Computational protein design

Arthur G Street[1] and Stephen L Mayo[2]*

A 'protein design cycle', involving cycling between theory and experiment, has led to recent advances in rational protein design. A reductionist approach, in which protein positions are classified by their local environments, has aided development of an appropriate energy expression. The computational principles and practicalities of the protein design cycle are discussed.

Addresses: [1]Division of Physics, Mathematics and Astronomy, California Institute of Technology, MC 147-75, Pasadena, CA 91125, USA and [2]Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, MC 147-75, Pasadena, CA 91125, USA.

*Corresponding author.
E-mail: steve@mayo.caltech.edu

Structure May 1999, 7:R105–R109
http://biomednet.com/elecref/09692126007R0105

 ISSN 0969-2126

Introduction

There are many reasons to pursue the goal of protein design. In medicine and industry, the ability to precisely engineer protein hormones and enzymes to perform existing functions under a wider range of conditions, or to perform entirely new functions, has tremendous potential. Furthermore, in the case of rational protein design, the knowledge obtained is likely to be linked to a more complete understanding of the forces underlying protein folding, enabling more rapid interpretation of the wealth of genomic information being amassed. Advances in protein design may also make possible the construction of a range of other self-organizing macromolecules.

Although some steps have been taken towards the rational design of functional enzymes [1], such a goal lies some distance away. Currently, attention is focused on redesigning portions of proteins to insert particular motifs, increase stability or modify function. Examples include the engineering of metal-binding centers, reviewed recently by Hellinga [2], and the introduction of disulfide bonds [3–5]. Theoretical work in the context of lattice models has also led to important insights. This work has been recently reviewed [6,7].

Attempts to design entire proteins *de novo* have been increasingly successful over the past decade. Early design efforts typically led to poorly characterizable states or molten globules, instead of a single target fold [8]. Other difficulties became apparent when a designed α-helical dimer [9] was shown to actually form a trimer [10]. This and subsequent studies relied on largely qualitative examinations of the target molecule [11], making generalization to other targets difficult.

This review focuses on the advances made in computational approaches to protein design. In particular, we examine those atomistic approaches that involve cycling between experiment and theory in a 'protein design cycle'.

Energy expression

Atomistic protein design requires an energy expression or force-field to rank the desirability of each amino acid sequence for a particular backbone structure. Over the past decade, elements of a suitable energy expression for atomistic protein design have been suggested and explored. To avoid over-fitting and to focus on only the most important contributors, the energy expression should contain as few terms as possible while maintaining predictive power. Communication between theory and experiment is required to determine which energy terms to include, and the relative importance of the included terms. In a protein design cycle, an energy expression is used to generate sequences that are subsequently made in the laboratory. Alterations and additions to the energy expression are then considered which improve the correlation between the computed and experimentally determined properties of the sequences. The improved energy expression is then used to generate new sequences, completing the cycle.

Energy minimization

In order to experimentally test the energy expression, the minimum-energy sequence of the target backbone must be determined. In the simplest implementation, the energy of every possible sequence is calculated using the energy expression, and the lowest energy sequence is reported. The size of most problems of interest renders this exhaustive approach impractical. Ignoring the possibility of multiple conformations of each amino acid, allowing the 20 naturally occurring amino acids at every position of a 100 amino acid protein yields 10^{130} possible sequence solutions. Clearly, ingenious energy minimization techniques are necessary.

Published search algorithms, including self-consistent mean-field approaches [12–14], Monte Carlo techniques [15,16], neural networks [17] and genetic algorithms [18,19], share the advantage of being able to sample large combinatorial space, but the disadvantage of not being guaranteed to find the global optimal solution. By contrast, dead-end elimination [20–23] and branch-and-terminate (DB Gordon and SLM, unpublished data) are search algorithms that give a final solution that is guaranteed to be the global optimum, but which require the discretization of

sidechain conformations into rotamers [24,25]. Such requirements will be discussed below. Search algorithms have been recently reviewed [26].

Discretization of sidechain conformations

To place a reasonable limit on the complexity of the computation, the allowed sidechain conformations are typically chosen from a library of discrete possibilities, known as rotamers. This discretization is necessary for some efficient search algorithms to be applicable — in particular, the dead-end elimination theorem.

Discretization of the sidechain conformations increases the likelihood of 'false negative' results. To be useful, atomistic protein design has only to output a subset of the sequences leading to the target fold, with simulation energies that correlate with their experimental stabilities. The simulation does not need to predict how well externally supplied sequences will fit the target fold. For example, the crystallographic structure of the Streptococcal protein G B1 domain (GB1) [27] shows Leu7 in an unusual conformation that does not appear in standard rotamer libraries [25]. Therefore, an atomistic algorithm using such a library may not suggest leucine at position 7 in the top ranked sequences.

The effect of the size of the rotamer library has also been considered [28,29]; in general, the larger the library the better. If the library contains too many similar conformations of each amino acid, however, the energy landscape is flattened and energy minimization can be slow.

Residue classification

A reductionist approach to protein design, in which subsets of a protein are designed independently, has proven fruitful. Computational attempts to design protein cores date back many years. More recently, there have been attempts to design surfaces and boundary positions as well.

The size of the design problem is reduced if only a subset of amino acid types need be considered in each of these three classes of residue positions. Protein cores are typically composed of hydrophobic amino acids, and protein surfaces are largely composed of hydrophilic amino acids, but the boundary residues must be selected from the full range of amino acids as these positions are observed to be both hydrophobic and hydrophilic. An automated way to classify residue positions is desirable, and a number of approaches have been described [30,31].

The important components of the energy expression relevant to the core, surface and boundary will be discussed in the following sections.

The core

Early attention on the protein design problem focused on the generally hydrophobic cores of proteins. It is believed that the folding process is driven principally by hydrophobic collapse of the polypeptide, implying that a well designed hydrophobic core is crucial to the structure and stability of the protein [32].

As might be expected, van der Waals forces (i.e., packing constraints) are crucial when designing the protein core. Models in which packing constraints are the only element of the energy expression are able to predict the stabilities of core mutations with high accuracy, when polar substitutions are not allowed [15,19,33–35]. The importance of packing constraints can be determined by scaling the atomic van der Waals radii by a factor α. When α is varied to very high (>105%) or very low (<85%) values, implying too little or too much volume being packed into the available space, respectively, the resulting proteins exhibit unfolded or molten globule-like behavior [34]. This is not surprising. Too much volume clearly requires the backbone to shift to accommodate the excess [36]. Too little volume would either leave cavities in the core, which have been shown to destabilize proteins [37], or again force the backbone to shift to fill the cavity. When the protein backbone is significantly different from the model backbone, the model can no longer accurately predict the stability of the protein, and there may cease to be a single stable folded state. The optimal value of α was found to be 90%, implying that a slight over-packing of hydrophobic residues in the core can actually stabilize a designed protein [34]. The benefit of using slightly diminished van der Waals radii can also be interpreted in terms of accommodating some backbone and rotamer flexibility (discussed later).

Consistent with the belief that the hydrophobic effect is a dominant cause of protein folding, the protein design cycle has been used to show that solvation effects also have an important role in the design of protein cores [33]. The hydrophobic effect is usually approximated as an energy benefit proportional to the amount of solvent-accessible hydrophobic surface area that is buried upon folding [38]. A penalty for burying polar area may also be included. Calculation of solvation energies is complicated by the need to construct the energy expression as a sum of two-body interactions [39,40].

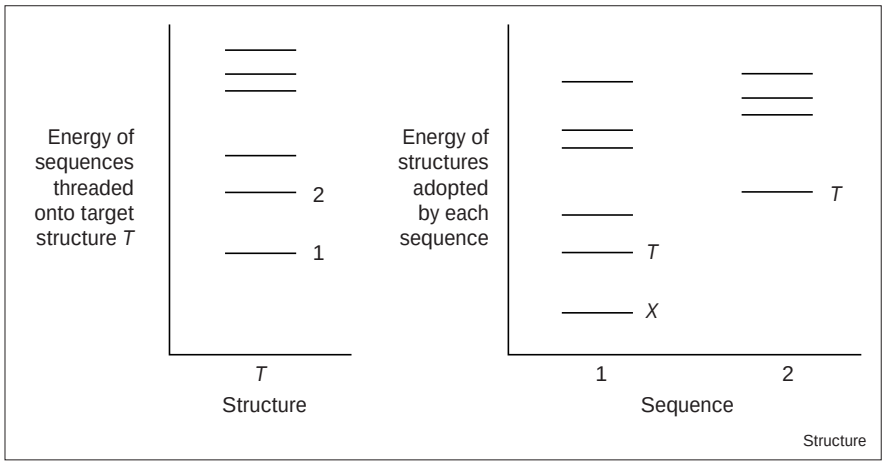
An entropic term has been tested [41], which may improve the correlation between predicted energy and biological activity [15]. Such a term should in particular penalize methionine, as the loss of rotational freedom upon burial of this residue in a protein core can lead to destabilized proteins [42].

The surface

With the successful redesign of a range of protein cores, it is natural to consider the redesign of protein surfaces. Despite the incontrovertible role of the hydrophobic core in folding, the surface is also crucial to a protein's structure and stability.

Figure 1

The role of negative design. Using a thermodynamic energy expression, a protein design algorithm computes that sequence 1 is the lowest energy sequence when threaded onto the target structure *T*. The ground-state structure of sequence 1, however, is an alternative structure *X*. In this case, the design algorithm would ideally return sequence 2, the lowest energy sequence with ground-state structure *T*.



The protein design cycle has been utilized to design surface sites, using as a starting point the energy expression determined from studies of protein cores. These studies showed the importance of electrostatics and hybridization-dependent hydrogen bonds [43]. In the case of α-helical surfaces, no further energy terms are necessary to achieve good predictive ability. This is possibly because the sidechains that are better hydrogen-bond formers are also good α-helix formers, as quantified by α-helical propensity [43,44].

The above energy terms are not sufficient to design β-sheet surfaces, however [45]. It may be necessary to directly bias the energy expression towards those sidechains with good β-sheet propensities. This is physically justifiable because common energy expressions do not otherwise include sidechain self-energies, which must at some level lead to propensities.

It is also possible that a main source of β-sheet stability is to be found elsewhere, for example, in the hydrogen bonds that cause alignment with neighboring β strands. In the case of antiparallel β strands, the turn joining the two strands has an important role. Modifying the component residues of the turn can seriously affect protein stability [46–48]. In the case of noncontinuous strands, it has been suggested that small clusters of hydrophobic area on the surface may help to set the register [49]. The hydrophobic effect may drive neighboring strands to align in such a way as to bury as much of the exposed hydrophobic area as possible, for example, by covering it with long amphiphilic sidechains.

The boundary

Some residues cannot be easily classified as core or surface constituents. Depending on the sidechain orientation they can interact with either the core of the protein or with the solvent. One such example is Trp43 of GB1 [34], which is predicted by modeling to rotate out into the solvent when nearby core residues are replaced with larger sidechains. Such unfavorable behavior can be attenuated by a hydrophobic exposure penalty [31,34].

Recent work has shown that the design of boundary residues can lead to impressively enhanced stability [50]. Just four boundary-site mutations in the 56-residue GB1 improve the stability from 3.3 kcal/mol to 7.1 kcal/mol at 50°C, converting a mesophilic protein into a hyperthermophilic protein.

Full de novo *sequence design*

To date there exists only a single example of a complete sequence calculation in which the structure of the designed protein was experimentally shown to achieve the design target [30]. This calculation included one core position, seven boundary positions and 18 surface positions, leading to a total of 10^{27} possible sequence solutions. The success of this design effort underscores the power of computational approaches.

Backbone

Most atomistic protein design efforts require a fixed backbone. The calculation is performed under the assumption that the target backbone is precisely the backbone that will be achieved by the computed sequence. Fortunately, alterations in the backbone do not necessarily lead to large changes in the accessible sequence space [51]. In one study, a 2 Å root mean square deviation (rmsd) in the backbone led to only a 0.5 Å rmsd in predicted sidechain conformations [29]. Backbone flexibility can be modeled by using a softer van der Waals potential — in other words, giving the modeled atoms a fuzzy edge. This effect can be obtained by using reduced atomic radii, which has been shown to improve the stability of designed proteins [34].

Protein backbone movements may be incorporated if the backbone is parameterizable [51,52], although to keep the calculation tractable, the number of sidechain rotamer combinations may be limited. A coiled-coil with right-handed superhelical twist, the backbone of which was necessarily designed *de novo*, has recently been reported [53], where 216 amino acid sequences were considered.

Negative design

The importance of negative design is the subject of much discussion. Recent work by Hellinga [54] highlights the importance of this issue in computational protein design. The inverse-folding design method determines the sequence of amino acids with the lowest energy when threaded onto the target backbone. It is conceivable that in some cases the computed sequence may actually prefer to fold to a different target structure, and that a sequence with a slightly higher computed energy would fold to the desired target (Figure 1). Unfortunately, knowledge of which structure will be adopted by the computed sequence requires a solution to the protein folding problem. Lattice models consisting of only two amino acid types can, however, be used to perform both sequence design and fold prediction. In this context, proposals to include non-thermodynamic potential functions aimed at addressing negative design issues have been developed [55–57]. The hydrophobic exposure penalty is one example of negative design that improves predictive power [31,34]. Despite the power of lattice model simulations, it has been suggested that the design procedure may be qualitatively different in such binary patterned systems [58].

Conclusions

The design of proteins that fold to a specified target backbone structure is becoming possible. Future advances are likely to follow from a tight coupling of experimental and computational work in a protein design cycle, with ever larger protein sequences designed *de novo* being revealed in the near future. Discovering the forces critical to the determination of backbone conformation and their coupling to sequence selection will rise as the major challenge in solving the 'complete' protein design problem. A general ability to design specific protein structures will pave the way towards the goal of rationally designing novel functional molecules.

References

1. Wilson, C., Mace, J.E. & Agard, D.A. (1991). Computational method for the design of enzymes with altered substrate specificity. *J. Mol. Biol.* **220**, 495-506.
2. Hellinga, H.W. (1998). The construction of metal centers in proteins by rational design. *Fold. Des.* **3**, R1-R8.
3. Yan, Y.B. & Erickson, B.W. (1994). Engineering of betabellin 14D: disulfide-induced folding of a β-sheet protein. *Protein Sci.* **3**, 1069-1073.
4. Matsumura, M. & Matthews, B.W. (1991). Stabilization of functional proteins by introduction of multiple disulfide bonds. *Methods Enzymol.* **202**, 336-356.
5. Pabo, C.O. & Suchanek, E.G. (1986). Computer-aided model-building strategies for protein design. *Biochemistry.* **25**, 5987-5991.
6. Dill, K.A., *et al*., & Chan, H.S. (1995). Principles of protein folding – a perspective from simple exact models. *Protein Sci.* **4**, 561-602.
7. Shakhnovich, E.I. (1998). Protein design: a perspective from simple tractable models. *Fold. Des.* **3**, R45-R58.
8. Betz, S.F., Raleigh, D.P. & DeGrado, W.F. (1993). *De novo* protein design: from molten globules to native-like states. *Curr. Opin. Struct. Biol.* **3**, 601-610.
9. O'Neil, K.T. & DeGrado, W.F. (1990). A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* **250**, 646-651.
10. Lovejoy, B., Choe, S., Cascio, D., McRorie, D.K., DeGrado, W.F. & Eisenberg, D. (1993). Crystal structure of a synthetic triple-stranded α-helical bundle. *Science* **259**, 1288-1293.
11. Bryson, J.W., *et al*., & DeGrado, W.F. (1995). Protein design: a heirarchic approach. *Science* **270**, 935-941.
12. Koehl, P. & Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* **6**, 222-226.
13. Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* **236**, 918-939.
14. Vasquez, M. (1995). An evaluation of discrete and continuum search techniques for conformational-analysis of sidechains in proteins. *Biopolymers* **36**, 53-70.
15. Hellinga, H.W. & Richards, F.M. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl Acad. Sci. USA* **91**, 5803-5807.
16. Lee, C. & Levitt, M. (1991). Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* **352**, 448-451.
17. Kono, H. & Doi, J. (1996). A new method for sidechain conformation prediction using a Hopfield network and reproduced rotamers. *J. Comp. Chem.* **17**, 1667-1683.
18. Pedersen, J.T. & Moult, J. (1996). Genetic algorithms for protein structure prediction. *Curr. Opin. Struct. Biol.* **6**, 227-231.
19. Desjarlais, J.R. & Handel, T.M. (1995). *De novo* design of the hydrophobic cores of proteins. *Protein Sci.* **4**, 2006-2018.
20. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein sidechain positioning. *Nature* **356**, 539-542.
21. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1994). The dead-end elimination theorem: a new approach to the sidechain packing problem. In *The Protein Folding Problem and Tertiary Structure Prediction.* (Merz, K., Jr. & Le Grand, S., eds), pp. 307-337, Birkhauser, Boston, MA.
22. Goldstein, R.F. (1994). Efficient rotamer elimination applied to protein sidechains and related spin-glasses. *Biophys. J.* **66**, 1335-1340.
23. Gordon, D.B. & Mayo, S.L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comp. Chem.* **19**, 1505-1514.
24. Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). Conformation of amino acid sidechains in proteins. *J. Mol. Biol.* **125**, 357-386.
25. Ponder, J.W. & Richards, F.M. (1987). Tertiary templates for proteins – use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791.
26. Desjarlais, J.R. & Clarke, N.D. (1998). Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.* **8**, 471-475.
27. Gronenborn, A.M., *et al*., & Clore, G.M. (1991). A novel, highly stable fold of the immunoglobin binding domain of streptococcal protein G. *Science* **253**, 657-661.
28. DeMaeyer, M., Desmet, J. & Lasters, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.* **2**, 53-66.
29. Tufféry, P., Etchebest, C. & Hazout, S. (1997). Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Eng.* **10**, 361-372.
30. Dahiyat, B.I. & Mayo, S.L. (1997). *De novo* protein design: fully automated sequence selection. *Science* **278**, 82-87.
31. Sun, S.J., Brem, R., Chan, H.S. & Dill, K.A. (1995). Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.* **8**, 1205-1213.
32. Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.
33. Dahiyat, B.I. & Mayo, S.L. (1996). Protein design automation. *Protein Sci.* **5**, 895-903.

34. Dahiyat, B.I. & Mayo, S.L. (1997). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA* **94**, 10172-10177.
35. Lazar, G.A., Desjarlais, J.R. & Handel, T.M. (1997). *De novo* design of the hydrophobic core of ubiquitin. *Protein Sci.* **6**, 1167-1178.
36. Baldwin, E.P., Hajiseyedjavadi, O., Baase, W.A. & Matthews, B.W. (1993). The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science* **262**, 1715-1718.
37. Lim, W.A. & Sauer, R.T. (1989). Alternative packing arrangements in the hydrophobic core of λ repressor. *Nature* **339**, 31-36.
38. Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199-203.
39. Street, A.G. & Mayo, S.L. (1998). Pairwise calculation of protein solvent accessible surface areas. *Fold. Des.* **3**, 253-258.
40. Kurochkina, N. & Lee, B. (1995). Hydrophobic potential by pairwise surface area sum. *Protein Eng.* **8**, 437-442.
41. Kono, H., Nishiyama, M., Tanokura, M. & Doi, J. (1998). Designing the hydrophobic core of *Thermus flavus* malate dehydrogenase based on sidechain packing. *Protein Eng.* **11**, 47-52.
42. Gassner, N.C., Baase, W.A. & Matthews, B.W. (1996). A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc. Natl Acad. Sci. USA* **93**, 12155-12158.
43. Dahiyat, B.I., Gordon, D.B. & Mayo, S.L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333-1337.
44. Chakrabartty, A., Kortemme, T. & Baldwin, R.L. (1994). Helix propensities of the amino-acids measured in alanine-based peptides without helix-stabilizing sidechain interactions. *Protein Sci.* **3**, 843-852.
45. Hecht, M.H. (1994). *De novo* design of β-sheet proteins. *Proc. Natl Acad. Sci. USA* **91**, 8729-8730.
46. Ybe, J.A. & Hecht, M.H. (1996). Sequence replacements in the central β-turn of plastocyanin. *Protein Sci.* **5**, 814-824.
47. Blanco, F., Ramirez-Alvarado, M. & Serrano, L. (1998). Formation and stability of β-hairpin structures in polypeptides. *Curr. Opin. Struct. Biol.* **8**, 107-111.
48. Garrett, J.B., Mullins, L.S. & Raushel, F.M. (1996). Are turns required for the folding of ribonuclease T1? *Protein Sci.* **5**, 204-211.
49. Tisi, L.C. & Evans, P.A. (1995). Conserved structural features on protein surfaces: small exterior hydrophobic clusters. *J. Mol. Biol.* **249**, 251-258.
50. Malakauskas, S.M. & Mayo, S.L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**, 470-475.
51. Su, A. & Mayo, S.L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* **6**, 1701-1707.
52. Harbury, P.B., Tidor, B. & Kim, P.S. (1995). Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc. Natl Acad. Sci. USA* **92**, 8408-8412.
53. Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T. & Kim, P.S. (1998). High-resolution protein design with backbone freedom. *Science* **282**, 1462-1467.
54. Hellinga, H.W. (1998). Construction of a blue copper analogue through iterative rational protein design cycles demonstrates principles of molecular recognition in metal center formation. *J. Am. Chem. Soc.* **120**, 10055-10066.
55. Deutsch, J.M. & Kurosky, T. (1996). New algorithm for protein design. *Phys. Rev. Lett.* **76**, 323-326.
56. Shakhnovich, E.I. & Gutin, A.M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA* **90**, 7195-7199.
57. Chiu, T.L. & Goldstein, R.A. (1998). Optimizing potentials for the inverse protein folding problem. *Protein Eng.* **11**, 749-752.
58. Micheletti, C., Seno, F., Maritan, A. & Banavar, J.R. (1998). Design of proteins with hydrophobic and polar amino acids. *Proteins* **32**, 80-87.