Adding backbone to protein folding: why proteins are polypeptides

Barry Honig[1] and Fred E Cohen[2]

It is argued that the chemical nature of the polypeptide backbone is the central determinant of the three-dimensional structures of proteins. The requirement that buried polar groups form intramolecular hydrogen bonds limits the fold of the backbone to the well known units of secondary structure while the amino acid sequence chooses among the set of conformations available to the backbone. 'Sidechain-only' models, based for example on hydrophobicity patterns, fail to account for the properties of the backbone and thus will have difficulty capturing essential features of a folding pathway. This is evident from the incorrect predictions they make for the conformations of the limiting cases of all-hydrophobic or all-polar sequences.

Addresses: [1]Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA. [2]Department of Cellular and Molecular Pharmacology, Box 0450, San Francisco, CA 94143-0450, USA.

Correspondence to: Barry Honig
E-mail address: Honig@Bass.Bioc.Columbia.Edu

Folding & Design 01 Feb 1996, 1:R17–R20

Anticipating a protein's fold presents a classic challenge: the problem is easy to state, the existence of an answer is clear, yet a solution remains elusive. A solution to the folding problem, particularly an elegant one, might be comparable in impact to the discovery of the structure of the double helix. The implications for biology (and for the scientists who solve the problem) are difficult to minimize. One can conceive, for example, of a day when a three-dimensional protein structure (even an approximate one) will be associated with each newly sequenced gene and when these structures will be exploited in generating new disease models, in discovering new cellular pathways, and in the design of new pharmaceuticals. The more esoteric and purely intellectual excitement associated with protein folding is evident from its ability to attract researchers from the biology, chemistry, physics and even mathematics communities. In recent months, moreover, protein folding has been featured in a leading newspaper [1] as well as in scientific publications oriented to the more general audience [2,3].

Given the widespread fascination with protein folding and the hazards of unrealistic expectations, it would appear important for the field as a whole to agree on objective standards for structure prediction. The standards we refer to include statistical tests, the use of adequate sample size to assess success and failure, rms deviations from known structures, and the requirement that 'predicted' structures be clearly identifiable with a criterion that does not require knowledge of the final answer. Even when such standards are applied, it is important to ensure that the criteria have been applied properly [4] and that the result might not have been derived with a nonsense 'control' sequence [5]. Indeed, the Asilomar Conference [6] on structure prediction provides an amusing yet valuable means for maintaining our modesty.

Given the complex molecular structure of proteins, researchers have for many years sought simplified descriptions that still maintain the essential features required for a proper description of folding [7]. Recently, there has been a resurgence of simplified models that have been used as a basis for structure prediction as well as for discussions of the general features of folding pathways. In this commentary, we describe a 'gedanken control experiment' that allows us to evaluate the validity of one class of simplified models. We also offer a perspective on the unique ability of proteins to form a wide variety of stable three-dimensional structures. In particular, we consider the relative roles of the polypeptide backbone and sidechain–sidechain interactions as crucial determinants of protein structure.

Two central issues that arise in protein folding concern the origin of the thermodynamic stability of the native state given the enormous number of other possible conformations [8] and the mechanism by which the chain locates the native structure on a biologically appropriate time scale [9] (the Levinthal paradox). Stated simply: how and why does a polypeptide sequence arrive at its unique three-dimensional structure? A focus on the role of the backbone echoes the work of Pauling [10] and Ramachandran [11] and emphasizes the importance of the polypeptide backbone, with a sidechain R group viewed as an appendage (R can represent one of the 20 naturally occurring amino acids, or any other chemical group). The physico-chemical properties of the backbone alone led to the correct prediction of the role of α-helices and β-sheets in protein architecture [10].

More recent simplifications of protein structure have accentuated the importance of sidechains, and sidechain–sidechain interactions, reducing the polypeptide backbone's role to mere connectivity, as for example in 'sidechain-only' lattice models [12–14]. Sidechain-only

models implicitly neglect the fact that the protein backbone is a string of amide units that are chiral at the C^{α} position. Indeed, the backbone representation would be the same for a repeating sequence of polyethylene units, or nucleic acids. These models have led to intriguing, but often controversial, proposals about the nature of protein folding pathways [12,15].

One of the fundamental goals of recent theoretical studies has been to understand why proteins are 'special' in the sense that ordinary polymers do not form unique three-dimensional structures, whereas polypeptides do. As described in a recent review [14], sidechain-only models imply that proteins are special among polymers "because the amino acids in proteins are linked in specific sequences." Taken literally, if two identical sidechain sequences were attached to two chemically distinct backbones, similar three-dimensional architectures would result.

In this commentary, we suggest that proteins are special because they are polypeptides [16] and that sidechain-only models fail to capture essential features of protein folding. We note in this regard that the rate-limiting step in protein unfolding appears to be due to the breaking of hydrogen bonds rather than to the loss of sidechain interactions [17]. The importance of the chemical nature of the backbone is also evident from the different properties of DNA and RNA, in that the structural and biological impact of a single hydroxyl group in the backbone is profound.

We begin by considering the limiting case of amino acid homopolymers, a classic topic of biophysical and statistical mechanical studies. The simplest representation of a sidechain-only model for a polypeptide exploits a lattice with only two classes of amino acids, non-polar (H for hydrophobic) and polar (P) [14]. Homopolymers consisting of n polar or non-polar amino acids will be referred to a P_n and H_n respectively. In the energy functions typically used, H–H contacts are always treated as attractive, whereas P–P contacts are destabilizing or of no consequence. In Figure 1a, we depict the standard representation of an α-helix on a cubic lattice (left) and contrast it with a more compact, cubic arrangement of the chain (right). For the helix, the number of contacts is linear with chain length (n) and is exactly n–3 when n is divisible by 4. When n segments can be placed in a perfect cubic arrangement, the total number of contacts is $2n–3n^{2/3} + 1$.

Figure 1b plots the number of contacts for a helix and a cube. It is clear that for H_n, where incremental stabilization is imparted by each H–H contact, helices on lattices will always be less stable than their compact counterparts. Moreover, the distinction will grow as the chain lengthens. A different behaviour is expected for P_n sequences. These would be predicted by sidechain-only models to form heterogeneous extended structures, as they lack stabilizing sidechain–sidechain interactions.

In contrast to these predictions, early spectroscopic studies of the conformation of poly-L-amino acids demonstrated that these materials can form periodic structures in aqueous solution [18]. Some polar sidechains form α-helices in water; for example, polylysine is α-helical at high pH when its amino group is neutral [19]. Solubility problems limit the study of non-polar polyamino acids, but helical propensities can be studied in block copolymers where solubilizing groups are included in the polypeptide chain. Studies of block copolymers have shown that non-polar amino acids form helices in aqueous solution, some of them with quite high helical propensities [20]. Moreover, helix stability increases with the length of the polypeptide chain. Although simple one-dimensional Ising models [21,22] can describe helix formation, it is clear that sidechain-only models will not reproduce this behaviour. Polyalanine, H_n, forms an α-helix, but is predicted by sidechain-only models to form an array of compact structures. Polylysine, P_n, also forms an α-helix, but is predicted to form an array of extended structures. Thus, models stabilized strictly by sidechain contacts fail to account for these two limiting cases.

The failure of sidechain-only models can be attributed directly to the neglect of the chemical nature of the polypeptide backbone. The essential feature of the backbone is usually described in terms of its ability to form intermolecular hydrogen bonds, but the role of the peptide group is somewhat more subtle. As has been pointed out recently [23,24], a crucial property of the polypeptide backbone is that it contains polar NH and CO groups whose removal from water involves a significant energetic penalty. Sidechain-only models with non-polar residues that fail to account for this penalty will, as shown in Figure 1b, necessarily favour compact conformations, as these can be formed at no energetic cost to the backbone. In reality, many compact conformations that appear stable in sidechain-only lattice models will be energetically inaccessible due to burial of polar groups.

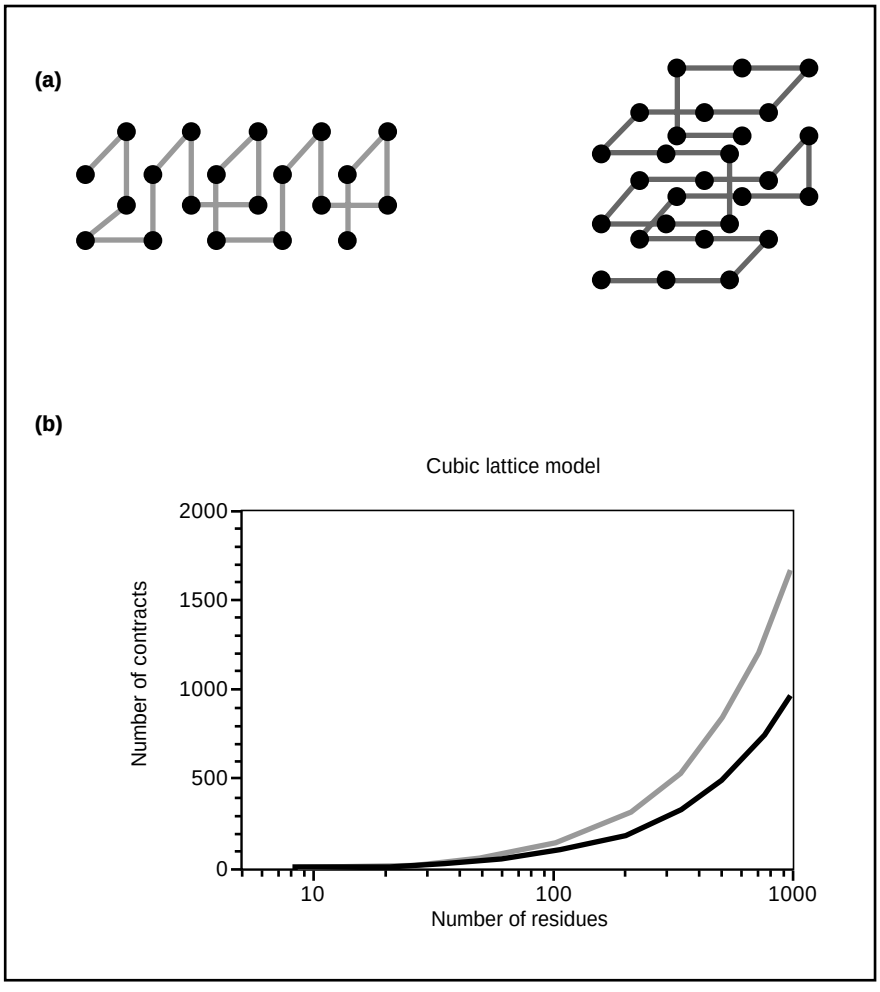
In our view, a realistic description of protein folding must account for the fact that polypeptides minimize the effects of removing polar groups from water by forming intramolecular hydrogen bonds. Buried hydrogen bonds may also be unstable relative to isolated donors and acceptors in the unfolded state, but the energetic penalty is small enough to allow the hydrophobic effect, which provides a driving force for compactness, to dominate [23,24]. These energetic principles suggest that the requirement that polar groups must either form intramolecular hydrogen bonds, or form hydrogen bonds with water, provides a strong structural constraint which excludes many of the confor-

mations that would be predicted to be stable by sidechain-only models. This point is nicely illustrated by α-helix formation, which appears driven in part by hydrophobic interactions but where the requirement of forming hydrogen bonds leads to the formation of a helical rather than a compact conformation [23–25]. A number of lattice-based models do take into account the structural and energetic constraints due to the backbone, suggesting that these features can, in principle, be incorporated into simplified models of protein folding [26–31].

As α-helices appear to solve the Levinthal paradox through a nucleation/propagation mechanism [21,22,32], there is every reason to believe that folded proteins solve the conformational search problem based in part on the same mechanism. This suggests that secondary structure is a component of the earliest stages of protein folding. However, this is not equivalent to stating that helices in proteins are stabilized only by local interactions. Rather, in keeping with the classical view [8], the picture that emerges is that early stages in protein folding involve the transient formation or 'flickering' of units of secondary structure which are primarily stabilized by a combination of long-range and local interactions that are primarily hydrophobic. This is because the best way for polypeptides to satisfy the hydrophobic effect is to form units of secondary structure. In this sense, secondary and tertiary structure are expected to appear simultaneously.

The central conclusion of this commentary is that proteins are 'special' primarily because they are polypeptides. As stated elegantly over 30 years ago by Doty and Gratzer [16], "it would seem extraordinary that no other polymer structures exist in which internal hydrogen bonding can give rise to periodically ordered conformations, but no others have been found thus far. We are therefore forced to recognize the uniqueness of this capacity in polypeptide chains, one which enables them to meet the exacting and sophisticated demands of protein structure and function." Of course, this 'backbone-centric' view of polypeptide conformation does not negate the obvious importance of amino acid sequence in determining protein tertiary

Figure 1



(a) Depiction of structures formed by a 20-mer on a cubic lattice. Dark circles correspond to amino acids placed at lattice points. Left, α-helix; right, an example of a compact state. (b) Plot of number of contacts versus chain length for the two structures shown in (a). The upper curve is for the compact structure and the lower curve is for the α-helix.

structure. However, rather than argue that sequence determines structure, it might be more accurate to state that sequence chooses among the limited set of secondary structure possibilities available to the polypeptide backbone and determines how they are combined to produce a unique three-dimensional structure.

The relative importance of sidechain and backbone contributions to protein stability and the stability of polymers in the aqueous phase has yet to be tested in detail. However, recent advances in solid phase synthetic organic chemistry have yielded sequence-specific heteropolymers with alternative backbones (e.g. poly N-substituted glycines [33], poly carbamates [34] and peptide backbones with nucleic acid sidechains [35]) that should facilitate such tests. Given that specific RNA sequences fold to form unique structures, for example, sidechain-only models would predict that peptide nucleic acid versions of these RNAs should adopt structures that are similar to their RNA cognates whereas a perspective that emphasizes the role of the backbone would predict that these molecules would behave more as polypeptides. Both predictions can be tested.

References

1. Brown, D. (1995). *Washington Post October 1.* p. A1.
2. Holden, C., ed. (1995). Folding proteins fast. *Science* **269**, 1821.
3. Borman, S. (1995). Researchers advance ability to predict structures of folded proteins. *Chem. Eng. News November 6.* pp. 44–49.
4. Cohen, F.E. & Sternberg, M.J.E. (1980). On the prediction of protein structures: the significance of the root-mean-square deviation. *J. Mol. Biol.* **138**, 321–333.
5. Hagler, A.T. & Honig, B. (1978). On the formation of protein tertiary structures on a computer. *Proc. Natl. Acad. Sci. USA* **75**, 554–558.
6. Lattman, E.E., ed. (1995). Protein structure prediction issue. *Proteins* **23**, 295–462.
7. Levitt, M. & Warshel, A. (1975). Computer simulation of protein folding. *Nature* **253**, 694–698.
8. Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223–230.
9. Levinthal, C. (1966). Molecular model-building by computer. *Sci. Am.* **214**, 42–52.
10. Pauling, L., Corey, R.B. & Branson, H.R. (1951). The structure of proteins: two hydrogen bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* **37**, 205–211.
11. Venkatchalam, C.M. & Ramachandran, G.N. (1969). Conformation of polypeptide chains. *Annu. Rev. Biochem.* **38**, 45–82.
12. Sali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold? *Nature* **369**, 248–251.
13. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. & Wolynes, P.G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195.
14. Dill, K.A., *et al*., & Chan, H.S. (1995). Principles of protein folding – a perspective from simple exact models. *Protein Sci.* **4**, 561–602.
15. Chan, H.S. (1995). Kinetics of protein folding [letter]. *Nature* **373**, 664–665.
16. Doty, P. & Gratzer, W.B. (1962). Some recent observations on polypeptides in solution. In *Polyamino Acids, Polypeptides and Proteins.* (Stahman, M.A., ed.), pp. 111–118, University of Wisconsin Press, Madison, WI.
17. Kiefhaber, T., Labhardt, A.M. & Baldwin, R.L. (1995). Direct NMR evidence for an intermediate preceding the rate-limiting step in the unfolding of ribonuclease A. *Nature* **375**, 513–515.
18. Auer, H.E. & Doty, P. (1966). The synthesis, structure, and optical properties of some copolypeptides containing nonpolar amino acid residues. *Biochemistry* **5**, 1708–1715.
19. Applequist, J. & Doty, P. (1962). α-Helix formation in poly-ε-carbobenoxy-L-lysine and poly-L-lysine. In *Polyamino Acids, Polypeptides and Proteins.* (Stahman, M.A., ed.), pp. 161–177, University of Wisconsin Press, Madison, WI.
20. Ingwall, R.T., Scheraga, H.A., Lotan, N., Berger A. & and Katchalski, E. (1968). Conformational studies of poly-L-alanine in water. *Biopolymers* **6**, 331–368.
21. Zimm, B.H. & Bragg, J.K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys*. **31**, 526–535.
22. Lifson, S. & Roig, A. (1961). On the theory of helix–coil transition in polypeptides. *J. Chem. Phys.* **34**, 1963–1974.
23. Honig, B. & Yang, A.S. (1995). Free energy balance in protein folding. *Adv. Protein Chem.* **46**, 27–58.
24. Yang, A. & Honig, B. (1995). Free energy determinants of secondary structure formation. 1. Alpha-helices. *J. Mol. Biol.* **252**, 351–365.
25. Bixon, M., Scheraga, M.A. & Lifson, S. (1963). Effect of hydrophobic bonding on the stability of poly-L-alanine helices in water. *Biopolymers* **1**, 419–429.
26. 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.
27. Hao, M.-H. & Scheraga, H.A. (1994). Statistical thermodynamics of protein folding–sequence dependence. *J. Phys. Chem*. **98**, 9882–9893.
28. Sun, S., Thomas, P.D. & Dill, K. (1995). A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng.* **8**, 769–778.
29. Wolynes, P.G., Onuchic, J.N. & Thirumalai, D. (1995). Navigating the folding routes. *Science* **267**, 1619–1620.
30. Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z. & Socci, N.D. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. USA* **92**, 3626–3630.
31. Luthey-Schulten, Z., Ramirez, B.E. & Wolynes, P.G. (1995). Helix–coil, liquid crystal and spin glass transitions of a collapsed heteropolymer. *J. Phys. Chem.* **99**, 2177–2185.
32. Zwanzig, R., Szabo, A. & Bagchi, B. (1992). Levinthal's paradox. *Proc. Natl. Acad. Sci. USA* **89**, 20–22.
33. Simon, R.J., *et al*., & Huebner, V.D. (1992). Peptoids–a modular approach to drug discovery. *Proc. Natl. Acad. Sci. USA* **89**, 9367–9371.
34. Cho, C.Y., *et al*., & Sundaram, A. (1993). An unnatural biopolymer. *Science* **261**, 1303–1305.
35. Egholm, M., Buchardt, O., Nielsen, P.E. & Berg, R.H. (1992). Peptide nucleic acids (PNA). Oligonucleotide analogues with an ahiral peptide backbone. *J. Am. Chem. Soc.* **114**, 1895–1897.