protein may play a role in structure and function, visualization typically focuses on the use of reduced models. For example, a structure is often rendered as a cartoon with the elements of secondary structure represented abstractly and connected by loops that follow

TABLE 10.2
Automated Web Services for Homology Modeling

Service	Web Location (URL)
Swiss-Model	http://swissmodel.expasy.org/SWISS-MODEL.html
Geno3D	http://geno3d-pbil.ibcp.fr/
ESyPred3D	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
What-If	http://swift.cmbi.kun.nl/WIWWWI/
CPHModels	http://www.cbs.dtu.dk/services/CPHmodels/
MODELLER*	http://salilab.org/modeller/modeller.html

** Downloadable software, not a Web service*

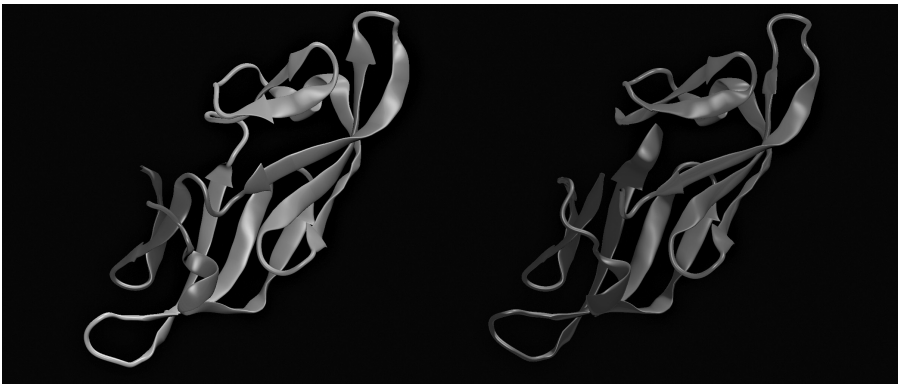


FIGURE 10.2 (see color insert following page 178) Homology building can allow a model structure to be built from the structure of a related sequence. Shown here is a homology model of a *de novo* designed dimeric complex (right), based on the structure of a monomeric protein (cyanovirin-N, left). The sequence of the dimer is 62% identical to that of the reference structure. Notice that all the key structural features are duplicated in the model, despite significant differences in sequence. Model was generated with MODELLER (Sali and Blundell 1993; Fiser and Sali 2003); figures were generated with VMD (Humphrey et al. 1996).

TABLE 10.3
Freely Available Software for Structural Analysis

Program	Web Location (URL)	Description
VMD	http://www.ks.uiuc.edu/Research/vmd/	Structural visualization program
PyMol	http://pymol.sourceforge.net/	Structural visualization program
GRASP	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:GRASP	Visualization of protein surfaces and electrostatic potential maps
Residual Potential	http://web.mit.edu/tidor/www/residual/index.html	GRASP scripts for computing the residual potential
Probe	http://kinemage.biochem.duke.edu/software/probe.php	Software for computation of steric complementarity

the backbone only. These cartoon models are useful for understanding the overall architecture of a protein or protein complex, and thus in identifying particular regions to include as variables in a combinatorial engineering application (either through directed-evolution or computational search). However, for the engineering of a small number of rational modifications, consideration of the chemical properties of each group can be essential. These properties are explicitly considered in the detailed force field-based calculations that will be discussed in later chapters, but can also be examined more qualitatively through visualization. The most straightforward approach involves mapping a property of interest, such as electrostatic potential or hydrophobicity, onto a surface representation of the protein.

Modulation of Stability and Affinity through Steric Complementarity

While chemical complementarity may be assessed simply by visualization of surface maps, steric complementarity is more difficult for the eye to capture. Both the cores of proteins and the interface of protein–ligand and protein–protein complexes are generally well-packed with atoms forming intricate complementarity. While not the only consideration, the overall size and shape of particular amino acids play an important role. As a result, protein stability and binding affinity can be perturbed by relatively small changes targeted at modulating this steric complementarity. The addition of even a single methyl unit to an underpacked region can stabilize a protein or complex, as can removal of steric bulk from an overpacked region. The Probe program (Word et al. 1999), accessible through the MolProbity Web service (Lovell et al. 2003), is designed specifically to address this issue. Comparison of the molecular surface generated with a small probe sphere with that generated with a water-sized sphere reveals regions of suboptimal packing. Steric clashes can be detected by the overlap of two or more atoms, and regions of near-optimal packing (close van der Waals contact between two atoms) may also be defined. This allows for a direct visualization of regions of suboptimal packing in the given structure.

The rational modulation of steric complementarity has been used by Jasanoff and coworkers to create variants of calmodulin (CaM) and the M13 CaM-binding peptide (from rabbit skeletal muscle myosin) with altered specificity relative to wild type (Green et al. 2006). As seen in Figure 10.3, the interface contains a key hydrophobic interaction involving a Trp on M13, which packs into a pocket formed by CaM Phe 92 and several aliphatic hydrophobic groups. Replacement of the M13 Trp by a smaller aromatic group (either Phe or Tyr) resulted in an underpacked interface, with subsequent loss of affinity; the W→F mutant binds eightfold worse than wild type, and the W→Y mutant binds 14-fold worse. This loss of affinity was then

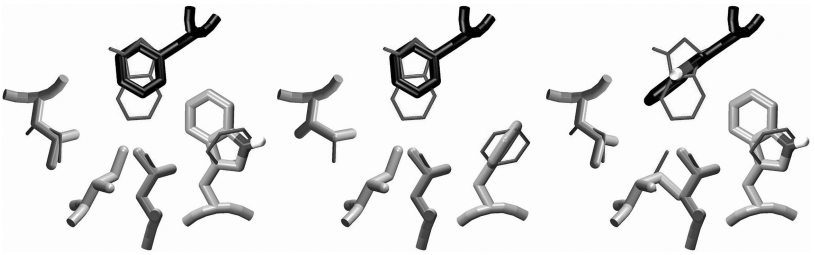


FIGURE 10.3 (see color insert following page 178) Affinities of protein complexes can be modulated by introducing perturbations in steric complementarity guided by visual analysis. Here we show a portion of the binding interface between calmodulin (gray) and the M13 CaM-binding peptide (black). **Left:** Reversing the positions of the Trp and Phe from the wild-type structure (shown in magenta) is expected to preserve near-optimal packing. **Center:** Changing only the Trp on M13 (to Phe) is expected to leave an unsatisfied void in the interface. **Right:** Changing only the Phe on CaM (to Trp) is expected to produced steric clashes and unfavorable structural rearrangements. Figures generated with VMD (Humphrey et al. 1996).

complemented by a corresponding mutation of Phe 92 on CaM to the larger Trp, as well as a conservative variation of Ile 125 to Leu. The mutant-mutant complex containing a W→F variation on M13 and a F→W (and I→L) substitution on CaM thus has the same total atom count as the wild type. The loss of affinity between the variant M13 and wild-type CaM is largely recovered in the complex with variant CaM. The affinity of the W→F mutant M13 for the variant CaM is only threefold different from WT and that of M13 W→Y differs by fourfold.

This work also reveals some of the challenges of taking advantage of steric complementarity in an engineering context. The site was chosen through visual analysis, with the motivation of engineering an orthogonal binding pair. That is, the goal was to create variants of both CaM and M13 that would bind to one another with affinity similar to that of the wild-type complex, but that would both have reduced affinity for the corresponding wild-type binding partner. As discussed previously, one direction of this specificity was achieved: The M13 W→F/Y variants bind preferentially to CaM FI→WL. However, specificity in the reverse direction was not achieved; in fact, the CaM FI→WL variant bound with twofold higher affinity to wild-type M13 than to the M13 mutants. This behavior is contrary to visual analysis—the pocket observed in the NMR solution structure seems well packed, and thus the replacement of Phe by the larger Trp would be expected to create steric clashes and reduce affinity. However, structural rearrangements in the pocket are able to accommodate this change, and the increase in buried hydrophobic surface in the F→W variant thus leads to increased affinity. While detailed packing calculations of the type that will be discussed in later chapters may be able to capture this behavior to some degree, consideration of the wild-type complex alone is inadequate.

Modulation of Affinity through Electrostatic Complementarity

Electrostatic interactions play essential roles in protein stability, binding affinity, and catalytic activity. As a result, modification of the electrostatic properties of a protein provides a particularly useful tool for structure-based protein design.

The energetic contributions of electrostatic interactions are complicated by solvent effects; not only can solvent reduce the strength of an interaction in a single state through screening effects, but differential interactions with solvent between two states (folded and unfolded, bound and unbound) directly contribute to their relative energies. In molecular association, favorable interactions between water and polar groups at the binding interface are lost upon binding. These are replaced with direct interactions between the two binding partners, but the degree of compensation varies from complex to complex. In many cases, the net electrostatic contribution may be unfavorable. Theoretical work by Tidor and colleagues has resulted in a framework to describe the degree of electrostatic complementarity at an interface using the Poisson-Boltzmann continuum model of solvent (Lee and Tidor 1997; Kangas and Tidor 1998). While the theory provides methods for the detailed consideration of atom-by-atom contributions through intensive calculations, a simple graphical assessment of complementarity can also be defined. Briefly, the electrostatic potential due to desolvation costs is mapped onto the surface of a target protein (desolvation potential). Additionally, the electrostatic interaction potential of the protein's

binding partner is similarly mapped onto the surface (interaction potential). It has been shown that in an electrostatically optimal complex the interaction potential must equal the negative of the desolvation potential across the entire protein surface. The residual potential is thus defined as the sum of the interaction potential and the desolvation potential. Deviations of the residual potential from zero highlight regions of the surface that deviate from optimal complementarity (see Figure 10.4)—a positive residual potential indicates that a reduction in positive charge (or increase in negative charge) would likely enhance affinity, and vice versa. Scripts for computing the residual potential with the GRASP software (Nicholls et al. 1993) are available online (Table 10.3).

An approach based on electrostatic analysis can also be applied to solve problems other than affinity optimization. For example, Lauffenburger and colleagues used this approach to design a modified variant of granulocyte-colony stimulating factor (GCSF) with enhanced lifetime (Sarkar et al. 2002). Cellular trafficking models had predicted that the rate of degradation depends on whether GCSF remains bound to its target receptor following internalization. In the late endosome, unbound ligands are recycled to the cell surface, while bound ligand-receptor complexes are retained for lysosomal degradation. Thus, recycling could be enhanced if a GCSF variant dissociates from the receptor in the low-pH environment of the late endosome. As the GCSFR binding surface has a positive charge, placement of histidines at the interface could produce the desired effect—at low pH, the protonation of histidine side chains on GCSF would create an electrostatic repulsion with the receptor, leading to reduced affinity. However, it is equally important to maintain affinity in the neutral

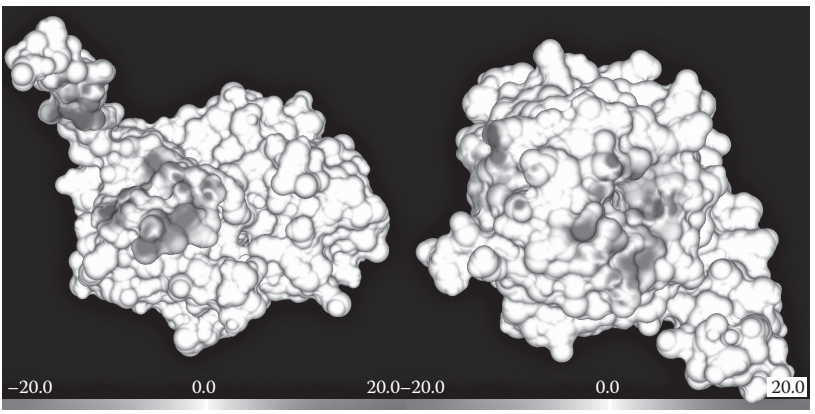


FIGURE 10.4 (see color insert following page 178) The residual potential graphically displays deviations from perfect electrostatic complementarity. Displayed is the interface between the α (left) and βγ (right) subunits of a heterotrimeric G-protein. Regions of white indicate complementary surfaces (including surfaces not involved in the interface), while colored regions indicate the deviation from optimal complementarity. Blue indicates that the chemical groups underlying that region are either too positive or not negative enough, while red indicates groups that are either too negative or not positive enough. In the left-hand figure, it is clear that there are both regions that are excessively positive or excessively negative on this surface.

environment of the cell-surface. Thus, the residual potential of GCSF for binding its target receptor was plotted on the GCSF surface and regions of excess negative potential were identified. These correspond to positions where GCSF is overly negative for optimal binding to the receptor, and thus may tolerate substitution with a neutral histidine without destabilizing the complex. Three acidic and three neutral residues were identified in this manner. Additional analysis suggested two aspartates as ideal candidates for mutation; the third acidic residue (a glutamate) was deemed essential to binding at neutral pH, and the substitution at the neutral residues did not provide adequate discrimination at low pH. Experimental characterization of the two D→H mutants demonstrated that (1) the mutants had affinities at pH 7.4 within threefold of wild type; (2) the mutants had over fourfold difference in affinity at pH 7.4 and at pH 5.5, compared with a difference of less than twofold for wild type; and (3) the mutants had significantly increased lifetime in cellular proliferation assay, with a 50% increase in recycling at each cycle, leading to between a 50% and 100% increase in effectiveness after 6–8 days.

FAST METHODS FOR MUTATIONAL EVALUATION

Visual analysis of protein structure, either with or without an energetic guide such as the residual potential, can suggest sites of modification; chemical intuition can then motivate particular amino-acid substitution. However, additional analysis is often needed due to the intricate networks of interactions typically found in proteins. In the cores of proteins, for example, a substituted residue must fit within the three-dimensional packing of residues; the substitution may, however, lead to rearrangements in this packing while maintaining stability. At binding interfaces, the same concerns with geometric packing exist, but as there are more polar groups, one must also consider balancing the desolvation costs with intermolecular interactions. The following chapters include a discussion of calculations that aim to address some of these problems. However, in some cases, it may be possible to evaluate the likely effect of a variation in a much simpler, and faster, manner.

Affinity Enhancement through Peripheral Electrostatic Interactions

Recently, it has been suggested that protein–protein binding affinities may be perturbed in a predictable manner without resorting to detailed calculations by targeting a particular class of modifications—electrostatic interactions made by surface residues at the periphery of a binding interface (Selzer and Schreiber 2001; Joughin et al. 2005; Shaul and Schreiber 2005). Charged residues on the surface of a protein can make significant interactions with a binding partner even when located at a moderate (5 to 10 Å) distance from the interface. These have been referred to as *action-at-a-distance interactions*. As surface residues in general do not form the same intricately packed networks as core and interface groups, mutations can be introduced at surface positions with less detailed modeling. Web-based interfaces for the identification of such mutations have been developed (see Table 10.4)

These interactions seem to work through two nonexclusive mechanisms. Schreiber and coworkers have suggested a kinetic mechanism, by which these

TABLE 10.4
Web Services for Structural and Energetic Analysis

Service	Web Location (URL)	Description
MolProbity	http://molprobity.biochem.duke.edu/	Web service for visualization and analysis of protein structure, including H-bonding and steric complementarity.
HyPARE	http://bip.weizmann.ac.il/hypareb/main	Web service for "Predicting Association Rate Enhancement" mutations.
AAAD	http://groups.csail.mit.edu/tidor/aaad/	Web service for prediction of thermodynamic "Action-at-a-Distance" interactions.

peripheral residues enhance the on-rate k_{on} of binding (Selzer and Schreiber 2001; Shaul and Schreiber 2005). Peripheral residues may make energetically significant interactions in the binding transition state to contribute to the rate of association, while such interactions are clearly absent in the unbound state. However, if the same interactions exist in the bound state as in the transition state, then they would have little effect on the rate of dissociation. Tidor and colleagues (including the author) have suggested an alternate mechanism, based purely on thermodynamic considerations (Joughin et al. 2005). Peripheral residues remain largely solvent exposed in the bound state, and thus pay only a very low (if any) desolvation penalty. However, the screening effects of solvent are such that small, but significant, nonspecific intermolecular interactions can persist at up to 10 Å separation. Fast methods of predicting long-distance interactions have been developed, using different methods designed to address different mechanisms of interaction. The kinetic and thermodynamic models lead to some overlap in predictions (residues that are expected to improve affinity also accelerate the kinetics of association), but many differences are seen as well. That is, some mutations are expected to increase affinity with minimal effect on the association rate, while others may increase the kinetics of association and disassociation without perturbing the overall affinity. These differences suggest that two distinct mechanisms are at play—peripheral action-at-a-distance interactions may independently be involved in modulating the affinity of an interaction as well as the kinetics of complex formation. Clearly both have applications in the engineering of protein complexes.

Mutations that add favorable interactions or remove an existing unfavorable interaction can be used to enhance binding affinity, as has been demonstrated. However, the same class of interactions could also be harnessed for additional modulations of affinity. For example, while introducing a mutation to severely reduce affinity is generally simple, destabilizing the complex by a desired degree is more challenging. Such an ability would be useful when engineering protein complexes for use as *in vivo* sensors or designing reagents with carefully tuned sensitivities. The introduction of an unfavorable action-at-a-distance interaction (or the elimination of an existing favorable interaction) could achieve this goal. The effect of individual residues in this type of interaction is relatively small, and thus engineering these interactions would be more effective in subtly modulating

the stability of a complex. Similarly, the introduction of titratable groups (such as histidine) could be used to modify the pH dependence of binding in a relatively straightforward manner.

SUMMARY

Information that is directly applicable to protein engineering can be found in the sequences and structures of known proteins. This knowledge can then be applied to new design objectives through the use of relatively simple computational approaches. Analysis of sequence conservation across a family of related proteins can be used to create hyperstable proteins through consensus-based design. Visualization of protein structure is an efficient way to identify key regions of interest, either by proximity to a site of known importance or by specifically targeting regions of suboptimal complementarity in packing or electrostatic interactions. Finally, fast, approximate methods are available for estimating the effects of mutations and have been shown to be highly successful in some cases, such as residues at the periphery of a protein–protein binding interface.

These approaches together provide a toolbox that complements the rational insight of a protein engineer in the process of design. Yet, they are not a solution to the design process in and of themselves. Rather, they act as a guide, suggesting a small number of mutations to consider or highlighting essential, conserved residues that should not be changed. Because they are simple, fast, and intuitive, they have been successfully adopted by experimental protein engineers in building a design or selection strategy.

REFERENCES

Binz, H. K., M. T. Stumpp, P. Forrer, P. Amstutz and A. Pluckthun. 2003. "Designing repeat proteins: Well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins." *J Mol Biol* 332: 489–503.

Fiser, A. and A. Sali. 2003. "Modeller: Generation and refinement of homology-based protein structure models." *Methods Enzymol* 374: 461–91.

Gouet, P., E. Courcelle, D. I. Stuart and F. Metoz. 1999. "ESPript: Analysis of multiple sequence alignments in PostScript." *Bioinformatics* 15: 305–308.

Green, D. F., A. T. Dennis, P. S. Fam, B. Tidor and A. Jasanoff. 2006. "Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide." *Biochemistry* 45: 12547–59.

Humphrey, W., A. Dalke and K. Schulten. 1996. "VMD: Visual molecular dynamics." *J Mol Graph* 14: 33–38.

Joughin, B. A., D. F. Green and B. Tidor. 2005. "Action-at-a-distance interactions enhance protein binding affinity." *Protein Science* 14: 1363–9.

Kangas, E. and B. Tidor. 1998. "Optimizing electrostatic affinity in ligand-receptor binding: Theory, computation, and ligand properties." *J Chem Phys* 109: 7522–45.

Knappik, A., L. M. Ge, A. Honegger, P. Pack, M. Fischer, et al. 2000. "Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides." *J Mol Biol* 296: 57–86.

Kohl, A., H. K. Binz, P. Forrer, M. T. Stumpp, A. Pluckthun, et al. 2003. "Designed to be stable: Crystal structure of a consensus ankyrin repeat protein." *Proc Natl Acad Sci U S A* 100: 1700–05.

Krizek, B. A., B. T. Amann, V. J. Kilfoil, D. L. Merkle and J. M. Berg. 1991. "A consenus zinc finger peptide—design, high-affinity metal-binding, a pH-dependent structure, and a His to Cys sequence variant." *J Am Chem Soc* 113: 4518–23.

Lee, L. P. and B. Tidor. 1997. "Optimization of electrostatic binding free energy." *J Chemical Phys* 106: 8681–90.

Lehmann, M., D. Kostrewa, M. Wyss, R. Brugger, A. D'Arcy, 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.

Lockless, S. W. and R. Ranganathan. 1999. "Evolutionarily conserved pathways of energetic connectivity in protein families." *Science* 286: 295–299.

Lovell, S. C., I. W. Davis, W. B. Adrendall, P. I. W. de Bakker, J. M. Word, et al. 2003. "Structure validation by C alpha geometry: Phi, psi and C beta deviation." *Proteins-Structure Function and Genetics* 50: 437–50.

Magliery, T. J. and L. Regan. 2004. "Beyond consensus: Statistical free energies reveal hidden interactions in the design of a TPR motif." *J Mol Biol* 343: 731–45.

Main, E. R. G., Y. Xiong, M. J. Cocco, L. D'Andrea and L. Regan. 2003. "Design of stable alpha-helical arrays from an idealized TPR motif." *Structure* 11: 497–508.

Nicholls, A., R. Bharadwaj and B. Honig. 1993. "GRASP—Graphical representation and analysis of surface-properties." *Biophys J* 64: A166.

Sali, A. and T. L. Blundell. 1993. "Comparative protein modeling by satisfaction of spatial restraints." *J Mol Biol* 234: 779–815.

Sarkar, C. A., K. Lowenhaupt, T. Horan, T. C. Boone, B. Tidor, et al. 2002. "Rational cytokine design for increased lifetime and enhanced potency using pH-activated 'histidine switching.'" *Nat Biotechnol* 20: 908–13.

Selzer, T. and G. Schreiber. 2001. "New insights into the mechanism of protein–protein association." *Proteins-Structure Function and Genetics* 45: 190–8.

Shaul, Y. and G. Schreiber. 2005. "Exploring the charge space of protein–protein association: A proteomic study." *Proteins-Structure Function and Bioinformatics* 60: 341–52.

Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, et al. 2005. "Evolutionary information for specifying a protein fold." *Nature* 437: 512–8.

Steipe, B., B. Schiller, A. Pluckthun and S. Steinbacher. 1994. "Sequence statistics reliably predict stabilizing mutations in a protein domain." *J Mol Biol* 240: 188–92.

Word, J. M., S. C. Lovell, T. H. LaBean, H. C. Taylor, M. E. Zalis, et al. 1999. "Visualizing and quantifying molecular goodness-of-fit: Small-probe contact dots with explicit hydrogen atoms." *J Mol Biol* 285: 1711–33.