Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4028–4030, June 1985
Biochemistry

Is there a single pathway for the folding of a polypeptide chain?

(native-like structure/microdomains/jigsaw-puzzle analogy/protein conformation/protein renaturation)

STEPHEN C. HARRISON AND RICHARD DURBIN*

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, MA 02138

Communicated by Frederic M. Richards, February 22, 1985

ABSTRACT **We argue that folding of the compact domains of proteins can occur with adequate rapidity in the absence of a unique directed mechanism, provided that native-like local structure dominates the folding process. We further suggest that the evolution of amino acid sequences should favor multiple paths to the folded state. Existing physicochemical and mutational data are not inconsistent with a many-pathway model. The analogy of a jigsaw puzzle, with multiple routes to a unique solution, appears to be particularly apt.**

A calculation discussed originally by Levinthal (1) shows that a folding polypeptide chain cannot in reasonable time assume all possible conformations. It is therefore often concluded that proteins fold along a defined pathway. Various experimental approaches to discover such a pathway have been applied to staphylococcal nuclease, bovine pancreatic trypsin inhibitor, ribonuclease, and other proteins (see ref. 2 for review), but present methods do not give structural details of the small number of distinguishable kinetic intermediates. We suggest here that a requirement for a defined pathway does not follow from the large number of available conformational states. We emphasize that the argument applies only to acquisition of the native folded structure by proteins with a single compact domain or by the individual domains of more complex structures. This restriction indeed defines what we mean by "folding." The additional levels of organization in multidomain chains and in multisubunit structures clearly require a distinction between folding and assembly, and definite temporal sequences at this higher level of structure do indeed occur. Moreover, it is likely that long polypeptide chains can fold domain-by-domain during biosynthesis, and this "quantized" *in vivo* polarity may account for difficulty often experienced in refolding multidomain structures. But numerous experiments with fragments of small proteins (e.g., ref. 3) rule out ordered amino-to-carboxyl-terminal folding within a single domain.

The following argument makes it plausible to think that proteins fold by large numbers of quite different, parallel pathways rather than by a single defined sequence of events. The argument is based on demands placed on the amino acid sequence by a requirement for particular intermediates. Suppose that a protein folds according to a unique sequential pathway. The intermediates in this pathway need not be very stable or very well ordered, but their structure(s) will be defined by properties of the amino acid sequence. The stability of these intermediates will therefore be subject to mutational variation, and relatively small changes might be expected to have large effects on folding kinetics. Now suppose, in contrast, that quite different pathways are possible in going from the denatured to the native state. Particular mutations, unless they strongly destabilize the native structure, are much less likely to disturb the kinetics of folding, since appearance of a barrier in one pathway need not imply a barrier in another. Extension of this argument suggests that multiple pathways would appear rapidly during evolution of a protein sequence. Mutation leading to appearance of a new route to the native state would make effective folding significantly more robust to further mutation, giving such a sequence strong selective advantage. The native structure could evolve toward optimal function or control, without interference from the effects such changes would have on the kinetics or process of folding.

Several examples show that a disorder-to-order transition with a defined end point and an astronomically large number of states can readily occur in finite time without there being a unique pathway. Consider growth of a single, already nucleated crystal from a supersaturated solution. For example, suppose that the crystallizing molecule is a spherical protein 100 Å in diameter. If we divide the total volume into 100 Å cubes, 1 cm^3 of a 10 mg/ml solution contains about 1000 times as many cubes (10^{18}) as protein molecules (10^{15}), and the number of possible translational configurations is about $e^{10^{16}}$. Growth of the crystal clearly does not occur by a search of all these states, but neither does it occur by a particular pathway. The process occurs in reasonable time because once a molecule has added to the crystal, there is only a small chance that it will dissociate. The dimensionality of the next step in the search is therefore decreased.

An example somewhat closer to protein folding, because it involves a nonrepetitive structure, is the solution of a jigsaw puzzle. The route to the completed puzzle will be different each time it is solved (especially if the rim is not uniquely defined by a smooth edge). Again, a key feature is the unlikelihood of disassembling a correct bit of local structure once it has been made. The rate of local disassembly need not be zero: you can complete a puzzle even if someone else keeps taking the pieces apart, provided that you work faster than your opponent. Since local, native-like structures smaller than a protein domain are unstable, this competition picture is a better analogy.

The diffusion–collision model of Karplus and Weaver (4) corresponds fairly closely to the jigsaw-puzzle-with-competition analogy. Their calculation shows that if local, native-like structures ("microdomains") have a lifetime comparable to or greater than segmental diffusion time and if coalescence of these microdomains gives structures with somewhat longer lifetimes, then folding of a 100- to 200-residue protein can occur in 1 sec or less. Their model does not require a particular sequence in which microdomains coalesce, nor does it require a uniquely defined set of microdomains for a given protein.

These two questions—are there defined microdomains for a given structure and, if so, is there sequential assembly—can be illustrated by the α-helix-assembly model for myoglobin put forward by Ptitsin and Rashin (5). In this model, preformed α-helical units pack sequentially to form the folded structure. The argument offered here suggests that even if

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*Present address: Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH U.K.

such a picture were roughly correct, the order in which particular helices formed and packed need not be determined and that the packing of some helices might occasionally occur before others have formed. Given the very large number of known sequences for different globins, it is indeed difficult to imagine that there can be a unique rank-order for the stability of the given helices. If various helix-packing sequences are possible, a mutation that somewhat destabilizes a particular helix but still permits the native structure will yield a globin that folds at an acceptable rate. That is, with all proteins—not just globins—a particular folded structure corresponds to many possible sequences. Even if one subset of folding routes dominated with one such sequence, a different subset might characterize folding of another. The myoglobin example also shows that even if structurally unique intermediates do not occur, kinetic intermediates can be detected: if α-helices are the microdomains of myoglobin, amide protons in α-helices will be protected from exchange "early" in Ptitsin-type folding (5). Likewise, if one averaged over the many paths actually taken in a series of efforts to assemble a given jigsaw puzzle, "blue sky" might appear earlier than "trees" or "grass" because of easily recognized characteristics. But there would be no absolute requirement to start with the sky, only a statistical tendency.

Few structures are as regular as myoglobin. The relative stability and cooperativity of α-helices focus attention on these elements as possible folding units, just as in some puzzles, certain subpatterns are easily detected and frequently assembled as "islands" en route to the final solution. It is reasonable to ask, for structures more complicated than myoglobin, how we might recognize the microdomains or, indeed, whether variant sequences that fold to the same structure might have different significant substructures. In terms of the jigsaw-puzzle analogy, proteins of a given folded conformation (e.g., globins) correspond to puzzles cut from a given template, whereas variant sequences correspond to different pictorial patterns. In solving a puzzle, partial solutions generally involve recognizable portions of the picture (e.g., blue sky) rather than conveniently shaped pieces. If the same were true of proteins, then different sequences forming the same folded structure might not have similar substructures while folding. Since one sequence evolves from another, however, there is a correlation between the "picture" and the "shape of the pieces," and we might expect variants of a given protein indeed to have related microdomains. Commonly found motifs, such as the $\beta\alpha$-$\beta\alpha\beta$ "super-secondary structure" first described by Rossmann *et al.* (6), and compact subdomains, such as those revealed by the procedures of Rose (7), Crippen (8), and Wodak and Janin (9), are plausible candidates for such substructures. The correlation between exon boundaries and various sorts of structurally defined demarcations has suggested another way to look for subdomains (10–12). None of these approaches appears to be uniquely compelling at present.

Some observed structures contain hints that they fold by multiple pathways. Richardson (13) showed that the most commonly occurring β-sheet topologies are those that maximize the number of alternative ways in which the sheet can form. For example, it is quite common that neighboring strands are adjacent in the sequence. Indeed, Levitt and Chothia (14) noticed that arrangements of secondary structural elements in proteins are often such that elements adjacent along the polypeptide chain are in three-dimensional contact as well. Lesk and Rose (15) have analyzed in more detail the structural hierarchy of compact subdomains. They have shown that in myoglobin and RNase, there are a number of ways to assemble the native structure by association of these units. They suggest that these alternatives correspond to accessible folding pathways.

An important feature of any picture of protein folding that involves coalescence of transiently formed substructures is that it allows for constant editing. Incorrect local folds (an incorrectly extended α-helix, for example) will dissolve, because they cannot be stabilized by other structures. In this respect, the instability of any part of a domain is important, since it prevents the locking-in of wrongly folded pieces. It is this internal editing, rather than a pathway defined by the amino acid sequence, that makes folding nonrandom.

The essential characteristic of protein folding in the picture we argue for here is that it can proceed by many pathways, all involving accretion of native-like structure. Can this view be reconciled with experimental evidence for folding pathways and for non-native-like intermediates? The most direct attempts to work out a pathway have dealt with bovine pancreatic trypsin inhibitor and with ribonuclease. Creighton (16) has studied the kinetics of formation of the three disulfide bonds in bovine pancreatic trypsin inhibitor, taking advantage of the rapidity of trapping a given configuration by low pH, iodoacetic acid, etc. There is a defined order with which different species are found, including several non-native combinations. The kinetic significance of "incorrect" pairings has been confirmed by selective sulfhydryl-blocking experiments. The apparently obligatory non-native arrangements seem at first to suggest that native-like structure appears only at the end of the pathway. States (17) has examined a number of these forms by proton NMR, and the results indicate that structures fall into two classes—those with spectra having a number of native-like features and those with spectra showing no evidence for chemical shifts due to local tertiary structure. The former category includes some species with a non-native disulfide bridge (between residues 5 and 14 or 5 and 38), suggesting that a reasonable part of the structure, especially the "core" around disulfide 30-51 and some of the β-sheet, is essentially as in native trypsin inhibitor. Such forms all have a defined two-state melting transition. No "new" recognizable structure is seen in forms with incorrect disulfides. The disulfide (30-51, 14-38) species, kinetically blocked from forming the last disulfide, appears to be nearly native in all regions of the chain. States *et al.* (18) have also shown that a form originally thought to be a conformational isomer of the protein (19) is in fact a species not in the originally determined pathway—it has two of the three native disulfide bridges (5-55 and 14-38), and it is entirely native-like in structure except just in the vicinity of the missing disulfide. It is a predominant form at low pH and temperature, and it apparently has buried-SH groups at positions 30 and 51, just opposite each other but unable to oxidize to -S-S- because they are buried in an extremely stable folded structure. These results suggest that a sequence of formation of disulfide bridges does not report overall structural events, since the results show that two species with different subsets of native-like bridges can have nearly identical structures and that much of the native structure can be present even with an incorrect pairing.

In the case of ribonuclease, initial attention focused on fast- and slow-folding forms of the unfolded enzyme. The barrier in the case of the slow-folding species has now been shown to be one or more proline isomerization events (20, 21). An early kinetic intermediate in the folding of the slow species has been detected by protection of amide protons from proton-exchange (22). The existence of intermediates in refolding of RNase S is indicated by the dependence of its kinetics on S-peptide concentration and by the broad, multistate transition of S-protein (23, 24). If refolding of the slow species of RNase A is carried out under "strongly native" conditions, kinetic studies suggest that a refolded, inhibitor-binding species forms with one or more incorrect prolines, which then isomerize (25). This interpretation implies that under conditions of marked stability of the native structure, essentially correct refolding may occur despite some bad "kinks," which work themselves out in due course. An alternative view of the refolding of RNase A from urea has been presented by Lin and Brandts (26), who do not believe that the kinetic data require any populated intermediate states. Either picture is consistent with the conclusion, drawn from States' work on bovine pancreatic trypsin inhibitor, that native structural features dominate the folding process.

It is to be emphasized that the experimental identification of intermediates is so far largely a kinetic rather than a structural description (2). As suggested above, there is nothing in the multiple-pathway, jigsaw puzzle description of folding that rules out kinetically defined intermediates, corresponding to classes of structures rather than to a single set of interactions (e.g., "secondary structural elements are present" or "hydrophobic core tends to form"). An important prediction of the argument at the beginning of this article is that mutants kinetically blocked at a particular stage are unlikely to exist: the native state is stable enough that there are a number of ways around particular barriers. There is a class of bacterial mutants sometimes described as folding lesions. These are the so-called tss (temperature sensitive for synthesis) mutants (27). The phenotype is that no activity occurs at the nonpermissive temperature but that when a culture is transferred from permissive to nonpermissive conditions, activity is retained until diluted out by cell division. A frequently cited interpretation is that once folded, such polypeptides are stable, but that the process of folding is impaired. The gene for which such analysis has been most thoroughly carried out is the *I* gene in *Escherichia coli*, which codes for the *lac* repressor. This protein is active as a tetramer, and an alternative interpretation is that once oligomerized, the chains are stable, but that the individual folded chains are of marginal stability and are readily degraded. This is indeed the interpretation originally put forth by Sadler and Novick (27). A similar phenotype occurs in some ts mutants of phage P22 tail-spike protein (28). The assembly of a trimeric tail spike is blocked at nonpermissive temperature, but the mutant spikes are not thermolabile. Since the mutants are all defective in trimer formation, a plausible explanation is that trimer contacts stabilize a folded structure that is unstable at nonpermissive temperatures in the mutant forms. This structure might, for example, be a domain whose folding is essential for trimerization.

Another class of mutations sometimes described as yielding incorrectly folded proteins are those whose products are particularly sensitive to cellular proteases. An alternative explanation for such properties is that the variant proteins are less stable than wild type and that the conditions leading to degradation are relatively close to their unfolding transition. Changes at spatially distant positions in the structure might then suppress the effect of the original mutation, if they produced a compensating enhancement of stability. Beginning with a large number of apparently proteolytically sensitive variants of staphylococcal nuclease, Shortle and Lin (29) have isolated just such second-site revertants. One particular change suppresses most of the unstable variants examined, which correspond to alterations widely distributed over the folded structure. These results are consistent with the hypothesis that mutations affecting a given local structure do not lead to large kinetic barriers. Were the effect of substitutions on protease-sensitive phenotype mediated by loss of local, non-native interactions, then a single second-site change would not be expected to suppress them all.

Summary. We have argued (*i*) that the evolution of a protein sequence should favor multiple routes to the folded structure of a compact domain; (*ii*) that folding by growth of native-like structure can occur with adequate rapidity in the absence of a unique directed mechanism; and (*iii*) that existing physicochemical and mutational data are not inconsistent with a multipathway model. One important consequence of such a model is that mutant sequences cannot be used to define a folding pathway. Another is that aspects of the stability of the native structure, and of native-like substructures, determine what selection from the ensemble of possible pathways actually occurs in the folding of a protein. Thus, in order to understand how a protein folds, it would appear particularly useful to try to define contributions to the stability of the native conformation.

Discussions with Don Wiley, Martin Karplus, David States, and Chris Dobson have been very helpful in formulating these arguments. S.C.H. is supported by National Institutes of Health Grant CA13202, and R. Durbin held a Fulbright Maintenance and Travel Award while a special student in biophysics at Harvard.

1. Levinthal, C. (1968) *J. Chim. Phys.* **65**, 44.
2. Kim, P. S. & Baldwin, R. L. (1982) *Annu. Rev. Biochem.* **51**, 459–489.
3. Taniuchi, H. (1973) *J. Biol. Chem.* **248**, 5164–5175.
4. Karplus, M. & Weaver, D. L. (1976) *Nature (London)* **260**, 404–406.
5. Ptitsin, O. B. & Rashin, A. A. (1975) *Biophys. Chem.* **3**, 1–20.
6. Rossmann, M., Moras, D. & Olsen, K. W. (1974) *Nature (London)* **250**, 194–199.
7. Rose, G. D. (1979) *J. Mol. Biol.* **134**, 447–470.
8. Crippen, G. M. (1978) *J. Mol. Biol.* **126**, 315–332.
9. Wodak, S. & Janin, J. (1981) *Biochemistry* **20**, 9–17.
10. Gilbert, C. O. (1978) *Nature (London)* **271**, 501.
11. Blake, C. C. F. (1978) *Nature (London)* **273**, 267.
12. Go, M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 1964–1968.
13. Richardson, J. (1977) *Nature (London)* **268**, 495–500.
14. Levitt, M. & Chothia, C. (1976) *Nature (London)* **261**, 552–558.
15. Lesk, A. M. & Rose, G. D. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4304–4308.
16. Creighton, R. E. (1978) *Prog. Biophys. Mol. Biol.* **33**, 231–297.
17. States, D. (1983) Dissertation (Harvard University, Cambridge, MA).
18. States, D. J., Dobson, C. M., Karplus, M. & Creighton, T. E. (1984) *J. Mol. Biol.* **174**, 411–418.
19. States, D. J., Dobson, C. M., Karplus, M. & Creighton, T. E. (1980) *Nature (London)* **286**, 630–632.
20. Brandts, J. F., Halvorsen, H. R. & Brennan, M. (1975) *Biochemistry* **14**, 4953–4963.
21. Cook, K. H., Schmid, F. X. & Baldwin, R. L. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 6157–6161.
22. Schmid, F. X. & Baldwin, R. L. (1979) *J. Mol. Biol.* **135**, 199–215.
23. Labhardt, A. M. & Baldwin, R. L. (1979) *J. Mol. Biol.* **135**, 231–244.
24. Labhardt, A. M. & Baldwin, R. L. (1979) *J. Mol. Biol.* **135**, 245–254.
25. Schmid, F. X. (1981) *Eur. J. Biochem.* **114**, 105–109.
26. Lin, L.-N. & Brandts, J. F. (1983) *Biochemistry* **22**, 573–580.
27. Sadler, J. & Novick, A. (1965) *J. Mol. Biol.* **12**, 305–327.
28. Goldenberg, D., Smith, D. H. & King, J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 7060–7064.
29. Shortle, D. & Lin, B. (1985) *Genetics*, in press.