Biochimie 119 (2015) 218–230

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Review

Evolution, energy landscapes and the paradoxes of protein folding

Peter G. Wolynes

Department of Chemistry and Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, USA

ARTICLE INFO

Article history:
Received 15 September 2014
Accepted 11 December 2014
Available online 18 December 2014

Keywords:
Folding landscape
Natural selection
Structure prediction

ABSTRACT

Protein folding has been viewed as a difficult problem of molecular self-organization. The search problem involved in folding however has been simplified through the evolution of folding energy landscapes that are funneled. The funnel hypothesis can be quantified using energy landscape theory based on the minimal frustration principle. Strong quantitative predictions that follow from energy landscape theory have been widely confirmed both through laboratory folding experiments and from detailed simulations. Energy landscape ideas also have allowed successful protein structure prediction algorithms to be developed.

The selection constraint of having funneled folding landscapes has left its imprint on the sequences of existing protein structural families. Quantitative analysis of co-evolution patterns allows us to infer the statistical characteristics of the folding landscape. These turn out to be consistent with what has been obtained from laboratory physicochemical folding experiments signaling a beautiful confluence of genomics and chemical physics.

Contents

1. Introduction

Paradoxically, protein folding has turned out to be easy. How? Why? What's paradoxical? This review is aimed at answering these questions, beginning with the last one [1]. Ever since Anfinsen's groundbreaking experiments, understanding the spontaneous nature of protein folding has been widely viewed as being a difficult problem [2]. Delbrück is quoted by Gunther Stent [3] as saying about protein folding "…the reduction in dimensionality from three dimensional continuous to one dimensional discrete in the genesis of proteins is a new law of physics and one nobody could have pulled out of quantum mechanics without first seeing it in operation". According to Stent, after saying this, Delbrück also immediately quoted Bohr as telling about a man who, upon seeing a magician saw a woman in half, shouts out "It's all a swindle". In a similar spirit, I think it is fair to say that a majority of people even in the last decade have looked with skepticism at claims that protein folding was becoming understood [4].

Yet today, a variety of computer algorithms can indeed translate, for the simpler systems, one dimensional sequence data into three dimensional structure albeit at moderate resolution [5–7]. The easiest to use algorithms rely ultimately on recoding existing

E-mail address: pwolynes@rice.edu.

http://dx.doi.org/10.1016/j.biochi.2014.12.007
0300-9084/

biological information into the form of an energy landscape for computer simulations [5] and thus ultimately these algorithms rely on having “seen it in operation” rather than deriving their results directly from quantum mechanics. Delbrück's demand to pull folding out of quantum mechanics, has, however, almost been satisfied by using very powerful computers to simulate fully atomistic models, the forces in which are only lightly parametrized by protein structural data [7]. So we see folding proteins really is easier than many people (including me!) thought. Nevertheless, there is something at least a bit strange, perhaps even paradoxical about this outcome, because theory has shown that protein folding indeed could have been much harder, almost as hard as Levinthal envisioned in his famous paradox in which it was pointed out that it would take a cosmological time to fold a protein by searching through all its possible structures [8].

Anfinsen's refolding experiment itself suggests that folding can be thought of as the search for a minimum free energy arrangement in space of a heteropolymeric chain of residues. The solvent averaged interactions between amino acids depend on the residue types; some residues will interact more strongly with particular other ones than they interact with others. Thus folding is a kind of matching problem: we must find which amino acid will finally make contact with which other one in the native structure in order to reach the lowest energy three dimensional structure. The matching problem is in general NP (non-polynomial) complete [9]. To say a problem is NP complete is a mathematician's way of saying there exist instances of such a problem (i.e. there should be certain sequences) for which no known algorithm could find the solution in a time scaling as a polynomial in the size of the problem. The exhaustive search through minima envisioned in Levinthal's paradox scales exponentially with protein length but even an algorithm that takes a time exponentially scaling with a lower power of length (as happens in some theories based on capillarity ideas [10,11]), also would satisfy strict NP completeness so Levinthal's estimate is certainly an exaggeration. Of course the key words in the mathematical statement are “there exist instances”. Not all instances of matching problems are expected to be equally hard. Some instances of matching problems can in fact be quite easy. A familiar biological example where search difficulty varies with the task is provided by the matching of base pairs in nucleic acid double helices as they assemble. The minimum energy matching is easily found if the two strands are exactly complementary, as in most nuclear DNA giving rise to the standard rules of replication. Finding a match is also easy if only a small fraction of the bases are not conjugate to their partners just as happens when we try to fish out functionally related sequences using primers or use modified primers for site directed mutagenesis via PCR. Comparing two divergent sequences for homology is easy, the difficulty scaling only with a polynomial in the chain length [12]. Most functional RNA sequences have easy matches also, if no pseudo-knots are formed in forming their secondary structures, but in principle the problem of pairing the bases in an arbitrary nucleic acid strand with itself can be quite difficult. This difficulty is also reflected in the cell where some messenger RNA's turn out actually to be metastable [13] – they first find an active conformation transiently but later rearrange to an inactive but more stable form after sufficient protein has been translated by the mRNA so that the messenger is no longer needed. In contrast, it appears from the Anfinsen experiment, along with its thousands of descendants, that monomeric protein folding is usually thermodynamically controlled, with functional metastability being the exception [14] rather than the rule.

Are the easy instances of matching an amino acid sequence to itself in order to fold into a three-dimensional structure exceptional or are the tough cases the weird examples? This, of course, may depend on the details of the interactions, but the simplest arguments suggest that for a sufficiently long chain, among those chosen at random, the easy examples should be the rarities. Bryngelson and Wolynes suggested that the energy landscape of a typical random amino acid sequence would be rough, riddled with deep metastable minima of widely differing structures and resemble the rugged landscape of a glass [15,16]. They analyzed this situation with the random energy model [17]. The idea was that in a sufficiently long random sequence conflicts between different choices of the favorable interactions would inevitably arise, a phenomenon known as frustration [18]. The low energy states of such a system are all highly compromised, satisfying some interactions very stably, perhaps, but with other interactions remaining unsatisfied and in conflict giving rise to many near degenerate configurations. The mathematical validity of this idea that the energy landscape of a completely random heteropolymer would be rugged, has been buttressed both by more sophisticated statistical mechanics using the replica method [19,20] and by numerous computer simulation studies on highly simplified models that capture the essential features of self-matching chains [21–23].

So proteins seem to be the rare, easy instances of energy minimization. What is it that makes folding protein-like sequences easy? What is it that allows proteins to pair residues properly through Brownian motion in order to find their lowest free energy

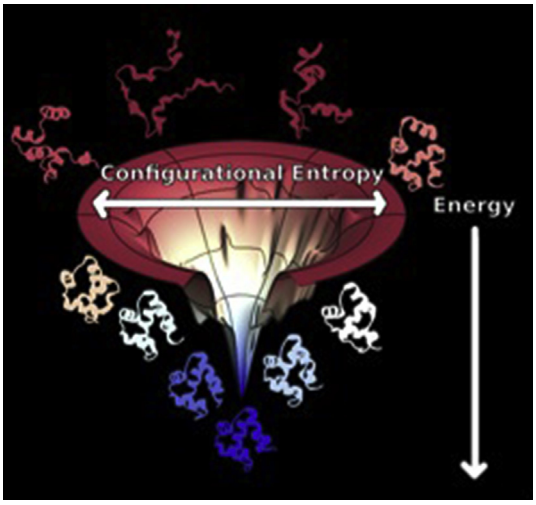

Fig. 1. The funnel diagram. A schematic diagram of the energy landscape of a protein, here illustrated with the PDZ domain whose native structure is shown at the bottom of the funnel. The energy landscape exists in a very high dimensional space. The diagram can only give a sense of this through its representation of two dimensions. The radial coordinate measures the configurational entropy which decreases as the protein takes on a more fully folded structure. The energy of individual configurations is represented by the vertical axis. The values of the energy indicated on this axis are strongly correlated with the fraction of native structures that has formed which is often measured by the fraction of correct native-like contacts called Q. Q also typically increases as the structures descend in the funnel. The energy and entropy oppose each other so that at high temperature the protein is found in an ensemble of states near the top of the funnel. Structures of denatured configurations thus are shown near the top of the funnel. At low temperature, in contrast, an ensemble clustered around the native structures becomes thermally occupied at the bottom of the funnel. The imperfect matching of entropy and energy leads typically to a free energy barrier that separates these two ensembles of states. Surmounting this barrier limits the folding rate. The small mini-funnels on the sides of the funnel represent trap states. These traps typically possess some native structure but also they contain energetically favorable alternative non-native contacts. Because the non-native contacts are not consistent with each other, rarely are such mini-funnels competitive in an energetic sense with the native basin. The stability of non-native interactions in any one of these traps is an unusual rare accident while the interactions that are formed in native structure have evolved in order for the individual natively folded structure to be especially stable.

states without the molecule getting trapped in metastable states of wildly different structure as sometimes happens with RNA? The answer is that natural proteins correspond with those special sequences that have been selected to have more consistently stabilizing interactions throughout the natively structured molecule than does a typical sequence which inevitably makes many compromises in its low energy structures. Proteins are not so "frustrated" as a usual heteropolymer sequence is – we say they are "minimally frustrated" [15].

Bryngelson and Wolynes showed that, in contrast to random sequences, such unusual, easy-to-fold minimally frustrated sequences have two potential structural thermodynamic phase transitions. One of the transitions is the folding transition to a well-organized highly stable structure with perhaps some defects. Nevertheless if this correct folding would be prevented from happening, a protein still would settle down to an ensemble of structurally disparate low energy states when cooled below a characteristic temperature T_g. We would find these misfolded states in the low energy trapped ensemble to have similar energies to each other, but to be structurally quite distinct from each other. Finding one given energy minimum by a simulated annealing protocol would yield a structural prediction for that sequence, but that prediction would be quite unreliable until all the low energy arrangements of the chain were found and examined individually. Which structure out of this wide ranging ensemble prevails for a particular molecule would depend on the history of the protein synthesis and annealing, i.e. the protein folding outcome would be kinetically controlled not thermodynamically controlled. This situation is much like what happens for a liquid which becomes a glass when it is cooled if the liquid fails to crystalize [24]. The detailed atomic structure of the resulting glass varies from sample to sample but the macroscopically averaged properties of the glass, like its energy, are nearly the same from sample to sample.

For easy folding sequences, i.e. minimally frustrated sequences, the folding transition at T_F resembles a crystallization transition while the history dependent transition at T_g for poor folders is like the glass transition for a liquid. Protein folding in this picture resembles nucleation of a crystal from a small fluid drop. The kinetics of nucleation is impeded by the multiplicity of trapped states that become prominent near the glass transition temperature of the fluid, T_g. The driving force to the native structure is impeded in its effectiveness by the energetic ruggedness reflected in T_g [16,20,21].

Nucleation, folding and self assembly differ from the more familiar multi-step chemical transformations of organic synthesis or intermediary metabolism in that collective self assembly has a multiplicity of possible mechanisms–no specific sequence of steps is absolutely needed in order to achieve successful folding. Nevertheless a dominant small set of pathways can emerge as the most important ways a particular sequence assembles under some thermodynamic conditions [25,26]. The dominant routes, as well as the activation barrier determining the folding rate, depend on the way in which the loss of entropy upon ordering the chain is compensated by the additional stabilization energy that properly assembled parts of the chain achieve when the energy is compared with the weaker stabilization that occurs when the chain transiently samples improperly folded states [27]. The landscapes for protein folding or for crystallization can be described as rugged funnels [28]. See Fig. 1.

The rate of attempting to escape from misfolded traps decreases as the stability of these traps increases and thus the kinetics of search depends on the glass transition temperature T_g. Easy-to-fold sequences can fold without getting trapped because they have a high T_F to T_g ratio. This ratio, in turn, depends on the comparison of the energy of the fully folded state E_F and the typical stabilization energy of a trap E_g which monotonically increases as T_g increases [29].

The minimal frustration or "funnel" [21,28] scenario as the explanation for the ease of protein folding has been confirmed, in concept, by numerous computer studies that simulate stylized stripped down models of self-associating polymers. The most comprehensive such studies simulate heteropolymers on lattices [30,31] where the exact enumeration of possibilities ensures that a complete survey of the landscape can be made so that even for bad folders the ground state can be found and certified as being correct. Other models that realistically describe the protein backbone geometry so that they give structures more easily recognized as being real proteins concur in supporting the idea that it is the magnitude of the T_F/T_g ratio that determines the ease of folding to a first approximation [32–34]. A discussion of these studies can be found in many early reviews [21,22,35]. Yet the funnel story has seemed too simple to many people [36–38]. They ask "Could that really be all there is to it?" Some of the skepticism is perhaps justified because the success of the funnel landscape picture raises other questions: How does the energy landscape theory apply to real proteins (in a quantitative sense)? Exactly how easy is it to fold the proteins of Nature? These questions form the impetus for much of the last two decades' work on protein folding theory.

In this paper I plan to first touch on the experimental evidence that proteins are indeed minimally frustrated polymers, i.e. that the energy landscape of naturally occurring proteins is actually a rugged funnel as hypothesized. The evidence for the funnel idea is actually quite overwhelming in volume. It is much too extensive to do justice to it here. I therefore refer the reader to an earlier comprehensive review on the experimental survey of protein folding landscapes [21] using the tools of protein engineering, fast kinetics, theory and simulation. That review documents the usefulness of the rugged funnel model in understanding the kinetics of a range of systems. Atomistic simulations done recently [7] also support the funneled energy landscape picture for a large number of examples, as I will describe below.

Having achieved minimal frustration is the "swindle" (to use Bohr's term) behind the ease of protein folding. Minimal frustration has apparently come about as the result of evolution and is encoded in existing protein structures and sequences. The mathematical instantiation of the minimal frustration principle via the T_F/T_g ratio can therefore be used as a learning tool for inferring the nature of forces within and between protein molecules – "reverse engineering" the folding problem [29,39–42]. This strategy has led to predictive algorithms for determining protein tertiary structure from sequence alone. Again the topic of protein structure prediction algorithms based on energy landscape theory has also recently been reviewed by us [5], so I will just touch on this question here. These practical successes, in my view, also should buttress our confidence in the main ideas of folding energy landscape theory.

The bulk of this review will focus on understanding precisely how easily do proteins fold – that is, what have we learned through trying to quantify the smoothness of the landscape by comparing the strength of the guiding forces to the ruggedness of the landscape. To a first approximation this quantification may be summarized in a single number, again, the T_F/T_g ratio. Several estimates of T_F/T_g have been made in different ways over the years, using laboratory data on the thermodynamics and kinetics of folding as well as experimental observations on the reconfigurational motions and residual structure in molten globules as input [43–45]. I will review those arguments.

As I have mentioned, the ease of protein folding must ultimately be the result of evolution. The "swindle" of easy protein folding was set up by natural history. Can we also check the idea of the evolutionary origin of minimal frustration in a quantitative sense? Is there a sign in the sequence data alone that landscapes have been funneled through natural selection? Recently it has become

possible to determine the degree of funneling of entire protein families as measured by their typical T_F/T_g ratio by carrying out information theoretic analyses of large families of protein sequences and comparing the detectable coevolutionary patterns to the sequence patterns that would be expected based on the physics contained in energy landscape derived transferable force fields that are able successfully to predict tertiary folds from individual protein sequences [46]. I will review the resulting confluence of the evolutionary genomic and the physics perspectives on the energy landscape. One finds a picture of the folding energy landscape from the genomic data alone that matches pretty well with what has been previously deduced from purely physico-chemical ideas and experiments. Finally I will return to the paradoxes of easy protein folding commenting on my current view of them.

2. General evidence that folding landscapes are funneled

The experimental study of protein structure and folding mechanism has yielded many general sorts of observations about the ways proteins fold. These patterns suggest protein energy landscapes, in the main, are funneled and that proteins are minimally frustrated heteropolymers. In addition, the details of the folding mechanisms of many specific proteins have been reproduced using highly idealized models of protein energy landscapes that have no frustration at all: so called "structure based" [47,48] or Gō models [49]. These models ignore any possibility of stabilizing misfolding interactions that would lead to kinetic traps nevertheless they give solid and surprisingly accurate predictions. The landscapes of such structure based models can be thought of as being "perfect funnels". Like the perfect gas model in elementary chemistry, the perfect funnel model, while working reasonably well, does not always have quantitative perfection. Major qualitative deviations from perfect funnel predictions are, however, quite unusual. Apparent failures of the funnel model have sometimes led us to think outside the box and to our delight have turned out to confirm that funneling is still preserved for the greater part of the folding process [50]. Many seeming exceptions to perfect funnel prediction can be accounted for by using models that have small perturbations from perfectly funneled landscapes. These models take into account weak trapping owing to residual frustration [51] or energetic heterogeneity [52] or other sources of landscape degeneracy such as the symmetry of designed protein sequences [53] or the symmetry intrinsic to oligomerization through domain swapping [54]. The success of the general predictions about folding that have come from funneled, minimally frustrated landscapes along with the multitude of detailed specific successes on many systems should give us a lot of confidence that the minimal frustration principle has more than a core of truth. The fact that we can understand and quantify deviations from the model itself is also very encouraging that we have begun to understand protein folding.

What general observations support the minimal frustration hypothesis? The first observation is the well-known robustness of protein structure to changes in sequence. Single mutations, and even sometimes dramatic overall changes in sequence like those found in "twilight zone" homologues still give remarkably similar folded native structures [55]. It is commonplace now for protein engineers to assume that making only a single mutation in a protein will leave its structure intact, once that protein has successfully folded, under the right solvent conditions. This robustness is not what would be expected for folding on a typical highly frustrated rugged landscape where an ensemble of structurally disparate competing low energy structures is the norm. Yes, a single mutation sometimes does cause a protein to unfold but this unfolding reflects an "unfair" competition between one single folded structure against myriads of unfolded alternatives, not the competition between specific structured individual alternatives. Only recently have sequences been found that adopt fixed structures but that also change drastically their structure when individual mutations are made [56]. These examples typically involve also changing of the oligomeric state of the proteins, however. Such sequences might be "missing links" in protein evolution.

For a minimally frustrated protein, the overall stability typically does not come from just a few residues but rather is spread throughout the structure. Owing to this delocalized character of the guiding forces to the native state, proteins with widely different sequences not only have similar final structures but often follow the same basic folding mechanism and sometimes have even quantitatively similar rates when tuned to the same stability [57,58]. This generalization that "topology" controls the folding mechanism (as well as its exceptions) has been documented for several protein families notably by Jane Clarke [59] but also by numerous others [60].

The most general powerful consequence of proteins having minimally frustrated landscapes is that folding kinetics almost always changes smoothly and monotonically as protein stability is changed. Studying the effect of such changes on the kinetics is the basis of modern systematic mechanistic folding studies [26,61]. The stability can be changed by changing temperature (up or down!), changing solvent conditions or by making site mutations. The smooth variation of rate with stability is expected for landscapes that are dominated by forming successively the contacts found finally in the native structure [62]. In contrast such a smooth variation is quite unexpected if wildly different structures involving specific non-native structures were to play a significant role in the folding mechanism as would be expected to happen for typical random sequences. Such unexpected variations do sometimes appear. The ROP dimer system seemingly violates the funnel concept because mutations were found that changed both the folding and unfolding rates without changing the protein stability [63]. To accommodate this anomaly, energy landscape thinking however led us to suggest the structure of the dimer had, in fact, been changed upon mutation [50]. It was later confirmed by single molecule FRET experiments that the ROP dimer indeed had, owing to its high symmetry, two different but similarly stable 4 helix bundle structures sharing a common set of interactions [64,65]. One form is dominant for the original protein but the other form dominates for the mutant. In the absence of the symmetry of the dimer system, only a few proteins show any evidence at all of degenerate trap states in which non-native contacts play a role in the key intermediates, as we would find for a random sequence. The most notorious of these strong exceptions to the perfect funnel model is the folding of Im7, a molecule that functions by binding a toxin [66]. It turns out to be "the exception that proves the rule". Computational analysis of the Im7 landscape shows that this rather small protein possesses a large region that is frustrated in the monomer but that takes part in the binding function of the molecule where the additional interactions with its target form that are finally favorable and the frustration is relieved upon binding [51]. This frustrated part of the molecule leads to a re-packing in an intermediate when Im7 is required to fold on its own without having a partner present [67].

A more detailed test of the ideal funnel model comes from making quantitative predictions using perfectly funneled landscape calculations and comparing them with accurate experimental measurements of kinetic changes on a residue by residue basis [68–72]. The ideal funnel models allow us to predict the fraction of the native stability change that becomes translated into the rate change – the so-called Φ value analysis. The Φ value essentially gives the fraction of the time that a residue participates energetically in the crucial transition states for folding [62]. Very good agreement between predictions of perfect funnel models and experiment has been found in several cases for extensive sets of Φ-

values measured for many systems. These include the Φ values for the folding of azurin, UIA, λ repressor and very large lactose repressor module [73].

Fully atomistic simulations also converge on confirming the validity of the funnel concept for the folding of natural proteins at least when correctly tuned all-atom force fields are used. Many early simulation studies did find some evidence for non-native intermediates in natural proteins [74–76]. It was suggested that these non-native traps simply had been missed by experimenters. Certainly some misfolded intermediates may be real and probably do exist in the laboratory but a poor force field also typically leads to transient misfolding (since the protein did not evolve with that man-made force field!) and this seems to be the problem in many of the early simulations of folding. In any event the importance of misfolded intermediates in the laboratory would be a quantitative question. Non-native traps may also be relatively more important in the smaller protein systems that first became accessible to simulation than they are in the more typical large proteins. In large, more slowly folding proteins, non-native traps would give bigger barriers and thus they would interfere much more dramatically with folding *in vivo*. This suggests that selection against misfolds should be stronger for longer proteins than shorter ones. As force fields used in the simulations have improved and the computational capabilities have also advanced so that the bigger systems can be studied [7], it has now become clear that the finally formed native contacts do, in fact, generally provide the dominant interactions in guiding the folding process [77,78]. Other non-native contacts may help collapse and also provide an overall frictional influence on the rate as expected by landscape theory [16] but they do not greatly modify the sequence of events. Eaton et al. have analyzed the heroic simulations of the Shaw group on a large number of systems. Their analysis shows that only for the artificial designed protein αD3 can one find evidence for specific non-native interactions playing any significant role in the dominant folding paths. Indeed even for αD3 the non-native interactions that form correspond to a simple sliding of helices upon each other. Such symmetrically related interactions would not be immediately apparent to the casual observer.

So proteins, in the main are minimally frustrated. Quantitatively, how much frustration is there?

3. How rugged is the folding landscape – physicochemical approach

The first attempt to quantify the energy landscape of real proteins through the T_F/T_g ratio was made by Onuchic et al. [43]. Their goal was to see whether the then existing toy model simulations of protein folding using simplified lattice models could be used to think about real proteins. The earlier lattice calculations already had dovetailed with the analytical energy landscape ideas of Bryngelson and Wolynes [15,16] that looked at the folding process of a small protein as a random walk of an order parameter that measures the native-like character of a polymer configuration. Folding is a biased random walk because entropy favors the myriad of polymer configurations unconstrained by similarity to the native state while energy, on the other hand, favors more native-like configurations. Energy and entropy combine to give a free energy that depends on the order parameter. The gradient of the free energy then describes the bias of the walk towards folding or unfolding and depends on the temperature. The diffusion rate for the order parameter depends on the energetic ruggedness, slow rates of diffusion corresponding to rugged glassy landscapes, fast rates to smoother landscapes. The resulting description via a Kramers-like diffusion equation was shown to describe the kinetics of lattice models quite well [16,79]. In recent years the biased diffusion model has also been shown to describe the dynamics of more realistic off-lattice models [80] and has been used directly to interpret experimental data [81–83].

As we see, the first key quantity needed to make the connection between landscape theory and real proteins is the entropy change of the polymer when it unfolds. Calorimetry is unfortunately only of modest help in getting that the part of the entropy change relevant to the polymer chain alone because the surrounding water is structured by hydrophobic side chains in the unfolded state. We cannot easily separate out the entropy change of the water environment from the overall calorimetric change. This solvent entropy change is a non-trivial phenomenon, ultimately being the origin of cold denaturation. The folding temperature itself is directly observable so we know the energy loss must exactly compensate the configurational entropy change so $T_F\Delta S = \delta E$. While many folding processes occur directly from a random coil state, the collapsed but disordered state of the protein is not too far away. Apparently proteins are near a triple point between the native, compact globule and random coil phases. The entropy of the collapsed molten globule state is key because ruggedness can, in any event, only arise to a significant extent in the collapsed ensemble. Searching through collapsed states can be slow but searching through the expanded states with few stabilizing contacts is not. Collapsed state ruggedness is what could make reconfigurational search difficult.

Onuchic et al. [43] estimated the configurational entropy of the collapsed, molten globule by noting that its helical content is quite high. Both collapse and helix formation lower the molten globule entropy. The helical entropy loss can be quantitatively estimated by considering the ease of nucleation in the uncollapsed phase as measured by the σ and s parameters of the helix coil transition for uncollapsed peptides. Luthey-Schulten et al. showed how to couple

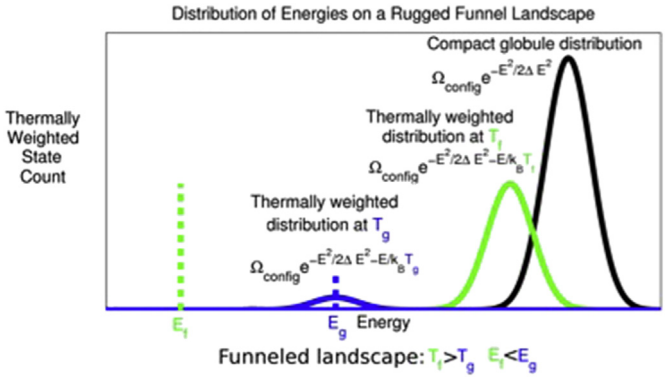

Fig. 2. The Distribution of energies on a funneled folding landscape. A schematic spectrum showing the density of states of a minimally frustrated protein. Compact alternative or decoy states are distributed with a nearly Gaussian distribution of energies through the random addition of conflicting contributions. At a temperature T, the thermally occupied decoys will be diminished in number but they will still have a Gaussian distribution of energies that is shifted downwards. At the glass temperature T_g only a very small number of such trap states would be thermally occupied. The energy E_g at T_g can be estimated from the width of the unbiased decoy distribution ΔE. For a minimally frustrated protein an evolved sequence fits the target structure quite well so that at a folding temperature T_F the Boltzmann weight for the target structure competes with the entire collection of states in the unfolded ensemble. For most random heteropolymers no significant gap in the spectrum exists. As the extra stability of the target δE_F increases, relative to the width of decoy distribution ΔE, the folded structure can be more and more easily picked out from the alternatives. By maximizing $\delta E_F/\Delta E$ over a set of sequences one finds more and more stable "well-designed" sequences. Conversely if many sequence/structure pairs are known the parameters in the energy function can be varied so as to maximize the energy of $\delta E_F/\Delta E$ for the set. The resulting energy function summarizes the structural sequence correlations in the training set. In this way structural data allow us to learn transferable energy functions. Energy landscape theory provides us a theoretical "license to do bioinformatics".

the quasi-one dimensional phase transition of helix formation with the three dimensional collapse transition to relate both the entropy of the globule and its helix content to the hydrogen bond strength [84]. This theory combined with experimental data on the structural content of molten globules gives a configurational entropy of about $0.6k_B$ per amino acid in the molten globule. This estimate is based on the observation that molten globules typically have about 65% α helical structure. This is quite a bit less than the $2.3k_B$ per monomer unit estimated for free chains and indeed is even smaller than the $1.0k_B$ per unit of the toy model lattice polymers.

Saven et al. arrive at similar numbers when the propensity of particular sequences to form helices of peptides or the thermodynamics of signals to start or stop helices is used as input to a theory of the collapsed globule [85].

According to this estimate, entropically a 60 amino acid helical protein maps pretty well on to the popular cubic lattice model of proteins that has 27 beads [43].

While the entropy of the collapsed ensemble follows from the study of residual structure in collapsed states, estimating the ruggedness of the collapsed ensemble requires knowing about the dynamics within molten globules. In 1995 Onuchic et al. [43] used the millisecond time scale value for the reconfiguration time τ_R that could be inferred from existing NMR measurements on molten globules [86] to estimate the ruggedness. Surprisingly it is still not easy to come by better values for molten globule dynamics in the laboratory. According to the Bryngelson–Wolynes random energy model estimate, τ_R is related to the ruggedness described by an energy variance ΔE^2 and a microscopic reconfiguration time τ_0:

$$\tau_R = \tau_0 e^{\Delta E^2/2T^2} \tag{1}$$

Onuchic et al. took a rather short time for $\tau_0 \approx 10^{-9}$ s, so that millisecond rates in the globule give $\Delta E^2/2T^2 \approx 15$.

The ruggedness ΔE^2 also determines the temperature of the glass transition which occurs when the temperature is low enough so that only deepest energy states out of the $e^{Sc/kB}$ possible ones are thermally sampled. This condition on the entropy gives the relation

$$k_B T_g = \sqrt{\Delta E^2/N} \Big/ \sqrt{2S_c/N} \tag{2}$$

where N is the chain length.

Onuchic et al. thus arrived at a value for the glass transition temperature which is much lower than the folding temperature $T_g \approx .6T_F$. This low value is consistent with a moderately strongly funneled landscape corresponding to a roughly three letter protein folding code.

The random energy model also gives an estimate for the energy of the deepest traps that would be thermally sampled at T_g. This competitive trap energy is $E_g = \sqrt{\Delta E^2 Sc}$.

The minimum frustration principle can also be formulated in terms of the condition that the energies δE_F are less than E_g. In this form then we see minimal frustration is equivalent to there being a gap in the spectrum of folded energy states, as in Fig 2. In reality of course, this gap is filled by partially folded configuration in which some of the native contacts have formed that correspond to the structures on the paths for nucleation of properly folded protein structure. Important as they are, these partially folded configurations are sufficiently small in number that they would hardly show up on the plot.

The T_F/T_g ratio of 1.6 inferred by Onuchic et al. clearly indicates the folding landscape is quite smooth. It is probably an underestimate of the smoothness. One reason the estimate may be on the low side is that the polymer chain motions in the globule are now thought to be intrinsically slower than the τ_0 estimate they used – more in the time scale of a microsecond. At the same time, escape from traps may not require rearranging the whole chain. In this regard, Plotkin et al. showed that correlations in the landscape cause only a fraction of the ruggedness to be translated into configurational slowing [87,88]. This error would go in the opposite direction, but correlations in the landscape also change the corresponding T_g estimate, so that correlations in the folding landscape turn out essentially to preserve the estimated T_F/T_g ratio at the value found for the Onuchic et al. model.

Chan has argued that the T_F/T_g ratio is actually much larger than 1.6, perhaps reaching a ratio as large as $T_F/T_g \approx 10$. This very high value would correspond to a very highly funneled energy landscape. His reasoning is based on trying to match the observed high cooperativity of folding [44]. He shows that a heteropolymer model with $T_F/T_g = 1.6$ would give denaturation curves that are much broader than typically seen, suggesting the smoother landscape.

Clementi and Plotkin also suggest that T_F/T_g is greater than 1.6. They base their suggestion on the quantitative success that perfect funnel models have in predicting Φ-values [45]. By adding random non-native interactions, to a heterogeneous structure based model they suggest that the variance of non-native energies must be quite a bit less than what the 1.6 value for the T_F/T_g ratio would indicate; otherwise specific non-native contacts the misfolded state would greatly modify the Φ values. They suggest that T_F/T_g ratios as large as 2 or 3 would better fit the small deviations from a perfect funnel that are in fact experimentally observed.

4. Reverse engineering the folding energy landscape

The minimal frustration hypothesis provides a strategy for "reverse engineering" the folding problem–thereby energy landscape theory provides an organized way of learning the folding landscape by "having seen it in action". The strategy which is rooted in the minimal frustration hypothesis is to tinker with the force field so as to ensure that a transferable form for the energy function, one that can be applied to any sequence, actually leads to minimally frustrated, funneled landscapes for a training set of proteins with natural sequences that we already know fold to specific known structures. If all folding landscapes are funneled and the parameters in the force field are universal and are thus transferrable, the predictive force field that results should work well for proteins not in the training set.

Testing the landscape for proper folding would be a hard reverse engineering task if it had to be done simply by trial and error. There are many parameters in even the simplest coarse-grained molecular force field. Testing even one set of parameters for the force field to see if it yields a funneled folding landscape takes quite a bit of computer time, so combinatorial search through parameter space would be prohibitively expensive. The good news is the minimum frustration principle provides a quick zeroth order check on whether a protein sequence can be folded with a given force field. All we have to do is to check that the energy of the folded state is much lower than the typical value of the energy of lowest thermal accessible misfolded state. How do we find the energy of misfolded states? The latter typical trap energy, E_g as we have seen can be estimated directly from the variance of the energy over a set of candidate decoy structures, the quantity we call the ruggedness ΔE^2. This variance can easily be calculated from a fixed set of representative decoys – this makes it unnecessary to find specifically the absolutely lowest energy misfolded state for a given force field. Finding the actual lowest energy misfolded would be an NP hard task. Sampling to find ΔE^2, however, is much easier and robust.

Taking this strategy one step further mathematically leads to an explicit optimization scheme: Maximize the ratio of $\delta E/\Delta E$. This quantity is monotonically related to the T_F/T_g ratio. If the

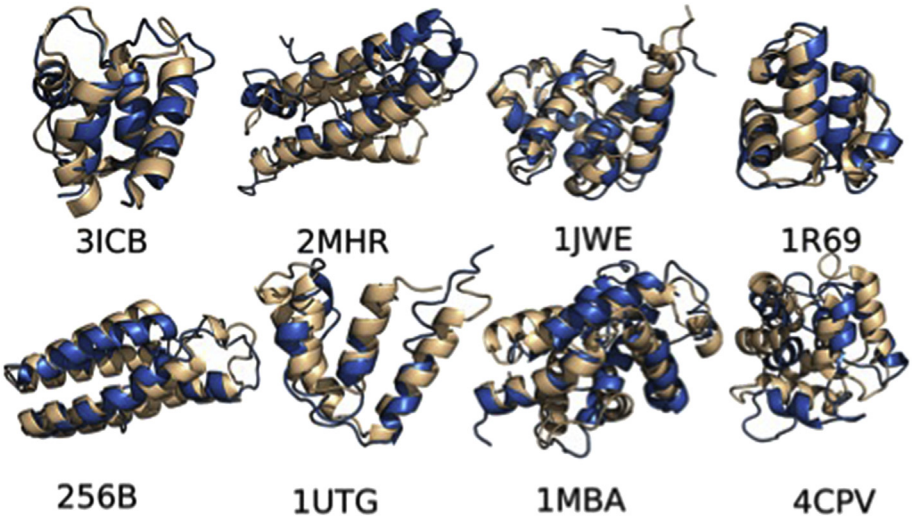

Fig. 3. Predictions of globular protein tertiary structure. A gallery of globular protein structures predicted by the AWSEM energy landscape optimized force field are shown overlapped with the correct x-ray structures. The agreement is comparable to what can be found via homology modeling but no homology information was used in making these predictions. The examples are 3ICB: vitamin D-dependent calcium-binding protein; 2MHR: myohemerythrin; 1JWE: N-terminal domain of *E. coli* DNAB helicase; 1R69: amino-terminal domain of phage 434 repressor; 256bB: cytochrome B562; 1utg: uteroglobin; 1MBA: aplysia limacina myoglobin; and 4CPV: carp parvalbuminCARP. For details please see Ref 90.

parameters in the force field enter into the energy function linearly then optimizing this ratio actually becomes a problem in linear algebra, as first noted by Goldstein et al. [29]. Even better funneled energy landscapes emerge when one employs still more elaborate nonlinear self-consistent optimization schemes in which the decoys are all iteratively re-computed by sampling misfolded structures for the already partially optimized force fields [89,90]. This strategy (illustrated in Fig 2) has over the years led to a series of

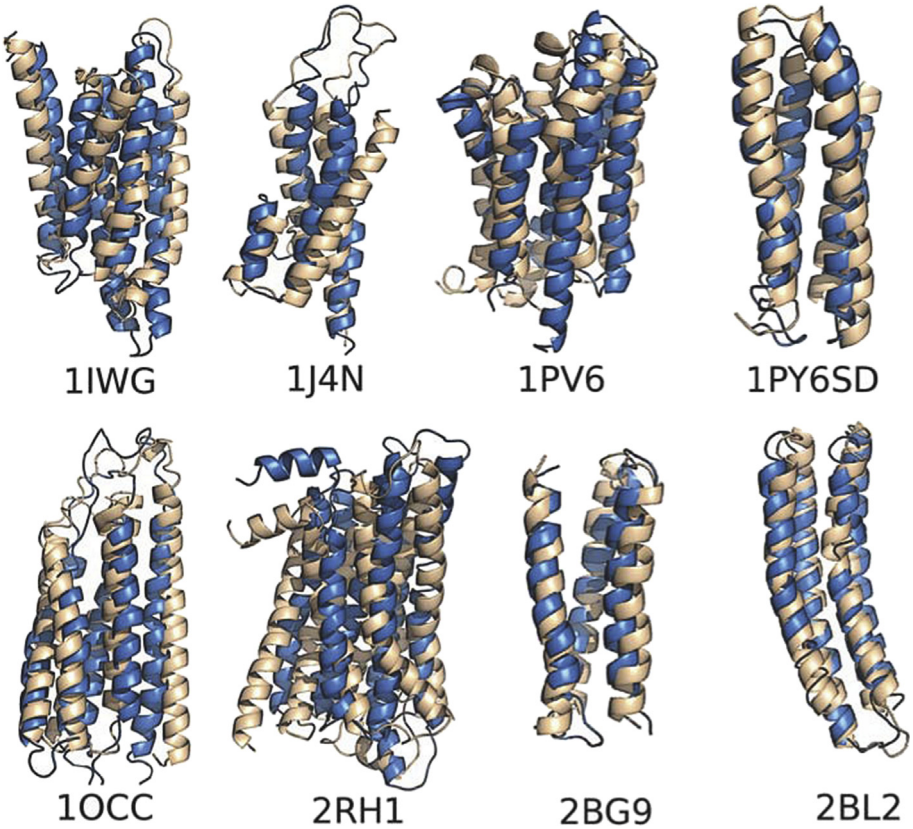

Fig. 4. Predictions of membrane protein structures. A gallery of membrane proteins structure predicted by the AWSEM-Membrane force field shown overlapped with the correct x-ray structures. The examples are 1IWG: subdomain of multidrug efflux transporter; 1J4N: subdomain of aquaporin water channel AQP1; 1PV6: subdomain of lactose permease transporter; 1PY6SD: subdomain of bacteriorhodopsin; 1OCC: subdomain of cytochrome C oxidase aa3; 2RH1: 2-Adrenergic GPCR; 2BG9: subdomain of nicotinic acetylcholine receptor; and 2BL2: subdomain of V-type Na^{+}-ATPase. For details please see Ref 91.

ever improving force fields that now can successfully fold many globular and membrane proteins using their sequence information alone.

An historical survey of the progress made using this energy landscape theory based strategy as well as the details of its underlying mathematics have been recently presented by us [5]. Nowadays, the quality of predictions made without homology information from such force fields is almost as high as what can be achieved by tweaking the structures of known homologs. We can see the comparison of the structures predicted by the latest force field (called AWSEM, the Associative Memory Water Mediated Structure and Energy Model) to the X-ray structures for several globular proteins in Fig. 3. The same strategy has even proved a successful route to developing force fields for studying membrane proteins where the funnel minimal frustration idea has been less tested. Indeed the very basis of the funnel hypothesis may be questioned for membrane proteins because of the kinetic control that may occur *in vivo* through the action of the translocon. Some examples of membrane protein structure prediction [91] using such a transferable force field based on energy landscape reverse engineering are shown in Fig 4. It appears that once the translocon has placed the protein in the membrane spontaneous folding à la Anfinsen ensues.

It is probably not an accident that these force fields which have been reverse engineered using the minimal frustrated landscape idea yield results with a resolution only comparable to what can be obtained using homology models. They are not as precise as fully determined atomistic x-ray structures. This is reasonable because the minimal frustration principle must be a result of evolution. Evolution works at the residue level not at the atomistic levels: only a limited set of side chains has been tried over the course of natural history and any new side chains beyond the standard 20 that would be needed to completely test the all-atom models haven't been tried by Nature. This evolutionary origin of the funneled landscape can be tested in another way, which I describe in the next section.

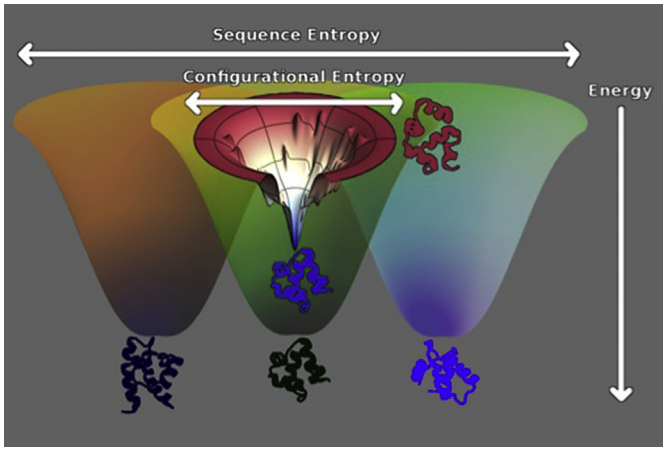

Fig. 5. The folding super landscape. The super-energy landscape of proteins is pictured here. Ideally this would be shown as a function of both sequence and conformation simultaneously. The large funnels are pictured as a function of sequence space with the radial sizes connoting sequence entropy. Energy is again the vertical axis. Natural proteins are not necessarily the lowest energy designs. These would be found at the bottom of the super funnel. For each target the configuration space landscape is funneled, but only to an energy E_F. This structural energy landscape is, however, shown superimposed on the sequence space landscape. Disordered compact structures and sequence scrambled decoys have comparable energy statistics. They are shown near the top of the landscape. The funnels to other structures start from these same high energy states but again would finally reach energies near E_F if they have sufficiently evolved under the minimal frustration selection constraint. The energy of the traps E_g can be estimated by scrambling sequences within the native structure. This diagram shows how the evolutionary and physical configurational landscapes are related to each other. Notice that sequence space is cosmologically bigger than the structure space is as reflected by the large sequence entropy at E_F. This excess coding space allows minimally frustrated landscapes to be found through the random processes of natural selection.

5. The evolutionary landscapes of proteins

Folding of proteins is important to their function and the proper function of proteins is needed for an organism's survival, therefore the folding landscape is a key part of natural selection. Evolution works, however, both by selection and by random experimentation through mutation and recombination. Functional proteins in widely divergent organisms therefore have wildly different sequences. There is much evidence that structure evolves more slowly than sequence does [92]. Slow structural evolution would be expected if the folding energy landscape is funneled. Can we quantitatively relate the diversity of protein sequences all of which fold to a largely common structure to the physical energy landscapes for each of these sequences? One way to address this question is first to assume that folding to the proper structure is actually the only selection constraint on molecular evolution. The funnel hypothesis would then summarize the folding constraint by saying the energy of the folded structure E_F must be far below E_g, the typical energy of a structurally distinct trap. As we have seen the latter depends on ΔE^2, which itself depends foremost on the protein sequence composition (but not primarily on the order of amino acids). This suggests we can picture the evolutionary energy landscape of proteins in a way much like the folding landscape of a single protein [93]. In Fig. 5 we show such a super landscape which describes how energy varies both in sequence and structure. There will be numerous funnels in the landscape corresponding to every possible structure of a protein. We can concentrate on the landscape in sequence space for evolving to a given target structure and compare it with the physical energy landscape for folding to that structure. For the sequence space landscape, the energy coordinate represents the energy that a particular sequence has when it takes on the particular target structure. One would find the most “well-designed” sequence at the bottom of the sequence space funnel. (Unfortunately many early folding theory papers used the “designed” terminology, which can be easily misinterpreted as being an endorsement of creationism!)

It is important to remember, however, that because evolution is a random process, there is no need for selection to have given the “best designed” or most stable sequence, it is only necessary for evolution to have found a sequence that folds to a functional structure sufficiently often. To find the folded structure quickly and reliably the physical energy landscape must have a low energy E_F below some threshold for at least two reasons. One of these reasons is to avoid kinetic trapping, as we see indeed natural sequences do but also it is necessary for survival of the species to avoid being unstable against next generation mutations that would inevitably occur in a finite population of organisms [91]. The first selection constraint based on kinetic trapping would act on the organism possessing the protein itself. Kinetic trapping (if E_F would be close to E_g, the typical energy of a trap) will create long-lived misfolded proteins. Such long lived species could lead to aggregation and problems with protein trafficking. It is often thought that this is an issue in giving rise to the numerous neurodegenerative diseases. The second selection constraint really acts at the population level not at the level of the presently living individual organism. Some fraction of the offspring of a viable organism which has a protein with energy E_F will, after a mutation has occurred now become unviable. Again the closer E_F is to E_g the more this poor performance in the offspring will be a problem for the species, as a whole. Whatever is the actual origin of the selection constraint on E_F, if this

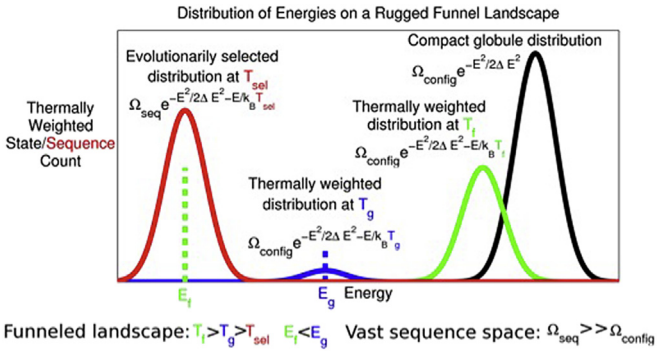

Fig. 6. The distributions of energies in sequence and configuration space. The schematic spectrum in sequence space is shown superimposed on the configurational energy spectrum. Notice that there are many sequences that fold to the same target structure because the selection temperature T_{sel} is greater than the sequence space glass transition temperature. This temperature in turn is lower than the structural space glass transition at T_g.

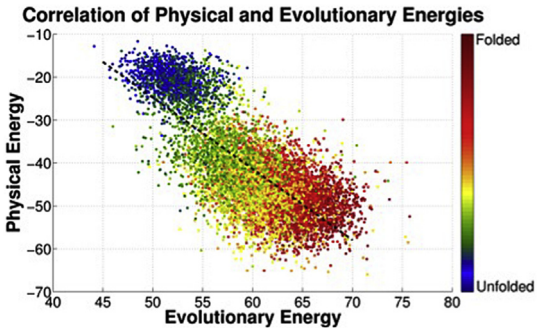

Fig. 8. The correlation between the evolutionary and physical folding landscape. We evaluate the energies of structures using both a physical and a genomic energy function. The pairs shown in the figures correspond to partially folded protein structures generated via molecular simulation using the AWSEM energy function for the Ir69 family. Again the two landscapes turn out to be funneled and strongly correlated as would be expected from the minimal frustration principle. The colors of the points correspond to the fraction of native contacts formed in the sampled structure. For details please see Ref. [46].

energy constraint is the only constraint, we can argue that after sufficient sequence variation has occurred, one would end up with an ensemble of proteins whose sequences are random apart from their satisfying a threshold constraint on their energy in the native structure E_F. Since sequence space is cosmologically large, the selected sample would be equally well described by a Boltzmann distribution for the energy $E(a_1,\ldots a_N)$ of a particular sequence of amino acids $(a_1\ldots a_N)$ in the target structure [94–97]. This Boltzmann distribution would be characterized by an effective temperature of selection T_{sel}. The lower the selection temperature T_{sel} the lower is E_F and the more stable or "better designed" or better said minimally frustrated the sequence would be. Thus as T_{sel} goes down, T_F gets larger relative to the glass transition temperature T_g. If one applies a random energy model ansatz to the sequence space, just as is done for the molten globule configuration space, one ends up with an elegant relation between the evolutionary funnel characterized by E_F (and characteristic evolutionary temperature T_{sel}) and the physical folding funnel characterized by the same E_F (but with different physical temperatures T_F and T_g):

$$\frac{2}{T_F T_{sel}} = \frac{1}{T_g^2} + \frac{1}{T_F^2} \tag{3}$$

The key point is that the two "funnels" one in physical space, the other in sequence space have the same energies for native structures and also would have similar distributions of decoy energies. See Fig 6. An elementary derivation of this relation may be found in the supplementary material of Ref. [46]. This relation between evolutionary and physical scales was first stated in this way by Pande, Grosberg and Tanaka through more sophisticated replica arguments [97].

The assumption of a Boltzmann distribution over sequences has been taken as the basis of several new information theoretic analyses of the genomic data for protein families which maximize sequence entropy given known correlations in aligned sequences [98]. Such analyses have recently become quite useful because of the extensive genome sequencing data now available. Gene sequencing has made known to us, for many structural families, thousands of sequences all of which presumably fold to sensibly the same structure in order to retain their functions. Using the sequence data, by quantifying co-evolution at distinct sites in the protein, one can do "reverse statistical mechanics" to find the form of the energy function $E_F(a_1,\ldots a_N)$ that would give such a sample of sequences. Typically the information theoretic energy function is parametrized via site energies and pair interactions. One algorithm for doing reverse statistical mechanics is known as direct coupling analysis (DCA). This algorithm yields an energy function H_{DCA}

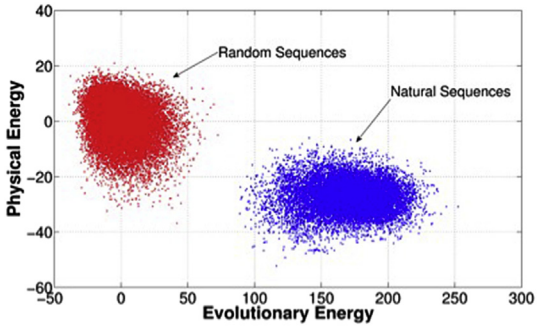

Fig. 7. The correlation between physical and evolutionary folding landscapes. We evaluate the energies, on using the physics based AWSEM energy function, the other using the direct contact approximation genomic based energy function both for scrambled sequences and for natural sequences in the 1r69 repressor family. These pairs of energies are then plotted. We see that both the physical and evolutionary energy landscapes have sizable gaps showing the minimally frustrated nature of the proteins. For details please see Ref 46.

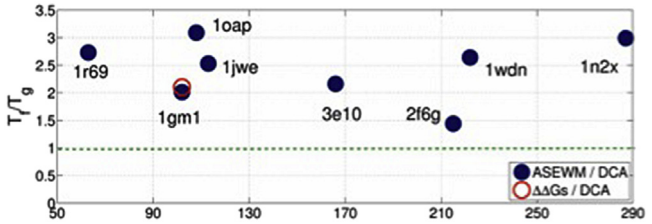

Fig. 9. The smoothness of the folding funnel quantified by coevolutionary information. The T_F/T_g ratios for several protein families are inferred using genomics and landscape theory shown. The families are denoted by the PDB ID codes of the representative structures which are described in Ref. [46]. T_F/T_g measures the smoothness of a folding landscape. Higher values correspond to more ideal funnels. The red circle is an alternate way of making an estimate by comparing changes in evolutionary energies and experimentally measured stability changes. The estimated T_F/T_g ratios for all the natural protein families studied are larger than one so the folding landscape is confirmed to be a funnel. The estimates are clustered around the value of $T_F/T_g = 2.5$ that was estimated by Clementi and Plotkin through a comparison of measured Φ values with simulated ones. For details please see Ref. [46].

which describes the evolutionary constraints at a given site and the coevolution of different positions in the sequence. H_{DCA} should be strongly correlated with the physical energy function. Indeed it has been shown that H_{DCA} correlates well with the energy functions developed for structure prediction. Two plots showing the correlations are found in Figs. 7 and 8. One of these plots shows the correlations of the in sequence space contrasting random sequences threaded on the native structure to the energies for natural sequences known to fold to the target. The other plot evaluates both the physical and evolutionary information theoretic energy for partially folded structures generated by a molecular dynamics simulation using the structure prediction force field near the folding temperature. Computing the DCA energy for differing configurations involves the inclusion only of contact pair energy terms.

You can see in both plots that both energy functions generate a gap between correct sequence–structure pairs and incorrect ones: the evolutionary funnel landscape and the physical funneled landscape are highly correlated.

This analysis also allows us to access the ratio T_F/T_{sel}. We can then use Equation (3) to find the corresponding T_F/T_g ratios. These are shown for several families in Fig 9.

It is heartening that evolutionary data give rise to estimates for T_F/T_g in the same range as those predicted by physical arguments using data from laboratory experiments. The values of the T_F/T_g ratio inferred in this way are larger than those first estimated by Onuchic et al., and are more in line with the values estimated by Clementi and Plotkin, discussed earlier.

6. Frustration and function

Protein function requires specificity of interaction as well as sufficient stability to live in the cell and in subsequent generations of cells. In the protein world, standing there and looking pretty is not always enough, however. Proteins act as nonlinear elements in cellular networks in order to process environmental information. This nonlinearity necessitates motion and thus often an energy landscape with multiple distinct stable states. Because stability of a limited number of specific structures is so important to prevent promiscuous interactions, most of the individual interactions in proteins have evolved to give collectively strongly funneled landscapes but some strategic parts of the sequences located at specific sites in the structure have been selected to be frustrated in order to allow both motion and interaction with partners. To quantify this phenomenon requires tools for localizing the sources of frustration [18,99,100]. Localization of frustration can be detected by examining the energy changes that would occur if only the local environment of a site were to be changed through sequence mutation or allowed to change by reconfigurational physical motion. Minimally frustrated native regions will have energies in the more stable part of the distribution of local energies; regions that are unstable in comparison with the bulk of the distribution of local energies are frustrated. Checking for local frustration is relatively easy once good structures and adequate energy functions are known [99], as is now the case. It has been shown that regions where highly frustrated interactions cluster often map onto sites of allosteric change [100] or can identify binding sites [99] which then

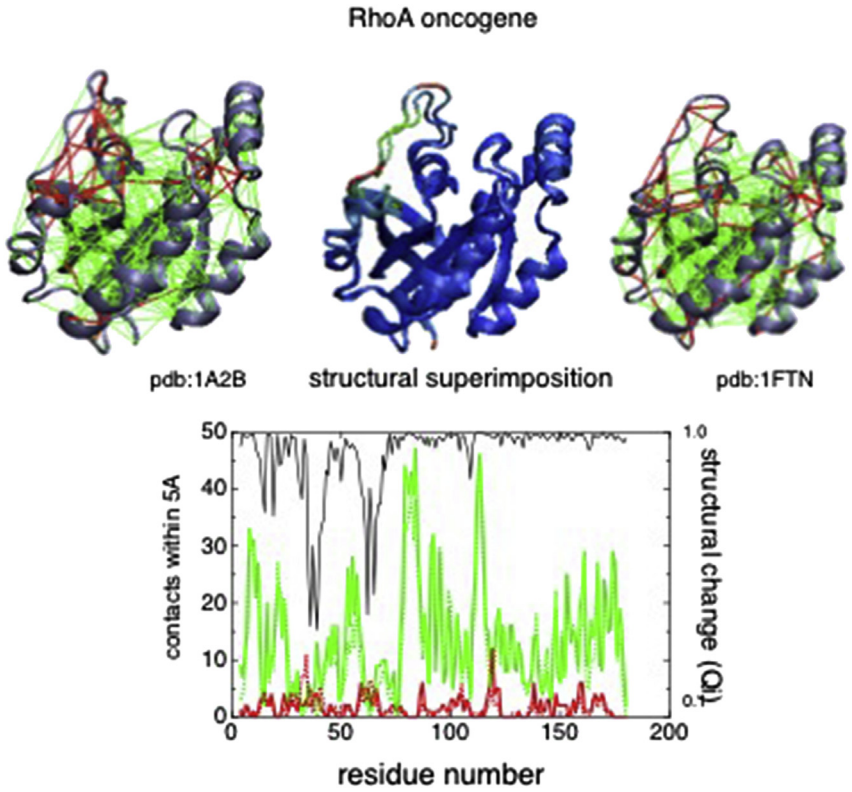

Fig. 10. Frustration serves a functional purpose. A diagram showing the minimally frustrated web of interactions in two structural forms of – RhoA an allosteric protein. This web is indicated in green. The frustrated interactions in the regions shown in red lead to alternate nearly energetically degenerate configurations that allow these regions to function as hinges. The lower panel shows the frustration levels at different sequence locations, the red line indicating the number of frustrated contacts, green the number of minimally frustrated contacts. The black line indicates the local overlap of the two interconnecting structures. Notice the regions that move (and have low Q) correspond to the most energetically frustrated regions. For details please see Ref. [103].

have their frustration relieved once a binding partner is found and then docked to the site. In contrast to the clustered, frustrated, functional regions, minimally frustrated interactions typically form a connected web throughout the protein that keeps subunits of the protein relatively rigid. The combination of the protein having modules with rigidity along with specifically frustrated local regions that act as moveable elements justifies considering many large proteins as being nearly macroscopic "machines". A key difference of these molecular machines from their fully macroscopic counterparts is that the minimally frustrated nature of most of the molecule allows the molecular protein to move by breaking or "cracking" locally and then re-assembling into a new configuration [101,102]. Local frustration analysis is easily automated and a Web server showing both the web of minimally frustrated interactions and the highly frustrated sites is available [103].

An example of the frustration patterns in an allosteric protein is shown for the RhoA oncogene in Fig 10.

7. Folding paradoxes revisited

Protein folding has turned out to be easy and that seems paradoxical. As we have seen Levinthal's original paradox can be avoided in several ways, the most important turning out to be that evolution has led to funneled energy landscapes. It is clear that there are other ways of simplifying the search through configuration by parsing out the configurational entropy piece meal such as capillarity effects so that only part of the chain folds at any time [10,11] or by utilizing local structural signals that increase the probability of the chain bending in the right places [85]. Nevertheless the minimal frustration hypothesis has proved to be a most fruitful tool for visualizing the folding mechanism and addressing protein design and structure prediction. The quantitative form of the minimal frustration principle has been confirmed in several ways through detailed kinetic predictions. In addition we have seen the minimal frustration principle is confirmed to be the result of natural selection and random variation through the comparison of the landscape characteristics inferred from sequence co-evolution and through folding physics. But doesn't this resolution of the Levinthal just raise another paradox? How did the random process of evolution find the minimally frustrated sequences?

Searching sequence space would seem to be a daunting task in the same sense that the searching structure space seemed to be in the Levinthal paradox. Fred Hoyle raised just such an issue in his science fiction work "The Black Cloud" [104]. Sequence space is indeed cosmologically larger than structure space so how does evolution find minimally frustrated sequences? The answer seems to be that while minimally frustrated sequences are exponentially rare, they are dense enough in sequence space that connected paths of mutation can find minimally frustrated sequences if there is sufficient selection pressure. Indeed we can see that there is a cosmologically large number of sequences that fold even to a specific given structure. In Fig. 6 we show the density of both available structural states measuring the configurational entropy and sequences at a given energy measuring the sequence entropy. The minimal frustration principle is quantitatively summarized by saying there is a gap between the folded energy and the glassy structural traps. Precisely because sequence entropy is so much larger than structural entropy the deepest most "well designed" sequence – the energy corresponding to a glass transition in sequence space will be much below the typically selected folding energy E_F. This excess of states signals the expected high connectivity of the sequence space.

As the diversity of amino acid type interactions reduces one will encounter in sequence space a glass transition that will occur now at a higher energy. Indeed it appears that such a coding crisis occurs when we are limited to 2 letter protein folding codes [105]. Finding foldable sequences for some structures becomes problematic even with 3 letter folding codes, in the computer [53]. The idea that interaction diversity is necessary to resolve the Levinthal paradox has been confirmed in the laboratory where 5 distinct amino acid types have been shown to be sufficient for design [106] but fewer numbers do not suffice.

We see that energy landscape analysis suggests that finding minimally frustrated sequences is not too hard for a random process like evolution. This has also been confirmed via simulation studies. Again when evolutionary searches are made in the computer using simplified models it has proved straightforward reach by trial and error sequences to fold into even quite complex structures [107–109]. Evolving for specific function at a specific locus turns out to entail evolving to fold globally as a prerequisite, according to the work of Sasai [108,109].

A rich problem, still unsolved however is how and when did the minimally frustrated sequences come into being in our world? Completely resolving this puzzle of natural history may be hard because folding appears as a phenomenon even in the earliest days of life that we can now probe through evolutionary sequence analysis. Studies of the last universal common ancestor suggest this organism had a full complement of foldable proteins [110]. So presently there is a veil over the early steps of protein biogenesis. There is hope, however, that the process of evolving foldability is still going on today. Eukaryotes encode protein with split genes. Gilbert has proposed that these exons might be themselves exchangeable units [111] and Gō made the case for this for hemoglobin [112]. Using energy landscape analysis Panchenko et al. showed that indeed many exons seemed to be minimally frustrated folding units. We called these units "foldons" [113,114] nearly twenty years ago. The case that foldons were exons was equivocal in 1995. The much larger amount of sequence data and better energy functions that are available today should lead to a re-examination of the foldon–exon correspondence.

There is also evidence that we can see minimal frustration evolving in real time today. During the somatic evolution of antibodies [115], Romesberg and co-workers have examined the dynamics of antibodies that have been made against fluorescent dyes. The motions of the dyes bound to the antibodies can then be probed spectroscopically. His investigations suggest that conformational substates of the antibody–dye complex disappear as the antibody evolves and that the energy landscape gets smoother and smoother as successive rounds of selection occur so that the antibody becomes a better binder. This is very much consistent with the mechanism envisioned by Sasai for evolving folding through selection pressure on specific binding at a given locus.

Thus while there is no new paradox to the trick of having evolved funneled landscapes it seems there is much more that we should be able to learn about evolution and folding in the near future. The union of energy landscape theory and modern genomics should be very fruitful.

Conflict of interest

There is no conflict of interest.

Acknowledgments

The author gratefully acknowledges support from NIH Grants R01 GM044557 and P01 GM071862 from the National Institute of General Medical Sciences, and the Center for Theoretical Biological Physics sponsored by the NSF (PHY-0822283). Generous support has also been provided by the D.R. Bullard-Welch Chair at Rice

University (Grant #C-0016). I am also happy to have the opportunity to thank all of my collaborators over the years who have contributed to the success of energy landscape theory described in this review. I give extra thanks to Nick Schafer, Bobby Kim and Diego Ferreiro who developed the figures in this manuscript.

References

[1] P.G. Wolynes, Three paradoxes of protein folding, in: H. Bohr, S. Brunak (Eds.), Proceedings on Symposium on Protein Folds: a Distance-based Approach, Symposium Distance-based Approaches to Protein Structure Determination II; Copenhagen, November 1994, CRC Press, Boca Raton, FL, 1995, pp. 3–17.

[2] C.B. Anfinsen, Studies on the Principles that Govern the Folding of Protein Chains, Nobel Lecture, National Institutes Of Health, Bethesda, MD, December 11, 1972, http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1972/anfinsen-lecture.pdf.

[3] G. Stent, That was the molecular biology that was, Science 160 (1968) 390–395.

[4] R.F. Service, Problem-solved* (*sort of), Science 321 (2008) 784–786.

[5] N.P. Schafer, B.L. Kim, W. Zheng, P.G. Wolynes, Learning to fold proteins using energy landscape theory, Isr. J. Chem. 53 (2013) 1–28.

[6] C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, D. Baker, Protein structure prediction using Rosetta, Methods Enzymol. 383 (2004) 66–93.

[7] S. Piano, K. Lindorff-Larsen, D.E. Shaw, Protein folding kinetics and thermodynamics from atomistic simulation, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 17845–17850.

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

[9] R. Unger, J. Moult, Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications, Bull. Math. Biol. 55 (1993) 1183–1198.

[10] A.V. Finkelstein, A.Y. Badretdinov, Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable *chain* field, Fold. Des. 2 (1997) 115–121.

[11] P.G. Wolynes, Folding funnels and energy landscapes of larger proteins within the capillarity approximation, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 6170–6174.

[12] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.

[13] D. Thirumalai, N. Lee, S. Woodson, D.K. Klimov, Early events in RNA folding, Ann. Rev. Phys. Chem. 52 (2001) 751–762.

[14] D. Baker, J.L. Sohl, D.A. Agard, A protein folding reaction under kinetic control, Nature 356 (1992) 263–265.

[15] J.D. Bryngelson, P.G. Wolynes, Spin glasses and the statistical mechanics of protein folding, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 7524–7528.

[16] J.D. Bryngelson, P.G. Wolynes, Intermediates and barrier crossing in a random energy model (with applications to protein folding), J. Phys. Chem. 93 (1989) 6902–6915.

[17] B. Derrida, Random-energy model: an exactly solvable model of disordered systems, Phys. Rev. B 2 (1981) 2613.

[18] D.U. Ferreiro, E.A. Komives, P.G. Wolynes, Frustration in biomolecules, Q. Rev. Biophys. 47 (2014) 285–363.

[19] E.I. Shakhnovich, A.M. Gutin, Formation of unique structure in polypeptide chains: theoretical investigation with the aid of a replica approach, Biophys. Chem. 34 (1989) 187–199.

[20] M. Sasai, P.G. Wolynes, Molecular theory of associative memory Hamiltonian models of protein folding, Phys. Rev. Lett. 65 (1990) 2740–2743.

[21] J. Bryngelson, J. Onuchic, N. Socci, P.G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: a synthesis, Proteins Struct. Funct. Genet. 21 (1995) 167–195.

[22] E.I. Shakhnovich, Theoretical studies of protein folding thermodynamics and kinetics, Curr. Opin. Struct. Biol. 7 (1997) 29–40.

[23] K.A. Dill, H.S. Chan, From levinthal to pathways to funnels, Nat. Struct. Biol. 4 (1997) 10–19.

[24] V. Lubchenko, P.G. Wolynes, Theory of structural glasses and supercooled liquids, Ann. Rev. Phys. Chem. 58 (2007) 235–266.

[25] J.N. Onuchic, P.G. Wolynes, Theory of protein folding, Curr. Opin. Struct. Biol. 14 (2004) 70–75.

[26] M. Oliveberg, P.G. Wolynes, The experimental survey of protein folding energy landscapes, Q. Rev. Biophys. 38 (2005) 245–288.

[27] B.A. Shoemaker, J. Wang, P.G. Wolynes, Structural correlations in protein folding funnels, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 777–782.

[28] P.E. Leopold, M. Montal, J.N. Onuchic, Protein folding funnels: a kinetic approach to the sequence–structure relationship, Proc. Natl. Acad. Sci. U. S. A. 89 (18) (1992) 8721–8725.

[29] R. Goldstein, Z. Luthey-Schulten, P.G. Wolynes, Optimal protein-folding codes from spin-glass theory, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 4918–4922.

[30] A.R. Dinner, V. Abkevich, E. Shakhnovich, M. Karplus, Factors that affect the folding ability of proteins, Proteins Struct. Funct. Genet. 35 (1999) 34–40.

[31] R. Mélin, H. Li, N.S. Wingreen, C. Tang, Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study, J. Chem. Phys. 110 (1999) 1252.

[32] M.S. Friedrichs, P.G. Wolynes, Toward protein tertiary structure recognition by means of associative memory Hamiltonians, Science 246 (1989) 371–373.

[33] P.G. Wolynes, Spin glass ideas and the protein folding problems, in: D. Stein (Ed.), Spin Glasses and Biology, World Scientific Press, Singapore, 1992, pp. 225–259.

[34] N.D. Socci, J.N. Onuchic, P.G. Wolynes, Protein folding mechanisms and the multidimensional folding funnel, Proteins Struct. Funct. Genet. 32 (1998) 136–158.

[35] J.N. Onuchic, Z. Luthey-Schulten, P.G. Wolynes, Theory of protein folding: the energy landscape perspective, Ann. Rev. Phys. Chem. 48 (1997) 545–600.

[36] M. Karplus, Behind the folding funnel diagram, Nat. Chem. Biol. 7 (2011) 401–404.

[37] G.R. Bowman, V.S. Pande, Protein folded states are kinetic hubs, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 10890–10895.

[38] C.R. Schwantes, V.S. Pande, Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9, J. Chem. Theory Comput. 9 (2013) 2000–2009.

[39] R.A. Goldstein, Z.A. Luthey-Schulten, P.G. Wolynes, Protein tertiary structure recognition using optimized Hamiltonians with local interactions, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 9029–9033.

[40] C. Hardin, M. Eastwood, Z. Luthey-Schulten, P.G. Wolynes, Associative memory Hamiltonians for structure prediction without homology: alpha-helical proteins, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 14235–14240.

[41] G.A. Papoian, J. Ulander, M.P. Eastwood, Z. Luthey-Schulten, P.G. Wolynes, Water in protein structure prediction, Proc. Natl. Acad. Sci. U. S. A. 101 (2014) 3352–3357.

[42] G.A. Papoian, J. Ulander, P.G. Wolynes, The role of water mediated interactions in protein-protein recognition landscapes, J. Am. Chem. Soc. 125 (2003) 9170–9178.

[43] J.N. Onuchic, P.G. Wolynes, Z. Luthey-Schulten, N.D. Socci, Toward an outline of the topography of a realistic protein-folding funnel, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 3626–3630.

[44] H. Kaya, H.S. Chan, Polymer principles of protein calorimetric two-state cooperativity, Proteins 40 (2000) 637–661.

[45] C. Clementi, S.S. Plotkin, The effects of nonnative interactions on protein folding rates: theory and simulation, Prot. Sci. 13 (2004) 1750–1766.

[46] F. Morcos, N.P. Schafer, R.R. Cheng, J.N. Onuchic, P.G. Wolynes, Coevolutionary information, protein folding landscapes and the thermodynamics of natural selection, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 12408–12413.

[47] H. Kenzaki, N. Koga, N. Hori, R. Kanada, W.F. Li, K. Okazaki, X.Q. Yao, S. Takada, CafeMol: a coarse-grained biomolecular multitop for simulating proteins at work, J. Chem. Theory Comput. 7 (2011) 1979–1989.

[48] J.K. Noel, P.C. Whitford, K.Y. Sanbonmatsu, J.N. Onuchic, SMOG@ ctbp: simplified deployment of structure-based models in GROMACS, Nucleic Acids Res. 38 (Web server issue) (Jul 2010) W657–W661, http://dx.doi.org/10.1093/nar/gkg498. Epub 2010 Jun 4.

[49] N. Gō, Theoretical studies of protein folding, Ann. Rev. Biophys. Bioeng. 12 (1983) 183–210.

[50] Y. Levy, S.S. Cho, T. Shen, J.N. Onuchic, P.G. Wolynes, Symmetry and frustration in protein energy landscapes: a near-degeneracy resolves the ROP dimer folding mystery, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 2373–2378.

[51] L. Sutto, J. Lätzer, J. Hegler, D. Ferreiro, P.G. Wolynes, Consequences of localized frustration for the folding mechanism of the IM7 protein, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 19825–19830.

[52] S.S. Cho, Y. Levy, P.G. Wolynes, Quantitative criteria for native energetic heterogeneity influences in the prediction of protein folding kinetics, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 434–439.

[53] H.H. Truong, B.L. Kim, N.P. Schafer, P.G. Wolynes, Funneling and frustration in the energy landscapes of some designed and simplified proteins, J. Chem. Phys. 139 (2013) 121908.

[54] S.C. Yang, S.S. Cho, Y. Levy, M.S. Cheung, H. Levine, P.G. Wolynes, J.N. Onuchic, Domain swapping is a consequence of minimal frustration, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 13786–13791.

[55] J.U. Bowie, R. Luthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, Science 253 (1991) 164–170.

[56] P.A. Alexander, Y. He, Y. Chen, J. Orban, P.N. Bryan, A minimal sequence decode for switching protein structure and function, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 21149–21154.

[57] J.R. Telford, P. Wittung-Stafshede, H.G. Gray, J.R. Winkler, Protein folding triggered by electron transfer, Acc. Chem. Res. 31 (1998) 755–763.

[58] K.W. Plaxco, K.T. Simons, D. Baker, Contact order transition state placement and the refolding rates of single domain proteins, J. Mol. Biol. 277 (1998) 985–994.

[59] A.A. Nickson, B.G. Wensley, J. Clarke, Take home lessons from the studies of related proteins, Curr. Opin. Struct. Biol. 23 (2013) 66–74.

[60] C. Clementi, H. Nymeyer, J.N. Onuchic, Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins, J. Mol. Biol. 298 (2000) 937–953.

[61] A.R. Fersht, Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding, W.H. Freeman, New York, 1999.

[62] J.N. Onuchic, N.D. Socci, Z. Luthey-Schulten, P.G. Wolynes, Protein folding funnels: the nature of the transition state ensemble, Fold. Des. 1 (1996) 441–450.

[63] M. Munson, R.L. Regan, R. O'Brien, J.M. Sturtevant, Redesigning the hydrophobic core of a four-helix-bundle protein, Prot. Sci. 3 (1994) 2015–2022.
[64] Y. Gambin, A. Schug, E.A. Lemke, J.J. Lavinder, T.J. Magliery, J.N. Onuchic, A.A. Deniz, Direct single-molecule observation of a protein living in two opposed native structures, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 10153–10158.
[65] A. Schug, P.C. Whitford, Y. Levy, J.N. Onuchic, Mutations as trapdoors to two competing native conformations of the Rop-dimer, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 17674–17679.
[66] A.P. Capaldi, C. Leanthous, S.E. Radford, Im7 folding mechanism: misfolding on the path to the native state, Nat. Struct. Biol. 9 (2002) 209–216.
[67] A.M. Figueiredo, S.B.-M. Whitaker, S.E. Knowling, S.E. Radford, G.R. Moore, Conformational dynamics is more important than helical propensity for the folding of the all α-helical protein Im7, Prot. Sci. 22 (2013) 1722–1738.
[68] C. Zong, C.J. Wilson, T. Shen, P.G. Wolynes, P. Wittung-Stafshede, *n*-Value analysis of apo-azurin folding: comparison between experiment and theory, Biochemistry 45 (2006) 6458–6466.
[69] C. Zong, C.J. Wilson, T. Shen, P. Wittung-Staffshede, S. Mayo, P.G. Wolynes, Establishing the entatic state in folding metallated pseudomonas *Aeruginosa azurin*, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 3159–3164.
[70] T. Shen, C.P. Hoffman, M. Oliveberg, P.G. Wolynes, Scanning malleable transition state ensembles: coupling theory and experiment for folding protein UnA, Biochemistry 44 (2005) 6433–6439.
[71] J. Portman, S. Takada, P.G. Wolynes, Microscopic theory of protein folding rates. I. Fine structure of the free energy profile and folding routes from a variational approach, J. Chem. Phys. 114 (2001) 5069–5081.
[72] S. Takada, Gō-ing for the prediction of folding mechanism, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 11698–11700.
[73] P. Das, C.J. Wilson, G. Fosatti, P. Wittung-Stafshede, K.S. Mathews, C. Clementi, Characterization of the folding landscape of monomeric lactose repressor: quantitative comparison of theory and experiment, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 14569–14574.
[74] V. Krivov, M. Karplus, Hidden complexity of free energy surfaces for peptide (protein) folding, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 14766–14770.
[75] D.L. Ensign, P.M. Klassen, V.S. Pande, Heterogeneity even at the speed limit of folding: large scale molecular dynamics study of fast folding variant of the Villin head piece, J. Mol. Biol. 374 (2007) 806–816.
[76] V.A. Voelz, G.R. Bowman, K. Beauchamp, V.S. Pande, Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(1-39), J. Am. Chem. Soc. 132 (2010) 1526.
[77] E.R. Henry, R.B. Best, W.A. Eaton, Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 17880–17885.
[78] R.B. Best, G. Hummer, W.A. Eaton, Native contacts determine protein folding mechanisms in atomistic simulations, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 17874–17879.
[79] N.D. Socci, J.N. Onuchic, P.G. Wolynes, Diffusive dynamics of the reaction coordinate for protein folding funnels, J. Chem. Phys. 104 (1996) 5860–5868.
[80] R.G. Best, G. Hummer, Coordinate depleted diffusion in protein folding, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 1088–1093.
[81] M. Gruebele, The fast protein folding problem, Ann. Rev. Phys. Chem. 50 (1999) 485–516.
[82] B.G. Wensley, S. Batey, F.A.C. Bone, Z.M. Chan, N.R. Tumelty, Experimental evidence for a frustrated energy landscape in a three-helix bundle protein family, Nature 463 (2010) 685–688.
[83] H.S. Chung, W.A. Eaton, Single molecule fluorescence probes dynamics of barrier crossing, Nature 502 (2013) 685–688.
[84] Z. Luthey-Schulten, B.E. Ramirez, P.G. Wolynes, Helix-coil, liquid crystal and spin glass transitions of a collapsed heteropolymer, J. Phys. Chem. 99 (1995) 2177–2185.
[85] J. Saven, P.G. Wolynes, Local conformational signals and the statistical thermodynamics of collapsed helical proteins, J. Mol. Biol. 57 (1996) 199–216.
[86] J. Baum, C.M. Dobson, P.A. Evans, C. Hanley, Characterization of a partly folded protein by NMR methods–studies of the molten globule state of guinea-pig alpha lactalbumin, Biochemistry 28 (1989) 7–13.
[87] J. Wang, S. Plotkin, P.G. Wolynes, Configurational diffusion on a locally connected correlated energy landscape; application to finite, random heteropolymers, J. Phys. I 7 (1997) 395–421.
[88] S.S. Plotkin, J. Wang, P.G. Wolynes, Statistical mechanics of a correlated energy landscape model for protein folding funnels, J. Chem. Phys. 106 (1997) 2932–2948.
[89] K.K. Koretke, Z. Luthey-Schulten, P.G. Wolynes, Self-consistently optimized energy functions for protein structure prediction by molecular dynamics, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 2932–2937.
[90] A. Davtyan, N.P. Schafer, W. Zheng, C. Clementi, P.G. Wolynes, G.A. Papoian, AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing, J. Phys. Chem. 116 (2012) 8494–8850.
[91] B.L. Kim, N.P. Schafer, P.G. Wolynes, Predictive energy landscapes for folding α-helical transmembrane proteins, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 11031–11036.
[92] C. Chothia, A.M. Lesk, The relations between divergence of sequence and structure in proteins, EMBO J. 5 (1986) 823–826.
[93] E. Bornberg-Bauer, H.S. Chan, Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 10689–10694.
[94] N.V. Dokholyan, E.I. Shakhnovich, Understanding hierarchical protein evolution from first principles, J. Mol. Biol. 312 (2001) 289–307.
[95] S. Ramanathan, E. Shakhnovich, Statistical mechanics of protein with evolutionarily selected sequences, Phys. Rev. E 1994 (50) (1994) 1303–1312.
[96] J.G. Saven, P.G. Wolynes, Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules, J. Phys. Chem. B 101 (1997) 8375–8389.
[97] V.S. Pande, A.Y. Grosberg, T. Tanaka, Statistical mechanics of simple models, Biophys. J. 73 (1997) 3192–3210.
[98] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, M. Weigt, Direct coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) E1293–E1301.
[99] D. Ferreiro, J. Hegler, E. Komives, P.G. Wolynes, Localizing frustration in native proteins and protein assemblies, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 19819–19824.
[100] D.U. Ferreiro, J.A. Hegler, E.A. Komives, P.G. Wolynes, On the role of frustration in the energy landscapes of allosteric proteins, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 3499–3505.
[101] O.M. Miyashita, J.N. Onuchic, P.G. Wolynes, Nonlinear elasticity, protein quakes and the energy landscapes of functional transitions in proteins, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 1679–1684.
[102] O. Miyashita, P.G. Wolynes, J.N. Onuchic, Simple energy landscape model for the kinetics of functional transitions in proteins, J. Phys. Chem. B 109 (2005) 1959–1969.
[103] M. Jenik, R. Gonzalo Parra, L.G. Radusky, A. Turjanski, P.G. Wolynes, D.U. Ferreiro, Protein frustratometer: a tool to localize energetic frustration in protein molecules, Nucleic Acids Res. 40 (2012) 348–351.
[104] F. Hoyle, The Black Cloud, William Heinemann Ltd, London, 1957.
[105] P.G. Wolynes, As simple as can be? Nat. Struct. Biol. 4 (1997) 871–874.
[106] D.S. Riddle, J.V. Santiago, S.T. Bray-Hall, N. Doshi, V.P. Grantcharova, Q. Yi, D. Baker, Functional rapidly folding proteins from simplified amino acid sequences, Nat. Struct. Biol. 4 (1997) 805–809.
[107] A.M. Gutin, V.I. Abkevich, E.I. Shakhnovich, Evolution–the selection of fast folding model protein, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 1282–1286.
[108] S. Saito, M. Sasai, T. Yomo, Evolution of folding ability of proteins through functional selection, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 11324–11328.
[109] C. Nagao, T.P. Terada, T. Yomo, M. Sasai, Correlation between evolutionary structural development and protein folding, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 18950–18955.
[110] C.G. Kurland, B. Canbäck, O.G. Berg, The origin of modern proteomes, Biochimie 89 (2007) 1454–1463.
[111] W. Gilbert, Why genes in pieces? Nature 271 (1978) 501.
[112] M. Gō, Correlation of DNA exonic regions with protein structural units in haemoglobin, Nature 291 (1981) 90–92.
[113] A.R. Panchenko, Z. Luthey-Schulten, P.G. Wolynes, Foldons, protein structural modules and exons, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 2008–2013.
[114] A.R. Panchenko, Z. Luthey-Schulten, R. Cole, P.G. Wolynes, The foldon universe: a survey of structural similarity and self-recognition of independently folding units, J. Mol. Biol. 272 (1997) 95–105.
[115] M.C. Thielges, J. Zimmermann, W. Yu, M. Oda, F.E. Romesberg, Exploring the energy landscapes of antibody antigen complexes: protein dynamics, flexibility and molecular recognition, Biochemistry 47 (2008) 7237–7247.