Published in final edited form as:

Arch Biochem Biophys. 2008 January 1; 469(1): 1–3. doi:10.1016/j.abb.2007.10.007.

Protein folding: Are we there yet?

A. Clay Clark

Department of Molecular and Structural Biochemistry, North Carolina State University, 128 Polk Hall, CB7622, Raleigh, NC 27695-7622, USA, clay_clark@ncsu.edu

There are known knowns. There are known unknowns. There are also unknown unknowns. [1]

After nearly one hundred years of examining forces that stabilize protein structure, protein chemists still cannot predict the native conformation of a protein from a known amino acid sequence, which is the nature of the protein folding problem. This problem is a bottleneck in many wide-ranging fields because the sequencing of genomes has accelerated in recent years and new fields of study are emerging as a result [2]. What have we learned in the hundred-or-so years of study?

We know that the amino acid sequence of a protein dictates the native conformation. In 1961, Anfinsen et al. showed that the folding of RNaseA was driven spontaneously by the free energy of folding [3], results that generally are credited with the beginning of protein folding as a research discipline. For the first time, Anfinsen demonstrated that the amino acid sequence of a protein contained all of the information needed for the protein to reach the native conformation. Subsequently it has been shown that some proteins need modulators, such as chaperones, to fold, but Anfinsen's results have been confirmed with numerous single and multi-domain proteins. Many years later, Khorana and colleagues showed that bacteriorhodopsin could be reconstituted into a lipid membrane [4]. Thus, Anfinsen's principle holds for membrane proteins as well. Indeed, the implications of Anfinsen's work are far reaching, to the point that the central dogma of biology, that DNA transcription provides RNA and RNA translation provides protein, had to be modified because translation gives an unfolded protein that must then fold to the native conformation.

Interest in protein structure and folding did not begin in the 1960s, however. In 1938, protein chemists learned about the hydrophobic factor from Langmuir [5], and his work was based on earlier studies of surface tension by Traube in the late 19th century [6]. The hydrophobic concept as a factor in collapsing the protein polymer is a consequence of the strong attraction of water molecules for one another, so contact between apolar surfaces and water is avoided. Although this was recognized in the early 20th century [7,8], the hydrophobic principle became widely known to protein chemists in 1959 due to a review by Kauzmann [9]. One important question at the time was what held together globular proteins? The idea of protein denaturation was known, and it was established that denaturation correlated with unfolding of the globular structure to a random polymer [10]. The prevailing view, based on work by Mirsky and Pauling [11], was that hydrogen bonds were a critical force in maintaining the native structure, and the internal hydrogen bond theory was prevalent. Before Kauzmann's review, it was thought that internal hydrogen bonds were responsible for creating collapsed protein molecules. It is important to note that much of the work concerning the nature of protein collapse occurred before Avery's experiments

demonstrating that DNA is the genetic material [12]. Prior to that time, proteins were widely viewed as the carriers of genetic information.

Following the seminal work of Anfinsen, Brandts in 1964 proposed the two-state model for protein unfolding, where the native and unfolded conformations are in equilibrium ($N \rightleftharpoons U$), based on data for chymotrypsinogen [13]. The model was used later by Lumry et al. [14] and Tanford [15] and remains today the standard model for interpreting the equilibrium folding reactions for many single-domain proteins. Indeed, measures of protein conformational free energy, entropy, and enthalpy have been important for the past one hundred (or so) years and have led to an understanding of the native state as nearly unique in structure, whereas the unfolded state has much higher entropy [11]. The importance of the two-state model is that folding occurs without populating stable intermediates, or partially folded structures, and it led to the musings of Cy Levinthal. In 1968, Levinthal speculated on how long the process would take if a protein must fold by a random search of all possible conformations in the absence of folding intermediates. For a small protein of one hundred residues, Levinthal estimated that the time for folding would exceed the age of the universe by many orders of magnitude [16]. Thus, real proteins do not fold by a random search.

Levinthal's paradox, as it is known today, suggested that folding intermediates and folding pathways must exist. Shortly thereafter, Wong and Tanford [17] and Kuwajima et al. [18] independently discovered equilibrium folding intermediates. And, in a series of classic experiments, Creighton [19] demonstrated methods for isolating intermediates using disulfide trapping. Later, the significance of folding intermediates was generalized by Ptitsyn and co-workers [20] and named "molten globules" by Ohgushi and Wada [21]. Molten globules from different proteins were shown to share common properties, such as a high content of native-like secondary structure, a compact conformation, little tertiary structure, and non-rigid side chains. Although the presence of folding intermediates generally is accepted today, their role in protein folding remains controversial. Do they guide the folding reaction toward the native state? Or do they represent misfolded structures that are off-pathway? This issue remains to be resolved.

The search for intermediates also led to an explosion of kinetic folding studies in the 1970s. Baldwin and colleagues [22] showed that fast kinetic techniques are useful for detecting folding intermediates, and they used temperature jump experiments to demonstrate a kinetic intermediate in the millisecond time range for the unfolding of RNaseA. Similarly, Ikai and Tanford used stopped-flow techniques to demonstrate kinetic intermediates in the guanidine hydrochloride induced unfolding transition of cytochrome *c* [23]. Marion and Wuthrich later revolutionized studies of kinetic intermediates by developing two-dimensional NMR methods [24,25], leading to hydrogen-deuterium exchange studies to explore conformational changes in proteins. The questions that arose for kinetic intermediates were similar to those for the equilibrium intermediates; that is, are they on- or off-pathway? To what extent do they guide folding to the native conformation? The issue was complicated further when Garel and Baldwin showed that complexity in the unfolded conformation could result in kinetic folding intermediates [26], and Brandts et al. demonstrated that this was due to isomerization around prolyl bonds in the unfolded state [27]. More recently, with ensemble approaches to folding, the unfolded conformation is considered an ensemble of states with similar free energies. This is important conceptually because the molecular biology revolution of the 1970s and 1980s also was embraced by protein chemists. That is, it has become routine today to use site-directed mutagenesis to change the amino acid sequence of a protein and to measure changes on the folding properties. However, one should consider effects of the mutations on the unfolded as well as on the native ensemble [28,29].

By the early 1980s, kinetic folding data led to the question of how protein folding starts. The two prevailing views at that time were that folding is hierarchic and that folding begins with a hydrophobic collapse. In the former, folding begins by forming backbone structures that persist and are stabilized by packing against one another (framework model), whereas in the latter, folding begins by a hydrophobic collapse to a molten globule state, from which the search for the native protein occurs from a smaller number of collapsed states. The models later were expanded to include the possibility that specific interactions direct folding [30], although the issue remains to be solved and still is relevant today. Current ideas on how folding starts appear promising [31,32].

During the search for folding intermediates, researchers became interested in proteins that failed to fold, usually involving large oligomeric proteins where concentration dependent protein aggregation dramatically decreased the yield of native protein. However, in examining how the proteins misfold and aggregate, one obtains an understanding of how chaperones interact with proteins to facilitate folding [33,34] as well as an understanding of how misfolded proteins are involved in neurodegenerative and other diseases [35,36]. These results have led to a wealth of drug development research to mitigate the effects of protein misfolding [37].

In the mid-1980s, Dill [38] and Bryngelson and Wolynes [39] published landmark papers describing the energy landscape of a protein as a funnel. In this model, the free energy of a given structure on the three-dimensional landscape decreases as the protein gets closer to the native ensemble, which is at the bottom of the funnel. The funnel landscape has profound consequences for structure prediction, folding kinetics, and biological function [40,41]. The difference between the forces stabilizing the native ensemble versus those associated with chain entropy determines the conformational free energy of a protein, which is small compared to those forces. The difference also gives rise to kinetic barriers to folding. Proteins that rapidly fold are thought to have a smooth funnel landscape so that the folding rate is dependent on the speed at which secondary and tertiary structures can form, so-called downhill folding. On the other hand, proteins that fold more slowly, or those with one or more kinetic intermediates, are thought to have rugged landscapes with multiple activation barriers [42].

As described by Dill and colleagues [43], solving the protein folding problem will require a concerted effort on several fronts: defining how the native structure is thermo-dynamically stable based on a given amino acid sequence, determining how to predict the native structure computationally, and determining how proteins fold fast (the Levinthal paradox). This special issue on protein folding contains 11 articles on current topics in protein folding. The questions have expanded since Anfinsen's studies to include more than the information content in an amino acid sequence. Of interest today is the nature of the transition state for folding, the unfolded ensemble, molecular crowding effects, the role of water in folding reactions, forces that influence protein aggregation, new classes of proteins such as repeat proteins, the effect of protein topology on folding rates, the folding of membrane proteins, and many, many others. It is not possible to cover all topics of interest in the field of protein folding in a single issue. The reviews focus on what is known about the topic (known knowns) and what remains to be elucidated (known unknowns). The "unknown unknowns" from the previous quote are left to future researchers. Readers are referred to other excellent reviews for topics such as the function of molecular chaperones [34,36,44,45], methods for studying protein folding [46–48], protein design [49,50], and protein structure prediction [51,52].

References

1. Rumsfeld, D. Press conference at NATO headquarters; June 6, 2002; Available from: <http://www.defenselink.mil/transcripts/transcript.aspx?transcriptid=3490>
2. Gross M. *Curr. Biol.* 2001; 11:R77–R78. [PubMed: 11231160]
3. Anfinsen CB, Haber E, Sela M, White FW Jr. *PNAS.* 1961; 47:1309–1314. [PubMed: 13683522]
4. Lind C, Hojeberg B, Khorana HG. *J. Biol. Chem.* 1981; 256:8298–8305. [PubMed: 7263654]
5. Langmuir I. *Proc. Roy. Inst. Grt. Brit.* 1938; 30:483–496.
6. Traube J. *Liebig's Ann. Chem.* 1891; 265:27–55.
7. Latimer WM, Rodebush WH. *J. Am. Chem. Soc.* 1920; 42:1419–1433.
8. Bernal JD, Fowler RH. *J. Chem. Phys.* 1933; 1:515–548.
9. Kauzmann W. *Adv. Prot. Chem.* 1959; 14:1–62.
10. Simpson RB, Kauzmann W. *J. Am. Chem. Soc.* 1953:5139–5152.
11. Mirsky AE, Pauling L. *PNAS.* 1936:439–447. [PubMed: 16577722]
12. Avery OT, MacLeod CM, McCarty M. *J. Exp. Med.* 1944; 79:137–158. [PubMed: 19871359]
13. Brandts JF. *J. Am. Chem. Soc.* 1964; 86:4291–4301.
14. Lumry R, Biltonen R, Brandts JF. *Biopolymers.* 1966; 4:917–944. [PubMed: 5975643]
15. Tanford C. *Adv. Prot. Chem.* 1968; 23:121–282.
16. Levinthal C. *J. Chim. Phys. Physico-Chimie Biol.* 1968; 65:44–45.
17. Wong K-P, Tanford C. *J. Biol. Chem.* 1973; 248:8518–8523. [PubMed: 4202778]
18. Kuwajima K, Nitta K, Yoneyama M, Sugai S. *J. Mol. Biol.* 1976; 106:359–373. [PubMed: 10445]
19. Creighton TE. *J. Mol. Biol.* 1974; 87:579–602. [PubMed: 4475110]
20. Dolgikh DA, Gilmanshin RI, Brazhnikov EV, Bychkova VE, Semisotnov GV, Yvenyaminov SY, Ptitsyn OB. *FEBS Lett.* 1981; 136:311–315. [PubMed: 7327267]
21. Ohgushi M, Wada A. *FEBS Lett.* 1983; 164:21–24. [PubMed: 6317443]
22. Tsong TY, Baldwin RL, Elson EL. *PNAS.* 1971; 68:2712–2715. [PubMed: 5288248]
23. Ikai A, Tanford C. *Nature.* 1971; 230:100–102. [PubMed: 4927005]
24. Marion D, Wuthrich K. *Biochem. Biophys. Res. Commun.* 1983; 113:967–974. [PubMed: 6307308]
25. Strop P, Wider G, Wuthrich K. *J. Mol. Biol.* 1983; 166:641–665. [PubMed: 6306250]
26. Garel JR, Baldwin RL. *PNAS.* 1973; 70:3347–3351. [PubMed: 4519627]
27. Brandts JF, Halvorson HR, Brennan M. *Biochemistry.* 1975; 14:4953–4963. [PubMed: 241393]
28. Raleigh DP, Plaxco KW. *Prot. Pep. Lett.* 2005; 12:117–122.
29. Bowler BE. *Mol. Biosys.* 2007; 3:88–99.
30. Baldwin RL. *TIBS.* 1989; 14:291–294. [PubMed: 2672452]
31. Dill KA, Fiebig KM, Chan HS. *PNAS.* 1993; 90:1942–1946. [PubMed: 7680482]
32. Voelz VA, Dill KA. *Proteins.* 2007; 66:877–888. [PubMed: 17154424]
33. Macario AJL, Conway de Macario E. *FEBS Lett.* 2007; 581:3681–3688. [PubMed: 17475257]
34. Lin Z, Rye HS. *Crit. Rev. Biochem. Mol. Biol.* 2006; 41:211–239. [PubMed: 16849107]
35. Merlin AB, Sherman MY. *Int. J. Hyperthermia.* 2005; 21:403–419. [PubMed: 16048838]
36. Bosl B, Grimminger V, Walter S. *J. Struct. Biol.* 2006; 156:139–148. [PubMed: 16563798]
37. Estrada, LD.; Yowtak, J.; Soto, C. *Protein Design: Methods and Applications.* Guerois, R.; de la Paz, M., editors. New York: Springer-Verlag; 2006. p. 277-294.
38. Dill KA. *Biochemistry.* 1985; 24:1501–1509. [PubMed: 3986190]
39. Bryngelson JD, Wolynes PG. *PNAS.* 1987; 84:7524–7528. [PubMed: 3478708]
40. Dill KA, Chan HS. *Nat. Struct. Biol.* 1997; 4:10. [PubMed: 8989315]
41. Leopold PE, Montal M, Onuchic JN. *PNAS.* 1992; 89:8721–8725. [PubMed: 1528885]
42. Wolynes PG. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 2005; 363:453–457.
43. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. *Curr. Opin. Struct. Biol.* 2007; 17:342–346. [PubMed: 17572080]

44. Pearl LH, Prodromou C. *Annu. Rev. Biochem.* 2006; 75:271–294. [PubMed: 16756493]
45. Yoshida H. *FEBS J.* 2007; 274:630–658. [PubMed: 17288551]
46. Cornish PV, Ha T. *ACS Chem. Biol.* 2007; 2:53–61. [PubMed: 17243783]
47. Uversky VN, Kabanov AV, Lyubchenko YL. *J. Proteome Res.* 2006; 5:2505–2522. [PubMed: 17022621]
48. Zagrovic B, Snow CD, Shirts MR, Pande VS. *J. Mol. Biol.* 2002; 323:927–937. [PubMed: 12417204]
49. Butterfoss GL, Kuhlman B. *Annu. Rev. Biophys. Biomol. Struct.* 2006; 35:49–65. [PubMed: 16689627]
50. Goodman CM, Choi S, Shandler S, Degrado WF. *Nat. Chem. Biol.* 2007; 3:252–262. [PubMed: 17438550]
51. Elofsson A, von Heijne G. *Annu. Rev. Biochem.* 2007; 76:125–140. [PubMed: 17579561]
52. Moult J. *Curr. Opin. Struct. Biol.* 2005; 15:285–289. [PubMed: 15939584]