Proc. Nat. Acad. Sci. USA
Vol. 68, No. 9, pp. 2293–2297, September 1971

Folding of Polypeptide Chains in Proteins: A Proposed Mechanism for Folding

PETER N. LEWIS, FRANK A. MOMANY, AND HAROLD A. SCHERAGA*

Department of Chemistry, Cornell University, Ithaca, N.Y. 14850

Contributed by Harold A. Scheraga, July 14, 1971

ABSTRACT **A mechanism is proposed for the folding of protein chains. On the basis of short-range interactions, certain aminoacid sequences have a high propensity to be, say, α-helical. However, these short helical (or other ordered) regions can be stabilized only by long-range interactions arising from the proximity of two such ordered regions. These regions are brought near each other by the directing influence of certain other aminoacid sequences that have a high probability of forming β-bends or variants thereof, also on the basis of short-range interactions. An analysis is made of the tendency of various amino acids to occur in β-bends, and it is possible to predict the regions of a chain in which a β-bend will occur with a high degree of reliability.**

In this series of papers, we will present a specific mechanism for the folding of a polypeptide chain into the native structure of a globular protein. In this presentation, we will attempt to demonstrate that specific backbone conformations such as the right-handed α-helix (α_R), the β-structure, and the β-bend, found to varying extents in the native structures of most globular proteins, are not only essential for the structural integrity of the protein but also are remnants of structures that play a key role in the folding process.

In this initial paper, we give a general description of the proposed mechanism, as well as some illustrative correlations between the aminoacid sequence and native structure of a protein that provide support for this mechanism. In subsequent papers in this series, we will discuss the energetics of the folding process.

PROPOSED MECHANISM

The protein molecule, under sufficiently denaturing conditions (or even, perhaps, directly after synthesis), behaves essentially as a random coil. Since the number of states accessible to the polypeptide chain in the random-coil condition is immense, it is reasonable to assume that (*a*) the folding of the chain into its most stable (native) conformation is *not* the result of a random event, and (*b*) a *specific* pathway exists for the folding process.

It was previously suggested (1) that one of the initial steps (which might be considered a nucleation step) during the folding process is the fortuitous meeting of two distant sections of the protein chain to form a stabilized pair of α-helices (or, for that matter, any other ordered structure), around which the rest of the polypeptide chain could fold. This idea developed from the demonstration (1) that, for most proteins, those portions of the chain that have a high helical probability in the denatured condition are found to be in the α_R conformation in the native structure. Further, it was shown (2) that, for the cytochrome *c* proteins of 27 species, the regions of high helical probability were, for the most part, conserved from species to species; this result is consistent not only with the proposed invariance of the native conformation of these proteins (3), but also with our proposal that these regions of high helical probability aid in directing the folding to the native structure.

While the above conclusion about one of the initial steps of the folding process seems warranted, it does not seem reasonable for the distant α-helical (or other ordered) conformations to rely on a *random* encounter to achieve a mutual stabilization (by means of long-range interactions) of the specific structures that have a propensity to be α-helical (because of short-range interactions) (1). Instead, it appears much more likely that two such distant helix-tending regions of the polypeptide chain are *directed* toward each other. The assumption of such a directing influence naturally introduces the proposition that certain regions of the chain function as "directing" sections. The role of these "directing" sections would be to provide the proper mutual orientation of distant (or near) ordered segments of the polypeptide chain so that the latter could interact with each other and, at the same time, serve as a substrate for interaction with still other chain segments. As an example, a β-bend or β-turn (defined in the next section) would "direct" the formation of an antiparallel β-structure which, in turn, might provide a surface for interaction (e.g., by means of hydrophobic bonds) with, and stabilization of, an α_R helix.

Our proposed mechanism for protein folding can be described as follows. A certain "directing" section promotes (in the manner described above) the formation of some small

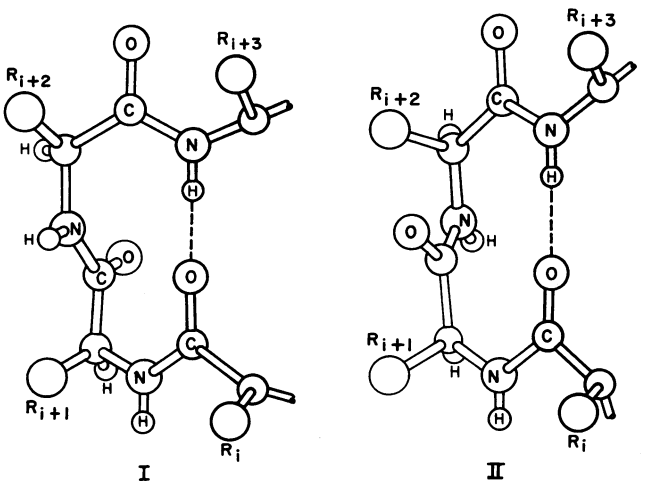

FIG. 1. Type I β-bend (*I*), with *any* L-residue at positions (i + 1) and (i + 2), and Type II β-bend (*II*), with only glycine (5) being possible at position (i + 2). Adapted from Fig. 7 of ref. 3.

* To whom requests for reprints should be addressed.

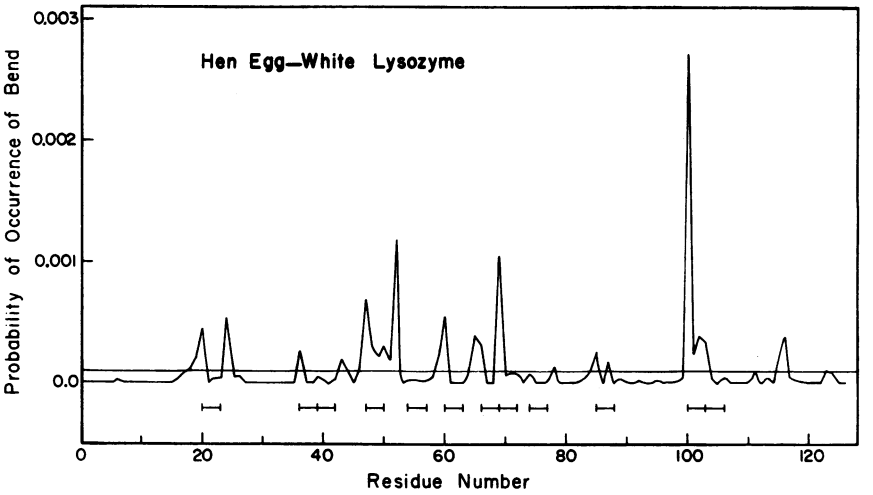

FIG. 2. Probability that a tetrapeptide bend begins at site j of the hen egg-white lysozyme chain. The horizontal line is an arbitrary cut-off probability. The horizontal bars indicate the positions of the observed bends (D. C. Phillips, personal communication), starting at $j = i$.

initial structure (e.g., two interacting α_R helical or other ordered structures). This ordered collection of aminoacid residues (backbone and side-chain groups) then serves as a substrate to direct other sections of the chain to either stabilize still other ordered structures (depending on the propensity of these additional segments to form such ordered structures) or simply to wrap around the original substrate†. If this mechanism is valid, then many of the distinctive features of the native structure (i.e., α_R helix or β-structure) would have been intimately involved in the folding pathway.

In the next section, the β-bend is shown to satisfy the conditions required for a "directing" section. It will be demonstrated that an analysis of the frequency of occurrence of particular aminoacid residues at the various loci of these bends in a sample of three proteins of known structure (namely, lysozyme, ribonuclease S, and α-chymotrypsin) provides sufficient information to enable us to predict, with a reliability of around 80%, the positions of bends that occur in the native conformations of other proteins.

THE β-BEND

In the previous section, it was argued that certain segments of the polypeptide chain are responsible for bringing distant portions of the chain into close proximity during the folding process. The simplest examples of such segments are the so-called β-bends, involving residues i to $(i + 3)$, as shown in Fig. 1. Venkatachalam (5) has considered the steric constraints in these bends for the formation of a hydrogen bond between the CO group of residue i and the NH group of residue $(i + 3)$; he showed that glycine must occur at position $(i + 2)$ for the bend in Fig. 1 (which is designated as a type-II bend) to involve such a hydrogen bond. However, for L-residues, there is no such restriction in the other bend shown in Fig. 1 (which is designated as a type-I bend). Both types of bends fulfill the requirements for being "directing" sections very nicely by (*a*) providing a possible 180° reversal of the chain direction, (*b*) having a high probability of occurrence, compared to a larger loop structure requiring long-range interactions, because the formation of the bend [and the i to $(i + 3)$ hydrogen bond] depends on only two pairs of dihedral angles (short-range interactions), and (*c*) involving only a small number (namely, four) of residues, thereby providing much conformational information in a small region.

Actually, the native structures of many proteins contain an abundance of β-bends, or β-like bends [distorted β-bends, which are similar to β-bends but lack the i to $(i + 3)$ hydrogen bond]. The distorted β-bend might arise by relaxation to satisfy the newly created interactions, once it has fulfilled its directing function during folding. In light of our earlier discussion of the possible importance of these bends for protein folding, it is of interest to consider the distribution of aminoacid residues at positions i, $(i + 1)$, $(i + 2)$, and $(i + 3)$ (see Fig. 1) in the actual bends found in the native conformations of some proteins. Toward this end, the coordinates of hen egg-white lysozyme‡, bovine ribonuclease S,(6) and the B and C chains of bovine α-chymotrypsin (7) were analyzed for bends. A bend was considered to exist if (*a*) the calculated $C^\alpha(i)$ to $C^\alpha(i + 3)$ distance was less than 7 Å (0.7 nm) and (*b*) the $(i + 1)$ or $(i + 2)$ residue was not in an α_R helix. The bends determined by the above criteria are given in Table 1§.

† It should be pointed out here that the notion that some initial structure (substrate) participates in the folding of a protein is not new and has been proposed by others (see, for example, ref. 4), although the emphasis placed in this paper on a "directing" section, such as a β-bend, has not to our knowledge been envisaged as an initial structure.

‡ D. C. Phillips, personal communication.

§ Criteria (*a*) and (*b*) in some cases are not sufficient to define the exact location of a bend. Molecular models of bovine ribonuclease S and bovine α-chymotrypsin, built in this laboratory, provided additional information concerning the existence and location of bends in these two proteins. The C-terminal (106–129) section of hen egg-white lysozyme is particularly difficult to analyze for bends by criteria (*a*) and (*b*), because of the presence of some helix in that part of the protein chain; therefore, for this region, no bends are given in Table 1, although it may be that some exist. It should be emphasized that the list of bends given in Table 1 may well be subject to revision and is presented here *only* to illustrate the possible role of "directing" sections. Further work is being performed, in this laboratory, to better characterize these bends.

TABLE 1. *β-bends and variants found in the native structures of hen egg-white lysozyme, bovine ribonuclease S, and bovine α-chymotrypsin (B and C chains)*

Hen egg-white lysozyme		Bovine ribonuclease S		Bovine α-chymotrypsin	
Number	Sequence	Number	Sequence	Number	Sequence
20–23	Tyr-Arg-Gly-Tyr	16–19	Ser-Thr-Ser-Ala	23–26	Val-Pro-Gly-Ser
36–39	Ser-Asn-Phe-Asn	36–39	Thr-Lys-Asp-Arg	27–30	Trp-Pro-Trp-Gln
39–42	Asn-Thr-Gln-Ala	65–68	Cys-Lys-Asn-Gly	35–38	Asp-Lys-Thr-Gly
47–50*	Thr-Asp-Gly-Ser	75–78	Ser-Tyr-Ser-Thr	48–51	Asn-Glu-Asn-Trp
54–57	Gly-Ile-Leu-Gln	87–90	Thr-Gly-Ser-Ser	56–59	Ala-His-Cys-Gly
60–63	Ser-Arg-Trp-Trp	91–94	Lys-Tyr-Pro-Asn	61–64	Thr-Thr-Ser-Asp
66–69	Asp-Gly-Arg-Thr	112–115	Gly-Asn-Pro-Tyr	72–75	Asp-Gln-Gly-Ser
69–72	Thr-Pro-Gly-Ser			91–94	Asn-Ser-Lys-Tyr
74–77	Asn-Leu-Cys-Asn			96–99	Ser-Leu-Thr-Ile
85–88	Ser-Ser-Asp-Ile			99–102	Ile-Asn-Asn-Asp
100–103*	Ser-Asp-Gly-Asp			108–111	Leu-Ser-Thr-Ala
103–106	Asp-Gly-Met-Asn			115–118	Ser-Gln-Thr-Val
				125–128	Ser-Ala-Ser-Asp
				131–134	Ala-Ala-Gly-Thr
				152–155	Pro-Asp-Arg-Leu
				172–175	Trp-Gly-Thr-Lys
				177–180	Lys-Asp-Ala-Met
				185–188	Ala-Ser-Gly-Val
				191–194	Cys-Met-Gly-Asp
				194–197	Asp-Ser-Gly-Gly
				203–206	Lys-Asn-Gly-Ala
				217–220	Ser-Ser-Thr-Cys
				221–224	Ser-Thr-Ser-Thr

* The $C^{\alpha}(i)$ to $C^{\alpha}(i+3)$ distances for these two bends exceeded 7 Å by 0.1 Å and 0.4 Å for residues 47–50 and 100–103, respectively. Nevertheless, these bends were counted because, in both cases, each was the region of a significant chain reversal. See ref. 4 for stereo drawings of hen egg-white lysozyme.

TABLE 2. *Frequency of occurrence of amino acid residues in β-bends and variants found in hen egg-white lysozyme, bovine ribonuclease S, and bovine α-chymotrypsin (B and C chains)*

Amino acid	Total occurrence*	i	$i+1$	$i+2$	$i+3$	$i \rightarrow (i+3)$ (total†)	$\frac{i \rightarrow (i+3) \text{ total}}{\text{(total occurrence)}}$
Ala	45	3	2	1	4	10	0.22
Asp	22	5	4	2	5	16	0.73
Cys	25	2	0	2	1	5	0.20
Glu	12	0	1	0	0	1	0.08
Phe	12	0	0	1	0	1	0.08
Gly	36	2	4	11	4	21	0.58
His	7	0	1	0	0	1	0.14
Ile	18	1	1	0	2	4	0.22
Lys	30	3	3	1	1	8	0.27
Leu	27	1	2	1	1	5	0.19
Met	8	0	1	1	1	3	0.38
Asn	36	4	4	3	4	15	0.42
Pro	13	1	4	1	0	6	0.46
Gln	19	0	2	1	2	5	0.26
Arg	18	0	2	2	1	5	0.28
Ser	51	11	6	6	5	28	0.55
Thr	39	5	4	6	4	19	0.49
Val	36	1	0	0	2	3	0.08
Trp	14	2	0	2	2	6	0.43
Tyr	13	1	2	1	2	6	0.46

* The numbers in this column represent the total occurrence of each residue in the three-protein sample.

† The $i \rightarrow (i+3)$ totals for Asn, Thr, and Ile are each larger by 1 than their actual occurrence, because, for example, Thr simultaneously occupies positions i and $(i+3)$ in bends 69–72 and 66–69, respectively, in lysozyme (see Table 1). Similarly, the total for Asp is larger by 2 than its actual occurrence.

TABLE 3. *Comparison of predicted and observed locations of bends in some globular proteins**

Protein		Location of *i*th bend position
Hen egg-white lysozyme, *N* = 129, 75/86†	*P*	20 24 36 43 47 52 60 65 69 78 85 100 103 116, 123
	O	20 36 39 47 54 60 66 69 74 85 100 103
Bovine ribonuclease *S*, *N* = 124, 86/78	*P*	1 15 21, 24 27, 31 66 70 75 80 87 91 98, 110 113
	O	16 36 65 75 87 91 112
α-Chymotrypsin B and C chains, *N* = 228, 87/89	*P*	23 27 32 35 42 48 61 72 92 96 100 115 125 128 132 138 151 164, 170 178 185 191 194 203 215 218 221 224
	O	23 27 35 48 56 61 72 91 96 99 108 115 125 131 152 172 177 185 191 194 203 217 221
Horse cytochrome *c*, *N* = 104, 67/91	*P*	22 28 39 47 54 75 78
	O	21 27 35 non-β bend 42–46 53 75
Carboxypeptidase A, *N* = 307, 74/85	*P*	3 19, 39 42 53 65 73, 77 89 101 112 128 133 144 153 159 162 165 185 197 205 208 212 233 239 251 254 257 260 273 287, 294
	O	3 41 53 56 65 89 108 112 133 144 150 159 162 167 184 non-β bend 196–200 206 213 232, 213 242, 242 250 273 277
Subtilisin BPN′, *N* = 275, 84/76	*P*	5 18 21 32 35 38 45 60 63 78 85 95 98 101, 104, 108 117, 125, 129 144, 152 155 158, 161 167 171 181 184 200 209 217, 220 239 245, 252 256 259 262
	O	31 39 52 60 77 85 94 98 156 non-β bend 159–163 166 171 181 187 202 210 238 260 263
Staphylococcal nuclease, *N* = 149, 100/94	*P*	18 24 27 48 77 84 94 116 141 146
	O	18 27 47 78 83 94 116 (undetermined)

* *P* and *O* designate the predicted and observed, respectively, locations of bends that start at the *i*th position. Also, see footnote ¶.

† The numerator is the percent of observed bends predicted correctly, and the denominator is the percent of the *N*/3 maximum possible number of bends in the chain that are predicted correctly.

The distribution of amino acid residues located at positions i, $(i+1)$, $(i+2)$, and $(i+3)$ in the bends given in Table 1 is shown in Table 2, together with the overall number of each aminoacid residue in the three-protein sample. It is interesting to note that, in 11 of the 42 bends shown in Table 1, glycine is located at position $(i+2)$. Presumably, most of these 11 bends are of type II. For the bends given in Table 1, no distinction is made between types I and II bends.

The data of Table 2 enable us to evaluate an *a priori* probability that a certain type of residue is located at the jth site in a β-bend (where $j = i, i+1, i+2$, or $i+3$), irrespective of its neighbors, by simply dividing the number of times the residue in question occurred in the jth site by the overall frequency of occurrence of that residue in the total protein sample; e.g., the *a priori* probability for Ala in site $(i+3)$ is 4/45. From the foregoing, and assuming that the residues are independent of each other, it follows that the probability of occurrence of a β-bend is simply the product of the four individual *a priori* probabilities for each amino acid residue type in each site j. The validity of the assumption that the residues in the bends behave independently (i.e., that side-chain to backbone interactions dominate in bend stabilization) is based on the observation (see below) that tetrapeptide bends of highest probability (calculated in this manner) correlate very well with the appearance of these bends in many native proteins. For illustrative purposes, the probability of occurrence of a β-bend starting at chain site j (computed by the procedure described above) is plotted in Figs. 2 and 3 for hen egg-white lysozyme and horse ferricytochrome *c*, respectively. There is a good correlation between the positions of those peaks lying above an arbitrary cut-off probability line (determined by observation from Fig. 2) at 10^{-4} and the observed positions of the β-turns in Figs. 2 and 3. This same procedure and criterion were applied to several other proteins, and the results for the start (the ith position) of the tetrapeptide β-bend are shown¶ in Table 3. If we allow an error of ±1 residue in the location of a bend [there being approximately $N/3$ (i.e. 1–4, 4–7, 7–10, ...) tetrapeptides that will include all possible bends], then the overall per cent of bends predicted correctly (not including the original set of three proteins, on which the *a priori* probabilities were based) is 80%. This high degree of predictability suggests that the role assigned here to residues in β-bends may be correct.

The data of Table 2 indicate that no particular aminoacid residue is associated exclusively with bends. Since most of the observed bends appear to be composed mainly of polar residues (see the high values for the *a priori* probability of occurrence in a β-bend of polar residues such as Asp, Asn, Ser, Thr, and Tyr in the last column of Table 2), it is not surprising that nearly all the bends are located at the surface of the native globular structures, presumably to solvate the polar side chains.

If these bends are as important to the native conformation as suggested here, then mutations that lead to changes in the residues in the bend regions should provide further informa-

¶ The observed positions of the bends listed in Table 3 for the proteins other than the original set of three were taken from the stereo drawings in refs. 4 and 8. Since the coordinates of these proteins are not now available to us, the positions listed for the bends are tentative, and may have to be revised when the coordinates become available.

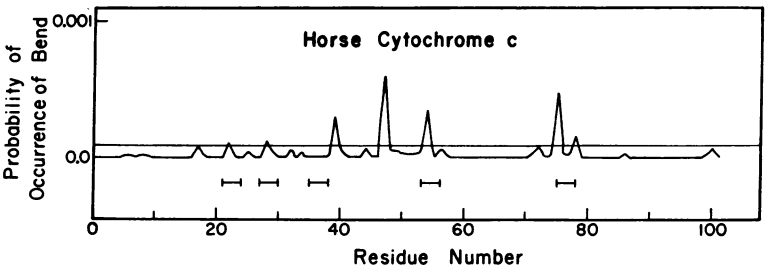

FIG. 3. Probability that a tetrapeptide bend begins at site j of the horse ferricytochrome c chain. The horizontal line is an arbitrary cut-off probability. The horizontal bars indicate the positions of the observed bends (3), starting at $j = i$. A bend (not shown here) occurs (3) between residues 42–46; this is not a β-bend (since it contains more than four residues) even though it results in a reversal of the chain.

tion as to the "directing" role of the bends in folding. This question is currently under investigation.

It should be pointed out that the β-bends discussed in this paper are thought to exist in specific sequences of amino acids. However, other β-bends (not considered here as necessarily likely to occur in globular proteins) can exist in *any* sequence of amino acids if the chain is short and constrained to form a ring, as in gramicidin S and oxytocin.

CONCLUSIONS

It is proposed that certain sections of a protein chain must play a role in bringing distant parts of the chain together to enable long-range interactions to stabilize those structures (i.e., α_R helix, β-structure, etc.) that have a propensity to form because of short-range interactions. The β-bend and its variants were shown to fulfill the requirements of a "directing" section. Further, it was shown that, to a good approximation, the distributions of aminoacid types in these bends are independent of each other; hence, the locations of a high percentage of the bends in proteins not included in the initial set of three can be predicted.

Since these bends in proteins are very localized, it would be interesting to determine (e.g., by NMR measurements) whether they occur in smaller structures, i.e., in isolated noncyclic tetra- or larger oligopeptides (with appropriate end groups).

In subsequent papers, we will consider the energetics of the β-bends and their variants.

P. N. L. was a National Research Council of Canada Postgraduate Fellow, 1971–1972. F. A. M. was a Special Fellow of the National Institute of General Medical Sciences, National Institutes of Health, 1968–1969. This work was supported by research grants from the National Science Foundation (GB-28469X and GB-17388), from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312), from the Eli Lilly, Hoffmann-LaRoche, and Smith Kline and French Grants Committees, and from Walter and George Todd.

1. Lewis, P. N., N. Gō, M. Gō, D. Kotelchuck, and H. A. Scheraga, *Proc. Nat. Acad. Sci., USA*, **65**, 810 (1970).
2. Lewis, P. N., and H. A. Scheraga, *Arch. Biochem. Biophys.*, **144**, 576 (1971).
3. Dickerson, R. E., T. Takano, D. Eisenberg, O. Kallai, L. Samson, A. Cooper, and E. Margoliash, *J. Biol. Chem.*, **246**, 1511 (1971).
4. Dickerson, R. E., and I. Geis, *The Structure and Action of Proteins* (Harper and Row, New York, 1969), p. 73.
5. Venkatachalam, C. M., *Biopolymers*, **6**, 1425 (1968).
6. Wyckoff, H. W., D. Tsernoglou, A. W. Hanson, J. R. Knox, B. Lee, and F. M. Richards, *J. Biol. Chem.*, **245**, 305 (1970).
7. Birktoft, J. J., B. W. Matthews, and D. M. Blow, *Biochem. Biophys. Res. Commun.*, **36**, 131 (1969).
8. Arnone, A., C. J. Bier, F. A. Cotton, V. W. Day, E. E. Hazen, Jr., D. C. Richardson, J. S. Richardson, and, in part, A. Yonath, *J. Biol. Chem.*, **246**, 2302 (1971).