Methods in
Molecular Biology 1414

Springer Protocols

Barry L. Stoddard *Editor*

Computational Design of Ligand Binding Proteins

Humana Press

Computational Design of Ligand Binding Proteins

Edited by

Barry L. Stoddard

Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

Humana Press

Editor
Barry L. Stoddard
Division of Basic Sciences
Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-3567-3 ISBN 978-1-4939-3569-7 (eBook)
DOI 10.1007/978-1-4939-3569-7

Library of Congress Control Number: 2016937968

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC New York

METHODS IN MOLECULAR BIOLOGY

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

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

Preface

Introduction: Design and Creation of Ligand-Binding Proteins

The appropriate balance of ligand binding affinity and specificity is a fundamental feature of most if not all biological processes, including immune recognition, cellular metabolism, regulation of gene expression, and cell signaling. The ability to accurately predict and recapitulate the physical basis for ligand binding behavior is therefore a crucial part of understanding and manipulating such biological phenomena. It also represents a critical technical requirement in the reciprocal fields of drug design and protein engineering.

This book provides a collection of protocols and approaches, compiled and described by many of today's leaders in the field of protein engineering, that they apply to the problem of creating ligand-binding proteins that display desirable combinations of target affinity and specificity. The descriptions provided by each chapter's authors also provide a snapshot of their current "belief system" regarding the challenging problem of protein engineering and design, as it is applied to the creation of novel ligand binding functions.

The problem of how to effectively engineer novel binding properties onto protein scaffolds, and how to do so while exploiting the information that is provided by high-resolution protein structures, has been under investigation for almost 40 years if not longer. Such efforts date back at least to the design of small folded peptides and proteins capable of binding individual nucleosides and single-stranded DNA, followed by subsequent attempts to generate additional ligand binding functions using various protein scaffolds (*see* Refs. [1, 2] for early examples of such work). By the early 1990s, some of the first computational algorithms intended to design novel ligand binding sites into proteins of known structure had been described [3], and the field of structure-based protein engineering as it is known today was underway.

Although the field of protein engineering, including the specific problem of designing novel ligand binding capabilities onto engineered protein folds, now comprises an extensive and growing publication record, significant challenges regarding the accurate calculation or prediction of protein–ligand binding affinities (even when provided a high-resolution structure of the actual complex) still represent significant hurdles to the field's advancement. For example:

- Several recent studies have demonstrated that current methods for structure-based calculation of binding affinities display variable accuracies. At least three broad (and somewhat overlapping) classes of scoring functions for predicting binding affinities from high-resolution structures have been developed: *force-field* (formulated by calculating the individual energetic contributions of physical interactions between the protein and ligand) [4, 5], *knowledge-based* (produced by statistical mining of large databases of protein–ligand structures to deduce rules and models that govern binding affinity) [6–9], and *empirical* (in which binding energy is calculated to be a product of a collection of weighted energy terms fit to a training data set of known binding affinities, with the weighting coefficients calculated via linear regression analyses) [10–14]. Even with all these tools, the accuracy of many methods that are intended to calculate structure-based binding affinities (as well as the ability to identify and rank the most tightly bound ligands to a given protein) has been shown to often be somewhat poor

[15–17], leading to the conclusion by one group that "more precise chemical descriptions of the protein–ligand complex do not generally lead to a more accurate prediction of binding affinity" [17]. Therefore, the reliable prediction of affinity remains a significant challenge in biophysical chemistry [15].

- Even for the most thoroughly studied of ligand-binding proteins, the basis for tight, specific binding is not well understood. For example, avidin and streptavidin exhibit some of the highest known affinities to their cognate molecular ligand (Ka ~ 10^{15} M^{-1}). Over 20 years of studies on these proteins have produced a wide range of hypotheses regarding their high affinities, including exceptional shape complementarity across a stabilized network of hydrophobic side chains and precisely arranged hydrogen bond partners [18], the precisely tuned dynamic behavior of the protein [19], a large free energy benefit upon ligand binding due to the strengthening of noncovalent interactions within the protein scaffold [20], or the induction of polarized moieties within the bound complex that create a cooperative effect between neighboring hydrogen bonds [21]. Not surprisingly, attempts to engineer altered binding properties onto avidin or streptavidin have yielded constructs with unexpected and unpredictable properties [22].
- Attempts to computationally engineer novel ligand-binding proteins have either been unsuccessful [23, 24] or have produced computationally designed constructs that display low affinities. Optimization of those designed proteins has then required laborious rounds of random mutagenesis and affinity maturation [25, 26].

The sources of error in calculating and modeling protein–ligand binding interactions and affinities are myriad, and their relative importance is still not entirely clear. These include: (1) Inaccuracies in the treatment of solvent and desolvation effects during binding [27–29]. (2) Limited consideration of protein dynamics [30–32]. (3) Difficulties incorporating the contribution of entropic changes into calculations of binding energies, leading to examples where modifications of ligand binding sites that lead to favorable enthalpic gains are confounded by substantial losses in entropy, with no improvement in overall binding affinity (recently reviewed extensively in Ref. [33]). Even for the most straightforward aspect of a protein–ligand interface (i.e., the observation of direct interatomic interactions and corresponding estimation of their enthalpic contributions to binding), uncertainties exist regarding interatomic distance cutoffs [17] and best strategies for estimating charge and protonation states [34].

Therefore, the creation of novel ligand-binding proteins that display tight binding affinity to their desired target and that also can discriminate between closely related targets remains an important goal, but is plagued by rather poor understanding of how to accurately calculate binding affinities or predict binding specificity, even when armed high structural information of protein–ligand complexes. As a result, the creation of highly specific ligand-binding proteins with high affinity remains extremely challenging and generally requires a substantial investment of time and effort to identify designed protein scaffolds that are actually active, and then to manually optimize their behavior. Nevertheless, studies from groups around the world have recently demonstrated that engineered proteins can, with considerable effort, be created that perform as desired, even in highly demanding in vivo applications. In this book, a series of 21 author groups present individual chapters that describe, in considerable detail, the types of overall thought processes and approaches, as well as very detailed computational and/or experimental protocols, that are used in their research groups as they attempt to address and resolve the difficulties associated with the design and creation of engineered ligand-binding proteins.

The reader will find a wide variety of technical issues and variables described in this volume. The first three chapters are largely concerned with a fundamental challenge that precedes actual protein engineering: identifying, characterizing, and modeling protein–ligand binding sites and predicting their corresponding modes and affinities of molecular interaction. Various strategies are shown to rely on both sequence-based and structure-based methods of analysis, and often utilize evolutionary information to determine the relative importance of positions within individual protein scaffolds that are important for form and function. With the development of controlled, blind binding site prediction challenges within the protein informatics and design community, the number of methods available to perform such analyses has exploded, as summarized in Chapter 2. Virtually all structure-based methods for binding site evaluation rely on accurate modeling of protein–ligand conformational sampling and scoring of individual docked solutions, which is further discussed in Chapters 3 and 4.

Beyond the basic ability to identify and model protein–ligand binding sites and their interactions, the field of protein engineering also now has at its disposal a number of increasingly powerful and robust computational platforms for structure-based engineering, including the widely used and rapidly evolving ROSETTA program suite as well as other programs such as POCKETOPTIMIZER and PROTEUS. Many of the fundamental features of these computational program suites, as well as individual examples of their utility and application for the design of a protein binding site for a defined small molecular ligand, are found in Chapters 5 through 7.

The output of even the most powerful structure-based computational design algorithms is usually augmented by considerable experimental time and effort, generally consisting of the preparation of combinatorial protein libraries or the systematic generation of large numbers of individual protein mutants on top of designed protein constructs, which are then subjected to selections or screens for optimal activity. While the ultimate goal of protein design is to eliminate the need for such manual intervention and effort, at this time many strategies for protein design involve combining information from computational design to the subsequent creation and screening of protein mutational libraries. Several examples of such approaches, which have resulted in particularly notable recent successes in protein engineering and the creation of designed ligand-binding proteins', are outlined and described in Chapters 8–10 and can then be found at various points within the remaining chapters.

Finally, the exact technical hurdles and necessary approaches required for the creation of ligand-binding proteins obviously are dependent upon the chemical and structural nature of the ligand to be recognized and bound with high affinity and specificity. The remaining 12 chapters describe a variety of specific scenarios and methodological approaches, ranging from the design of metal-binding proteins and light-induced ligand-binding proteins, to the creation of binding proteins that also display catalytic activity, to binding of larger peptide, protein, DNA, and RNA ligands.

The continued development of approaches to design and create ligand-binding proteins, beyond enabling the creation of unique protein-based reagents and molecules for biotechnology and medicine, will continue to test and refine the ability of modern biophysical chemistry to fundamentally understand and exploit the forces and principles that drive molecular recognition. The behaviors and properties of designed ligand-binding proteins resulting from the types of methods described in this book (including the "failures"—those constructs that fail to bind their intended targets and those that bind to unintended ligands) will eventually be explained by systematically examining their structures and properties. As has been famously attributed to Richard Feynman, "That which I cannot create, I do not understand." The following volume provides detailed (although by no means complete and total) examples of the current approaches and methods by which the protein engineering and design community attempt to do both.

Seattle, WA, USA *Barry L. Stoddard*

References

1. Gutte B, Daumigen M, Wittschieber E (1979) Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids. Nature 281:650–655
2. Moser R, Thomas RM, Gutte B (1983) Artificial crystalline DDT-binding polypeptide. FEBS 157:247–251
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. Huang N, Kalyanaraman C, Bernacki K et al. (2006) Molecular mechanics methods for predicting protein-ligand binding. Phys Chem Chem Phys 8: 5166–5177
5. Ewing T, Makino S, Skillman A et al. (2001) DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. J Comut Mol Des 15:411–428
6. Gehlhaar DK, Verkhivker GM, Rejto PA et al. (1995) Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. Chem Biol 2: 317–324
7. Muegge I, Martin Y (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem 42: 791–804
8. Mooij W, Verdonk M (2005) General and targeted statistical potentials for protein-ligand interactions. Proteins 61: 272–287
9. Hohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 295: 337–356
10. Bohm H (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput Mol Des 8: 243–256
11. Eldridge M, Murray C, Auton T et al. (1997) Empirical scoring functions: the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aid Mol Des 11: 425–445
12. Friesner R, Al E (2004) Glide: a new approach for rapid, accurate docking and scoring. J Med Chem 47: 1739–1749
13. Krammer A, Kirchhoff P, Jiang X et al. (2005) LigScore: a novel scoring function for predicting binding affinities. J Mol Graphics Model 23: 395–407
14. Wang R, Lai L, Wang S (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Mol Des 16: 11–26
15. Ross G, Morris G, Biggin P (2013) One size does not fit all: the limits of structure-based models in drug discovery. J Chem Theory Comput 9: 4266–4274
16. Ashtawy H, Mahapatra N (2012) A comparative assessment of ranking accuracies of conventional and machine-learning-based scoring functions for protein-ligand binding affinity prediction. IEEE/ACM Trans Comput Biol Bioinform 9: 1301–1312
17. Ballester P, Schreyer A, Blundell T (2014) Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? J Chem Inform Model 54: 944–955
18. Livnah O, Bayer EA, Wilchek M et al. (1993) Three-dimensional structures of avidin and the avidin-biotin complex. Proc Natl Acad Sci U S A 90: 5076–5080
19. Trong I, Wang Z, Hyre D et al. (2011) Streptavidin and its biotin complex at atomic resolution. Acta Crystallogr D Biol Crystallogr 67:813–821
20. Williams D, Stephens E, O'brien D et al. (2004) Understanding noncovalent interactions: ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. Angew Chem Int Ed Engl 43:6596–6616
21. Dechancie J, Houk K (2008) The origins of femtomolar protein–ligand binding: hydrogen bond cooperativity and desolvation energetics in the biotin–(strept)avidin binding site. JACS 129: 5419–5429
22. Aslan FM, Yu Y, Mohr SC et al. (2005) Engineered single-chain dimeric streptavidins with an unexpected strong preference for biotin-4-fluorescein. Proc Natl Acad Sci U S A 102: 8507–8512
23. Schreir B, Stumpp C, Wiesner S et al. (2009) Computational design of ligand binding is not a solved problem. Proc Natl Acad Sci U S A 106: 18491–18496
24. Looger L, Dwyer M, Smith J et al. (2003) Computational design of receptor and sensor

proteins with novel functions. Nature 423: 185–190

25. Procko E, Berguig G, Shen B et al. (2014) A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. Cell 157: 1644–1656
26. Tinberg CE, Khare SD, Dou J et al. (2013) Computational design of ligand-binding proteins with high affinity and selectivity. Nature 501: 212–216
27. Leach A, Shoichet B, Peishoff C (2006) Prediction of protein-ligand interactions. Docking and Scoring: successes and gaps. J Med Chem 49: 5851–5855
28. Schneider G (2010) Virtual screening: an endless staircase? Nat Rev Drug Discov 9: 273–276
29. Huang S, Grinter S, Zou X (2010) Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. Phys Chem Chem Phys 12: 12899–12908
30. Michel J, Esses J (2010) Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. J Comput Aid Mol Des 24: 639–658
31. Mobley D (2012) Let's get honest about sampling. J Comput Aid Mol Des 26: 93–95
32. Guvench O, Mackerell A (2009) Computational evaluation of protein-small molecule binding. Curr Opin Struct Biol 19: 56–61
33. Chodera J, Mobley D (2013) Entropy-enthalpy compensation: role and ramification in biomolecular ligand recognition and design. Ann Rev Biophys 42: 121–142
34. Rocklin GJ, Boyce SE, Fischer M et al. (2013) Blind prediction of charged ligand binding affinities in a model binding site. J Mol Biol 425: 4569–4583

Contributors

BRITTANY ALLISON • *Department of Chemistry, Vanderbilt University, Nashville, TN, USA; Center for Structural Biology, Vanderbilt University, Nashville, TN, USA*

GEORGIOS ARCHONTIS • *Theoretical and Computational Biophysics Group, Department of Physics, University of Cyprus, Nicosia, Cyprus*

MINKYUNG BAEK • *Department of Chemistry, Seoul National University, Seoul, Republic of Korea*

BRIAN M. BAKER • *Department of Chemistry and Biochemistry and the Harper Cancer Research Institute, University of Notre Dame, South Bend IN USA*

KYLE A. BARLOW • *Graduate Program in Bioinformatics, California Institute for Quantitative Biomedical Research, and Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA*

BRIAN J. BENDER • *Department of Chemistry, Vanderbilt University, Nashville, TN, USA; Department of Pharmacology, Vanderbilt University, Nashville, TN, USA*

STEVE J. BERTOLANI • *Department of Chemistry, University of California Davis, Davis, CA, USA*

JANUSZ M. BUJNICKI • *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland; Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland*

DYLAN ALEXANDER CARLIN • *Biophysics Graduate Group, University of California Davis, Davis, CA, USA*

MARINO CONVERTINO • *Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC, USA*

BRUNO E. CORREIA • *Institute of Bioengineering, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland*

WAYNE DAWSON • *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland*

NIKOLAY V. DOKHOLYAN • *Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC, USA*

KAREN DRUART • *Department of Biology, Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, Palaiseau, France*

SANJIB DUTTA • *Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA*

GEVORG GRIGORYAN • *Department of Biological Sciences, Dartmouth College, Hanover, NH, USA; Department of Computer Science, Dartmouth College, Hanover, NH, USA*

WILLIAM A. HANSEN • *Computational Biology and Molecular Biophysics Program, Rutgers State University of New Jersey, Piscataway, NJ, USA; Center for Integrative Proteomics Research, Rutgers State University of New Jersey, Piscataway, NJ, USA*

LIM HEO • *Department of Chemistry, Seoul National University, Seoul, Republic of Korea*

BIRTE HÖCKER • *Max Planck Institute for Developmental Biology, Tübingen, Germany; Lehrstuhl für Biochemie, Universität Bayreuth, Bayreuth, Germany*

DANIEL HOERSCH • *California Institute for Quantitative Biomedical Research and Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA; Fachbereich Physik, Freie Universität Berlin, Berlin, Germany*

TIM JACOBS • *University of North Carolina, Chapel Hill, NC, USA*

JOANNA M. KASPRZAK • *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland; Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland*

AMY E. KEATING • *Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA*

SAGAR D. KHARE • *Department of Chemistry and Chemical Biology, Rutgers State University of New Jersey, Piscataway, NJ, USA; Center for Integrative Proteomics Research, Rutgers State University of New Jersey, Piscataway, NJ, USA*

TANJA KORTEMME • *California Institute for Quantitative Biomedical Research and Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA*

BRIAN KUHLMAN • *Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC, USA*

HASUP LEE • *Department of Chemistry, Seoul National University, Seoul, Republic of Korea*

TOM LINSKEY • *University of Washington, Seattle, WA, USA*

MARK W. LUNT • *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO, USA*

BHARAT MADAN • *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland*

MARCIN MAGNUS • *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland*

LIAM JAMES MCGUFFIN • *School of Biological Sciences, University of Reading, Reading, UK*

JENS MEILER • *Department of Chemistry, Vanderbilt University, Nashville, TN, USA; Center for Structural Biology, Vanderbilt University, Nashville, TN, USA; Department of Pharmacology, Vanderbilt University, Nashville, TN, USA*

ELENI MICHAEL • *Theoretical and Computational Biophysics Group, Department of Physics, University of Cyprus, Nicosia, Cyprus*

DAVID MIGNON • *Department of Biology, Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, Palaiseau, France*

JEREMY H. MILLS • *Department of Biochemistry, University of Washington, Seattle, WA, USA*

ROCCO MORETTI • *Department of Chemistry, Vanderbilt University, Nashville, TN, USA; Center for Structural Biology, Vanderbilt University, Nashville, TN, USA*

MEHDI NELLEN • *Max Planck Institute for Developmental Biology, Tübingen, Germany*

VINCENT L. PECORARO • *Department of Chemistry, University of Michigan, Ann Arbor, MI, USA*

BRIAN G. PIERCE • *Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD, USA*

JEFFERSON S. PLEGARIA • *Department of Chemistry, University of Michigan, Ann Arbor, MI, USA*

SAVVAS POLYDORIDES • *Theoretical and Computational Biophysics Group, Department of Physics, University of Cyprus, Nicosia, Cyprus*

ERIK PROCKO • *Department of Biochemistry, University of Illinois, Urbana, IL, USA*

LOTHAR "LUTHER" REICH • *Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA*

TIMOTHY P. RILEY • *Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, USA; Harper Cancer Research Institute, University of Notre Dame, Notre Dame, IN, USA*

RYAN S. RITTERSON • *California Institute for Quantitative Biomedical Research and Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA*

DANIEL BARRY ROCHE • *Institut de Biologie Computationnelle, LIRMM, CNRS, Université de Montpellier, Montpellier, France; Centre de Recherche en Biologie cellulaire de Montpellier, CNRS-UMR 5237, Montpellier, France*

CHAOK SEOK • *Department of Chemistry, Seoul National University, Seoul, Republic of Korea*

JUSTIN B. SIEGEL • *Department of Chemistry, University of California Davis, One Shields Avenue, Davis, CA, USA; Genome Center, University of California Davis, One Shields Avenue, Davis, CA, USA; Department of Biochemistry and Molecular Medicine, University of California Davis, One Shields Avenue, Davis, CA, USA*

DANIEL-ADRIANO SILVA • *Department of Biochemistry, University of Washington, Seattle, WA, USA*

THOMAS SIMONSON • *Department of Biology, Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, Palaiseau, France*

NISHANT K. SINGH • *Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, USA; Harper Cancer Research Institute, University of Notre Dame, Notre Dame, IN, USA*

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

YIFAN SONG • *Department of Biochemistry, University of Washington, Seattle, WA, USA*

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

KRZYSZTOF SZCZEPANIAK • *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland*

SUMMER THYME • *Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA*

CHRISTINE E. TINBERG • *Department of Biochemistry, University of Washington, Seattle, WA, USA; Amgen, South San Francisco, CA, USA*

IRINA TUSZYNSKA • *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland; Institute of Informatics, University of Warsaw, Warsaw, Poland*

MENG WANG • *Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

ZHIPING WENG • *Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, USA*

HUIMIN ZHAO • *Departments of Chemical and Biomolecular Engineering, Biochemisry, and Chemistry and the Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL USA*

FAN ZHENG • *Department of Biological Sciences, Dartmouth College, Hanover, NH, USA*

Contents

Chapter 1

In silico Identification and Characterization of Protein-Ligand Binding Sites

Daniel Barry Roche and Liam James McGuffin

Abstract

Protein–ligand binding site prediction methods aim to predict, from amino acid sequence, protein–ligand interactions, putative ligands, and ligand binding site residues using either sequence information, structural information, or a combination of both. In silico characterization of protein–ligand interactions has become extremely important to help determine a protein's functionality, as in vivo-based functional elucidation is unable to keep pace with the current growth of sequence databases. Additionally, in vitro biochemical functional elucidation is time-consuming, costly, and may not be feasible for large-scale analysis, such as drug discovery. Thus, in silico prediction of protein–ligand interactions must be utilized to aid in functional elucidation. Here, we briefly discuss protein function prediction, prediction of protein–ligand interactions, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) and the Continuous Automated EvaluatiOn (CAMEO) competitions, along with their role in shaping the field. We also discuss, in detail, our cutting-edge web-server method, FunFOLD for the structurally informed prediction of protein–ligand interactions. Furthermore, we provide a step-by-step guide on using the FunFOLD web server and FunFOLD3 downloadable application, along with some real world examples, where the FunFOLD methods have been used to aid functional elucidation.

Key words Protein function prediction, Protein–ligand interactions, Binding site residue prediction, Biochemical functional elucidation, Critical Assessment of Techniques for Protein Structure Prediction (CASP), Continuous Automated EvaluatiOn (CAMEO), Protein structure prediction, Structure-based function prediction, Quality assessment of protein–ligand binding site predictions

1 Introduction

Proteins play an essential role in all cellular activity, which includes: enzymatic catalysis, maintaining cellular defenses, metabolism and catabolism, signaling within and between cells, and the maintenance of the cells' structural integrity. Hence, the identification and characterization of a protein binding site and associated ligands is a crucial step in the determination of a protein's functionality [1–3].

Barry L. Stoddard (ed.), *Computational Design of Ligand Binding Proteins*, Methods in Molecular Biology, vol. 1414,
DOI 10.1007/978-1-4939-3569-7_1, © Springer Science+Business Media New York 2016

1.1 Predicting Protein–Ligand Interactions

Protein–ligand interaction prediction methods can be categorized into two broad groups: sequence-based methods and structure-based methods [1, 3, 4]. Sequence-based methods utilize evolutionary conservation to determine residues, which may be structurally or functionally important. These methods include firestar [5, 6], WSsas [7], INTREPID [8], Multi-RELIEF [9], ConSurf [10], ConFunc [11], DISCERN [12], TargetS [13], and LigandRFs [14]. Structure-based methods can additionally be separated into geometric-based methods (FINDSITE [15], Surflex-PSIM [16], LISE [17], Patch-Surfer2.0 [18], CYscore [19], LigDig [20], and EvolutionaryTrace [21, 22]), energetic methods (SITEHOUND [23]), and miscellaneous methods that utilize information from homology modeling (FunFOLD [3], FunFOLD2 [2], COACH [24], COFACTOR [25], GalaxySite [26], and GASS [27]), surface accessibility (LigSite^{CSC} [28]), and physiochemical properties, utilized by methods including SCREEN [29].

1.2 The Role of CASP and CAMEO on the Development of Protein–Ligand Interaction Methods

In recent years, there has been an explosion in the development and availability of protein–ligand binding site prediction methods. This is a direct result of the inclusion of a ligand binding site prediction category in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition [30–32], along with the subsequent inclusion of ligand binding site prediction in the Continuous Automated EvaluatiOn (CAMEO) competition [33].

Ligand binding site residue prediction was first introduced in CASP8 [30], where the aim was to predict putative binding site residues, in the target protein, which may interact with a bound biologically relevant ligand. The top methods in CASP8 (LEE [4] and 3DLigandSite [34]) utilized homologous structures with bound biologically relevant ligands in their prediction strategies. In both CASP9 [31] and CASP10 [32], protein–ligand interaction methods converged on similar strategies; the structural superposition of models, onto templates bound to biologically relevant ligands [1].

After the CASP10 competition, the protein–ligand interaction analysis moved to the CAMEO [33] continuous evaluation competition. This was a direct result of a lack of targets for evaluation, over the 3-month prediction period of the CASP competition, although predictions were still accepted for the CASP11 competition. This also resulted in a change of prediction format, where methods not only have to predict potential ligand binding site residues, but also predict the probability that each residue binds to a specific ligand type: I, Ion; O, Organic ligand; N, nucleotide; and P, peptide. In addition, the most likely type that a protein may bind is also predicted [33]. The continuous weekly assessment of CAMEO allows for a much better picture, of how a method performs, on a large diverse data set, containing a wide diversity of ligand types [33].