Folding proteins: finding a needle in a haystack

Ken A. Dill

University of California, San Francisco, USA

I revisit the Levinthal paradox of protein folding kinetics. Efficient computational methods are now finding or approaching native states on model free energy landscapes for small molecules and chains up to about 60 monomers. Also, recent experiments show counterexamples to the 'thermodynamic hypothesis', i.e. they indicate that the global free energy minimum may not always identify biologically active conformations of proteins.

Current Opinion in Structural Biology 1993, **3**:99–103

Introduction

I assess the prospects for searching protein conformations by computer to find native states. I review recent experiments showing that proteins may sometimes fall into deep kinetic traps on the way to conformations of lower free energy. At the same time, several recent computational search strategies are becoming increasingly successful at reaching deeply into free energy landscapes towards finding globally optimal conformations.

Folding is remarkable

On the one hand, it would seem remarkable that a protein can fold up to find its unique native state in a matter of seconds. According to the Levinthal argument [1,2], this is like searching for a needle in a haystack. This argument is based on two incontrovertible facts: some native states are at global minima of free energy (see below); and it would take too long (by tens of orders of magnitude) to find global minima by random search. Hence the 'Levinthal paradox': how does a protein find its native state in such a short time?

Biological mechanisms such as proline isomerases, chaperone proteins, and ribosomes can assist in the folding of proteins. Biological mechanisms are ubiquitous. It is important to understand how they act. But biological mechanisms cannot resolve the Levinthal problem. Protein folding is all the more remarkable when it happens, as it commonly does, *in vitro*, without biological helpers. If helpers prevent aggregation (see Jaenicke, this issue, pp 104–112) or are catalysts that speed up folding without affecting reactant (denatured state) or product (native state), then we should ultimately be able to understand the folded state on its own, without having to deal with the details of the folding pathway. It remains a most remarkable and central feature of protein folding that it can happen without biological helpers.

Global minima

Some proteins can find their single global minimum without the assistance of biological mediators, according to the following evidence. Proteins such as ribonuclease find their native states by refolding from incorrect disulfide-bonded conformations [3]. Further, for at least some proteins, native activity is thermodynamically reversible [4], independent of the folding pathway, and independent of the folding rate [5–8], indicating that those native proteins are stable like crystals, rather than metastable like glasses. The structures of glasses are dependent on their preparation histories whereas the native structures of reversible proteins are not.

Why does this evidence imply that the native state is at a global minimum of free energy? Or put differently, what is the meaning of 'global minimum'? If a thermodynamic system starts in any of states 1,2,3,..., m and goes to state γ in a time τ, then it means that: on that time scale, γ is kinetically accessible from the starting conformations; and γ has a lower free energy than any other state accessible from the starting conformations. A protein denatured in 6 M guanidine hydrochloride will adopt an exceedingly large ensemble of conformations [9]. When the denaturant is removed, every molecule can begin its journey towards the native state from any one of a large number of different starting conformations, drawn broadly from throughout the conformational space. Moreover, if a protein is caused to start from a different denatured ensemble, say at low pH instead of high denaturant, and if all the molecules arrive at the same folded structure, then that structure must be at a minimum of free energy which is lower than any others accessible from those starting

Abbreviations
HP—hydrophobic/polar; **MC**—Monte Carlo.

© Current Biology Ltd ISSN 0959-440X

conformations for the degrees of freedom accessible on that time scale.

Time scale is always implicit in thermodynamics. Even though time does not appear in thermodynamics, the time scale of an experiment determines what degrees of freedom are relevant to the thermodynamic description of it. A protein may fold on a time scale of seconds. It may aggregate in days to weeks. It may hydrolyze into amino acids in months or years, and amino acids may covalently degrade over even longer times. This does not mean that a folded protein is not at equilibrium. It only means that this particular equilibrium (folding) happens on a time scale of seconds (or thereabouts), and that it involves conformational degrees of freedom, rather than intermolecular or covalent degrees of freedom. Thermodynamics is useful because it applies to processes that happen in real time. The time scale dictates the accessibilities of states: the logarithm of a time constant is proportional to the height of an energy barrier divided by temperature. Thermodynamics is not just applicable to the limit of infinite time, but rather to the limit of time that is long for the relevant set of degrees of freedom.

What are the relevant degrees of freedom, and what is the relevant time scale? Whereas some proteins fold in seconds, slower events and higher barriers are also well known. For example, proline isomerization can increase the time scale of folding from seconds to hours [10,11•]. Nevertheless, it has been commonly assumed that all the conformational degrees of freedom are accessible on the scale of hours.

Baker, Sohl and Agard [12••] have recently found a counterexample, indicating that folding can also involve much longer time scales. α-Lytic protease carries along its own catalyst, a pro region, which helps to carry the protein over a high kinetic barrier to folding. The pro region, even when it is covalently disconnected from the protein, is required to fold the chain into its active form. In the absence of the pro region, the chain is stable for weeks in an inactive form. Therefore, both active and inactive forms appear stable for weeks under the same conditions; one of them is undoubtedly metastable. In addition, the inactive form of plasminogen activator inhibitor is more stable than its active form [13,14]. These results imply that computer algorithms that aim to fold proteins by searching the whole conformational space might find an inactive form, the wrong answer from a biological perspective. In such cases, finding the right answer goes beyond thermodynamics: 'nativeness' will then be defined by a criterion other than stability alone.

On the other hand...

Hence, it is extraordinary that proteins fold so quickly without biological helpers. Or is it? It could be argued that protein folding is actually much simpler than many elementary processes of macroscopic thermodynamics. Macroscopic systems can be stable, i.e. at global minima, yet they have far more degrees of freedom than one folding protein molecule. One protein molecule may have hundreds or thousands of degrees of freedom corresponding to the conformations of backbone and side-chain bonds. If the number of degrees of freedom is m, then the number of conformations is approximately z^m, where z, the number of conformational isomers, has been estimated to be about 3.8 for polypeptide backbone conformations, or 1.4 if we consider only compact conformations [15]. But the properties of macroscopic materials involve not only the internal degrees of freedom of one molecule but also macroscopic numbers of molecules. Their total numbers of degrees of freedom can exceed Avogadro's number. Polymers, or small molecules such as sodium chloride, can crystallize in a short time to the global minimum on a free energy surface that involves many orders of magnitude more degrees of freedom than the folding of a protein molecule. Indeed, more than 170 different structures of crystalline polymers and 400 structures of crystallized proteins are known [16]. In each case, each single chain must find the right conformation at the same time that an exceedingly large number of other molecules is doing so and they must all come together to pack and align correctly.

What are the degrees of freedom? In addition to internal degrees of freedom of each molecule, there are degrees of freedom corresponding to the center-of-mass coordinates of the molecules. The free energy of transfer for localizing the centers of mass into a crystal is $nRT\ln(a_2/a_1)$, where n is the number of molecules, a_2 is the high activity of the solute in the crystal, and a_1 is the lower activity in the crystallization solution. Put in terms of microstates, rather than free energies, the relative number of configurations of the system is $(a_2/a_1)^{-n}$, i.e. a number around 10–10^3 raised to a power nearly equal to the magnitude of Avogadro's number! (This argument can be refined using molecular theories of crystallization [17•], but this conclusion remains unchanged.) If folding is remarkable, then the crystallization of materials is extraordinary. In this light, the fact that *anything* is thermodynamically stable would seem extraordinary. But these observations are only extraordinary from a particular perspective, i.e. that which assumes that the global minimum is sought by random search and, therefore, that the main determinant of success is the number of degrees of freedom.

But the problem really isn't the *size* of conformational space. The problem is the *shape* of conformational space. This is the fly in the ointment of the Levinthal argument. On a flat golf course, a golf ball rolling randomly is not likely to hit the hole. The bigger the golf course, the harder the problem. But if we had a golf course shaped like a funnel, downhill everywhere towards a hole at the bottom, a hole-in-one would happen every time, no matter how big the golf course. Searching for a needle in a haystack by random sampling would take forever, but with a big magnet it would be quick. Thermodynamics finds global minima because energetic landscapes have downhill grades, not because we wait for processes to happen on flat landscapes.

Search strategies

So to predict the native conformation of a protein, the problem is not how to exhaustively or randomly explore all the possible conformations. The problem is how to develop search strategies that can reach quickly and deeply into the free-energy landscape. This is a problem, unlike exhaustive search, that need not scale exponentially with chain length.

For example, an n-dimensional parabola:

$$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} (x_i - x_{i0})^2$$

is a landscape which has i = 1,2,3,...,n degrees of freedom x_i where x_{i0}s are constants. But no matter how big n is, it is evident that the single global minimum is $x_i = x_{i0}$, for all i. Hence, the size of this landscape is not important for locating the single global minimum, because the degrees of freedom are independent. It is clear from this example, that an n-dimensional fictitious landscape can always be constructed that will have a global minimum at any chosen point, and simple search strategies can find it quickly.

This approach has been the basis for some early and simple models of protein folding, which have 'native-like' propensities built into them. In these cases, a landscape is constructed that is designed to lead to the desired structure. In analogy, this is like re-shaping the natural landscape of a golf course by creating artificial hills and valleys so that the ball always runs downhill to the hole. One way of doing this is to specify many native-like interactions that are to be 'built into' the target structure. Skolnick and Kolinski [18–22] have taken a major step forward beyond this type of model. They have designed landscapes with only three to seven parameters. The parameters are adjustable but are based on physical driving forces. Relatively quickly, Monte Carlo (MC) dynamics finds the global minima of these landscapes, as determined by the thermodynamic reversibility test: if different simulations lead to the same structure starting from different denatured states, then confidence increases that it may be the native state for that landscape. Using different parameters for different folds, the global minimum structures in the Skolnick and Kolinski models closely resemble native proteins.

A major challenge is to find a single universal set of energetic parameters for all proteins, that will fold any protein correctly depending only on the amino acid sequence. Because the shape of energy landscape depends on the energy function, obtaining the correct folding kinetics will require a universal and correct set of energy parameters. During the past two years, rapid progress has been made in understanding the effectiveness of various search strategies in finding the global minima of landscapes constructed from physical potentials without native-like propensities. In some cases, the global minima for these landscapes are known by the thermodynamic reversibility test; in other cases (for short chains), the global minima are known by exhaustive simulation beforehand.

Covell and Jernigan [23] performed exhaustive searches of lattice conformations of short proteins, subject to the constraint that the conformations must fit within the known surface shape. Even though the surface shape constraint introduces information beyond that contained in the amino acid sequence, this study has been important in demonstrating how close to a known global optimum a computational strategy can come using simple universal potential functions. Hinds and Levitt [24••] have recently been able to avoid the use of the surface shape constraint for exhaustive simulations of relatively compact chains using a lattice model.

Two methods reported recently find global minima by perturbing the ϕ,ψ conformational energy surface of peptides using a force field. In one, the energy landscape is smoothed by a diffusion equation transformation [25,26,27•]. In the other, Head-Gordon *et al.* [28] eliminated local minima by introducing physical considerations, e.g. by favoring L over D amino acid isomers. The computer time in these methods scales better than for exhaustive searching.

Standard MC methods assume fixed types of elemental transitions, or 'move sets'. The dynamically optimized MC method of Bouzida *et al.* [29••] changes the move set as the simulation proceeds. So far, it has only been tested on searching the force field conformational spaces for small molecules, including adenosine, but it finds stable conformations much faster than traditional MC methods.

Another approach is to use a simpler model in which the global optimum is known from prior exhaustive simulation. In the HP (hydrophobic/polar) lattice model [30–32], native conformations can be found by exhaustive simulation, or by design [33]. For short chains, the native states have been found for many different monomer sequences. O'toole and Panagiotopoulos [34••] have studied the search efficiency of two different sampling algorithms using the HP model on the three-dimensional simple cubic lattice: Metropolis, and a variant of the method of Rosenbluth and Rosenbluth. They found the Rosenbluth strategy more effective than the Metropolis method. With the Rosenbluth strategy, they obtained thermodynamic reversibility for some sequences of 48 monomers. The same strategy falls slightly short of the global optimum for 80-mers. Unger and Moult [35••] have found a genetic algorithm to be an improvement by two to three orders of magnitude in speed over Metropolis in finding low-energy states for HP chains up to 64 monomers long on two-dimensional square lattices. Yue and Dill have developed a search strategy that uses a theory of constraints to find native states of HP chains five to eight orders of magnitude faster than brute force exhaustive searches, and which can locate global optima for chains up to about 36 monomers long (K Yue, KA Dill, unpublished data).

Yet another search strategy is based on a hypothesis for the kinetic pathways of protein folding, and has been tested using the HP lattice model. Assuming that proteins fold via 'hydrophobic zippers' [36•,37•], in which hydrophobic contacts are made opportunistically in the presence of existing contacts, searches have been performed using the HP lattice model for chains up to 58 monomers long on three-dimensional simple cubic lattices, for which a lower limit on globally optimal native energies may be computed from analytical theory (K Yue, KA Dill, unpublished data). This approach too finds conformations at or close to global minima [36•,37•]. But for some monomer sequences, this strategy never reaches global minima. It finds kinetically trapped states instead, offering a possible explanation for why some proteins might not fold to global minima.

Conclusion

The long-term goal is to predict native protein structures using only knowledge of the amino acid sequence. The examples above indicate that native states of low-resolution model landscapes can now be found, or at least closely approached, for chain lengths up to about 60 monomers, without the need for extra constraints or native propensities. High-resolution models are not far behind. To fold proteins by computer, the main hurdles are: to develop more accurate models and potential functions for free energy landscapes; to learn whether or not these computational search strategies can handle chains longer than bovine-pancreatic-trypsin-inhibitor-length molecules; and to learn when to seek the global minimum and when to follow a kinetic pathway. The needles in the haystacks of protein conformational space need no longer be sought by picking out one straw at a time.

Acknowledgements

I thank D Agard, D Baker, S Bromberg and CM Dobson for helpful comments.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Levinthal C: **Are There Pathways for Protein Folding?** *Chim Phys* 1968, **65**:44–45.

2. Wetlaufer DB: **Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins.** *Proc Natl Acad Sci USA* 1973, 70:697–701.

3. Anfinsen CB: **Principles that Govern the Folding of Protein Chains.** *Science* 1973, **181**:223–230.

4. Privalov PL: **Stability of Proteins: Small Globular Proteins.** *Adv Protein Chem* 1979, **33**:167–241.

5. Garel J-R, Nall BT, Baldwin RL: **Guanine-unfolded State of Ribonuclease A Contains Both Fast- and Slow-refolding Species.** *Proc Natl Acad Sci USA* 1976, **73**:1853–1857.

6. Kato S, Okamura M, Shimamoto N, Utiyama H: **Spectral Evidence for a Rapidly Formed Structural Intermediate in the Refolding Kinetics of Hen Egg-white Lysozyme.** *Biochemistry* 1981, **20**:1080–1085.

7. Denton JB, Koniski Y, Scheraga HA: **Folding of Ribonuclease A from a Partially Disordered Conformation. Kinetic Study under Folding Conditions.** *Biochemistry* 1982, **21**:5155–5163.

8. Lynn RM, Koniski Y, Scheraga HA: **Folding of Ribonuclease A from a Partially Disordered Conformation. Kinetic Study under Transition Conditions.** *Biochemistry* 1984, **23**:2470–2477.

9. Tanford C: **Protein Denaturation.** *Adv Protein Chem* 1968, **23**:121–282.

10. Nall B: **Proline Isomerization and Protein Folding.** *Comm Mol Cell Biophys* 1985, **3**:123–143.

11. • Hurle MR, Anderson S, Kuntz ID: **Confirmation of the Predicted Source of a Slow Folding Reaction: Proline 8 Bovine Pancreatic Trypsin Inhibitor.** *Protein Eng* 1991, **4**:451–455.

Replacing a proline in bovine pancreatic trypsin inhibitor speeds up folding by a few orders of magnitude.

12. •• Baker D, Sohl JL, Agard DA: **A Protein-folding Reaction under Kinetic Control.** *Nature* 1992, **356**:263.

A pro piece, normally a covalent piece of α-lytic protease, catalyzes folding of the protease, reducing the kinetic barrier by several orders of magnitude.

13. Katagiri K, Okada K, Hattori H, Yano M: **Bovine Endothelial Cell Plasminogen Activator Inhibitor. Purification and Heat Activation.** *Eur J Biochem* 1988, **176**:81–87.

14. Banzon JA, Kelly JW: **β-Sheet Rearrangements: Serpins and Beyond.** *Protein Eng* 1992, **5**:113–116.

15. Dill KA: **Theory for the Folding and Stability of Globular Proteins.** *Biochemistry* 1985, **24**:1501–1509.

16. Tadakoro H: *Structure of Crystalline Polymers.* New York: John Wiley & Sons; 1979.

17. • Berland CR, Thurston GM, Kondo M, Broide ML, Pande J, Ogun O, Benedek GB: **Solid–Liquid Phase Boundaries of Lens Protein Solutions.** *Proc Natl Acad Sci USA* 1992, 89:1214–1218.

Experiments and theory for the crystallization of lens proteins.

18. Kolinski A, Skolnick J, Yaris R: **Monte Carlo Simulations on an Equilibrium Globular Protein Folding Model.** *Proc Natl Acad Sci USA* 1986, **83**:7267–7271.

19. Skolnick J, Kolinski A: **Dynamic Monte Carlo Simulations of Globular Protein Folding/Unfolding Pathways. II. α-Helical Motifs.** *J Mol Biol* 1990, **212**:819–836.

20. • Skolnick J, Kolinski A: **Dynamic Monte Carlo Simulations of a New Lattice Model of Globular Protein Folding, Structure and Dynamics.** *J Mol Biol* 1991, **221**:499–531.

Introduces a high-resolution (2,1,0) lattice model with energy parameters for torsion angles hydrophobic and polar contacts, and cooperative interactions.

21. Skolnick J, Kolinski A: **Simulations of the Folding of a Globular Protein.** *Science* 1990, **250**:1121–1125.

22. Rey A, Skolnick J: **Comparison of Lattice Monte Carlo Dynamics and Brownian Dynamics Folding Pathways of the α-Helical Hairpins.** *Chem Phys* 1991, **158**:199–219.

23. Covell DG, Jernigan RL: **Conformations of Folded Proteins in Restricted Spaces.** *Biochemistry* 1990, **29**:3287–3294.

24. •• Hinds DA, Levitt M: **A Lattice Model for Protein Structure Prediction at Low Resolution.** *Proc Natl Acad Sci USA* 1992, 89:2536–2540.

An exhaustive search of conformations of lattice chains constrained to be relatively compact; this filters good from poor conformations.

25. Stillinger FH: **Role of Potential-energy Scaling in the Low-temperature Relaxation Behavior of Amorphous Materials.** *Phys Rev B* 1985, **32**:3134–3141.

26. Piela L, Kostrowicki J, Scherga HA: **The Multiple-minima Problem in the Conformational Analysis of Molecules. De-**

formation of the Potential Energy Hypersurface by the Diffusion Equation Method. *J Phys Chem* 1989, **93**:3339–3346.

27. Kostrowicki J, Scheraga HA: **Application of the Diffusion**
• **Equation Method for Global Optimization in Oligopeptides.** *J Phys Chem* 1992, **96**:7442–7449.

Describes a method for deforming an energy landscape by using the diffusion equation to 'smooth out' local minima. Application of the diffusion equation method of [26] to finding the global energy minimum for metenkephalin, a pentapeptide. The native structure is found in minutes on a workstation.

28. Head-Gordon T, Stillinger FH, Arrecis J: **A Strategy for Finding Classes of Minima on a Hypersurface — Implications for Approaches to the Protein Folding Problem.** *Proc Natl Acad Sci USA* 1991, **88**:11076–11080.

29. Bouzida D, Kumar S, Swendsen RH: **Efficient Monte**
•• **Carlo Methods for the Computer Simulation of Biological Molecules.** *Phys Rev [A]* 1992, **45**:8894–8901.

Two methods are presented for very fast searching of conformational energy landscapes, based on changing move sets as the MC procedure proceeds.

30. Lau KF, Dill KA: **A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins.** *Macromolecules* 1989, **22**:3986–3997.

31. Lau KF, Dill KA: **Theory for Protein Mutability and Biogenesis.** *Proc Natl Acad Sci USA* 1990, **87**:638–642.

32. Chan HS, Dill KA: **"Sequence Space Soup" of Proteins and Copolymers.** *J Chem Phys* 1991, **95**:3775–3787.

33. Yue K, Dill KA: **Inverse Protein Folding Problem: Designing Polymer Sequences.** *Proc Natl Acad Sci USA* 1992, **89**:4163–4167.

34. O'Toole EA, Pangiotopoulos AZ: **Monte Carlo Simulation of**
•• **Folding Conditions of Simple Model Proteins using a Chain Growth Algorithm.** *J Chem Phys* 1993, in press.

A new MC search strategy based on an algorithm by Rosenbluth and Rosenbluth versus standard Metropolis MC, is tested on the three-dimensional HP lattice model. This algorithm finds native states for chains of 48 monomers, and falls a few HH contacts short of global optima for chains of 80 monomers.

35. Unger R, Moult J: **Genetic Algorithm for Protein Folding**
•• **Simulations.** *J Mol Biol* 1993, in press.

A test of a 'genetic algorithm' developed by the authors versus standard Metropolis MC on the two-dimensional HP lattice model. The genetic algorithm finds native states for chains of up to 60 monomers and comes within five HH contacts of the native state for a 64-monomer chain.

36. Fiebig K, Dill KA: **Protein Core Assembly Processes.** *J Chem*
• *Phys* 1993, in press.

Outlines an approach for assembling hydrophobic cores of heteropolymers. Tests on the short-chain HP lattics model show that this non-exhaustive search method finds native states for about 70% of all HP monomer sequences.

37. Dill KA, Fiebig K, Chan HS: **Cooperativity in Protein Folding**
• **Kinetics.** *Proc Natl Acad Sci USA* 1993, in press.

Proposes that proteins fold by 'hydrophobic zippers' (defined in [36•]), a non-exhaustive process in which hydrophobic contacts are made opportunistically. Tests on the three-dimensional HP lattice model show that this search strategy finds conformations within about three HH contacts of the global minima for chains of 46 and 58 residues.

KA Dill, Department of Pharmaceutical Chemistry, University of California, 3333 California Street, Room 102, San Francisco, California 94118, USA.