

Computational protein design promises to revolutionize protein engineering

Oscar Alvizo¹, Benjamin D. Allen², and Stephen L. Mayo³

¹Biochemistry and Molecular Biophysics Option, ²Division of Chemistry and Chemical Engineering, and ³Howard Hughes Medical Institute, Division of Biology and Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA

Natural evolution has produced an astounding array of proteins that perform the physical and chemical functions required for life on Earth. Although proteins can be reengineered to provide altered or novel functions, the utility of this approach is limited by the difficulty of identifying protein sequences that display the desired properties. Recently, advances in the field of computational protein design (CPD) have shown that molecular simulation can help to predict sequences with new and improved functions. In the past few years, CPD has been used to design protein variants with optimized specificity of binding to DNA, small molecules, peptides, and other proteins. Initial successes in enzyme design highlight CPD's unique ability to design function de novo. The use of CPD for the engineering of potential therapeutic agents has demonstrated its strength in real-life applications.

Introduction

Proteins are linear heteropolymers, synthesized from 20 different amino acids. They participate in nearly all of the structural, catalytic, sensory, and regulatory functions that are associated with living systems; these myriad roles are made possible by their ability to self-assemble into well-defined structures specified by their sequences. Molecular evolution has shown that varying a protein sequence through mutation and recombination can generate new structures and functions. Drexler first noted that existing biological systems for protein fabrication could be harnessed to produce nanoscale molecular machines with designed functions (1). However, the potential of this approach can only be fully realized with reliable methods to predict sequences that perform the desired functions. Over the last decade, computational protein design (CPD) has clearly shown its potential as a solution to this problem.

CPD was first conceived as the inverse of the protein-folding problem, since its goal is to generate amino acid sequences that adopt a specific three-dimensional fold (2). CPD utilizes the main-chain coordinates of a known protein structure as a fixed scaffold. Various amino acid types are modeled at each designed position, and potential mutations are suggested based on their interactions with the scaffold and with each other (Figure 1A).

Although the main chain is held fixed during a CPD calculation, various conformations of each amino acid type at each position are sampled to find sequences expected to stabilize the fold and satisfy any additional functional requirements. The distribution of energetically accessible conformations available to each amino acid side-chain is approximated using a set of discrete, low-energy conformations called rotamers (3,4) (Figure 1, B–D). At the beginning of a typical CPD calculation, rotamers of the

user-specified amino acid types are assigned to residue positions from a predefined library. The problem is thus to find a choice of rotamer at each position such that the fold is stabilized and the desired function is achieved.

Energy Functions

Amino acid sequences and conformations are scored using a set of energy functions designed to reproduce the features of stable proteins. Although the specific energy functions used and their parameters vary between different CPD implementations, most implementations include a function that prevents atomic overlap and favors van der Waals interactions (5) and a function that benefits the formation of hydrogen bonds (6,7). Although interactions between a protein and its aqueous environment are crucial for stability, it would be prohibitively expensive to model water molecules explicitly in a CPD calculation. Therefore, solvation potentials are used to reward the burial of hydrophobic groups and penalize the burial of polar groups; energies are computed using surface area (8,9) or occluded volume models (10,11). Electrostatic interactions may be modeled using Coulomb's law with a constant dielectric (6,12), a statistical pair potential (13), or more sophisticated methods (14). These energy functions were designed to simulate different conformations of a single sequence and can give spurious results when used to choose between different sequences. Therefore, the scoring functions are typically supplemented with heuristic, statistical, and negative design terms to compensate for the limitations of the inverse folding model. These terms include heuristic estimates of side-chain entropy (15), penalties for non-polar exposure (5), statistical rotamer probabilities (13), and composition-based unfolded state energies (13,15). Proteins have been successfully designed with multiple

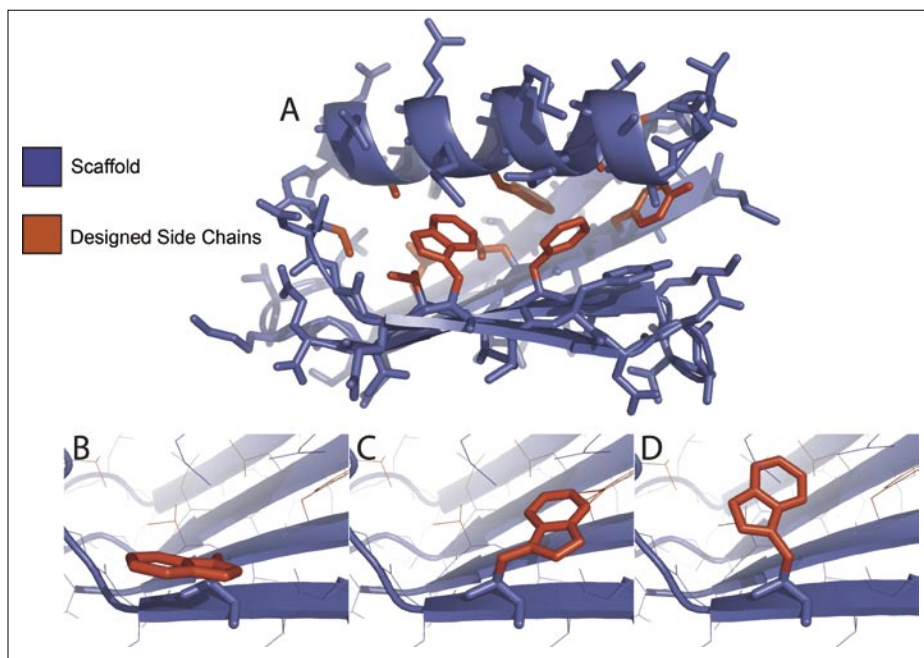


Figure 1. The inverse folding design model. (A) A protein partitioned into designed side-chains that are allowed to assume multiple rotamers (red) and a fixed scaffold (blue). (B–D) Different rotamers of the amino acid tryptophan at a particular position in the protein.

combinations of these functions, but no consensus has yet been reached on the ideal set of functions or the proper weight for each term.

Sequence Optimization

The energy functions in CPD are used to compute the interaction energies between each pair of rotamers at different designed positions (pair energies), as well as the interaction energy between each rotamer and the fixed scaffold (singles energies). Given a rotamer at each position, the total energy can be computed by summing the singles for each rotamer and pairs for each pair of rotamers. Typically, all singles and pair energies are precomputed, yielding a combinatorial optimization problem for which the minimum energy solution corresponds to the amino acid sequence and conformation expected to best stabilize the fold. Several optimization algorithms have been used to solve the design problem. Monte Carlo with simulated annealing (16–18), genetic algorithms (18,19), and fast and accurate side-chain topology and energy refinement (FASTER) (20,21) are stochastic optimization routines that sequentially improve an existing solution by making perturbations and accepting or rejecting them based on their energies. These algorithms are able to generate reasonable solutions quickly and will always give an answer, regardless of the difficulty of the design problem. Their running times can be extended indefinitely in an attempt to improve the quality of the prediction. However,

they can never guarantee that the optimal sequence has been found. Alternatively, strategies based on the dead-end elimination (DEE) theorem remove singles and pairs from the calculation that can be mathematically proven not to exist in the optimal solution (22,23). Although DEE can find optimal solutions to smaller problems, it does not scale well, and one must resort to the inexact algorithms described above when it fails to yield a definitive answer.

Design of Function

Although a designed protein sequence must adopt the desired fold, most design goals require that some useful function be achieved as well. To this end, the scoring functions described above are sometimes extended with additional function-specific constraints (24–27). For example, the geometric and chemical requirements for ligand or transition-state binding

in designed sensors or enzymes may be too difficult to model using standard potential functions. Thus, a theoretical model of the appropriate ligand contacts may be generated using quantum mechanics or chemical intuition, and the chemical and geometric features of this model can be enforced explicitly during the design. Typically, this type of model will specify the required contacts between the protein and the ligand, the types of protein groups allowed to make each contact, and a range of acceptable geometries for each interaction.

Applications of CPD

CPD was initially used as a tool to improve our understanding of the intrinsic properties of proteins and focused primarily on maintaining or enhancing protein stability. However, in recent years, the use of larger scaffolds and a focus on function have moved the CPD field into a new chapter. It is now more often used to produce proteins with novel or modified functions.

Ultimately, the CPD field would like to be able to rationally engineer any function into a given scaffold. A scaffold is usually obtained from the crystal structure of a naturally occurring protein; however, it is possible to computationally design one from the ground up (13). Recent work has focused on a handful of common biological functions, such as the interaction of proteins with small molecules, nucleic acids, and other proteins. In addition, major efforts toward the de novo design of catalytic activity are underway. CPD

of function is still in its infancy, and successful designs have taken their cues from biology.

One of the goals of CPD is to produce proteins that bind to other protein targets with high specificity. Initial investigations led to the design of protein mutants with increased specificity for both natural and novel peptide targets. For example, CPD of calmodulin led to a variant with eight mutations that maintained high affinity for the target peptide, while showing decreased binding affinity for nontargeted peptides (28,29). Specificity design was taken a step further when a PDZ domain mutant was engineered to bind a new sequence (30).

More recent work on protein-peptide and protein-protein specificity has revealed the value of negative design. Negative design is the process by which undesired properties are considered and designed against. A conceptually straightforward implementation of negative design is one in which the scoring function favors mutations that are stabilizing in the target structure while destabilizing in alternative structures. In the design of coil-coil dimers (31), four states were considered: (i) the homodimer state; (ii) the heterodimer state; (iii) the unfolded state; and (iv) the aggregated state. Successful designs selected amino acids that were predicted to favor the target dimer state over the rest. Further evidence of the usefulness of negative design was demonstrated in the redesign of the colicin E7 DNase-I_{m7} immunity protein interface (32).

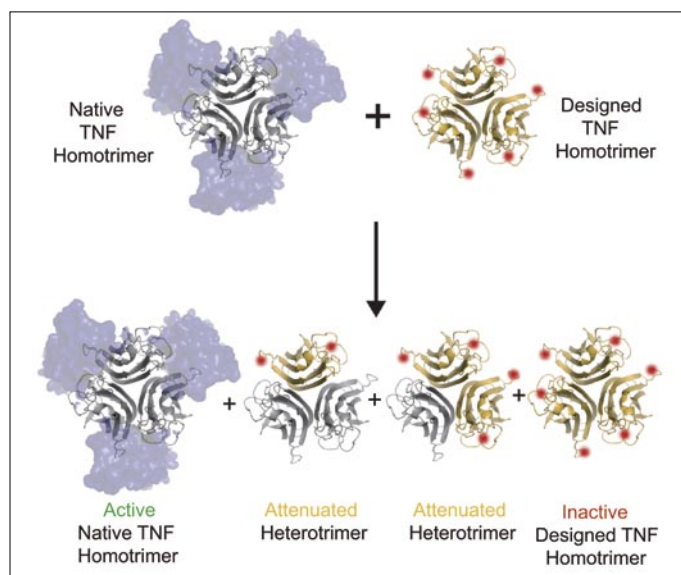


Figure 2. Computationally designed tumor necrosis factor (TNF) variant is depicted in gold. Mutations, highlighted in red, prevent designed variant from binding to TNF receptors, shown in blue. The addition of the mutated TNF results in a heterogeneous mixture of trimers, most of which have compromised activity. Adding an excess of mutated TNF insures low levels of active native TNF homotrimers. This figure is based on one provided in the supplementary material of Reference 39. Crystal structure 1TNR was used to generate the representative structures.

CPD has shown promise in the engineering of protein-DNA interactions as well. Currently, selective DNA cleavage is limited to those sequences for which naturally occurring specific endonucleases are available. CPD aims to change that by designing endonuclease variants with altered specificity (33). The first computational example focused on mutations that accommodated direct hydrogen bonds with the altered DNA base pairs. After identifying a suitable mutation, the surrounding residues were designed for optimal side-chain complementarity. This achievement may pave the way for the automated design of restriction endonucleases with therapeutic applications. However, CPD must first address water-mediated hydrogen bonds before it can be generally applicable to all DNA sequences. One possible solution that has been considered is to include solvated side-chains in the rotamer library (34). The expanded rotamer library contains explicit waters hydrogen-bonded to polar side-chain atoms. As a result, predictions can include rotamers that hydrogen bond to DNA through a bridging water molecule.

The design of protein-ligand interactions poses its own challenges. Currently, much of the design work on small ligand binding has been carried out using periplasmic binding proteins (PBPs) (35). Because wild-type PBPs can bind to a wide variety of small molecules, they have proven to be convenient scaffolds. In addition, ligand binding is easily detected due to the large conformational change associated with the binding event. Currently, PBPs have been computationally redesigned to bind metals, neurotransmitters, sugars, explosives, and nerve agents (26,36,37). The successes in protein-ligand design using PBPs suggest that scaffold selection is crucial. Focusing the design to a preexisting binding region on a scaffold shown to accommodate a wide variety of ligands appears to improve the chances of a successful design.

A similar trend seems to be applicable to the design of catalytic activity. Initial attempts to design an enzyme that catalyzes the hydrolysis of p-nitrophenyl acetate used thioredoxin as a scaffold and were modestly successful (25). The redesign of the binding region of a PBP successfully introduced triose phosphate isomerase activity (38). In both cases, the catalytic residues and transition states are geometrically constrained in an optimal conformation for catalysis. A design algorithm then identifies a position in the scaffold that accommodates both the catalytic residues and the transition state. The catalytic binding pocket is finally repacked for optimal complementarity with the substrate. The design of catalytic activity in various scaffolds is an excellent way to extend our understanding of enzyme catalysis. In addition, the ability to design novel enzymes that carry out difficult reactions would be of great use to the pharmaceutical industry.

The recent accomplishments of CPD clearly depict the field's ability to address biologically relevant applications. One of the best examples of this is the computational

design of a tumor necrosis factor (TNF) variant that is currently being developed as a therapeutic agent for the treatment of rheumatoid arthritis. The design focused on the binding interface between TNF and its receptor. Using a minimalist approach, mutations that disrupted the interface between TNF and its receptor were identified (Figure 2). Studies have shown that the designed variant inactivates native TNF by sequestering it, using a new dominant negative mechanism of action (39). The success of this potential therapeutic highlights the capabilities of CPD. With further improvements, CPD could revolutionize protein engineering.

As computational power improves, larger and more sophisticated designs will become possible. Furthermore, the predictive power of rational design is expected to grow as our knowledge of protein function advances. This will aid in improving the scoring functions and in identifying problems with the design model. It is clear from current results that further conformational sampling of the backbone and side-chains would be beneficial. A decrease in false positives coupled with high-throughput screening will ultimately lead to a rapid and reliable tool for the development of novel proteins with unique functions.

Acknowledgments

O.A. and B.D.A. contributed equally to this work.

References

- Drexler, K.E. 1981. Molecular engineering—an approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. USA* 78:5275-5278.
- Pabo, C. 1983. Molecular technology—designing proteins and peptides. *Nature* 301:200.
- Janin, J., S. Wodak, M. Levitt, and B. Maigret. 1978. Conformation of amino-acid side-chains in proteins. *J. Mol. Biol.* 125:357-386.
- Dunbrack, R.L. and F.E. Cohen. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* 6:1661-1681.
- Dahiyat, B.I. and S.L. Mayo. 1997. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* 94:10172-10177.
- Dahiyat, B.I., D.B. Gordon, and S.L. Mayo. 1997. Automated design of the surface positions of protein helices. *Protein Sci.* 6:1333-1337.
- Kortemme, T., A.V. Morozov, and D. Baker. 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* 326:1239-1259.
- Lee, B. and F.M. Richards. 1971. Interpretation of protein structures—estimation of static accessibility. *J. Mol. Biol.* 55:379-400.
- Street, A.G. and S.L. Mayo. 1998. Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.* 3:253-258.
- Colonna-Cesari, F. and C. Sander. 1990. Excluded volume approximation to protein-solvent interaction—the solvent contact model. *Biophys. J.* 57:1103-1107.
- Lazaridis, T. and M. Karplus. 1999. Effective energy function for proteins in solution. *Proteins* 35:133-152.
- Zollars, E.S., S.A. Marshall, and S.L. Mayo. 2006. Simple electrostatic model improves designed protein sequences. *Protein Sci.* 15:2014-2018.
- Kuhlman, B., G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, and D. Baker. 2003. Design of a novel globular protein fold with atomic-level accuracy. *Science* 302:1364-1368.
- Wis, M.S. and H.W. Hellinga. 2003. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. *Proteins* 51:360-377.
- Pokala, N. and T.M. Handel. 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.
- Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Physiol.* 21:1087-1092.
- Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220:671-680.
- Voigt, C.A., D.B. Gordon, and S.L. Mayo. 2000. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* 299:789-803.
- Desjarlais, J.R. and T.M. Handel. 1995. De-novo design of the hydrophobic cores of proteins. *Protein Sci.* 4:2006-2018.
- Desmet, J., J. Spriet, and I. Lasters. 2002. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* 48:31-43.
- Allen, B.D. and S.L. Mayo. 2006. Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* 27:1071-1075.
- Desmet, J., M. Demeyer, B. Hazes, and I. Lasters. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356:539-542.
- Gordon, D.B., G.K. Hom, S.L. Mayo, and N.A. Pierce. 2003. Exact rotamer optimization for protein design. *J. Comput. Chem.* 24:232-243.
- Hellinga, H.W. and F.M. Richards. 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.
- Bolon, D.N. and S.L. Mayo. 2001. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA* 98:14274-14279.
- Looger, L.L., M.A. Dwyer, J.J. Smith, and H.W. Hellinga. 2003. Computational design of receptor and sensor proteins with novel functions. *Nature* 423:185-190.
- Lassila, J.K., H.K. Privett, B.D. Allen, and S.L. Mayo. 2006. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* 103:16710-16715.
- Shifman, J.M. and S.L. Mayo. 2002. Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol.* 323:417-423.

29. **Shifman, J.M. and S.L. Mayo.** 2003. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc. Natl. Acad. Sci. USA* *100*:13274-13279.
30. **Reina, J., E. Lacroix, S.D. Hobson, G. Fernandez-Ballester, V. Rybin, M.S. Schwab, L. Serrano, and C. Gonzalez.** 2002. Computer-aided design of a PDZ domain to recognize new target sequences. *Nat. Struct. Biol.* *9*:621-627.
31. **Havranek, J.J. and P.B. Harbury.** 2003. Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* *10*:45-52.
32. **Joachimiak, L.A., T. Kortemme, B.L. Stoddard, and D. Baker.** 2006. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. *J. Mol. Biol.* *361*:195-208.
33. **Ashworth, J., J.J. Havranek, C.M. Duarte, D. Sussman, R.J. Monnat, Jr., B.L. Stoddard, and D. Baker.** 2006. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* *441*:656-659.
34. **Jiang, L., B. Kuhlman, T. Kortemme, and D. Baker.** 2005. A “solvated rotamer” approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* *58*:893-904.
35. **Dwyer, M.A. and H.W. Hellinga.** 2004. Periplasmic binding proteins: a versatile superfamily for protein engineering. *Curr. Opin. Struct. Biol.* *14*:495-504.
36. **Marvin, J.S. and H.W. Hellinga.** 2001. Conversion of a maltose receptor into a zinc biosensor by computational design. *Proc. Natl. Acad. Sci. USA* *98*:4955-4960.
37. **Allert, M., S.S. Rizk, L.L. Looger, and H.W. Hellinga.** 2004. Computational design of receptors for an organophosphate surrogate of the nerve agent soman. *Proc. Natl. Acad. Sci. USA* *101*:7907-7912.
38. **Dwyer, M.A., L.L. Looger, and H.W. Hellinga.** 2004. Computational design of a biologically active enzyme. *Science* *304*:1967-1971.
39. **Steed, P.M., M.G. Tansey, J. Zalevsky, E.A. Zhukovsky, J.R. Desjarlais, D.E. Szymkowski, C. Abbott, D. Carmichael, et al.** 2003. Inactivation of TNF signaling by rationally designed dominant-negative TNF variants. *Science* *301*:1895-1898.