Hypothesis

# A TESTABLE MODEL FOR PROTEIN FOLDING

George D. ROSE, Ronald H. WINTERS* and Donald B. WETLAUFER**

*Oregon State University Computer Center, Corvalis, Oregon 97331. Current address: Department of Chemistry, University of Delaware, Newark, Delaware 19711, USA, *School of Pharmacy, Oregon State University, Corvallis, Oregon 97331, USA, and **Department of Chemistry, University of Delaware, Newark, Delaware 19711, USA*

Received 8 December 1975

## Introduction

A testable, biphasic model for protein folding is formulated. In this model, linearly short and medium range interactions dominate early folding, causing the chain to assume independently nucleated modules of persisting structure termed LINCs. In a later stage of folding, the LINCs fold relative to each other, and it is only at this time that the protein assumes its characteristic interior and exterior and its overall globular structure.

In the perspective of this model, a computational approach is outlined, requiring first a systematic examination of steric and energetic constraints that can be calculated with some confidence by accepted means. To this end, we have begun by calculating the sterically allowed conformation for: (1) a post-helical residue situated at the carboxy-terminal end of a backbone-only helix, (2) various side-chains of an intra-helical residue, and (3) the constraints imposed on Lys and Arg side-chains if some accounting is made for hydration of the respective cationic side-chain moieties. We find substantial steric constraints engendered in all three cases.

The transition of a denatured protein into its native structure is defined to be a *global folding process*, whereas any linearly piece-wise folding that occurs in a nascent chain is a *local folding process*. Convincing instances of local folding have been demonstrated in various contexts [1,2]. In general, the folded end product is expected to be process dependent because conformational states adopted by partial chains will be deprived of any information that accrues with additional chain growth. That is, a nascent chain cannot foresee its future.

Cases are known, however, in which both local and global folding processes yield the same final structure [3,4]. One conception of how qualitatively differing initial states converge to the same final structure rests on the assumption that this structure is necessarily synonymous with a global free energy minimum for the molecule [3]. Another conception will also rationalize the directed emergence of a unique conformation from differing initial states. In particular, a biphasic model for protein folding is presented here. In this model, linearly short and medium range interactions dominate early folding from any state, and order the polypeptide chain into independently nucleated, persistent modular units of structure. Following this early assembly, linearly long-range interactions are then responsible for the further ordering of modular entities into the full three-dimensional configuration of the protein.

The general notion of a biphasic model is no longer novel inasmuch as elements thereof are to be found, either explicitly or implicitly, in several recent publications [5,6], and the concept of nucleation events proposed by Levinthal is, of course well known [7]. Our attempt here, however, has been to provide a highly specific model that both takes into account

the body of experimental evidence and includes sufficient detail to allow a quantitative examination of its consequences.

In detail, we propose that the polypeptide chain, dominated by linearly short and medium range interactions, folds initially into *L*ocal *I*ndependently *N*ucleated *C*ontinuous segments (LINCs). The ordering of the chain into LINCs is promoted during any local folding that takes place in a nascent chain, and LINC formation is also favored in a global folding process because the chain will fold into LINCs before it can fold into anything else.

LINCs are structurally persistent, separable, modular entities that are precursors to their counterparts in a folded protein. LINCs are usually, if not invariably, bounded by peptide chain turns [8,9] which are construed to be the conformationally permissive [10] hinges that allow an ensemble of LINCs to fold relative to each other.

In this model, a protein is comprised entirely of LINCs and interspersed hinges. Not until the occurrence of inter-LINCs folding does the protein take on its characteristic interior and exterior or its overall globular structure. It is at this latter stage in the folding pathway that linearly long-range forces come into play and the LINCs are disposed into their native conformation.

The LINCs and hinges model is consistent with the observation that both local and global processed can yield the same final configuration. The model is also consistent with the success of recent empirical efforts [11] to predict secondary structure based only upon correlations between local amino acid sequences. In the present model, alpha helices and extended chain are considered as particular instances of LINCs.

Viewed from a perspective prompted by this model, the problem of structure formation can be divided into two parts: prediction of LINCs conformation and prediction of inter-LINCs conformation. Some of the factors limiting inter-LINCs folding in the case of myoglobin suggest that packing constraints and hydrophobic interactions place major restrictions on any possible solution set [12,13].

Turning now to the question of LINC's conformation, a study by Gelin and Karplus [14] finds side-chain torsional angles in pancreatic trypsin inhibitor at or near their expected minima in the free amino acid. Such a result is consistent with the present model, for, within a LINC, short and medium range interactions direct the folding process for side-chains as well. Thus, when an independently nucleated oligopeptide 'jiggles' into a persisting conformational minimum, the side-chains are expected to populate their respective minima too, because the steric constraints at this stage in the folding process are not comparable to those imposed on a side-chain at the interior of a protein. It might be thought that when the LINCs subsequently fold relative to each other, displacement of the side-chain from a rotational minimum may find compensation in better inter-LINCs packing. In practice, this trade-off becomes less feasible because a side-chain displacement is no longer free to occur independently, but only in co-operation with other structural determinants in the LINC.

In the general case, the problem of predicting the conformation of only a single LINC by complete energy minimization [15] is still too complex to solve directly. In a recent attempt to reduce the computational complexity, each amino acid residue in the protein is represented by just two points [16]. While this approximation is presented as being highly successful, we believe that the information loss arising from a point representation of the side-chain must inevitably be too drastic a data reduction to permit predictive results.

The approach we adopt here is to compile a catalog of constraints limiting the conformational freedom of a LINC. The catalog can then be used to winnow conformation space to a limited set of energetically favorable conformations for a LINC. In this manner, the computational complexity will be suitably reduced without concomitant loss of information.

The remainder of this paper describes computations that reflect the stringent limitations inherent in LINCs packing, based primarily on steric restrictions.

Upon termination of a right-handed alpha helix at its C-terminus, the first residue no longer in a helical orientation will be termed a *post-helical* residue. The subspace of conformation space that can be occupied by selected post-helical residues is now explored.

Fig.1 is a Ramachandran (phi,psi) plot with peptide coordinates taken from Marsh and Donohue [17]. This (360 × 360) space was sampled every ten degrees and each 'x' marks a sample point where the dipeptide Gly—Ala is found to be sterically allowed. The contact distance criteria used to compute steric inhibition

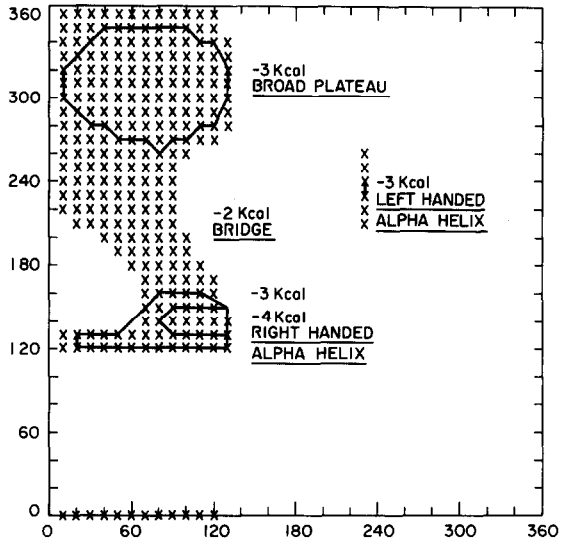


Fig.1. Allowed positions for the dipeptide Gly–Ala. Positions found to be sterically allowed are indicated by an X. Some favorable energy contours are outlined, and the regions are named.

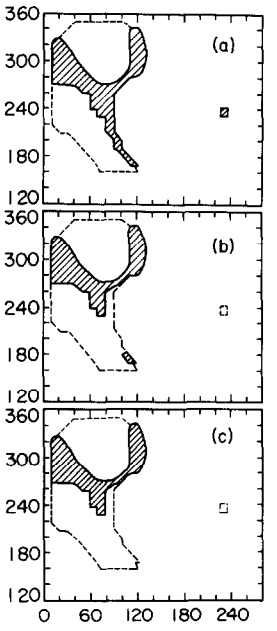


Fig.2. Sterically allowed positions for the first post-helical residue adjoining the C-terminus of a backbone-only α-helix. (a) Allowed positions for a backbone-only residue. Backbone-only residues are allowed only in the area shaded by diagonal lines. (b) Allowed positions for His. (c) Allowed positions for Trp.

were taken from Ramachandran and Sasisekharan [18]. Superimposed upon the 'hard-sphere' contact map in fig.1 are energy contours of a 'soft-sphere' function [19]. The good agreement between hard sphere and soft sphere functions is no longer surprising to us, as repulsive forces are known to play a dominant role in such functions. To facilitate discussion, dipeptide space is partitioned and named as shown in fig.1.

Inspection of fig.1 shows a narrow energy well in the map area corresponding to right-handed alpha helix. For helical residues populating this region of the map, narrowing of the well ought to be further enhanced by hydrogen bonding within the helix. This expectation appears to be borne out for the refined X-ray structure of lysozyme [15,20] by the apparent clustering of $(\phi,\psi)$ values in the neighborhood of $\phi = 120$, $\psi = 130$. This is the only high density cluster of points in the $(\phi,\psi)$ plot of lysozyme.

We first examine steric constraints resulting from backbone-only interactions between a post-helical residue at the carboxyl end of a right-handed α-helix and the four preceding residues; all five residues are backbone-only residues. A *backbone-only* residue is one without a side-chain; it can be viewed as a desmethyl L-alanyl residue. Steric constraints imposed on a backbone-only residue are the minimal constraints for any actual residue, regardless of the nature of the side-chain.

With one turn of backbone-only helix preceding a backbone-only post-helical residue, only the conformations shown in fig.2(a) are allowed. This restriction of conformation space is due to steric interference between the backbone atoms in the post-helical residue and the adjacent carbonyl oxygen from the preceding turn of the helix. Since the restriction involves only backbone atoms, every post-helical residue is at least this restricted.

When the side-chain in a post-helical residue is also taken into consideration, further structural limitations are seen. While a post-helical backbone-only residue is not distinguishable in this analysis from a post-helical alanine, differences do begin to appear with further increases in side-chain size. Corresponding diagrams for

the cases of histidine and tryptophan are shown in fig.2(b) and (c). In this computation, side-chain configurations arising from the domain $\chi^1 = 60°, 180°, 300°$ (±10°) and $\chi^2 = 0°, 90°, 180°, 270°$, were examined. It can be seen from the figure that the side-chains can impose significant additional constraints on the possible disposition of a post-helical residue.

The structural limitations shown for post-helical residues are based on the assumption of energetically well-formed helix [21]. When the helix used for these computations is appropriately distorted at a constraining locus, there is an accompanying relaxation of the observed constraints.

In addition, deviation from the ideal peptide geometry used here may tend to reduce the limitations shown in fig.2. However, we have attempted to compensate for this possibility by a conservative choice of contact distance criteria. Studies on steric hindrance show a sensitive dependence upon the choice of contact distance criteria [17], with the Ramachandran values being the most conservative set proposed.

A second example of stringent packing constraints is seen in the case of an intra-helical residue. The helix-breaking tendency of proline due to steric effects was observed some time ago [18,22]. In our second example, attention is focused on the converse steric effect, limitation of side-chain freedom by the helical backbone.

Each of the amino acids listed in table 1 was included as the middle residue between two turns of backbone helix (i.e. $(Gly)_4$-X-$(Gly)_4$ where X is the residue under inspection). The side-chains were then examined at configurations where side-chain groups

Table 1

$(Gly)_4$–X–$(Gly)_4$ in helix

Domain A ≡ position I = 60° ± 10°
II = 180° ± 10°
III = −60° ± 10°

Domain B ≡ position I = 0°
II = 90°
III = 180°
IV = −90°

The domains given for each residue are the domains of definition over which each side chain group was varied, listed in sequential order of increasing distance from the C-alpha along the side chain. For example, Tyr has two degrees of rotational freedom in its side chain arising at the Cα–Cβ bond and at the Cβ–Cγ bond. With two degrees of freedom, it is necessary to specify two domains of definition. These are listed in the table below as A, B where domain A pertains to the Cα–Cβ bond and domain B pertains to the Cβ–Cγ bond.

| Residue | Domain | Allowed positions | Hydrated form allowed positions |
|---|---|---|---|
| Lys | A, A, A, A | II, II, II, I-III | II, II, II, II |
| | | II, I, II, I-III | II, I, II, II |
| | | III, II, II, I-III | III, II, II, II |
| | | III, III, II, I-III | III, III, II, II |
| Cys | A | II | |
| | | III | |
| Glu | A, A, B | II, I, II or IV | |
| | | II, II, II or IV | |
| | | I, II, II or IV | |
| | | I, I, II or IV | |

| Residue | Domain | Allowed positions | Hydrated form allowed positions |
|---|---|---|---|
| His | A, B | II, II or IV | |
| Met | A, A, A | II, I, I or II | |
| | | II, II, I-III | |
| | | III, II, I or II | |
| | | III, III, II or III | |
| Asp | A, B | II, II or IV | |
| Thr | A | III | |
| Tyr | A, B | II, II or IV | |
| Ser | A | II | |
| | | III | |
| Val | A | III, III | |
| Ilu | A, A | III, II or III | |
| Leu | A, A | II, II | |
| | | III, III | |
| Phe | A, B | II, II or IV | |
| Trp | A, B | II, II or IV | |
| Arg | A, A, A, B | II, II, II or III, I or II or IV | II, II, II or III, I or II |
| | | II, II, I, I or IV | |
| | | II, I, II, I or II or IV | II, I, II, I |
| | | II, I, I, I or IV | II, I, I, I or IV |
| | | III, II, II, I or II | III, II, II, I |
| | | III, II, I, I or IV | III, II, I, I or IV |
| | | III, III, II, I or II or IV | III, III, II, I |
| | | III, III, III, I or II | III, III, III, I or II |

are in one of the conventionally observed torsional minima. Aliphatic groups were varied over the domain 60°, 180°, and 300° (±10°), while planar and aromatic groups were varied over the domain 0°, 90°, 180°, and 270°. Table 1 summarizes the positions found to be sterically allowed. Backbone helix is seen to strongly limit the accessible side-chain structures of several amino acid residues.

In the formation of a LINC, charged polar residues are probably hydrated. The attachment of a hydration shell to the terminal group of arginine or lysine, for example, will increase the packing constraints. To approximate hydration effects, X-ray data from salts of arginine and lysine [23–25] were examined and water molecules were attached to the terminal groups at loci where hydrogen bonding was observed in the crystal structures. The water was oriented so that its hydrogen atoms were symmetrically positioned above and below the plane of the side-chain group. The hydrated amino acid residues, Lys·$(H_2O)_3$ and Arg·$(H_2O)_5$ were then used in the intra-helical computation. In table 1, it can be seen that the inclusion of hydration tends to force both arginyl and lysyl side-chains towards extended chain configurations.

As a final experiment, sequentially adjacent lysyl and arginyl residues, both intra-helical, were inspected

Table 2

$(Gly)_4$-Lys-Arg-$(Gly)_4$ in helix

Domains are defined as in table 1. Any of the allowed positions listed for lysine are sterically compatible with any of the allowed positions listed for arginine. *All other pairwise positional arrangements are sterically incompatible.*

| Allowed positions for hydrated lysine | Allowed positions for hydrated arginine |
|---|---|
| II, II, II, II | II, II, II, I |
| II, I, II, II | II, II, III, I or II |
| III, III, II, II | II, I, II, I |
| | II, I, I, I or IV |
| | III, III, II, I |
| | III, III, III, I or II |

to see whether such a juxtaposition imposes constraints in addition to those experienced by these residues taken individually. Additional constraints were observed, as summarized in table 2.

The values obtained from the preceding computations were not compared to values available from X-ray studies since a correspondence between individual torsion angles will depend in part on factors not included here. These initial computations have employed an idealized moiety called backbone-only helix, and with it, the assumption of a completely regular geometry for a helix. While helical fibers of poly-L-alanine appear to be compatible with these assumptions [26], it is not expected that a heterogeneous collection of helical residues will exhibit equivalent regularity. For these reasons, we feel an appropriate test of the model must wait until predicted LINCs can be compared to their X-ray elucidated counterparts in solved structures.

In closing, it should be noted that the LINCs and hinges model is the simplest representative taken from a spectrum of related models. In the preceding paragraphs, we have emphasized the similarity in structural identity of a LINC from the onset of structure formation through folding to incorporation in the final globular assembly. The model is simple in a computational sense because, with these assumptions, the approximate structure of a given LINC can be calculated without regard for its neighbors and then treated as a single structural entity during subsequent computations. It is possible, however, that when the ensemble of LINCs is packed into a final globular assembly, a more extreme deformation of the original structures occurs. In the most extreme case, the original structure would be deformed beyond recognition, but for reasons given earlier, we think that this is unlikely. In the event that limited deformation takes place during inter-LINCs assembly, the initial conformation of the undeformed LINC would serve as a suitable starting structure.

In summary, we have proposed a testable biphasic model for the folding of globular proteins. In this model, linearly short and medium range interactions dominate early folding, causing the chain to assume independently nucleated, structurally persistent modular units of structure; these postulated entities are termed LINCs. In a later stage of folding, the LINCs fold relative to each other, forming a structure in which linearly long-range interactions also play a role. It is only at this time that the protein assumes its characteristic interior and exterior and its overall globular structure.

If these ideas about the folding process are valid, then demonstrable stabilizing forces must exist in oligopeptides of even moderate size. One strong source of structural stabilization is steric repulsion, and, to this end, some packing constraints for intra-helical and post-helical residues have been shown. Work is now in progress to further develop the catalog of structural determinants for a LINC. At the same time we are exploring the interfaces between LINCs and hinges. In the transition from a LINC to a hinge, steric constraints can no longer take such a key role, since by our working assumptions hinges are comparatively flexible. In order to predict the locations of these interfaces, it will be necessary to have some accounting of hydrogen-bonding and hydrophobic interactions.

**Acknowledgements**

We gratefully acknowledge the long-standing support and encouragement of Professor K. Van Holde, Professor Larry Hunter and the Oregon State University Computer Center. We are also indebted to Professor R. R. Becker, Dr Ted Hopkins, and Mr Rjay Murray for helpful discussions. The final stages of this work were supported by a University of Delaware Provost's Grant.

## References


[1] Brown, J. E. and Klee, W. A. (1971) Biochemistry 10, 470–476.
[2] Villarejo, M. R. and Zabin, I. (1973) Nature New Biol. 242, 50–52.
[3] Anfinsen, C. B. (1973) Science 181, 223–230.
[4] Saxena, V. P. and Wetlaufer, D. B. (1970) Biochemistry 9, 5015–5023.
[5] Baldwin, R. L. (1975) Ann. Rev. of Biochemistry 44, 453–475.
[6] Ptitsyn, O. B., Lim, V. I. and Finkelstein, A. V. (1972) Federation of European Biochemical Societies 25, 421–431.
[7] Levinthal, C. (1968) J. Chim. Phys. 65, 44–45.
[8] Kuntz, I. D. (1972) J. Am. Chem. Soc. 94, 4009–4012.
[9] Lewis, P. N., Momany, F. A., Scheraga, H. A. (1971) Proc. Nat. Acad. Sci. USA 65, 2293–2297.
[10] Wetlaufer, D. B. and Ristow, S. (1973) Ann. Rev. of Biochemistry 42, 135–158.
[11] Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, B. and Nagano, K. (1974) Nature 250, 140–142.
[12] Ptitsyn, O. B. (1975) Biophys. Chem. 3, 1.
[13] Lim, V. I. (1974) J. Mol. Biol. 88, 857–894.
[14] Gelin, B. R. and Karplus, M. (1975) Proc. Nat. Acad. Sci., USA 72, 2002–2006.
[15] Warme, P. K. and Scheraga, H. A. (1974) Biochemistry 13, 757–782.
[16] Levitt, M. and Warshel, A. (1975) Nature 253, 694–698.
[17] Marsh, R. E. and Donohue, J. (1967) Adv. Prot. Chem. 22, 235–256.
[18] Ramachandran, G. N. and Sasisekharan, V. (1968) Adv. Prot. Chem. 23, 326–438.
[19] Brant, D. A. and Flory, P. J. (1965) J. Am. Chem. Soc. 87, 2791–2800.
[20] Levitt, M. (1974) J. Mol. Biol. 82, 393–420.
[21] Ramachandran, G. N. (1972) Conformation of Biological Molecules and Polymers; The Jerusalem Symposium on Quantum Chemistry and Biochemistry 5, 1.
[22] Szent-Gyorgyi, A. G. and Cohen, C. (1957) Science 126, 697.
[23] Karle, I. L. and Karle, J. (1964) Acta Cryst. 17, 835–841.
[24] Wright, D. A. and Marsh, R. E. (1962) Acta Cryst. 15, 54–64.
[25] Ramachandran, G. N., Mazumdan, S. K., Venkatesan, K. and Lakshminarayanan, A. V. (1966) J. Mol. Biol. 15, 232–242.
[26] Arnott, S. and Dover, S. D. (1967) J. Mol. Biol. 30, 209–212.