REVIEW

What has de novo protein design taught us about protein folding and biophysics?

David Baker*

Department of Biochemistry, University of Washington, Box 357350, 1705 NE Pacific Street, Seattle, Washington, 98195-7350, USA

Received 11 February 2019; Accepted 11 February 2019
DOI: 10.1002/pro.3588
Published online 12 February 2019 proteinscience.org

Abstract: Recent progress in de novo protein design has led to an explosion of new protein structures, functions and assemblies. In this essay, I consider how the successes and failures in this new area inform our understanding of the proteins in nature and, more generally, the predictive computational modeling of biological systems.

Keywords: protein design; protein folding; computational modeling

Introduction

I started my research group at the UW 25 years ago focused on the protein folding problem. For the first several years, our approach was primarily experimental: we sought to use random library selection methods to generate new proteins only very distantly related to naturally occurring proteins, and then by comparing the folding rates and mechanisms of these non-biological proteins to those in nature, to determine the extent to which evolution had operated on protein folding kinetics. The results, to my surprise (and initially, chagrin), were quite clear—while the novel proteins we generated almost always were less stable than naturally occurring proteins, the folding rates were as often faster as they were slower. While we did not obtain (as I had hoped) extremely slowly folding proteins whose folding process could be studied in detail, we were drawn inescapably to the conclusion that the sequences of naturally occurring proteins are not optimized for rapid folding.[1,2] This then led us to ask what factors do determine folding rates, which led in turn to the discovery of the relationship between the contact order (the average sequence separation of the residue–residue contacts in the native structure) and folding rates,[3] and more generally, the considerable extent to which the rate and mechanism of protein folding are dictated by the topology of the folded state.[4]

As our experimental work provided insight into how proteins fold, we sought to develop computational approaches for modeling folding that incorporated this knowledge. Guided by experimental observations from our lab and others, for example that local amino acid sequence biases but does not uniquely specify the distributions of local structures sampled during folding, and the principle that proteins fold to their lowest free energy states, we developed the Rosetta program for ab initio protein structure prediction.[5] As we achieved some success in predicting protein structure from sequence by searching for the lowest energy state for the amino acid sequence, Brian Kuhlman, then a postdoc in the lab, realized we could use Rosetta to go backwards from a new computer generated protein

David Baker is the winner of the 2018 Hans Neurath Award.

**Correspondence to:* David Baker, Department of Biochemistry, University of Washington, Box 357350, 1705 NE Pacific Street, Seattle, Washington 98195-7350, USA. E-mail: dabaker@u.washington.edu

structure to an amino acid sequence encoding it by searching for the lowest energy sequence for the structure-that is, design new proteins.[6] In the last 15 years, through advances in understanding protein structure, folding, binding, and assembly, coupled with steady improvement in the Rosetta energy function and sampling methodology, we and our colleagues have succeeded in designing a whole new world of proteins far exceeding anything I imagined when I started my group.[7]

The purpose of this article is not to review progress in de novo protein design. Instead, here I return to consider the goal I had when starting my research group of using the study of laboratory generated novel proteins to shed light on the proteins in nature—albeit with the difference that the new proteins are produced by computational design rather than random selection.

Kinetic versus thermodynamic control of folding

The success in de novo design has immediate bearing on the question I originally set out to study. The de novo design process only considers the stability of the target structure, not any partially folded structures on a "pathway" to the target structure. Indeed, the final test for a de novo designed protein, before proceeding to experimental characterization, is to determine whether the lowest energy conformations sampled for the designed sequence in large scale protein structure prediction calculations are close to the designed target structure. The fact that robust stable proteins can be designed by exclusively focusing on the stability of the target structure shows that thermodynamics generally trumps kinetics in protein folding—if a sufficiently low free energy state exists, it deforms the surrounding free energy landscape for folding to this state to proceed efficiently. The dominance of thermodynamics in determining state likely reflects the large entropic cost of chain ordering, and the fact that the interactions which overcome this entropy loss and stabilize protein structures (for example, hydrogen bonds and van der Waals interactions) are individually relatively weak (~1 kT). Folding can only occur if there are very large numbers of such interactions adding up coherently to stabilize a specific folded structure, which is unlikely to occur by chance (indeed almost all randomly generated protein sequences fail to fold to a unique structure). This results in a funnel shaped energy landscape: conformations similar to the folded structure will necessarily share a subset of the many stabilizing interactions and, hence, also be relatively low in energy, and large energy barriers will generally be absent because there are likely few competing low energy states with very different structures.

Success in de novo design of protein–protein interactions and assemblies suggests that the primacy of thermodynamics in determining structure is a quite general principle. The assembly of viral capsids and other large naturally occurring protein assemblies has been extensively studied, but it has been unclear to what extent evolutionary optimization has shaped the assembly process (beyond the obvious constraint, in the cases of viruses, of co-assembling with nucleic acid genome). Designed tetrahedral, octahedral, and icosahedral protein assemblies spontaneously and rapidly adopt the designed target nanostructure upon synthesis. The largest of these assemblies—120 subunit designed icosahedral nanocages built from two different building blocks—form in minutes following mixing of the two subunits with little or no formation of alternative species or evidence for kinetic traps.[8] These results suggest any sufficiently low free energy state of a set of protein chains will likely be kinetically accessible. These designs also highlight the control afforded by computational protein design: despite being composed of hundreds of thousands of atoms: crystal structures have RMSDs to the design models between 0.8 and 2.7 Å.

Anfinsen's thermodynamic hypothesis of protein folding—that proteins fold to their lowest free energy states—is very difficult to prove (it is much easier to prove that a given state is not the lowest free energy state, because only one counterexample need be found). While not constituting a proof, success in protein design strongly supports the thermodynamic hypothesis, as it is the core principle that de novo protein design is based on. Similarly, success in de novo protein design bears on the question I get after every talk about the importance of the order of chain synthesis on the ribosome to protein folding; computational protein design calculations completely ignore the order of synthesis which hence cannot be critical to protein folding.

Before leaving this topic, I note caveats to the argument that folding and association are thermodynamically driven. First, a significant fraction of de novo designed proteins either fail to express or are insoluble, and it is possible that kinetic accessibility is an issue for these failures (although toxicity in *E coli* during expression and aggregation through self association are more likely contributors). Second, if there is sufficient selective pressure, large kinetic barriers can be generated by natural selection, and in such cases protein folding can clearly be under kinetic control (the folding of alpha lytic protease for example[9]). A more precise statement of the conclusion of this section hence is that in the absence of specific selection (or design) for kinetic barriers, protein folding and assembly will generally be under thermodynamic control.

Protein thermostability is the rule, not the exception

Protein stability is determined by the balance between the (unfavorable) loss of configurational entropy during folding and the (favorable) formation of attractive interactions in the folded state. The first term increasingly dominates as the temperature increases, and

most naturally occurring proteins unfold at high (~95°C) temperature. In contrast, most de novo designed proteins which can be solubily expressed in *Escherichia coli* remain folded at 95°C—they are generally quite a bit more stable than their naturally occurring counterparts.

What does the observation that high thermostability is relatively easy to attain by de novo protein design tell us about the lack of thermostability of naturally occurring proteins and computational modeling of the forces governing folding generally? First, either there has not been evolutionary pressure for protein thermostability, or selection for protein function has come at the expense of stability (in some cases where protein turnover is important, selection for function may have even have favored marginal stability). Second, our understanding of the forces stabilizing proteins must be at least qualitatively accurate for brand new proteins to consistently be so stable.

Why are de novo designed proteins so stable? Whereas native proteins almost always have idiosyncrasies such as irregular loops and non-ideal secondary structures, designed structures are closer to the Platonic ideal of a protein: they have well-packed exclusively polar surfaces and exclusively hydrophobic cores (with the exception of the hydrogen bond network cores described below), regular secondary structural elements, and canonical turns. Whether such ideal structures could actual exist was unclear before the advent of de novo protein design; it is now evident that there are an essentially unlimited number of them. There has been considerable discussion of the origins of the high stability of thermophilic proteins; the comparison with de novo designed proteins suggests that there may be nothing particularly remarkable to look for.

Insight into the balance of forces in folding: the importance of backbone strain

What has de novo protein design taught us about the balance of forces in folding? It has long been accepted that the hydrophobic effect is the dominant force favoring protein, and indeed, like most native proteins, de novo designed proteins generally have primarily hydrophobic cores. A less well-appreciated contribution to protein folding whose understanding was critical for success in protein design is local backbone strain. Much of our understanding of protein stability has come from investigation of the effect of amino acid substitutions—while this directly reports on the contribution of hydrophobic interactions, buried charge, etc., it does not report on the relative free energies of different backbone conformations. In contrast, success in de novo protein design, particularly of beta sheet containing structures, has required systematic analysis of the consequences of backbone stiffness and chirality on folding into compact globular structures.

De novo protein design proceeds in two steps: first, the generation of target protein backbones, and second, the design of sequences whose lowest energy states are the target backbones. Somewhat unintuitively, the first step is often the hardest—a target backbone must have sufficiently little strain that it is designable; i.e., that there exists an amino acid sequence for which it is the lowest energy state. Simply collapsing of the chain into a structure with a buried hydrophobic core almost always produces strained backbones. Understanding how to design compact backbones without strain has required a combination of geometric reasoning, detailed simulations of simple model polypeptide systems, and study of naturally occurring structures.[10,11]

The consideration of backbone strain and how to systematically relieve it has been critical in the design of all beta sheet structures. For example, key to success in designing beta-barrel structures was the realization that maintaining extensive hydrogen bonding between the strands without introduction of backbone strain required the breaking of cylindrical symmetry. Breaking symmetry and reducing strain through introduction of beta bulges and glycine residues in the middle of the curved beta strands to relieve steric clashes enabled the de novo design of fluorescent proteins.[12] The importance of strain is more clearly brought out by de novo protein design than by structural characterization of native proteins because nature has put little premium on economy and simplicity: in minimal protein systems, backbone strain is highlighted (long flexible loops dissipate strain), and the vast variety in nature masks fundamental constraints on the structures that can be encoded by amino acid sequences.

Unexplored regions of protein space

Has nature fully explored the space of folded protein structures? The many tens of thousands of protein structures that have been experimentally determined fall into a much smaller set of fold classes; is this because there are a limited number of possible protein folds, or because evolution has only sampled a subset of what is possible? Like the other questions addressed in this essay, this one is very hard to answer solely by study of naturally occurring proteins.

De novo protein design provides a route to address this question by direct construction. For example, in Nature, repeat proteins made from tandemly repeated identical ~30–50 amino acid structural units, such as ankyrin repeat proteins, leucine rich repeat proteins and others play many important functions. Do these represent the full set of what is possible for repeat protein structures or only a subset? It turns out that even this relatively narrow class of structures is only very sparsely sampled by Nature—by de novo design dozens of new repeat proteins with sequences and structures (beyond the individual repeat unit)

unrelated to any known natural proteins could readily be generated.[13]

De novo protein design has shown further that nature has not completely sampled the interaction modalities available to peptide chains. For example, the buried central hydrogen bond networks that confer the modular specificity of the DNA double helix do not have a clear analog in native proteins. The development of methods for designing extended hydrogen bond networks in proteins[14] has led to the design of a new world of protein structures with interaction specificity determined by buried hydrogen networks. This opens the door to design of protein allostery and protein logic; the hydrogen bond networks set the register of the interacting segments so the interaction strengths can be systematically tuned.

That nature has only explored a small subset of what is possible for proteins is not surprising given the vast size of sequence space. For even a relatively small protein of 100 residues, there are 20**100 = ~10**130 possible amino acid sequences (any 1 of the 20 amino acids at each of the 100 positions). For comparison, there are ~10 million species on earth today, with ~100,000 genes each; ~10**12 proteins in total. The total number of proteins sampled over evolutionary time is likely 10**3–10**5 larger than this (<<10**20); a tiny fraction of the 10**130 sequences possible for a 100 residue protein. Likewise, laboratory selection experiments are limited to libraries with a maximum of ~10**15 different sequences. De novo protein design now provides a route to explore the full range of amino acid sequence space for new and useful structures and functions.

Membrane proteins follow same principles

Membrane proteins play critical roles in biology, and, hence, have been extensively studied using X-ray crystallography and biophysical methods. The physics of folding of these proteins is more complex than native proteins as the hydrophobic environment of the membrane is very different than the aqueous environment outside the membrane. The design of complex multipass membrane proteins is hence a stringent test of our understanding of membrane protein biophysics. Crystal structures of de novo designed membrane proteins with up to 215 residues[15] demonstrate that we understand the fundamental driving forces and the determinants of structural specificity sufficiently well to design new stable membrane protein structures. As described below, considerably more accurate models will likely be necessary to model the delicate functions of naturally occurring membrane proteins.

Protein–protein interactions

De novo design has also informed our understanding of protein–protein interactions. By incorporating buried hydrogen networks, it has turned out to be easier than one might have thought to de novo design pairs of proteins which interact with each other with high affinity and specificity.[16] However, the individual partners in these designs are generally not monomeric on their own. The notable challenge solved by evolution that is not yet solved by de novo protein design is the generation of large sets of orthogonal and high affinity protein–protein interactions using protein monomers that are stable and monomeric in the absence of their binding partner(s). The latter requirement likely constrains the size and hydrophobicity of protein interaction surfaces, perhaps favoring smaller hydrophobic patches surrounded by polar groups disfavoring non-specific association.

Beyond the 20 naturally occurring amino acids

The principles and approaches developed for designing new protein structures have been found to hold well outside of the chemical domain sampled by naturally occurring proteins. By designing for very low energy ground states, it has been possible to design a wide variety of structured macrocycles with circular backbones using unnatural amino acids to favor the ground state.[17] Local sequence structure compatibility/lack of strain are particularly important in small macrocyclic systems as the hydrophobic effect plays a smaller role because there is little or no hydrophobic core. Hence, the ground states of these systems are determined more by torsional constraints imposed by the placement of L and D amino acids, in particular proline residues, which restrict the region of the Ramachandran map sampled at each position, and backbone–backbone and sidechain–backbone interactions, than by hydrophobic interactions. Together with the reduction in configurations accompanying backbone closure, these interactions are sufficient to enable the design of sufficiently large energy gaps to specify unique folded states. It is notable that the same forcefield and design principles used to create megadalton icosahedral nanocages hold for eight residue non-natural amino acid containing macrocycles.

Large leaps paradoxically can be easier than small ones

To what extent does progress in de novo protein design constitute a solution to the protein folding problem? On the one hand, successes in de novo design of a wide range of protein structures and folds suggest that the broad outlines of how proteins fold are reasonably well understood. On the other hand, while progress has been made in protein structure prediction, accurately computing structures de novo from a single amino acid sequence remains an outstanding problem (recent progress has relied heavily on co-evolution derived contact prediction which requires large numbers of evolutionarily related sequences[18]). Accurate structure predictions can be made for designed proteins—indeed, as noted above, comparison between the predicted structure and the design model is the obligatory final step before proceeding with gene synthesis and experimental characterization—but

much less regular and ideal native proteins are a greater challenge. As argued below, the problem of accurately computing small energy differences (between near native structures, loop conformations, etc.) will likely compromise accuracy of structure modeling for some time to come.

Despite the progress in de novo protein design, there are still considerable challenges to protein structure modeling. Perhaps surprisingly, some of the hardest problems are those relating to the redesign of the function of existing proteins. Likewise, the classical structure prediction problem of predicting the changes in protein structure accompanying sequence changes can be harder than de novo structure prediction. Non-intuitively, the bigger the leap, the more useful computational protein design becomes.

Why are big leaps better supported by computation than small steps? The answer has to do with the magnitude of the energy differences involved. Both computational protein design and protein structure prediction are based on energy calculations—the fundamental hypothesis is that proteins adopt their lowest energy states. To design a sequence that adopts a completely new structure, we sculpt a sequence for which the target structure has much lower energy than any other state. If the designed global minimum is according to the energy function 10 kcal/mol more favorable than any other state, the outcome is insensitive to energy calculation errors in the 1–2 kcal/mol range—the designed protein will fold to the desired conformation whether the energy gap is 8 or 12 kcal/mol. In contrast, suppose one has an already characterized protein–protein interaction, and the goal is to identify single substitutions which increase affinity by ~5-fold. Here, the relevant energy differences are <1 kcal/mol, and within the error of the energy function/force field. Hence, modeling the effect of individual substitutions on relaxation of a protein monomer, of a protein–protein complex, and on binding affinity can be beyond the precision of current computational models (free energy perturbation methods, which take advantage of cancelation of errors, are probably the best current approaches to such challenges). A further advantage of de novo design for computation currently is that one can focus on properties and interactions (ideal structural elements, hydrophobic packing, etc.) that are well understood: accurate computation of the effect of mutations on native systems can require consideration of non-ideal structural elements, structural waters, potential backbone changes, and extensive buried polar interactions, which are all difficult to model (even very low energy designed all polar interfaces generally fail to form) and can be avoided in de novo designed systems.

The accuracy of an energy function is analogous to the resolution of a microscope-one cannot accurately discern features below the available resolution. So for example, predicting the effect of sequence changes on enzyme activity is very difficult (this is why directed evolution is currently a more powerful way of improving and modifying enzyme activities than computation). Thus, rather unintuitively, de novo design of proteins from scratch is more accessible than computing single sequence changes that increase native enzyme/binding activity, and ab initio structure prediction (to moderate resolution) easier than predicting the subtle effects of single sequence changes on protein structures.

The limited energy function accuracy problem in modeling biological systems is compounded by the biological and evolutionary advantages of systems poised between multiple states. Indeed, biology has likely optimized key transitions to be maximally difficult for computation-to be maximally sensitive to inputs the protein assemblies involved in the transitions should be poised in a fine balance between alternative energy minima. Any decision point—advancing in cell cycle, transcription initiation, etc. at which multiple inputs are integrated—is optimally set hovering between the two competing energy minima such that small shifts in the balance of the inputs can shift the outcome. In this sense, biology may have evolved to be maximally inscrutable (predictably describable by computational models). The de novo computational design of allosteric switchable systems has a similar problem—but if the design is modular and tunable, with some experimental optimization a wide range of switchable behaviors can be obtained.

I would like to close with a few comments on scientific collaboration and creativity. Advances in protein design using Rosetta were made possible by improvements in the Rosetta energy function and sampling methodology made collaboratively by the many research groups in the Rosetta Commons. Free and open sharing of methodological improvements has been critical for progress, and our yearly (and ever growing) meetings have been fun and stimulating. Likewise, within my group, full connectivity and sharing of information has been key to advances. An archetypal model of modern molecular biology research consists of a single PI with a limited supply of magic dust (ideas, insights, knowledge, etc.) leading a research group which she/he distributes this among. Closer to my experience (and goals) is what I call the communal brain—just as the actual human brain has considerably greater cognitive power than a very large collection of several neuron invertebrates, communities of closely interacting scientists (both within and between labs as in the RC) can advance science at a remarkable rate. And just as in the physical brain, the higher the connectivity in the communal brain the more powerful it will be. It is not just the science that profits—my enjoyment of my job and my career thus far is the sum of the interactions with all of the wonderful scientists I have been most privileged to work with.

Acknowledgments

I thank Brian Kuhlman and Brian Matthews for very helpful comments on the manuscript, Ian Haydon for

help in preparing it, and my wonderful group members past and present for making this all possible.

References

1. Kim DE, Gu H, Baker D (1998) The sequences of small proteins are not extensively optimized for rapid folding by natural selection. Proc Natl Acad Sci U S A 95:4982–4986.
2. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D (1997) Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Mol Biol 4:805–809.
3. Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277:985–994.
4. Baker D (2000) A surprising simplicity to protein folding. Nature 405:39–42.
5. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268: 209–225.
6. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368.
7. Huang P-S, Boyken SE, Baker D (2016) The coming of age of de novo protein design. Nature 537:320–327.
8. Bale JB, Gonen S, Liu Y, Sheffler W, Ellis D, Thomas C, Cascio D, Yeates TO, Gonen T, King NP, Baker D (2016) Accurate design of megadalton-scale two-component icosahedral protein complexes. Science 353:389–394.
9. Baker D, Sohl JL, Agard DA (1992) A protein-folding reaction under kinetic control. Nature 356:263–265.
10. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D (2012) Principles for designing ideal protein structures. Nature 491:222–227.
11. Marcos E, Basanta B, Chidyausiku TM, Tang Y, Oberdorfer G, Liu G, Swapna FVT, Guan R, Silva D-A, Dou J, Pereira JH, Xiao R, Sankaran B, Zwart PH, Montelione GT, Baker D (2017) Principles for designing proteins with cavities formed by curved β sheets. Science 355:201–206.
12. Dou J, Vorobieva AA, Sheffler W, Doyle LA, Park H, Bick MJ, Mao B, Foight GW, Lee MY, Gagnon LA, Carater L, Sankaran B, Ovchinnikov S, Marcos E, Huang P-S, Vaughan JC, Stoddard BL, Baker D (2018) De novo design of a fluorescence-activating β-barrel. Nature 561:485–491.
13. Brunette TJ, Parmeggiani F, Huang P-S, Bhabha G, Ekiert DC, Tsutakawa SE, Hura GL, Tainer JA, Baker D (2015) Exploring the repeat protein universe through computational protein design. Nature 528:580–584.
14. Boyken SE, Chen Z, Groves B, Langan RA, Oberdorfer G, Ford A, Gilmore JM, Xu C, DiMaio F, Pereira JH, Sankaran B, Seelig G, Zwart PH, Baker D (2016) De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. Science 352:680–687.
15. Lu P, Min D, DiMaio F, Wei KY, Vahey MD, Boyken SE, Chen Z, Fallas JA, Ueda G, Sheffler W, Mulligan VK, Xu W, Bowie JA, Baker D (2018) Accurate computational design of multipass transmembrane proteins. Science 359:1042–1046.
16. Chen Z, Boyken SE, Jia M, Busch F, Flores-Solis D, Bick MJ, Lu P, VanAernum ZL, Sahasrabuddhe A, Langan RA, Bermeo S, Brunette TJ, Mulligan VK, Carter LP, DiMaio F, Sgourakis NG, Wysocki VH, Baker D (2019) Programmable design of orthogonal protein heterodimers. Nature 565:106–111.
17. Hosseinzadeh P, Bhardwaj G, Mulligan VK, Shortridge MD, Craven TW, Pardo-Avila F, Rettie SA, Kim DE, Silva D-A, Ibrahim YM, Webb IK, Cort JR, Adkins JN, Varani G, Baker D (2017) Comprehensive computational design of ordered peptide macrocycles. Science 358:1461–1466.
18. Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D (2017) Protein structure determination using metagenome sequence data. Science 355:294–298.