Cell

Leading Edge
Perspective

Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution

Sarel J. Fleishman[1,*] and David Baker[2,3,*]
[1]Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel
[2]Department of Biochemistry
[3]Howard Hughes Medical Institute
University of Washington, Seattle, WA 98199, USA
*Correspondence: sarel@weizmann.ac.il (S.J.F.), dabaker@uw.edu (D.B.)
DOI 10.1016/j.cell.2012.03.016

The folding of natural biopolymers into unique three-dimensional structures that determine their function is remarkable considering the vast number of alternative states and requires a large gap in the energy of the functional state compared to the many alternatives. This Perspective explores the implications of this energy gap for computing the structures of naturally occurring biopolymers, designing proteins with new structures and functions, and optimally integrating experiment and computation in these endeavors. Possible parallels between the generation of functional molecules in computational design and natural evolution are highlighted.

Introduction

The complexities of life arise from the marvelous and intricate functions carried out by the millions of precisely ordered macromolecules present in living systems. As suggested by Anfinsen (Epstein and Anfinsen, 1962), it is likely that these precisely ordered states are global free energy minima and that the precise ordering reflects folding of macromolecules to their lowest free energy states. For biomolecule native states to be at global free energy minima, the attractive interactions in the folded state must be strong enough to overcome the very large entropic cost to folding (Figure 1). Encoding the large energy gaps required for folding in the linear amino acid or nucleic base sequence is quite nontrivial given the weak and relatively unspecific noncovalent van der Waals, hydrogen bonding, and hydrophobic interactions operating within macromolecules.

Biological self-organization depends not only on the folding of biopolymers to precise structures, but also on the specificity of macromolecular interactions. Proper cell functioning requires that protein interaction networks have well-defined specificity, with each protein interacting with a small subset of the myriad of other biomolecules in a typical cell compartment. Temporal and spatial control of expression partly insulate proteins from interacting with one another (Kuriyan and Eisenberg, 2007; Shapiro and Losick, 2000), and chaperones encapsulate slowly folding proteins to ensure that their hydrophobic cores are not exposed to spurious binding; yet, at any point in time and space, most biological macromolecules are in close vicinity to millions of others with which they do not interact (Scott and Pawson, 2009). A simple consideration of the quantities involved illustrates the magnitude of the challenge of insulating protein interactions from one another: a bacterial cell is estimated to contain 4,000 different protein types totaling a million polypeptide chains with total protein concentration of 3 mM (Moran et al., 2010). Against the backdrop of such high density of competing partners, specificity requires an energy gap between the correct set of interactions and the very much larger set of nonspecific interactions (Figure 1) that, if overly populated, would interfere with proper signaling and other molecular transactions. Breakdowns of interaction specificity (Zarrinpar et al., 2003), like breakdown in folding stability (Dobson, 2003), can lead to reduced organism fitness and disease, providing the selection pressure to drive the evolution of these high energy gaps.

This Perspective explores the far-reaching implications of the requirement of a large energy gap for biomolecular organization. First, we describe how the necessity of the energy gap has opened up new approaches to macromolecule structure determination. Second, we describe the challenge that the energy gap requirement poses for both protein design and natural evolution and the approaches that have been taken in design, and the possibly related strategies in natural evolution, to overcome it.

Structure Prediction

The necessity of the energy gap has direct implications for macromolecule structure prediction. Because of the energy gap, structure prediction can be posed as a search for the lowest energy conformation of a protein or RNA polymer. Successful structure prediction should be possible if conformations close to the native structure are sampled and the errors in energy functions are smaller than the energy gap. In such cases, the lowest computed energy state will likely be close to the native state despite error in the energy calculations.

Over the past several years, structure prediction problems for a wide variety of biological systems have been found to exhibit remarkably similar properties. We have used extensive conformational sampling with the Rosetta structure modeling program to map out the energy landscapes and predict structures of globular proteins (Bradley et al., 2005), membrane proteins (Barth et al., 2009), homo-oligomers (André et al., 2007), RNA

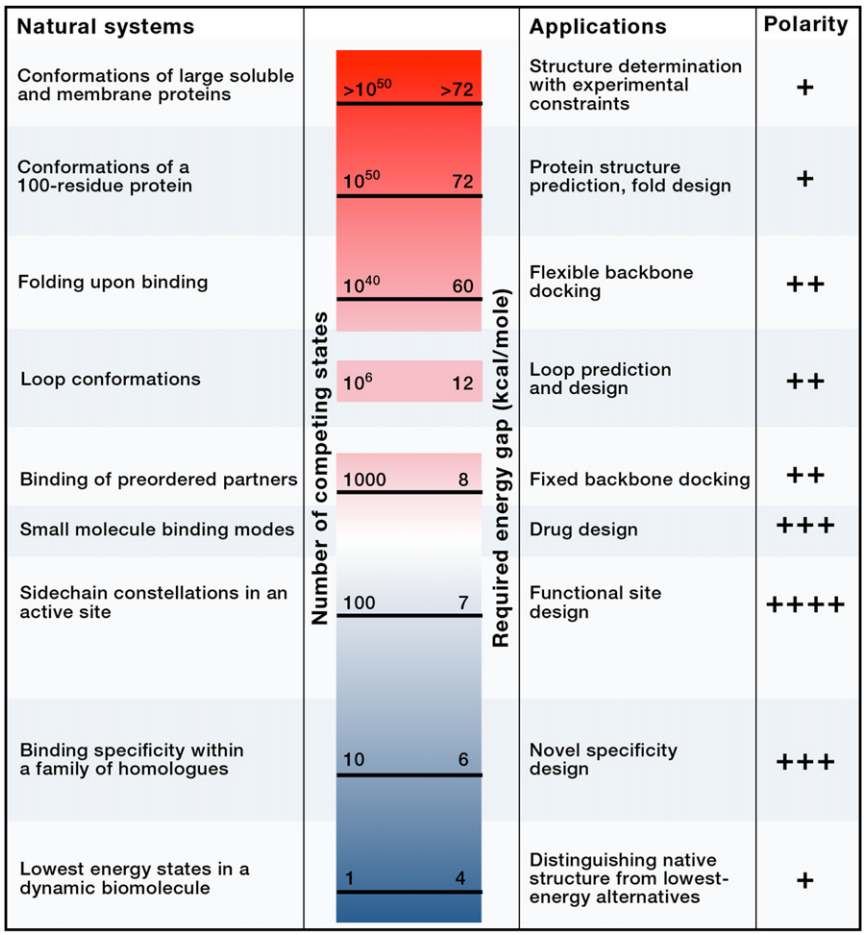

Figure 1. Modeling Difficulty Is Determined by Computational Complexity, the Energy Gap, and System Polarity

The figure illustrates connections between computation, thermodynamics, and applications for diverse molecular phenomena in biology. There are two major determinants of the difficulty of structure prediction and design problems. The first is the number of competing states (numbers on the left of bar). This ranges from a handful (bottom) to astronomically large (top). The second is the polarity of the interactions (column on right side). This ranges from almost completely nonpolar in the case of monomeric protein folding to largely polar in the case of functional site modeling. For structure prediction problems, though the search becomes more difficult, distinguishing the native structure from the alternatives becomes easier as the number of competing states increases, as the built-in energy gap (right of bar) must be larger for the native state to be highly populated. For design problems, in contrast, difficulty increases because design cannot rely on a built-in energy gap. For both structure prediction and design, nonpolar interactions can be more accurately modeled than charged interactions, and hence fold prediction and design can be easier than functional site prediction and design despite the larger number of competing state in the former. The energy gap ΔE required to give rise to 99.9% population of a desired state in equilibrium with N equal energy competitors is readily obtained from the Boltzmann expression: $p = \exp(-\Delta E/kT)/[\exp(-\Delta E/kT)+N]$, in which k is the Boltzmann constant and T is the absolute temperature. For protein folding (top), assuming three degrees of freedom per residue, the number of unfolded states of a 100 residue polypeptide chain is on the order of 3^{100} (Levinthal, 1968), and to obtain 99.9% of a single state in a population of this size would require on the order of 70 kcal/mole of attractive interactions (similar estimates are obtained from experimental data; Brady and Sharp, 1997). The numbers and polarity valuations are very coarse estimates and are for illustrative purposes only.

molecules (Das and Baker, 2007), and protein-DNA complexes (Ashworth et al., 2010). As illustrated in Figure 2, the energy landscapes mapped out by these calculations are qualitatively very similar for very diverse biomolecular systems. The results of calculations on all of these systems exhibit several common features: (1) independent structure calculation trajectories end in different local minima with widely varying conformations and energy, (2) the native conformation has lower energy than almost all nonnative minima, and (3) the energy drops only quite close (<2 Å root-mean-square deviation [rmsd] of the main-chain atoms) to the native state. A sharp drop in energy near the native state has also been observed in long molecular dynamics (MD) simulations of protein-small-molecule binding (Shan et al., 2011). Property 1 reflects the rugged nature of macromolecular free energy landscapes in which even closely related minima can be separated by large barriers from high-energy atomic clashes. Property 2 follows from the necessity of the energy gap for the existence of stable and unique conformations and suggests that the depth of the energy gaps is, in general, greater than the magnitude of the noise due to inaccuracies in current energy functions. Property 3 arises because the tight complementary jigsaw puzzle-like packing of side chains/bases that are responsible for the very low energy of the native state requires close to native state backbone geometry. The existence of the energy gap (property 2) can have unexpected consequences; because of the gap, nonscientists can contribute to structure prediction efforts through online games like FoldIt by using human intuition and problem-solving skills to improve search for the lowest-energy (highest-scoring) state (Cooper et al., 2010).

Because the energy gap from unfolded conformations to the native structure is so large, success in structure prediction and folding to the native structure does not indicate that forcefields have the ~1 kcal/mole accuracy that is necessary for successful small-molecule docking, discrimination between protein excited and ground states, and other applications that require fine energy discrimination (Figure 1). Indeed, despite success in ab initio structure prediction for very small proteins (Kinch et al., 2011), refinement of models based on structures of sequence homologs for larger proteins has been very challenging—there only need exist a small energy gap between the native structure and compact low-energy near-native states (as long as there are

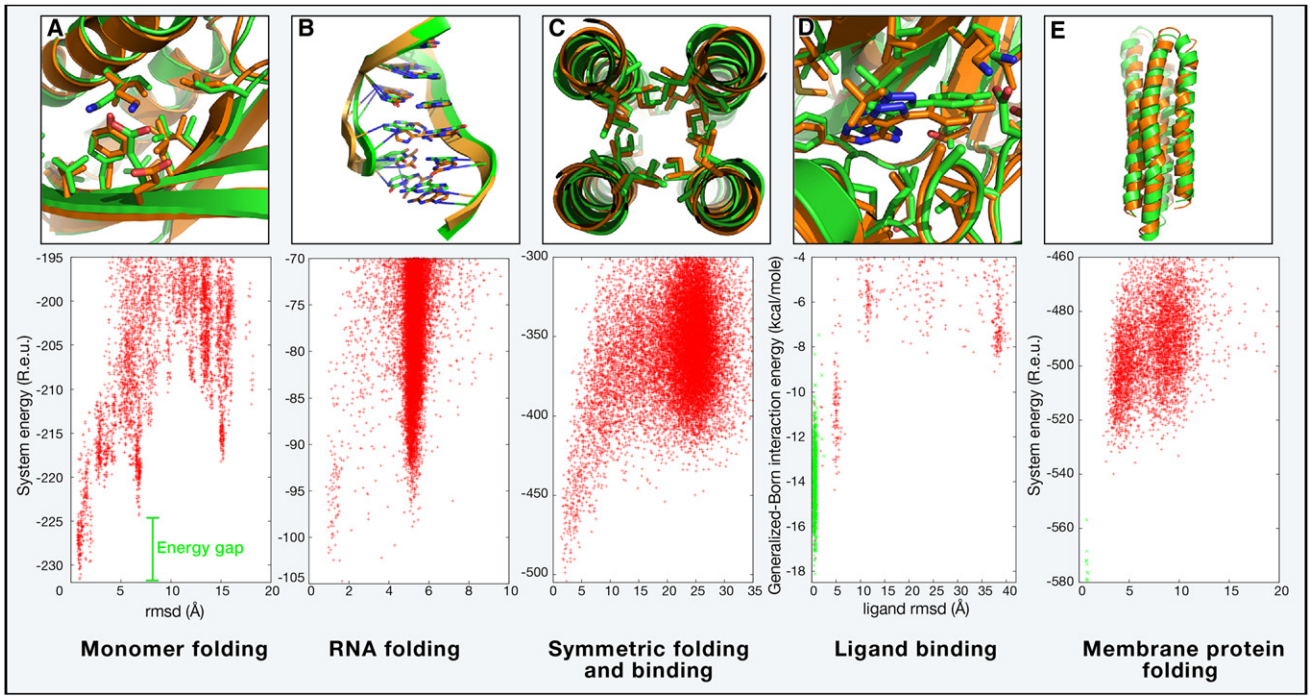

Figure 2. Energy Gaps in De Novo Structure Prediction Calculations for Diverse Macromolecular Systems

(A–E) The structures of small macromolecular systems can be predicted de novo because the evolutionarily encoded energy gap between the native state and the large number of nonnative states compensates for inaccuracies in current force fields. In each of these systems, the starting point for simulation is an extended chain or unbound monomers. In structure prediction (A–C and E), search is carried out by stochastic Monte Carlo sampling of the internal degrees of freedom (monomer folding) and, in the case of complexes, the rigid body degrees of freedom. To effectively sample the astronomically large conformation space in these systems (see competitor states in Figure 1), conformational search is biased to sample backbone and side-chain conformations that are observed in natural biopolymers. Each configuration is evaluated according to an energy function representing van der Waals interactions, hydrogen bonding, solvation, and electrostatic interactions and is selected if it is energetically more favorable than the preceding structure or only slightly worse. This process is repeated, iteratively isolating lower-energy structures. In molecular dynamics (MD) simulations (D), the system is similarly started away from the native state, and the physical forces operating between the atoms are deterministically simulated at very short temporal intervals, simulating the motion of biomolecular systems. In all cases, conformations with low rmsd from the native state have lower energies (y axis) than those with high rmsd.

(A) In simulations of protein folded states of ribosomal protein S6 based on sequence information alone, the native state is identified with high precision. Alternative conformations also score favorably but less so than the native.

(B) The folded state of the sarcin/ricin domain of 23S bacterial rRNA is precisely captured (Das and Baker, 2007). An alternative local energy minimum is also identified, but it is higher in energy.

(C) Although homo-oligomers are large, symmetric modeling reduces the search space and produces clear energy gaps that identify the native state of the *S. Aureus* tetrabrachion coiled coil.

(D) Long MD simulations starting from nonnative conformations of the PP1 inhibitor bound to Src end in a state that is essentially identical to the native state (Shan et al., 2011). Early phases of the simulation (<700 ns) have high rmsd from the native state and high energies (red points), and as the trajectory progresses (>1400 ns), the conformations converge on the native state (green points). Data for generating the panel were generously provided by Yibing Shan and David Shaw.

(E) The membrane-embedded vacuolar ATPase shows a clear energy gap between near native conformations (green) and far-from-native conformations (red), although, in this case, the prediction lacks atomic-level accuracy. Molecular representations were generated with PyMol (DeLano, 2002). Green and gold represent the prediction and the native state, respectively, and blue, red, and yellow represent nitrogen, oxygen, and sulfur atoms. R.e.u., Rosetta energy units. Protein data bank accession codes for the native states from (A)–(E) are: 1LOU (Otzen et al., 1999), 1Q9A (Correll et al., 2003), 1FE6 (Stetefeld et al., 2000), 1QCF (Schindler et al., 1999), and 2BL2 (Murata et al., 2005), respectively.

relatively few of them; Figure 1), and this small energy gap may, in some cases, be within the noise of current forcefields. Indeed, very intensive conformational sampling can reveal alternative minima, which might be due to energy function errors (Das, 2011; Mandell et al., 2009) or may correspond to alternative conformational states that are not seen in static crystal structures (Tyka et al., 2011). Due to the small magnitude of the energy gap, such alternative conformations are quite difficult for macromolecular forcefields to distinguish from the predominant state observed in experiment (Figure 1). Another important challenge is correctly accounting for electrostatics and interactions with solvent in both folding and binding. Flexible ligand docking and drug design are challenging for two reasons: first, ligands often interact with their targets through polar and charged interactions that are difficult to model accurately; second, the total free energy of binding is generally small, and hence the energy gap between the experimentally observed binding mode and alternatives is, in many cases, too small for current methods to detect reliably.

As the number of conformations accessible to a biopolymer increases very rapidly with chain length (Levinthal, 1968), the ability of unconstrained ab initio folding calculations or MD

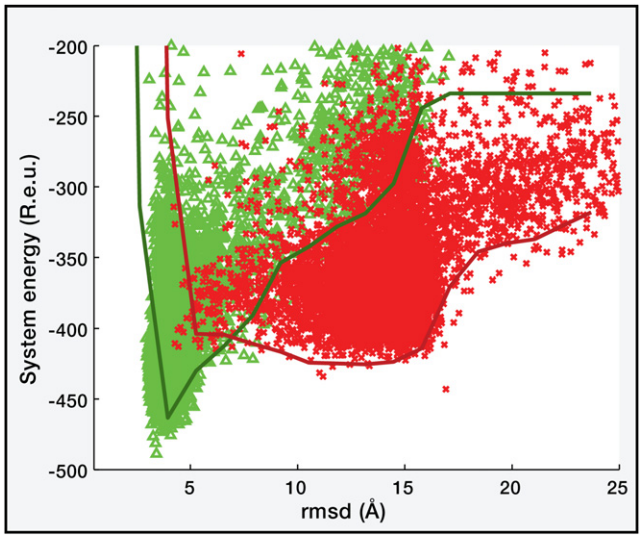

Figure 3. Structure Determination Using Sparse Experimental Constraints to Guide Conformation Research

Modeling of large macromolecules is limited by conformational sampling, but sparse experimental constraints can guide sampling toward the native state. In these simulations, the energy is computed solely based on a physical model, and the constraints are only used to bias sampling.

NMR chemical shift residual dipolar coupling (CS-RDC) data were used to constrain sampling of a 25 kD protein (green), yielding a clear energy gap between near and far from native states (Raman et al., 2010). Without the experimental constraints (red), no energy gap is seen, with conformationally very different structures showing equally low energies. Identifying energy gaps in the presence of experimental constraints, but not in their absence, can thus provide an inherent control of the prediction's veracity. Panel was adapted with permission from Raman et al. (2010).

trajectories to sample close enough to the folded state to detect the energy gap in large systems is quite limited (Figure 1). For this reason, ab initio structure prediction currently is not practically useful for structure determination of proteins larger than ~80 residues, including multidomain and membrane-embedded proteins, and important progress remains to be made in this arena (Cozzetto et al., 2008).

New Approaches to Structure Determination

Though current ab initio structure prediction methods are not of practical use in determining reliable macromolecular structures for all but the smallest proteins, the necessity of an energy gap has led to new areas of application for these methods, where they appear to have considerable utility. Traditional biomolecular structure determination methods use experimental nuclear magnetic resonance (NMR) spectroscopy or X-ray data to determine the detailed arrangements of atoms in protein and RNA structures. Large amounts of data are needed to unambiguously determine the positions of the atoms. However, because of the energy gap, the experimental data can be used in a quite different way. Rather than determining the detailed atomic positions, they can be used to guide the search process, and the correct structure can then be selected based on its very low energy.

A simple analogy illustrates the power of even a very limited amount of experimental data in locating a global minimum. Consider the problem of finding the lowest elevation point on the land-covered surface of the earth. Without experimental information, search may incorrectly converge on Death Valley in California. However, with the single datum that the lowest elevation point is not in North America, this can be immediately eliminated, and with the additional datum that the lowest elevation point is in the Middle East, search can much more rapidly hone in on the Dead Sea. As in the structural calculation case, the experimental data do not define the exact location of the minimum; they can serve, rather, to rule out large regions of space that would otherwise greatly slow down the search for the lowest-energy structure.

New methods that exploit the energy gap have been particularly successful for NMR structure determination. Traditional structure determination via NMR involves assignment of the backbone and side chain resonances and, subsequently, the interpretation of NOESY spectra, which report on distances between atoms. The assignment of backbone resonances is largely automated, but assigning NOESY spectra is time consuming and, for larger proteins, complicated by considerable spectral overlap. The distances obtained from NOESY spectra are critical for traditional approaches that seek to define the positions of all atoms based on the experimental data. By contrast, new approaches that utilize the data primarily to guide sampling can build reliable models, in some cases using just the chemical shift assignments for the backbone atoms, which provide information on local backbone structure. The use of these data to guide search dramatically increases the accuracy of the resulting structures, which for small proteins can be close to the accuracy of models determined using much larger data sets with conventional methods (Shen et al., 2008). For larger proteins, sampling again becomes problematic, but supplementing the backbone chemical shifts with sparse backbone RDC and H^N-H^N NOE data allows the search to hone in on the low-energy native structure for proteins up to 25 kD (~200 residues) (Raman et al., 2010). The backbone chemical shift data constrain which backbone torsion angles are sampled, whereas the long-range interactions bias the search to conformations with the correct overall topology. The necessity of the energy gap also enables a new structure validation criterion: because constraints focus search on the region where the native state lies, the energy should be lower when sampling with constraints than without (Figure 3) (Raman et al., 2010). This is only likely to be the case if the native energy minimum is sampled; otherwise, constraining sampling should result in higher rather than lower energies. In addition to extending the size range and reducing the time that is required for NMR structure determination, this method has considerable potential for determining the structures of transiently populated states, in which experimental data are often very sparse (Bouvignies et al., 2011; Korzhnev et al., 2010).

Structure Determination from Sparse X-Ray, Cryo-EM, and Proteomics Data

Macromolecular structures are being solved at a very rapid pace by X-ray crystallography, but when the resolution is low (>3.5 Å) or the starting phase information is poor, it becomes difficult to resolve the positions of the atoms and obtain an accurate

structure using traditional methods. In such cases, structure determination may still be possible by using the experimental data to guide the search for the lowest-energy structure. X-ray structure determination by molecular replacement utilizes phase information from homologs to initiate the structure refinement process but can fail when the homologous structures are too divergent. However, even this very noisy information can still help guide Rosetta energy-based search for low-energy structures (DiMaio et al., 2011); this has been found to considerably increase the radius of convergence of molecular replacement and has allowed the solution of many previously unsolved structures. The use of electron density to guide energy-based refinement also has promise for obtaining atomic models from cryo-EM data and low-resolution X-ray data sets. Refinement in these cases is quite challenging due to the larger size of the molecules typically studied and the lower resolution of the data, and consistent refinement of cryo-EM models to atomic accuracy remains an open research problem (Baker et al., 2010; DiMaio et al., 2009; Schröder et al., 2010). At still lower resolution, large-scale proteomic data can be used to guide modeling of large macromolecular complexes such as the nuclear pore (Alber et al., 2007). Looking forward, hybrid approaches that utilize experimental data from cryo-EM density maps, solid-state NMR, and proteomics experiments to guide energy-based search could provide much-needed information on the internal structures of other large complexes that defy conventional structure determination.

How to Encode the Energy Gap: Design and Evolution

Biomolecule design is a stringent test of our understanding of the principles underlying biomolecular organization and can, in principle, lead to a whole new world of molecules with novel and useful functions. Design and structure prediction are inverse problems: whereas in structure determination/prediction, the challenge is to find the lowest-energy structure for fixed sequence, in design, the challenge is to find the lowest-energy sequence for a specified structure or function. Because both prediction and design are fundamentally searches for low-energy states, closely related methods can be used to solve both problems; this duality has spurred the development of Rosetta and other prediction and design software (Kuhlman et al., 2003).

While both problems involve searches for low-energy states, unlike in structure prediction, in protein design, there is no built-in energy gap to favor the target conformation over the competitors, as there has been no evolutionary selection for function and conformational uniqueness. Design is thus much more susceptible to forcefield inaccuracies, particularly involving polar interactions, such as hydrogen bonding and electrostatics, which play important roles in catalysis and binding specificity (Sharp and Honig, 1990). In the perspective taken here, structure prediction probes the principles of biomolecular organization from within the confines of the thermodynamic hypothesis, which ensures the existence of sufficiently high energy gaps selected by evolution, whereas design probes biomolecular organization from outside of these confines. How to encode the necessary energy gap into designed biomolecules is a central challenge (Figure 1), which highlights the remarkable effectiveness of natural evolution in producing energy gaps and precisely folded native states for hundreds of thousands of diverse biomolecules. Indeed, protein design has often turned to nature for inspiration on how to encode energy gaps.

There are three broad classes of approaches to designing biomolecules with large energy gaps. The first class we will call "forward design." This strategy seeks to optimize the sequence such that the target folded structure is so low in free energy that any other folded structure is likely to be higher in energy and thus disfavored. The second class we will call "explicit negative design." This strategy explicitly considers a set of alternative structures and optimizes the sequence such that the desired state is lower in energy than any of the alternatives. The third class we will call "heuristic negative design." This strategy seeks to disfavor alternative energy minima by employing heuristics that increase the energy of most nontarget states. In the following, we will describe how these strategies have been employed to design new biomolecules and how they have drawn inspiration and guidance from nature.

Forward Design

In this approach, possible competing states are not considered explicitly, and the focus is instead on making the desired state as low in free energy as possible. This approach has been applied most successfully to protein fold design. The justification for this approach is that, to fold into a unique structure, a biopolymer must encode very many precise stabilizing atomic interactions, and so stable alternative structures are unlikely to arise by chance (Figure 1). Indeed, screens for random sequences that adopt folded structures suggest that they are extremely rare (Scalley-Kim et al., 2003) unless they have certain hydrophobic-polar patterns (Xu et al., 2001). The most stringent test of forward design applied to monomeric proteins is the computational design of a protein topology not observed in nature, which yielded a very stable protein named Top7 (Kuhlman et al., 2003). X-ray crystallographic studies of Top7 showed that the molecular structure was nearly identical to the computational model, demonstrating the sufficiency of current energy functions for creating new structures from scratch with atomic-level accuracy. Even in this case, however, there were elements of negative design, as structure prediction calculations were used to ensure that the native state was lower in energy than any alternatives.

Forward design has also been used to design small molecules to bind, inhibit, or induce the function of enzymes and proteins involved in signal transduction to target many traditional classes of drug targets, including kinases, proteases, and, more recently, protein-protein interactions (reviewed by Ekins [2006]). In high-throughput computer calculations, millions of small molecules can be docked into a target site and the most tightly binding compounds identified. However, because the docked molecules are small and the interactions often quite polar, it unfortunately is often the case that the desired bound state does not have a significant energy gap relative to other states, and the small molecules in practice bind in alternative modes to the same structure or to entirely different proteins. Indeed, the small energy gaps are perhaps the major issue confounding computer-based drug design.

Explicit Negative Design

In this approach, a number of ''competitor'' states are explicitly modeled, and design seeks to maximize the Boltzmann weight of the desired state relative to the competitors by both decreasing the energy of the desired state and increasing the energy of the competitors (Havranek and Harbury, 2003). Such multistate design has been used to generate specific coiled coils and DNA binding and cleaving enzymes (Ashworth et al., 2006; Grigoryan et al., 2009; Havranek et al., 2004; Havranek and Harbury, 2003).

There are interesting possible parallels to explicit negative design in nature. Cellular protein interaction networks in critical processes such as signaling often involve highly homologous binding components (Meenan et al., 2010; Newman and Keating, 2003; Zarrinpar et al., 2003). Due to this high homology, insulating interactions from one another is challenging but crucial for proper function. The importance of insulating interactions for organismal fitness is illustrated by a yeast SH3 domain-binding peptide that binds with high specificity to only one of the 27 SH3 domains in yeast but nonspecifically crossreacts with many non-yeast SH3 domains. (Zarrinpar et al., 2003). Sequence variants of this peptide, which bound additional yeast SH3 domains, conferred a fitness defect to yeast cells expressing them, suggesting a role for negative selection in interaction insulation. In vitro evolution studies have shown further that binding specificity does not arise simply as a byproduct of selection for higher binding affinity and that selection pressure against binding undesired targets must sometimes be explicitly enforced to get high-specificity binding (Collins et al., 2006; Levin et al., 2009).

Heuristic Negative Design

A problem with the explicit negative design strategy for both computational design and natural evolution is that the set of undesired alternatives must be enumerated for the calculations or present during selection. In the design case, this requires that the set of undesired structures/complexes be already known and not too large (otherwise the calculations become intractable). The heuristic negative design strategy, in contrast, builds up the energy gap not by explicitly disfavoring specific alternative competitors but by incorporating features that are likely to increase the energy of most undesired states, making them less favorable. An example of heuristic negative design is presented by edge strands in β sheets, which often are quite polar and somewhat irregular, which disfavors pairing with other strands and hence aggregation or nonspecific association with other β sheet-containing proteins (Richardson and Richardson, 2002); the resultant proteins avoid undesired association without requiring selection against binding to each and every undesired protein. Natural drug-like small molecules have, on average, more chiral centers, fewer rotatable bonds, and more rings than do molecules in chemical libraries used for screening and identification of drug candidates; these properties likely enhance binding specificity and are key features that drug design aims to emulate (Feher and Schmidt, 2003). Again, there are close parallels between strategies used in negative design calculations and strategies that nature has appeared to employ to achieve energy gaps required for function.

Though, as noted above, the forward design strategy appears to be sufficient to generate unique folded states, it appears likely that nature has employed heuristic negative design to increase folding cooperativity. Kinetic studies of the folding of the de novo designed protein Top7 revealed the population of several stable intermediate structures and overall low folding cooperativity (Scalley-Kim and Baker, 2004; Watters et al., 2007). These results suggested that cooperative folding is not a necessary feature of stable proteins but, rather, that cooperativity emerges by evolutionary selection. A likely explanation for natural selection of cooperatively folding polymers is that partially folded substructures of proteins are more prone to aggregation and amyloid formation with potentially catastrophic fitness consequences (Dobson, 2003; Eichner and Radford, 2011). The low sequence identity of homologous domains in large multidomain proteins may also reflect heuristic negative design to reduce interdomain misfolding and aggregation (Borgia et al., 2011; Wright et al., 2005).

Heuristic negative design is likely to be particularly important for biomolecular interactions. Interactions between biopolymers or between biopolymers and small molecules need only overcome the entropy loss of limiting six rotational and translational degrees of freedom (though conformational changes have an important role in molecular recognition), and because the entropic barrier to binding is much lower than that for folding, new macromolecular interactions can arise quite readily. For example, it has been estimated that more than 50% of the T cell receptors that are capable of undergoing positive selection for binding of peptide-MHC complexes are deleted due to undesired interactions with self antigens (van Meerwijk et al., 1997). It is unlikely that interaction specificity in biological systems arises solely from explicit negative design: a necessity for each protein or RNA molecule to be specifically selected not to bind to all coexisting biopolymers implies an improbable fragility of the biopolymer complement in every cellular compartment. Rather, it seems probable that there are general rules that, although they cannot prevent all crossreactivity, do minimize its likelihood. One clear trend that likely arises from heuristic negative design is the absence of large clusters of hydrophobic residues (which can nucleate protein-protein interactions) on most protein surfaces.

Design efforts can help to bring into focus principles underlying biological self-organization. A comparison of interaction sites on natural proteins to those on designed proteins suggested that the former were more conformationally restricted (Fleishman et al., 2011b, 2011d). Thus, one of the mechanisms that nature uses for heuristic negative design appears to be conformationally restricting potentially promiscuous sets of side chains and loop segments so that they are unable to form undesired interactions or result in protein misfolding. As described below, protein interface design methodology has had some success attempting to emulate this property, illustrating again how insights from nature can inform design.

Another biomolecule class in which nature has apparently used heuristic negative design is the intrinsically unfolded proteins (IUP) (Wright and Dyson, 1999). These proteins contain domains that lack structure in solution but often fold into a distinct three-dimensional structure when bound to their target proteins (Dyson and Wright, 2005). To ensure that the proteins do not fold in isolation, IUPs lack bulky hydrophobic residues

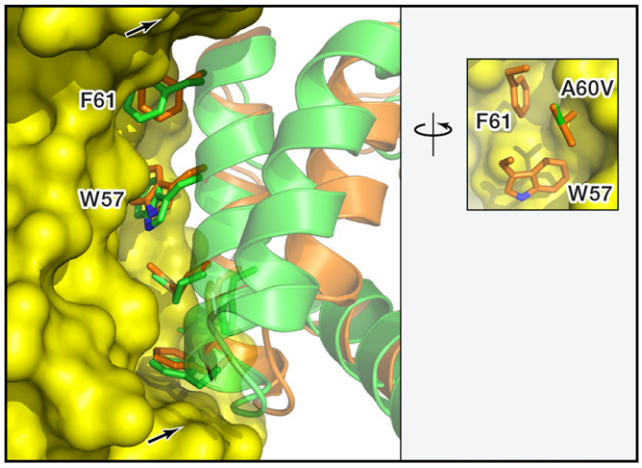

Figure 4. Increasing the Energy Gap in Designed Proteins through In Vitro Evolution

Design of energy gaps in macromolecules is limited by the accuracy of underlying energy functions but can be achieved by experimental iterative improvements of activity through in vitro selection. The experimentally determined molecular structure of a de novo designed binder of influenza hemagglutinin (gold) shows atomic-level agreement with the model (green) (Fleishman et al., 2011c). Starting binding affinity was low (Kd > 1uM), but affinity maturation through in vitro selection identified mutations that improved binding affinity, e.g., A60V, which increases the shape complementarity of the interacting surfaces (inset), but was not identified by the design calculations because of minor steric clashes with neighboring protein backbone atoms. Thus, affinity-increasing substitutions reveal missing elements in macromolecular modeling and design and drive improvements in design methodology. Arrows point to hemagglutinin surfaces that form a canyon around the critical site, likely to evade recognition by bulky immune antibodies (Rossmann, 1989). By utilizing small protein scaffolds, computational design can circumvent the constraints imposed by pathogens on binding surfaces vital to their reproduction. Hemagglutinin is rendered as a yellow surface.

and have high polar and charged residue propensity. These negative design rules are so prominent that they have been used quite successfully to predict the existence of IUPs on the basis of sequence information alone (e.g., Mizianty et al., 2011). Countering the vast entropic penalty of the unfolded to folded-and-bound transition requires a large contact surface area encompassing many favorable interactions with the target molecule, which results in high specificity, and the entropic penalty of folding-upon-binding reduces affinity: high specificity and low affinity are hallmarks of the regulatory processes in which IUPs are prominent, such as signaling and transcriptional regulation. In the perspective taken here, IUPs stand out in that the required energy gap only arises when their binding partner is present, ensuring that, though they do not adopt unique structures in isolation, they bind their targets with very high specificity.

Pathogens also appear to have utilized heuristic negative design to fend off the host immune system. For example, viral surface proteins such as influenza hemagglutinin have deep surface depressions that hide regions that participate in viral attachment and cellular invasion. These so-called structural "canyons" (Rossmann, 1989) significantly reduce access to immune antibodies and thus allow the maintenance of conserved sites free of immune system pressures (arrows in Figure 4). Rather than responding to selective pressures from individual antibodies, viruses thereby employ heuristics to prevent recognition by a vast majority of antibodies.

De Novo Design of Function

There has been considerable progress in designing proteins with novel functions. As is clear from the above considerations, success requires the existence of an energy gap between functional conformation(s) and the vastly larger number of nonfunctional conformations. We summarize recent progress in designing novel functions, emphasizing both approaches for generating the required energy gap and areas where more work is required to achieve such a gap. In the following, our focus is on de novo design driven by physically realistic modeling, but exciting progress has also been made in de novo design of a hydrogenase (Jones et al., 2007), an oxygen carrier (Koder et al., 2009), and cofactor binders (Cochran et al., 2005) using low-resolution modeling (reviewed by Samish et al., 2011).

Design of Protein-Protein Interfaces

The ability to design proteins that bind tightly to any desired surface on a target macromolecule of known structure would be of tremendous utility in biomedicine. Most methods for designing protein-protein interfaces have relied on forward design by generating sequences that are predicted to bind tightly to their targets. For example, two membrane-spanning peptides targeting two homologous human integrins were designed by utilizing the membrane-protein 5 residue dimerization motif small-xxx-small, in which small are Gly, Ala, or Ser residues, and x is any intervening residue (Yin et al., 2007). Low-affinity homo-oligomeric and hetero-oligomeric complexes have been designed by docking natural proteins and redesigning the residues at the interface (Huang et al., 2007; Jha et al., 2010). Higher-affinity interactions were designed between two normally noninteracting proteins by computational docking guided by specific hydrogen bonding interactions across the interface followed by design of the surrounding residues (Karanicolas et al., 2011). The computationally designed complex, PrbPdar, bound with a dissociation constant (Kd) of 150 nM, and in-vitro selection for higher-affinity variants identified mutations that increased affinity to the Kd < 1 nM range. However, a crystal structure of the evolved complex showed that, although the proteins interacted through the designed residues, conformational changes at the interface led to reorientation of the binding mode by 180°. The observed conformational changes underscored the pliability of protein surfaces and the importance of encoding heuristic elements of negative design to ensure that the desired binding mode is favored over alternatives.

To emulate the conformational restriction of binding patches noted above, which likely functions for heuristic negative design in native proteins, we developed a method that starts by computing clusters of disembodied amino acid side chains that interact favorably with one another and the protein target (Fleishman et al., 2011c). Next, this method identifies scaffolds that can accommodate one of these clusters and finally designs the remaining surface for high binding affinity. The requirement that core side chains form energetically favored spatial clusters reduces the conformational plasticity of the designed binding surfaces because alternative conformations are likely to have higher energies (Fleishman et al., 2011a); this method thus

encompasses elements of both forward design and heuristic negative design. The method was used to generate two proteins that, following sequence optimization by in vitro selection for high-affinity binders, interacted at low nanomolar dissociation constants with a spatially recessed surface on influenza hemagglutinin. One of the proteins inhibited the pH-dependent conformational changes in Spanish and avian influenza hemagglutinin, and the crystallographically determined molecular structure of the other with the Spanish influenza hemagglutinin revealed high accuracy in the modeled interaction (Figure 4). Such designed proteins could potentially serve as antiviral therapeutics and diagnostics.

Design efforts have some potential long-term advantages over evolution in devising new inhibitors. Nature recycles certain protein scaffolds such as the immunoglobulin, PDZ, and ankyrin repeat, and those recur as binders of diverse targets (Pawson and Nash, 2003). Though these scaffold proteins have favorable characteristics as binders, the reuse of scaffolds reflects evolutionary history rather than thermodynamic necessity. By contrast, protein design is unencumbered by evolutionary dynamics and can use any energetically appropriate scaffold; in the case of the hemagglutinin binders (Fleishman et al., 2011c), steric constraints imposed by the hemagglutinin surface that likely result from selection of surfaces that would avoid immune system recognition (Rossmann, 1989) favored the use of small helical protein scaffolds (Figure 4).

Enzyme Design

The challenge in computational enzyme design is to generate proteins that bind to a high-energy transition state and catalyze the chemical transformation. Enzyme design methodology could have wide application in chemical synthesis, creation of new metabolic pathways, bioremediation, and numerous other areas. Enzyme design has been approached by forward design methods: first finding a constellation of amino acid side chains that can catalyze the reaction and then stabilizing these side chains and transition state binding by sequence design (Zanghellini et al., 2006). Computational design has been used to generate several new enzymes, including unimolecular Kemp eliminases (Röthlisberger et al., 2008) and retroaldolases (Jiang et al., 2008) and bimolecular Diels-Alderases (Siegel et al., 2010). Kemp elimination has also been generated in calmodulin to produce an allosterically regulated enzyme by introducing a single glutamate residue in a hydrophobic pocket (Korendovych et al., 2011). Catalytic rates in all of de novo designed enzymes have been quite low (Kcat/Kuncat 10^5) compared to most natural enzymes. *Apo* crystal structures of some designed enzymes have shown good correspondence with the original designed models (Jiang et al., 2008; Röthlisberger et al., 2008; Siegel et al., 2010), but structures in complex with transition state analogs have only started to emerge in the case of de novo designed retroaldolases (Wang et al., 2011). The experimental complex structures broadly agree with the design conception: the catalytic lysine residue forms covalent interactions with the transition state analog, and other hydrophobic interactions are similarly well captured, but the fine details of the water structure surrounding the substrate as well as the positioning of the substrate often differ in the experimental structures.

Achieving catalytic rates, turnovers, and substrate selectivities approaching those of natural enzymes will likely require advances in understanding of the subtle interplay between structural stability and enzyme function (Tokuriki and Tawfik, 2009). Encoding significant energy gaps for (1) the catalytically competent arrangement of active site residues relative to the much larger number of nonfunctional arrangements, (2) the substrate binding mode relative to all other binding modes, and (3) the reaction transition state relative to the ground state is likely to be critical for increasing activity. The number of competitor states in each of these scenarios is small, but encoding the required energy gaps (Figure 1) is a particular challenge for enzyme design, as the catalytic residues are frequently charged and flexible, such as the lysine residue that forms a critical Schiff base in the retroaldolase design mechanism (Jiang et al., 2008). Whereas for native proteins, enzyme dynamics can contribute to catalysis (Eisenmesser et al., 2005), because the chemical step(s) are so well optimized that substrate binding and/or product release become rate limiting, for designed enzymes, at this stage of development of the field, achieving structural precision (by encoding the necessary energy gaps) and reducing sampling of nonproductive states is likely to be critical.

Experimental Optimization of the Energy Gap in Designed Proteins

As described above, directed evolution has been employed to increase the energy gap in designed interactions. This has resulted in orders of magnitude improvements in catalytic activity and binding and has underscored important areas for improvements in computational methods. In the design of influenza hemagglutinin inhibitors, the affinity-increasing mutations improved the shape and charge complementarity of the designed and target surfaces and relieved energetic strain in the designed binding surface (Figure 4) (Fleishman et al., 2011c). In de novo enzyme design, starting activities were improved by two to three orders of magnitude by plate-based activity assays of error-prone PCR (epPCR) libraries encoding variants of the computational designs (Jiang et al., 2008; Khersonsky et al., 2011; Röthlisberger et al., 2008). In the design of retroaldolases, activity-enhancing mutants packed more tightly around a catalytic lysine residue presumably to stabilize the catalytic geometry. Amino acid substitutions that increased the activity of de novo designed Kemp eliminases likely improve the electrostatic compatibility of the enzyme active site for the substrate as well as the stability of the catalytic sites (Khersonsky et al., 2011). Thus, directed evolution of designed proteins can provide insights into energetic aspects of protein function that are systematically missing from design calculations and can provide a clear guide to improving design calculations in future applications. On the flip side, evolution can increase the energy gap of alternative conformations, as may have occurred in the directed evolution of the de novo designed protein-binding pair Prb-Pdar (Karanicolas et al., 2011).

Contrasting Roles of Experiment and Computation in Prediction and Design

The requirement for the energy gap leads to a fundamental difference between the roles of experiment and computation in structure determination and design (Table 1). In structure determination, because of the omnipresent energy gap for

Table 1. Contrasting Roles for Experiment and Computation in Structure Prediction and Design

	Search Problem?	Large Energy Gap?	Solution
Structure calculation	yes	yes	experiment then computation
Function design	no	no	computation then experiment

The necessity of the energy gap leads to contrasting roles for experiment and computation in structure prediction and design. Structure prediction can rely on the naturally encoded energy gap between the native state and its alternatives to accurately identify the native state, but in large macromolecular systems, conformational search is limiting. The solution is to use experimental data to constrain sampling. In the design of function, the target state is arbitrarily chosen, avoiding the search problem, but the low accuracy of the energy function yields small initial energy gaps. The solution here is to iteratively improve designed functions by in vitro selection for higher-activity variants. Energy function accuracy is still an issue for structure prediction (in particular, distinguishing among a handful of lowest-energy states [Figure 1, inset]) but much less so than for design.

native folded structures, the main limitation for the methods described here is conformational sampling, and thus experiments precede computation to guide sampling toward the native state. Once the native state is sampled, it can be identified as such due to the evolutionarily encoded energy gap with respect to other states. In design, sampling is less of a problem because there is no single native state to be found and any conformation that satisfies the design specifications is adequate, but energy function inaccuracies entail small initial energy gaps and hence low initial activities. These small energy gaps can be increased by experimental activity optimization through sequence variation and selection, which has the great advantage of utilizing nature's own energy function. Thus, because of the existence of the energy gap in structure determination of naturally occurring biomolecules and the difficulty of designing energy gaps in new biomolecules, experiment is most effective preceding computation in structure determination and following computation in design efforts.

The Energy Gap, Specificity, and Promiscuity in Natural and Designed Systems

The challenges facing biomolecule designers echo those that constrain the evolution of natural enzymes and cellular interaction networks (Figure 1). Cells developed elaborate machinery to prevent misfolding and undesired associations through chaperones and the unfolded protein response (reviewed by Dobson, 2003). In addition, there are general strategies for avoiding the pitfalls of macromolecular folding in a high-concentration environment, as outlined above. When these rules are broken by mutation, dysregulation and disease can result. The case of sickle cell anemia illustrates the importance of interaction insulation: a surface Glu→Val mutation (Ingram, 1956) forms a hydrophobic surface that does not impair the hemoglobin oxygen-carrying capability but, rather, forms a new site for favorable interactions with other hemoglobin proteins (Pauling and Itano, 1949), leading, under certain conditions, to large oligomers that cripple erythrocytes and cause anemia.

Though this essay has focused on the large energy gaps required for folding and specificity, switching between alternative conformations or interactions requires small energy gaps between competing states. Allosteric regulation by effectors requires very small energy gaps: the free energy difference between the low- and high-affinity states of hemoglobin (for oxygen) states must be quite small (Perutz et al., 1998) to allow the population of the two states to be modulated by protons and CO2; this modulation is critical for vertebrate life. The cell cycle-governing cyclin-dependent kinase (CDK) is regulated by very similar interactions at a common site with different cyclins, which direct kinase activity to different substrates (Miller and Cross, 2001; Morgan, 2007); agonists, antagonists, posttranslational modifications, localization, expression levels, and degradation act to ensure precise patterns of activation, and for these to switch specificity, the energy gaps between the different states must be quite small. Though the energy gap between alternative conformations or interactions must be small, the energy gap relative to the vastly larger number of unfolded or nonspecific interactions must still be very large to ensure that the desired set of discrete states is populated at all. Designing allosteric switchable systems with two low energy states separated by a large energy gap from the sea of unfolded or noninteracting conformations is a current challenge for protein design (Ambroggio and Kuhlman, 2006).

Conclusions

Fascination with the ability of biological macromolecules to interact specifically and at high affinity with one another predates the elucidation of the first protein structures (Pauling, 1948). The energy gap determines self-organization in biological macromolecules and is maintained through evolution by specific and general rules. These rules have traditionally been uncovered through the biophysical examination of aberrant macromolecules (Pauling and Itano, 1949) and are increasingly being characterized through engineering and computational design, shedding light on evolutionary processes and enabling the design of novel protein functions. Although there has been considerable progress, design has far to go in matching naturally occurring binders and enzymes. As one concrete example, immune system antibodies are able to bind a bewildering diversity of biological macromolecules with exquisite specificity through the use of unstructured loops, but the ability to consistently predict loop conformations, let alone design them, remains elusive. Macromolecular prediction and design will continue to gain from better understanding of the mechanisms employed by nature to encode the energy gaps that are required for folding, function, and specific binding. Conversely, prediction and design efforts should highlight fundamental contributions to biological self-organization.

ACKNOWLEDGMENTS

The authors thank Yibing Shan, David E. Shaw, Vladimir Yarov-Yarovoy, Mike Tyka, Oliver Lange, Rhiju Das, and Ingemar Andre for providing the data used to produce Figure 2 and Ingemar Andre, Gira Bhabha, Rhiju Das, Sagar Khare,

Olga Khersonsky, Tanja Kortemme, Eva-Maria Strauch, and Dan S Tawfik for critical reading. S.J.F. is supported by a grant from the Geffen Trust, the Yeda-Sela Center for Basic Research, a donation from Sam Switzer and Family, and the Human Frontier Science Program. Research in the Baker lab was supported by grants from the National Institutes of Health, the Defense Advanced Research Program Agency, the Defense Threat Reduction Agency, and the Howard Hughes Medical Institute.

REFERENCES

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., et al. (2007). The molecular architecture of the nuclear pore complex. Nature *450*, 695–701.

Ambroggio, X.I., and Kuhlman, B. (2006). Computational design of a single amino acid sequence that can switch between two distinct protein folds. J. Am. Chem. Soc. *128*, 1154–1161.

André, I., Bradley, P., Wang, C., and Baker, D. (2007). Prediction of the structure of symmetrical protein assemblies. Proc. Natl. Acad. Sci. USA *104*, 17656–17661.

Ashworth, J., Havranek, J.J., Duarte, C.M., Sussman, D., Monnat, R.J., Jr., Stoddard, B.L., and Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. Nature *441*, 656–659.

Ashworth, J., Taylor, G.K., Havranek, J.J., Quadri, S.A., Stoddard, B.L., and Baker, D. (2010). Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. Nucleic Acids Res. *38*, 5601–5608.

Baker, M.L., Zhang, J., Ludtke, S.J., and Chiu, W. (2010). Cryo-EM of macromolecular assemblies at near-atomic resolution. Nat. Protoc. *5*, 1697–1708.

Barth, P., Wallner, B., and Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. Proc. Natl. Acad. Sci. USA *106*, 1409–1414.

Borgia, M.B., Borgia, A., Best, R.B., Steward, A., Nettels, D., Wunderlich, B., Schuler, B., and Clarke, J. (2011). Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. Nature *474*, 662–665.

Bouvignies, G., Vallurupalli, P., Hansen, D.F., Correia, B.E., Lange, O., Bah, A., Vernon, R.M., Dahlquist, F.W., Baker, D., and Kay, L.E. (2011). Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. Nature *477*, 111–114.

Bradley, P., Misura, K.M., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868–1871.

Brady, G.P., and Sharp, K.A. (1997). Entropy in protein folding and in protein-protein interactions. Curr. Opin. Struct. Biol. *7*, 215–221.

Cochran, F.V., Wu, S.P., Wang, W., Nanda, V., Saven, J.G., Therien, M.J., and DeGrado, W.F. (2005). Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. J. Am. Chem. Soc. *127*, 1346–1347.

Collins, C.H., Leadbetter, J.R., and Arnold, F.H. (2006). Dual selection enhances the signaling specificity of a variant of the quorum-sensing transcriptional activator LuxR. Nat. Biotechnol. *24*, 708–712.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., and Players, F. (2010). Predicting protein structures with a multiplayer online game. Nature *466*, 756–760.

Correll, C.C., Beneken, J., Plantinga, M.J., Lubbers, M., and Chan, Y.L. (2003). The common and the distinctive features of the bulged-G motif based on a 1.04 A resolution RNA structure. Nucleic Acids Res. *31*, 6806–6818.

Cozzetto, D., Giorgetti, A., Raimondo, D., and Tramontano, A. (2008). The evaluation of protein structure prediction results. Mol. Biotechnol. *39*, 1–8.

Das, R. (2011). Four small puzzles that Rosetta doesn't solve. PLoS ONE *6*, e20044.

Das, R., and Baker, D. (2007). Automated de novo prediction of native-like RNA tertiary structures. Proc. Natl. Acad. Sci. USA *104*, 14664–14669.

DeLano, W.L. (2002). The PyMol molecular graphics systems (Palo Alto, CA, USA: DeLano Scientific).

DiMaio, F., Tyka, M.D., Baker, M.L., Chiu, W., and Baker, D. (2009). Refinement of protein structures into low-resolution density maps using rosetta. J. Mol. Biol. *392*, 181–190.

DiMaio, F., Terwilliger, T.C., Read, R.J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H.L., et al. (2011). Improved molecular replacement by density- and energy-guided protein structure optimization. Nature *473*, 540–543.

Dobson, C.M. (2003). Protein folding and misfolding. Nature *426*, 884–890.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Eichner, T., and Radford, S.E. (2011). A diversity of assembly mechanisms of a generic amyloid fold. Mol. Cell *43*, 8–18.

Eisenmesser, E.Z., Millet, O., Labeikovsky, W., Korzhnev, D.M., Wolf-Watz, M., Bosco, D.A., Skalicky, J.J., Kay, L.E., and Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. Nature *438*, 117–121.

Ekins, S. (2006). Success stories of computer-aided design. In Computer applications in pharmaceutical research and development (Hoboken, NJ: John Wiley & Sons).

Epstein, C.J., and Anfinsen, C.B. (1962). The reversible reduction of disulfide bonds in trypsin and ribonuclease coupled to carboxymethyl cellulose. J. Biol. Chem. *237*, 2175–2179.

Feher, M., and Schmidt, J.M. (2003). Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. J. Chem. Inf. Comput. Sci. *43*, 218–227.

Fleishman, S.J., Corn, J.E., Strauch, E.M., Whitehead, T.A., Karanicolas, J., and Baker, D. (2011a). Hotspot-centric de novo design of protein binders. J. Mol. Biol. *413*, 1047–1062.

Fleishman, S.J., Khare, S.D., Koga, N., and Baker, D. (2011b). Restricted side-chain plasticity in the structures of native proteins and complexes. Protein Sci. *20*, 753–757.

Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.-M., Wilson, I.A., and Baker, D. (2011c). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science *332*, 816–821.

Fleishman, S.J., Whitehead, T.A., Strauch, E.M., Corn, J.E., Qin, S., Zhou, H.X., Mitchell, J.C., Demerdash, O.N., Takeda-Shitaka, M., Terashi, G., et al. (2011d). Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J. Mol. Biol. *414*, 289–302.

Grigoryan, G., Reinke, A.W., and Keating, A.E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature *458*, 859–864.

Havranek, J.J., and Harbury, P.B. (2003). Automated design of specificity in molecular recognition. Nat. Struct. Biol. *10*, 45–52.

Havranek, J.J., Duarte, C.M., and Baker, D. (2004). A simple physical model for the prediction and design of protein-DNA interactions. J. Mol. Biol. *344*, 59–70.

Huang, P.S., Love, J.J., and Mayo, S.L. (2007). A de novo designed protein protein interface. Protein Sci. *16*, 2770–2774.

Ingram, V.M. (1956). A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. Nature *178*, 792–794.

Jha, R.K., Leaver-Fay, A., Yin, S., Wu, Y., Butterfoss, G.L., Szyperski, T., Dokholyan, N.V., and Kuhlman, B. (2010). Computational design of a PAK1 binding protein. J. Mol. Biol. *400*, 257–270.

Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas, C.F., III., et al. (2008). De novo computational design of retro-aldol enzymes. Science *319*, 1387–1391.

Jones, A.K., Lichtenstein, B.R., Dutta, A., Gordon, G., and Dutton, P.L. (2007). Synthetic hydrogenases: incorporation of an iron carbonyl thiolate into a designed peptide. J. Am. Chem. Soc. *129*, 14844–14845.

Karanicolas, J., Corn, J.E., Chen, I., Joachimiak, L.A., Dym, O., Peck, S.H., Albeck, S., Unger, T., Hu, W., Liu, G., et al. (2011). A de novo protein binding pair by computational design and directed evolution. Mol. Cell *42*, 250–260.

Khersonsky, O., Röthlisberger, D., Wollacott, A.M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K.N., Baker, D., and Tawfik, D.S. (2011). Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. J. Mol. Biol. *407*, 391–412.

Kinch, L., Yong Shi, S., Cong, Q., Cheng, H., Liao, Y., and Grishin, N.V. (2011). CASP9 assessment of free modeling target predictions. Proteins *79* (*Suppl 10*), 59–73.

Koder, R.L., Anderson, J.L., Solomon, L.A., Reddy, K.S., Moser, C.C., and Dutton, P.L. (2009). Design and engineering of an O(2) transport protein. Nature *458*, 305–309.

Korendovych, I.V., Kulp, D.W., Wu, Y., Cheng, H., Roder, H., and DeGrado, W.F. (2011). Design of a switchable eliminase. Proc. Natl. Acad. Sci. USA *108*, 6823–6827.

Korzhnev, D.M., Religa, T.L., Banachewicz, W., Fersht, A.R., and Kay, L.E. (2010). A transient and low-populated protein-folding intermediate at atomic resolution. Science *329*, 1312–1316.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science *302*, 1364–1368.

Kuriyan, J., and Eisenberg, D. (2007). The origin of protein interactions and allostery in colocalization. Nature *450*, 983–990.

Levin, K.B., Dym, O., Albeck, S., Magdassi, S., Keeble, A.H., Kleanthous, C., and Tawfik, D.S. (2009). Following evolutionary paths to protein-protein interactions with high affinity and selectivity. Nat. Struct. Mol. Biol. *16*, 1049–1055.

Levinthal, C. (1968). Are there pathways for protein folding? J. Chim. Phys. *65*, 44–45.

Mandell, D.J., Coutsias, E.A., and Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat. Methods *6*, 551–552.

Meenan, N.A., Sharma, A., Fleishman, S.J., Macdonald, C.J., Morel, B., Boetzel, R., Moore, G.R., Baker, D., and Kleanthous, C. (2010). The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proc. Natl. Acad. Sci. USA *107*, 10080–10085.

Miller, M.E., and Cross, F.R. (2001). Cyclin specificity: how many wheels do you need on a unicycle? J. Cell Sci. *114*, 1811–1820.

Mizianty, M.J., Zhang, T., Xue, B., Zhou, Y., Dunker, A.K., Uversky, V.N., and Kurgan, L. (2011). In-silico prediction of disorder content using hybrid sequence representation. BMC Bioinformatics *12*, 245.

Moran, U., Phillips, R., and Milo, R. (2010). SnapShot: Key numbers in biology. Cell *141*, 1262–1262.e1.

Morgan, D.O. (2007). The Cell Cycle: Principles of Control (London: New Science Press Ltd).

Murata, T., Yamato, I., Kakinuma, Y., Leslie, A.G., and Walker, J.E. (2005). Structure of the rotor of the V-Type Na+-ATPase from Enterococcus hirae. Science *308*, 654–659.

Newman, J.R., and Keating, A.E. (2003). Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science *300*, 2097–2101.

Otzen, D.E., Kristensen, O., Proctor, M., and Oliveberg, M. (1999). Structural changes in the transition state of protein folding: alternative interpretations of curved chevron plots. Biochemistry *38*, 6499–6511.

Pauling, L. (1948). Nature of forces between large molecules of biological interest. Nature *161*, 707–709.

Pauling, L., Itano, H.A., et al. (1949). Sickle cell anemia a molecular disease. Science *110*, 543–548.

Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science *300*, 445–452.

Perutz, M.F., Wilkinson, A.J., Paoli, M., and Dodson, G.G. (1998). The stereochemical mechanism of the cooperative effects in hemoglobin revisited. Annu. Rev. Biophys. Biomol. Struct. *27*, 1–34.

Raman, S., Lange, O.F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T.A., Eletsky, A., Szyperski, T., et al. (2010). NMR structure determination for larger proteins using backbone-only data. Science *327*, 1014–1018.

Richardson, J.S., and Richardson, D.C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc. Natl. Acad. Sci. USA *99*, 2754–2759.

Rossmann, M.G. (1989). The canyon hypothesis. Hiding the host cell receptor attachment site on a viral surface from immune surveillance. J. Biol. Chem. *264*, 14587–14590.

Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., et al. (2008). Kemp elimination catalysts by computational enzyme design. Nature *453*, 190–195.

Samish, I., MacDermaid, C.M., Perez-Aguilar, J.M., and Saven, J.G. (2011). Theoretical and computational protein design. Annu. Rev. Phys. Chem. *62*, 129–149.

Scalley-Kim, M., and Baker, D. (2004). Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection. J. Mol. Biol. *338*, 573–583.

Scalley-Kim, M., Minard, P., and Baker, D. (2003). Low free energy cost of very long loop insertions in proteins. Protein Sci. *12*, 197–206.

Schindler, T., Sicheri, F., Pico, A., Gazit, A., Levitzki, A., and Kuriyan, J. (1999). Crystal structure of Hck in complex with a Src family-selective tyrosine kinase inhibitor. Mol. Cell *3*, 639–648.

Schröder, G.F., Levitt, M., and Brunger, A.T. (2010). Super-resolution biomolecular crystallography with low-resolution data. Nature *464*, 1218–1222.

Scott, J.D., and Pawson, T. (2009). Cell signaling in space and time: where proteins come together and when they're apart. Science *326*, 1220–1224.

Shan, Y., Kim, E.T., Eastwood, M.P., Dror, R.O., Seeliger, M.A., and Shaw, D.E. (2011). How does a drug molecule find its target binding site? J. Am. Chem. Soc. *133*, 9181–9483.

Shapiro, L., and Losick, R. (2000). Dynamic spatial regulation in the bacterial cell. Cell *100*, 89–98.

Sharp, K.A., and Honig, B. (1990). Electrostatic interactions in macromolecules: theory and applications. Annu. Rev. Biophys. Biophys. Chem. *19*, 301–332.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., et al. (2008). Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. USA *105*, 4685–4690.

Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St Clair, J.L., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L., et al. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science *329*, 309–313.

Stetefeld, J., Jenny, M., Schulthess, T., Landwehr, R., Engel, J., and Kammerer, R.A. (2000). Crystal structure of a naturally occurring parallel right-handed coiled coil tetramer. Nat. Struct. Biol. *7*, 772–776.

Tokuriki, N., and Tawfik, D.S. (2009). Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. *19*, 596–604.

Tyka, M.D., Keedy, D.A., André, I., Dimaio, F., Song, Y., Richardson, D.C., Richardson, J.S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. *405*, 607–618.

van Meerwijk, J.P., Marguerat, S., Lees, R.K., Germain, R.N., Fowlkes, B.J., and MacDonald, H.R. (1997). Quantitative impact of thymic clonal deletion on the T cell repertoire. J. Exp. Med. *185*, 377–383.

Wang, L., Althoff, E.A., Bolduc, J., Jiang, L., Moody, J., Lassila, J.K., Giger, L., Hilvert, D., Stoddard, B., and Baker, D. (2011). Structural analyses of covalent enzyme-substrate analog complexes reveal the strengths and limitations of de novo enzyme design. J. Mol Biol. *415*, 615–625.

Watters, A.L., Deka, P., Corrent, C., Callender, D., Varani, G., Sosnick, T., and Baker, D. (2007). The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. Cell *128*, 613–624.

Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. *293*, 321–331.

Wright, C.F., Teichmann, S.A., Clarke, J., and Dobson, C.M. (2005). The importance of sequence diversity in the aggregation and evolution of proteins. Nature *438*, 878–881.

Xu, G., Wang, W., Groves, J.T., and Hecht, M.H. (2001). Self-assembled monolayers from a designed combinatorial library of de novo beta-sheet proteins. Proc. Natl. Acad. Sci. USA *98*, 3652–3657.

Yin, H., Slusky, J.S., Berger, B.W., Walters, R.S., Vilaire, G., Litvinov, R.I., Lear, J.D., Caputo, G.A., Bennett, J.S., and DeGrado, W.F. (2007). Computational design of peptides that target transmembrane helices. Science *315*, 1817–1822.

Zanghellini, A., Jiang, L., Wollacott, A.M., Cheng, G., Meiler, J., Althoff, E.A., Röthlisberger, D., and Baker, D. (2006). New algorithms and an in silico benchmark for computational enzyme design. Protein Sci. *15*, 2785–2794.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature *426*, 676–680.