PROTEINS: Structure, Function, and Genetics 19:1–3 (1994)

FUTURE DIRECTIONS

Protein Folding: Predicting Predicting

George D. Rose and Trevor P. Creamer
Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110

Key words: protein folding, protein conformation, Paracelsus award

Prediction is very uncertain,
especially about the future
Niels Bohr

In the early 1960s, a well-known computer scientist made a very public wager that before the decade's end the world chess champion would be a computer. It was an optimistic bet, and he lost—more than once. Predictions have a way of coming back to twit would-be seers. Nonetheless, undeterred by sensible examples and egged-on by our gadfly editor, we are knowingly crawling out to the C-terminus of a precarious limb to predict that the problem of predicting protein conformation from the amino acid sequence will be solved by this decade's end, leading to both a chronological and intellectual millennium for protein folders.

What prompts such a prediction, apart from an errant daredevil tendency? After all, many first-rate minds have tangled with the folding problem during the preceding half century, yet we are still unable to reliably predict any single aspect of structure from sequence. So, is our optimism just another exercise in wishful thinking, spurred on perhaps by the human genome project?

Actually, we are emboldened by recent, dramatic advances in methodology: X-ray crystallography, NMR spectroscopy, mutagenesis, and computing. The number of new, high-resolution structures solved every year continues to grow exponentially,[1] along with the ability to visualize, probe, manipulate, and, in a very real sense, *understand* them.

However, over and above any specific advances, the morphology of the folding problem has changed in at least two important respects. One of these concerns the local/global conundrum. Global characteristics are conspicuous in globular proteins. The protein core is comprised of residues that are distant in sequence but close in space, suggesting a comprehensive architectural plan. Typically, this core is exquisitely packed, with few voids of atomic dimension, despite the lumpy, idiosyncratic shape of its constituent side chains. A ready analogy for such packing is the assembly of a 3-dimensional jigsaw puzzle, a familiar paradigm, and one that conveys an implicit message of global constraint. Indeed, the central verity of protein folding is the existence of a unique native conformation, a fact that underscores the puzzle analogy. Again, in numerous folding studies, all probes sense a single two-state folding ↔ unfolding transition. These and other examples are persuasive evidence for a global view of the folding process.

But surprisingly, a more local view of folding events is now emerging. Nowhere is this more evident than in mutational studies. Single-residue mutations in proteins generally result in highly localized structural change, demonstrating that the molecule is plastic enough to accommodate local deformation without global reorganization. A short while ago, before such studies were commonplace, it seemed plausible that the impact of a mutation, especially a core mutation, would ripple throughout the protein. To be sure, mutational studies of nature—in the globins, for example—did not support this view, but one could argue that these had been carefully vetted by evolution. That view is no longer tenable. Even outrageous mutations, like excavating a benzine-sized cavity within the core, fail to induce global rearrangement.[2]

Local folding represents an enormous simplification. If the primary determinants of conformation are realized locally, then the massive complexity of the folding problem can be reduced substantially by factoring it into smaller, more tractable problems. For example in secondary structure prediction, the validity of local folding translates into the extent to which helices are predictable from the local sequence, independent of their eventual three-dimensional microenvironment in the protein. If local fold-

Address reprint requests to Dr. George D. Rose, Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, Box 8231, 660 S. Euclid Avenue, St. Louis, MO 63110.

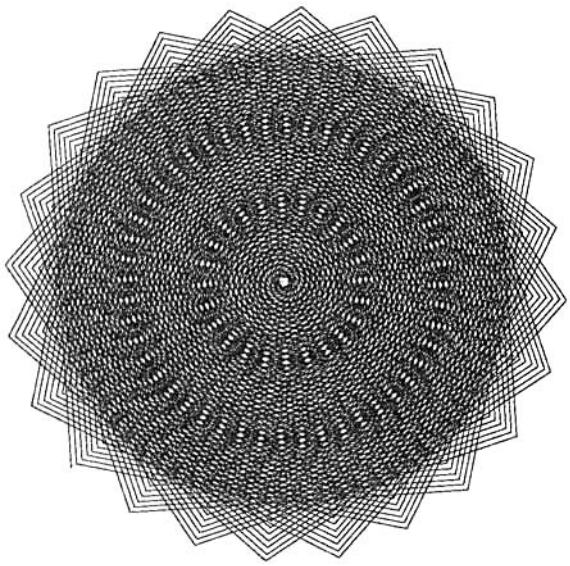

Fig. 1. An antimacassar generated by iteration of a regular, open polygon, with an angle of 75° and a smoothly increasing radius.

ing predominates, then local sequence effectively determines secondary structure which, in turn, fosters a suitable three-dimensional environment. If not, then the three-dimensional environment influences secondary structure to some significant degree, and a given sequence has the potential to toggle between alternative formats in different proteins. Indeed, Kabsch and Sander[3] have identifed sequences of up to five residues that are helical in one molecule but non-helical in some other molecule. However, in their examples, the residues in question may not be the ones that specify secondary structure.[4] The best current algorithms to predict secondary structure are accurate about 65% of the time, with ongoing debate over whether their intrinsic perfectibility is hampered by global limitations.

Global appearances can be deceiving. It is not hard to devise strictly local constructs with an ostensible global affect. The antimacassar in Figure 1. seems to consist of two different stitches, woven together in a global design. In fact, the pattern was created by repetition of a single, periodic rule.

How can the apparent global complexity of protein folding be engendered by a predominantly local folding process? It is known that protein molecules are organized as a structural hierarchy—at any level the structure can be successively decomposed into smaller, wholly contained units.[5,6] The molecular hierarchy of proteins traverses a now familiar route, ranging in size from the entire molecule through conventional domains and supersecondary structure to secondary structure and, ultimately, individual residues. This architecture immediately suggests a model of folding by hierarchic condensation[5–8] that recasts the observed hierarchy in a temporal sequence: neighboring chain sites interact to form small folding modules, with the accretion of structure resulting from stepwise association, leading, in turn, to secondary structure, supersecondary structure, domains and the folded protein. A folding process that proceeds by hierarchic condensation would rationalize the local/global conundrum because each folding event is a local one at some step in the hierarchy. A similar idea has been proposed by Dill.[9]

Another important area in which the current perspective is shifting might be called the stability/specificity conundrum.[10] That is, factors that stabilize a folded protein against denaturation are not synonymous with the factors that determine the fold itself. By way of macroscopic analogy, both your house and my house have similar stabilizing structural features to keep them from collapsing under the influence of gravity or the weather, but the houses are not alike. In general, work done to fortify my house does not change it to resemble your house. Extending the analogy to proteins, the question of why a lysozyme molecule remains folded at physiological kT is one of stability, while the question of why the lysozyme sequence adopts the lysozyme fold and not, for example, the ribonuclease fold is one of specificity.

The distinction between stability and specificity is well accepted in DNA studies.[11] There, stability factors that promote strand association (e.g., base stacking) are distinguished from the specificity factors that underlie base pair complementarity (e.g., hydrogen bonding). (Nevertheless, it should be emphasized that the issues remain coupled: Watson-Crick base pairing is imposed only in the context of double stranded DNA, and when this constraint is relaxed, the complementarity rules degenerate.)

Life is metastable, at best. The pen that writes these words will oxidize with time, together with the hand that guides it, achieving greater stability at the expense of biological function. However, the kinetics of oxidation are slow, and the drive toward ultimate thermodynamic stability is negligible in day-to-day biological processes.

In this vein, studies of protein stability have elucidated many important aspects of protein thermodynamics. But they do not reveal the underlying stereochemical code by which sequence determines conformation, any more than a computer's instruction repertoire can be deduced from the heat capacity of its components.

The issue of specificity is addressed more directly by efforts to design proteins de novo. For example, experiments aimed at devising sequences that fold into helical bundles[12,13] have met with remarkable success. A critic might argue that these initial efforts did not fully satisfy their design specifications, but, to paraphrase Dr. Johnson, the remarkable thing is that they fold at all.

Similarly, assumptions about specificity are implicit in recent work on the inverse folding problem.[14–17] Here the goal is to identify sequences that are compatible with a known fold. Presumably, among sequences inconsistent with the chosen fold are those that promote other folded alternatives.

The issue of specificity, as distinct from stability, can be cast in the following terms. Starting with two proteins of known structure, having similar composition and sequence length but differing folds—lysozyme and ribonuclease, for example—find the minimum number of residue substitutions needed to switch one conformation to the other. Demonstrably, a solution exists since, in the worst case, every residue could be changed. The interesting aspect of the question is whether the conformation can be switched by substitution of only a small subset of suitably chosen residues. To focus attention on this question, we have established the Paracelsus Challenge,[18] a one-time prize of $1000, to be awarded to the first individual or group that successfully transforms one globular protein's conformation into another by changing no more than half the sequence.[19]

Time and again, the history of science teaches us that what we see depends upon the perspective from which we look. For protein folders, it is a time of rapidly evolving perspective.

ACKNOWLEDGMENTS

We thank Rajeev Aurora, Eaton Lattman, Timothy Lohman, Jeffrey W. Seale, and Rajgopal Srinivasan for formative discussions and the NIH (GM29458) for support.

REFERENCES

1. Lattman, E.E. Protein crystallography for all. Proteins 18: 103–106, 1994.
2. Eriksson, A.E., Baase, W.A., Zhang, X.-J., Heinz, D.W., Blaber, M., Baldwin, E.P., Matthews, B.W. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. Science 255:178–183, 1992.
3. Kabsch, W., Sander, C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci USA 81:1075–1078, 1984.
4. Presta, L.G., Rose, G.D. Helix signals in proteins. Science 240:1632–1641, 1988.
5. Crippen, G.M. The tree structural organization of proteins. J. Mol. Biol. 126:315–332, 1978.
6. Rose, G.D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447–470, 1979.
7. Oas, T.G., Kim, P.S. A peptide model of a protein folding intermediate. Nature 336:42–48, 1988.
8. Rose, G.D., Wolfenden, R. Hydrogen bonding, hydrophobicity, packing and protein folding. Annu. Rev. Biophys. Biomol. Struct. 22:381–415, 1993.
9. Dill, K.A., Fiebig, K.M., Chan, H.S. Cooperativity in protein-folding kinetics. Proc. Natl. Acad. Sci. USA 90:1942–1946, 1993.
10. Lattman, E.E., Rose, G.D. Protein folding—what's the question? Proc. Natl. Acad. Sci. USA 90:439–441, 1993.
11. Spolar, R.S., Record, M.T., Jr. Coupling of local folding to site-specific binding of proteins to DNA. Science 263:777–784, 1994.
12. Regan, L., DeGrado, W.F. Characterization of a helical protein designed from first principles. Science 241:976–978, 1988.
13. Hecht, M.H., Richardson, J.S., Richardson, D.C., Ogden, R.C. De novo design, expression, and characterization of Felix: A four-helix bundle protein of native-like sequence. Science 249:884–891, 1990.
14. Bowie, J.U., Luthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
15. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature 358:86–89, 1992.
16. Sali, A., Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234:779–815, 1993.
17. Godzki, A., Skolnik, J. Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc. Natl. Acad. Sci. USA 89:12098–12102, 1992.
18. Names after Paracelsus, the 16th century Swiss (then Austria) physician, whose interest in alchemy led him to chemistry. Paracelsus is usually credited with being the father of pharmaceutical chemistry and modern medicine.
19. The prize will be held in escrow by one of us (G.D.R.) and judged by the current Editor-in-Chief of Proteins, Eaton E. Lattman, who will be the sole arbiter of success.