Modeling protein folding: the beauty and power of simplicity

Eugene I Shakhnovich

It is argued that simplified models capture key features of protein stability and folding, whereas more detailed models may be more appropriate for protein structure prediction. A brief overview of experimental and theoretical results is presented that corroborates these points. I argue that statistical models capture the key principle of protein stability – cooperativity – and therefore provide a reasonable estimate of protein free energy whereas more detailed but less physically transparent calculations fail to do so. I also explain that the previously published claim that simple models give predictions that are inconsistent with experiments on polypeptide block-copolymers is based on incomplete analysis of such experiments.

Address: Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA.
e-mail: eugene@diamond.harvard.edu

Electronic identifier: 1359-0278-001-R0050

Folding & Design 01 June 1996, 1:R50–R54

 ISSN 1359-0278

The recent commentary by Honig and Cohen [1] in *Folding & Design* discusses the role of different models in folding studies. Such discussion is timely, because protein folding is a fascinating cross-disciplinary field that attracts scientists with different background and scientific cultures. They bring to the protein folding field the models and ways of thinking that are accepted in their respective background fields. Such diversity of scientific cultures is a great virtue of the protein folding field, in which physics, chemistry, biology and mathematics meet. It is important for our cross-disciplinary field to discuss with balance both strong points and limitations of different approaches. To this end, the article by Honig and Cohen [1] can be seen as a 'nucleus' for such a discussion in *Folding & Design* both on paper and on the internet (see the discussion pages on BioMedNet at http://BioMedNet.com/). Unfortunately, as happens in many discussions, the commentary presents somewhat extreme views, minimizing the value of approaches based on simplified folding models. Such an assessment does not seem to be substantiated by facts and by consistent analysis of recent developments. The purpose of my response is to present an alternative viewpoint based on experimental facts and theoretical analyses which show that simplified models have been a valuable tool for the study of folding. It is my hope that this response will present an account of successes and limitations of simplified models that is more balanced and respectful to all points of view.

It is obvious that proteins are too complicated to allow exact modeling — some simplifications are necessary already at the stage of initial formulation of models. The crucial issue, common to any theoretical analysis, is where to draw 'a line in the sand', i.e. what degree of simplification is acceptable without 'throwing the baby out with the bath water'. A useful model must exhibit folding in an unbiased simulation and, moreover, should allow for hundreds of folding events to occur. The latter requirement follows from the fact that folding is an intrinsically statistical phenomenon and no conclusion can be derived from a single folding or unfolding trajectory. Currently, folding simulations satisfying these important requirements are feasible only in the context of simplified lattice and off-lattice models (most of these are 'sidechain-only', in the terminology of [1]; e.g. see [2–12] and references therein). It was suggested by Honig and Cohen [1] that such "sidechain-only models fail to capture essential features of protein folding."

The existing gap between foldable and detailed atomistic models leaves us no choice other than to study folding mechanisms using simplified models, benchmarking them against experiment. After all, how should theory be judged? By its ability to explain existing experiments and to predict the results of future experiments. In this regard, contrary to the assertions of Honig and Cohen [1], simplified models provide valuable insights into the basic principles of protein stability and folding.

First of all, the lattice model captures one of the most fascinating and important features of proteins — cooperativity of their structure [6,7,13–16]. Contrary to a popular but incorrect view, helix/coil transition is not truly cooperative (not a first-order type, e.g. see [11]). The reason for this is given by the Landau theorem, which prohibits stable phases (such as one long helix) and phase transitions in one-dimensional systems without long-range (in space) interactions [16].

This means that helix decay and formation cannot account for cooperativity of protein folding transition [17]. A clear example is the non-cooperative character of unfolding of the molten globule state of myoglobin which involves massive helix breakdown but is non-cooperative [18].

A virtue of simplified models is their intimate connection with statistical mechanics. This is very important, as it often allows us to compare simulation results with statistical-mechanical analytical theories. As the two approaches (analytical and numeric) are often complementary (e.g.

analytical theories require a large number of particles, whereas simulations often restrict it to modest values), the correlation between them brings more confidence in the validity of obtained results and allows us to get rid of artifacts. An important example in this regard is the convergence of analytical theories and numeric simulations on the explanation of the basic reason for cooperative behavior as connected with the non-randomness of protein sequences, their optimization for efficient folding and higher stability [7,14,19,20]. Random sequences were predicted to exhibit noncooperative, diffuse transitions [6,7,21]. This prediction has been confirmed by experiments [22,23].

Lattice models capture the cooperative character of protein folding and therefore, in spite of their simplicity, they provide a reasonable estimate of generic protein stability of 0.1 kcal M^{-1} $residue^{-1}$ [24], close to experimental values [25]. This is a better estimate than calculations within a much more detailed 'backbone-centric' model, which in spite of a large number of parameters involved in many cases gives estimates of the free energy of folding more than an order of magnitude off from experimental values (see Fig. 5c in [26]). For example, calculations in [26] estimate myoglobin stability to be greater than 150 kcal M^{-1} while actually it hardly exceeds 10 kcal M^{-1} at physiological conditions [25]. This is not surprising because the protein free energy represents a small difference between two large numbers — energetic contributions (roughly 1 kcal M^{-1} $residue^{-1}$) and the opposing entropic contributions of the same order of magnitude.

Calculations that involve many parameters but fail to incorporate the key physics of folding — cooperativity — are not likely to have sufficient precision to give a reliable estimate of folding free energy of an individual protein.

Another important issue is how simplified models treat protein folding kinetics. The simplification introduced in many — but not all — lattice models is that they do not include backbone hydrogen bonds explicitly. Is this factor crucial? Clearly the impact of backbone hydrogen bonds on folding kinetics is essential only if they are formed, to a considerable extent, before folding transition state is reached.

In this regard, Honig and Cohen [1] refer to two papers [27,28] which concern unfolding intermediates. In their view, the results of these papers suggest that the transition state of unfolding "appears to be due to the breaking of hydrogen bonds". I believe that this conclusion does not follow from the experiments reported by Baldwin and coauthors [27,28]. What they actually show is that all hydrogen bonds get disrupted at the rate-limiting step of unfolding, after the dry molten globule intermediate [29]. That does not imply that hydrogen bonds break at or near the transition state. In fact, as transition from the dry molten globule unfolding intermediate to the unfolded state is two state, all properties change with the same kinetic exponent. Therefore, it is impossible to judge from the experiments [27,28] at what point(s) of the unfolding trajectories overcoming the major barrier the disruption of hydrogen bonds occurs. Independently of where each hydrogen bond does actually break in the rate-limiting step of unfolding, it will be decaying with the same kinetic exponent.

Currently, in fact, only the protein engineering method can provide information about the structure of the folding transition state [30]. Keeping that in mind, let me turn to the analysis of experiments that aim specifically to study the folding transition state. These clearly point out that the hydrogen bond network, or secondary structure, is formed after the folding transition state is overcome. First, consider a series of conclusive experiments by Fersht and co-workers on CI2 [31,32]. They provide extensive data suggesting that mutations outside the nucleation core (but affecting the stability of secondary structure) do not significantly affect folding kinetics. A good example is Val19→Ala substitution in CI2. Mutation of two neighboring core residues belonging to the same α-helix, Ala16 and Ile20, strongly affect folding kinetics having ϕ-values of 1.0 and 0.48 respectively. Val19 does not participate in the core but belongs to the α-helix. Its mutation to alanine stabilizes the helix [33], but leads to only a 3% increase of the folding rate (practically within experimental error). Normalized to a small stability change, it yields a ϕ-value of –0.15 for Val19 [31,32]. This shows clearly that the contribution of formation of transient fragments of secondary structure to folding rate is a few percent.

The conclusive result of simulations and protein engineering experiments is that the major factors determining folding kinetics are long-range (in sequence) interactions forming the nucleus [10,32]. These results are corroborated in the work of Sauer and co-workers [34] with another protein, P22 arc repressor. Mutation Pro8→Leu increased the stability of the β-sheet, but the folding rate remained largely unchanged. The conclusion was made in this work [34] that "the β-sheet forms after the rate-limiting state in folding." Subsequent study from the same laboratory [35] included substitution of many helical sites in the P22 arc repressor by helix-propensity altering alanine with little change in folding rates also.

Similar results were obtained recently for another protein, CheY [36]. For example, mutation Gly39→Ala in this protein stabilizes α-helix 2 but does not affect the folding rate (ϕ-value for folding is 0.03).

Mutational analysis was also employed by Sosnick *et al.* [37] to determine when secondary structure forms in the folding of an α-helical dimer, GCN4. Residues at non-per-

turbing positions along the exterior length of the helices were substituted one at a time with alanine and glycine to vary helix propensity. For all variants the bimolecular folding rate remains largely unchanged; the change in stability appears largely in the unfolding rate. Sosnick *et al.* [37] conclude that "contrary to most folding models, widespread helix is not yet formed at the rate-limiting step in the folding pathway."

Theoretical efforts in the framework of simplified lattice [10] models lead, along with experiments [32,38], to the discovery of the nucleation-condensation mechanism via formation of a specific nucleus in the folding transition state. This mechanism is likely to capture important features of folding kinetics of at least small proteins. Further, analysis of lattice model kinetics allowed us to develop a successful approach to predict the residues which constitute the folding nucleus [39], i.e. making even very specific predictions of the theory testable experimentally.

Finally, I analyze the 'gedanken experiment' interpreted by Honig and Cohen [1] as a convincing proof of the serious limitations of lattice models. Honig and Cohen discuss experiments of Scheraga and co-workers [40] on conformational transition in polyalanine. Alanine is a moderately hydrophobic helix-stabilizing amino acid. According to Honig and Cohen [1] "polyalanine, H_n, forms an α-helix, but is predicted by sidechain-only models to form an array of compact structures" (H stands for hydrophobic). I was quite intrigued by this conclusion because it did not sound to be consistent with statistical mechanics. Indeed, helical conformation in this case is unfavorable energetically as well as entropically. Compaction into globular state with high helical content could result in both enhanced hydrophobic interactions and entropy increase due to a multitude of available compact conformations.

Some details of the experiments of Scheraga and co-workers [40] were not clearly explained by Honig and Cohen. Reading the original paper, I found the following. In order to make their polymers soluble, Scheraga and co-workers [40] attached long charged polylysine 'tails' to both ends of the polyalanine. In salt-free solution, where Coulomb repulsion between polylysines is unscreened and overwhelming, the central fragment, polyalanine, indeed adopts extended helical conformation. When salt is added, the repulsion between polylysine tails is screened and polyalanine adopts a compact globular conformation with high helical content. It is clear that the attached polylysine plays a crucial role in determining the conformation of the polyalanine fragment. It also explains why longer polyalanine helices in [40] are more stable. Short polyalanine fragments force polylysine tails to come close to each other. Repulsion between polylysine fragments destroys polyalanine helix forcing the chain into extremely extended conformation. Longer polyalanine sequences experience less stretching force from the polylysine tails and, although they still cannot become compact in salt-free solution, they can become helical.

Most lattice models of folding do not include long-range (in space) Coulomb forces, the reason being that in most cases folding occurs at physiological salt concentrations, at which electrostatic forces are screened. Long helices without chain compaction were observed in [40] in salt-free solution. A trivial modification of a lattice model to account for such conditions would be to introduce Coulomb interactions between charged monomers. Then a fair comparison with lattice models would be to consider a poly-H sequence with very strong unscreened Coulomb repulsion between the ends. Such a lattice chain, under the action of stretching force between the ends, will adopt an extended conformation. Simple addition of local $i,i+3$ attraction makes this extended conformation helical if the chain is long enough to keep the ends sufficiently far from each other. This corresponds to the salt-free case of the experiment [40]. When the repulsion between chain ends diminishes (upon addition of salt in [40]), hydrophobic attraction takes over and the chain becomes compact in the lattice model, in agreement with basic statistical mechanics and the experiment [40]. Moreover, when local $i,i+3$ attraction (modeling hydrogen bond) is added along with hydrophobic attraction in the lattice model, the resulting lattice conformations were shown to be compact and helical (see Fig. 2a in [10]), exactly like the ones in the high-salt case in the experiment [40]. Summarizing the discussion of the 'gedanken experiment', I note that, opposite to the claim of Honig and Cohen [1], the results of this elegant study of Scheraga and co-workers [40] are consistent with statistical mechanics and can be rationalized within simple lattice models. In fact, a statistical-mechanical rationale for the observed effects was given already in the original paper [40]. It will be a matter of interest in future study to model block-copolymer polypeptides like the ones studied by Scheraga and co-workers [40].

Lattice and other simplified analytical models are the statistical mechanician's contributions to protein folding. It is part of the culture in this field to appreciate that simplified models provide coarse-grained descriptions and as such they may be adequate to describe the effects taking place on longer (than microscopic) time and length scales. The most important example of this kind is folding, and I have presented evidence here that statistical-mechanical models do capture many essential features of folding.

Certainly such models have their limitations. Many important effects, such as protein function or ligand binding specificity, occur on microscopic length and time scales. Lattice models do not provide a sufficient degree of detail to study these phenomena. Their analysis requires all-atom models and molecular dynamics simulations [41].

Besides that, there are a number of aspects of folding that are more microscopic in nature for which lattice models, in their present form, can have problems. One example is proline isomerization, which is known to prevent folding if proline isomers are incorrect. Other issues are chirality of amino acids and tight packing of sidechains. These aspects of the folding problem can be addressed in the framework of more atomistic models, which are currently not feasible for folding simulations.

It is also possible that more detailed models are needed to address the second, more visible to the public, side of the 'holy grail' of protein folding — prediction of protein conformation. Conclusive successes in that direction may be for the distant future, despite considerable progress achieved in the elucidation of the folding mechanism. This situation has many analogies with other fields of physics or chemistry where models often adequately tackle fundamental large-scale behavior but do not capture all microscopic properties. One example of this kind is that although crystallization as a physical phenomenon is understood, we are still unable to predict crystal symmetry from the chemical structure of the constituting molecules.

I believe that it is indeed important to appreciate the real limitations of popular models. Detailed objective analysis of experiments which cannot be explained by existing models is a means to elucidate such limitations. This can be a powerful stimulus for further development of models and theory.

It is an exciting time in protein folding because theory and experiment have started to talk to each other and agreement is encouraging. Though the scope of simplified protein folding models is limited, they are proving to be a very useful tool to study fundamental principles of this fascinating phenomenon.

References

1. Honig, B. & Cohen, F. (1996). Adding backbone to protein folding: why proteins are polypeptides. *Folding & Design* **1**, R17–R20; 1359-0278-001-R0017.
2. Skolnick, J. & Kolinski, A. (1991). Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics *J. Mol. Biol.* **221**, 499–531.
3. Shakhnovich, E., Farztdinov, G.M., Gutin, A.M. & Karplus, M. (1991). Protein folding bottlenecks: a lattice monte-carlo simulation. *Phys. Rev. Lett.* **67**, 1665–1667.
4. Sali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold? *Nature* **369**, 248–251.
5. Miller, R., Danko, C., Fasolka, M.J., Balazs, A.C., Chan, H.S. & Dill, K.A. (1992). Folding kinetics of proteins and copolymers. *J. Chem. Phys.* **96**, 768–780.
6. Shakhnovich, E. (1994). Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Lett.* **72**, 3907–3910.
7. Hao, M.-H. & Scheraga, H. (1994). Monte-carlo simulation of a first order transition for protein folding. *J. Phys. Chem.* **98**, 4940–4945.
8. Hao, M.-H. & Scheraga, H. (1994). Statistical thermodynamics of protein folding: sequence dependence. *J. Phys. Chem.* **98**, 9882–9886.
9. Guo, Z., Thirumalai, D. & Honeycutt, J.D. (1992). Folding kinetics of proteins: a model study. *J. Chem. Phys.* **97**, 525–535.
10. Abkevich, V., Gutin, A. & Shakhnovich, E. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026–10036.
11. Karplus, M. & Shakhnovich, E. (1994). Protein folding: theoretical studies of thermodynamics and dynamics. In *Protein Folding.* (Creighton, T., ed.) pp. 127–196, W.H. Freeman and Company, New York.
12. Bryngelson, J., Onuchic, J.N., Socci, N.D. & Wolynes, P. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195.
13. Privalov, P.L. (1989). Stability of proteins. I. Small globular proteins. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 47–69.
14. Goldstein, R., Luthey-Schulten, Z.A. & Wolynes, P. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922.
15. Abkevich, V., Gutin, A. & Shakhnovich, E. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460–471.
16. Landau, L.D. & Lifshitz, E.M. (1980). *Statistical Physics.* Pergamon, London.
17. Poland, D. & Scheraga, H.A. (1970). *Theory of Helix-Coil Transitions in Biopolymers.* Academic Press, New York.
18. Griko, Y.V. & Privalov, P.L. (1994). Thermodynamic puzzle of apomyoglobin unfolding. *J. Mol. Biol.* **235**, 1318–1325.
19. Shakhnovich, E. & Gutin, A. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA* **90**, 7195–7199.
20. Ramanathan, S. & Shakhnovich, E. (1994). Statistical mechanics of proteins with "evolutionary selected" sequences. *Phys. Rev. E* **50**, 1303–1312.
21. Shakhnovich, E. & Gutin, A.M. (1989). Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of replica approach. *Biophys. Chem.* **34**, 187–199.
22. Chaffotte, A., Guillou, Y. & Goldberg, M. (1992). *Biochemistry* **31**, 9694–9702.
23. Davidson, A. & Sauer, R. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA* **91**, 2146–2150.
24. Gutin, A., Abkevich, V. & Shakhnovich, E. (1995). Is burst hydrophobic collapse necessary for rapid folding? *Biochemistry* **34**, 3066–3076.
25. Makhatadze, G. & Privalov, P.L. (1995). Energetics of protein structure. *Adv. Protein Chem.* **47**, 307–425.
26. Honig, B., & Yang, A.S. (1995). Free energy balance in protein folding. *Adv. Protein Chem.* **46**, 27–58.
27. Kiefhaber, T. & Baldwin, R. (1995). Kinetic of hydrogen bond breakage in the process of ribonuclease a measured by pulsed hydrogen exchange. *Proc. Natl. Acad. Sci. USA* **92**, 2657–2661.
28. Kiefhaber, T., Labhardt, A.M. & Baldwin, R. (1995). Direct NMR evidence for an intermediate preceding the rate-limiting step in the unfolding of ribonuclease A. *Nature* **375**, 513–515.
29. Shakhnovich, E. & Finkelstein, A.V. (1989). Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers* **28**, 1667–1681.
30. Matouschek, A., Kellis, J. Jr., Serrano, L., Bycroft, M. & Fersht, A.R. (1990). Transient folding intermediates characterized by protein engineering. *Nature* **346**, 440–445.
31. Jackson, S.E., elMasry, N. & Fersht, A.R. (1993). Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. *Biochemistry* **32**, 11270–11278.
32. Itzhaki, L., Otzen, D. & Fersht, A. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288.
33. Creighton, T. (1992). *Proteins. Structure and Molecular Properties.* W.H. Freeman & Co, New York.
34. Schildbach, J., Milla, M., Jeffrey, P., Raumann, B. & Sauer, R. (1995). Crystal structure, folding and operator binding of the hyperstable arc repressor mutant PL8. *Biochemistry* **34**, 1405–1412.
35. Milla, M., Braun, R., Walburger, C. & Sauer, R. (1995). P22 arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding. *Biochemistry* **34**, 13914–13919.
36. López-Hernández, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Folding & Design* **1**, 43–55.

37. Sosnick, T., Jackson, S., Englander, S.W. & DeGrado, W. (1996). The role of helix formation in the folding of a fully α-helical coiled-coil. *Proteins* 24, 427–432.
38. Sosnick, T.R., Mayne, L. & Englander, S.W. (1996). Molecular collapse: the rate-limiting step in two-state cytochrome C folding. *Proteins* 24, 413–426.
39. Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature* 379, 96–98.
40. Ingwall, R., Scheraga, H., Lotan, N., Berger, A. & Katchalski, E. (1968). Conformational studies of poly-L-alanine in water. *Biopolymers* 6, 331–368.
41. Brooks, C. III, Karplus, M. & Petitt, B. (1988). *Proteins: a Theoretical Perspective on Dynamics, Structure and Thermodynamics.* Wiley, New York.