Proc. Natl. Acad. Sci. USA
Vol. 78, No. 7, pp. 4304–4308, July 1981
Biochemistry

Folding units in globular proteins

(protein folding/domains/folding intermediates/structural hierarchy/protein structure)

ARTHUR M. LESK* AND GEORGE D. ROSE†‡

*Fairleigh Dickinson University, Teaneck, New Jersey 07666; and †Department of Biological Chemistry, The Milton S. Hershey Medical Center, The Pennsylvania State University, Hershey, Pennsylvania 17033

Communicated by Charles Tanford, April 20, 1981

ABSTRACT **We present a method to identify all compact, contiguous-chain, structural units in a globular protein from x-ray coordinates. These units are then used to describe a complete set of hierarchic folding pathways for the molecule. Our analysis shows that the larger units are combinations of smaller units, giving rise to a structural hierarchy ranging from the whole protein monomer through supersecondary structures down to individual helices and strands. It turns out that there is more than one way to assemble the protein by self-association of its compact units. However, the number of possible pathways is small—small enough to be exhaustively explored by a computer program. The hierarchic organization of compact units in protein molecules is consistent with a model for folding by *hierarchic condensation*. In this model, neighboring hydrophobic chain sites interact to form folding clusters, with further stepwise cluster association giving rise to a population of folding intermediates.**

We have been interested in analyzing the structural organization of globular proteins and in investigating how subunits of this structure might participate in the folding process. The spontaneous self-assembly of a protein from its components is a paradoxical process. On the one hand, the apparent stereochemical constraints seem insufficient to determine how arrangements of transient structural components can be thoroughly explored and the native one reliably selected while the traps of non-native but metastable conformations are avoided. On the other hand, the constraints appear to be too limiting when one tries to reconcile the search for native structure with the attractive notion that the protein collapses at an early stage under the influence of hydrophobic forces.

We propose a way to resolve this dilemma by identifying the complete collection of compact, contiguous-chain, structural units for a given protein and then describing a comprehensive set of hierarchic folding pathways for these units. The number of possible pathways is large enough to be plausible but small enough to be useful, and their hierarchic nature dramatically reduces the possibility of non-native chain folds.

Our analysis suggests a mechanism of protein folding by hierarchic assembly of structural intermediates. Fundamental to this view is the question of whether steps in the underlying folding dynamics can be inferred from the resultant native structure. Two principal lines of evidence seem suggestive.

(*i*) It has been shown that a local measure of hydrophobicity (1) partitions the amino acid sequence into structural segments such as helices and strands (2, 3). The segmentation of the molecule, predicted from the linear sequence, persists in the native structure. The simplest explanation for the success of these predictions is to assume that the pattern of segmentation is also maintained during intermediate stages in the folding.

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.

(*ii*) Analysis of protein structures elucidated by x-ray crystallography showed that protein molecules can be dissected into a succession of spatially compact pieces of graduated size (4). Each of these elements is formed from a contiguous stretch of the polypeptide chain. A related analysis was reported by Crippen (5). The spatial compartmentation of linear segments, seen in the final structure, is likely to be a feature of intermediate folding stages as well. Otherwise, mixing of chain segments occurring during intermediate stages would have to be followed by spontaneous unmixing as folding progresses.

These results suggest that remnants of structural intermediates in the folding process will still be discernible in the native structure.

Analysis of protein structures has often emphasized identification of a predefined set of elements: helices, strands, and their suprastructures (6–8). In contrast, our aim has been to recognize systematically all compact contiguous-chain units in the protein, independent of their secondary structure composition.

The choice of compactness as our main focus, rather than hydrogen-bonded secondary structure, was prompted in part by earlier evidence (3, 9, 10) suggesting a progression of events in the folding process: (*a*) formation of primitive folding nuclei—nearby hydrophobic chain elements interact to form small folding clusters of low stability; and (*b*) growth of intermediates—neighboring clusters coalesce in stepwise fashion, leading to successively larger intermediate structures.

Intermediates formed in this way are expected to be compact units. A helix or strand would arise in this process as one of a few energetically favorable alternatives for a given hydrophobic primitive.

Inspection of protein structures determined by x-ray analysis reveals hydrogen-bonded supersecondary structures that are evidently fashioned from noncontiguous chain segments. Intuitively, it may seem that a method to identify only the compact units comprised of contiguous-chain might overlook these noncontiguous suprastructures; but this need not be the case. From an analysis of the observed β-sheet topologies, Richardson (7) formulated a stepwise construction rule that accounts for sheets with nonconsecutive but adjacent strands by "taking either a β-strand or a prefolded unit and laying it down next to a prefolded part of the sheet with which it is also contiguous in sequence."

It is important to emphasize that spatial compactness need not occur at the expense of hydrogen-bonded structural elements. For example, the small compact units in myoglobin turn out to be the helices, as discussed later in this paper. Larger compact units include the usual supersecondary structures in addition to some novel supersecondary structures to be described elsewhere.

‡ To whom reprint requests should be addressed.

We distinguish between the set of residues comprising a compact unit and the conformation of this unit. It need not be assumed that a structural intermediate adopts its native conformation during all stages in the folding process; it need only be assumed that the intermediate has some compact conformation and is spatially segregated from its neighbors. It is possible that the region of chain comprising a particular unit can be in equilibrium between multiple conformers as folding progresses (3).

In recent years, Creighton (11, 12) has demonstrated the existence of an obligatory but non-native disulfide intermediate in the refolding of reduced pancreatic trypsin inhibitor. This result would appear to limit the prospects for deducing meaningful folding intermediates by analysis of the final structure. However, we note that a compact unit within the protein may possibly rearrange during folding while retaining its compact nature, as proposed by Lim (13).

In this analysis, a "structural subunit" or "domain" is defined as a contiguous region of the linear amino acid sequence that forms a spatially compact structure in the native protein. A method is presented to identify the complete set of structural subunits from x-ray coordinates and then to examine systematically all ways to combine these units in a tree of folding pathways, based upon the physical hypothesis of a mechanism of growth and accretion.

These domains are found to be arranged as a structural hierarchy (14), a form of organization in which each component of interest is wholly contained in a higher-level component. It turns out that there is more than a single potential pathway for assembling the protein from its domains in the cases we have studied, but it is always a number small enough to be exhaustively explored by a computer program.

In this report, we describe conclusions about protein folding that emerge from our analysis of the combinatorics of subunit assembly.

IDENTIFICATION OF COMPACT UNITS

To identify all compact contiguous-chain segments of any size from x-ray coordinates, every segment is represented by the inertial ellipsoid of its atoms other than hydrogen; the most compact ellipsoids are then selected.

The inertial ellipsoid of a set of atoms is a smooth, convex body superimposed upon the atoms and having the same rigid-body dynamics (15–17). That is, the ellipsoid has the same principal moments of inertia as the set of atoms. The area and volume of the ellipsoid are readily calculated. Richards (18) has shown that the ratio

$$\frac{\text{surface area of the protein}}{\text{area of the inertial ellipsoid}}$$

is approximately constant, at least for whole proteins.

To develop a complete set of compact units, we consider all chain segments of fixed length, n, as n increases uniformly: $n = 8,12,16,\ldots,L$, the length of the entire monomer. For each n there are $L - n + 1$ possible segments of that length; we refer to these as n-tuples. We have determined the inertial ellipsoid of each n-tuple and calculated its geometric properties.

Fig. 1 is a series of normalized curves for two proteins. Each curve shows the specific volume of the n-tuple of indicated size plotted as a function of the position in the sequence. For each protein the n-tuples range from 8-mers up to whole monomers in graduated steps.

The most compact units are taken to be those local minima in Fig. 1 for which the absolute value of the specific volume is within the lowest 10% of all values for n-tuples of equal length. Compact units recognized in this way are indicated by a circle at their central position and by a horizontal bar spanning their length.

The primitive compact units discovered by this procedure depend upon a parameterized acceptance level, which here is chosen to be within 10% of the absolute minimum. Relaxation of this parameter would yield a larger set of primitive units. The inclusion of an additional factor to correct for side-chain variation within the n-tuple would result in a more intelligent selection strategy, but we have postponed such refinement because the simple 10% rule does a clean job of selecting the Mb helices as primitive compact units.

We have examined other geometric criteria for compactness including minimal volume, minimal area, maximal density, and minimal ratio of area to volume. Although alternate criteria and different threshholds generate slightly different sets of compact units, these variations are minor and do not affect any of the

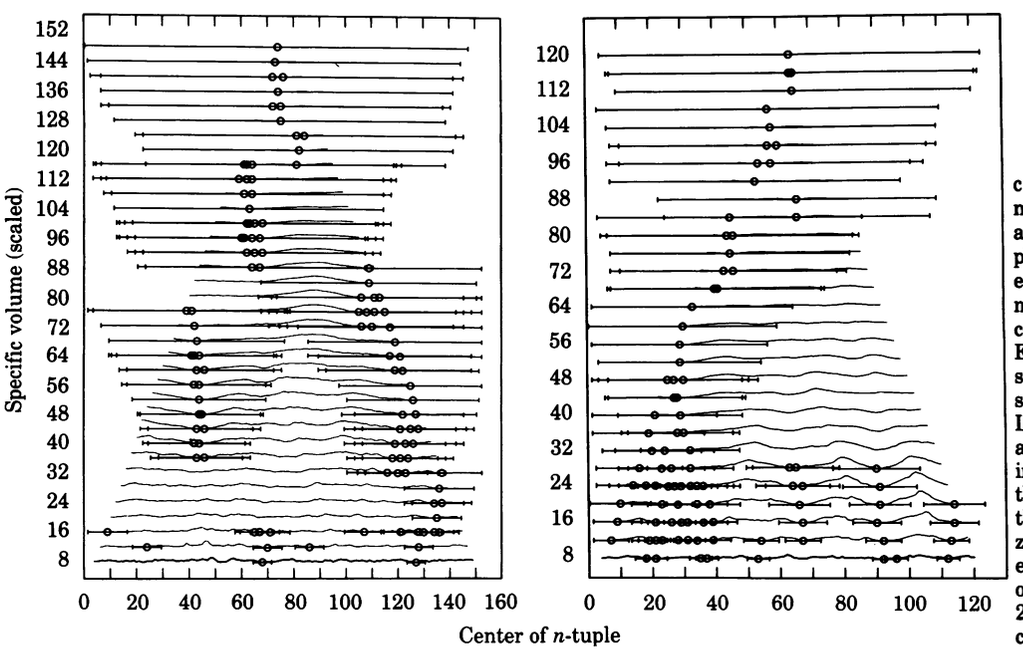

FIG. 1. The complete set of compact contiguous-chain segments for sperm whale Mb (*Left*), and RNase (*Right*). Each curve plots the specific volume for the inertial ellipsoid of the chain segment of indicated length versus the central position of that segment. Each curve is scaled to fit within a strip of fixed height and then assembled in order of segment size. Local minima in these curves that are within 10% of the absolute minimum are selected as compact units; these are marked with a circle at their central position and a horizontal bar that spans the unit. For example, Mb is found to have only one compact unit of segment length 28 residues (residues 122–149 with center at 135).

conclusions presented here. All of these measures are correlated with indices for compactness used by other groups (19, 20).

Casual inspection of the family of curves for a given protein in Fig. 1 suggests that a set of contiguous residues may be part of many compact units of different lengths; compact units at one level are often contained within larger compact units at the next level. These interunit relationships are examined in the next section. Critical to this stage of the analysis is assurance that no compact unit is overlooked; this was achieved by systematically testing every possible segment of contiguous chain.

PATHWAYS BETWEEN COMPACT UNITS

We turn now to the question of how compact units are packaged for inclusion in the native structure. In Fig. 1 it can be seen that a chain segment comprising a small unit is generally also part of larger units, thereby giving rise to the hierarchic organization discussed earlier. The inclusion trees of Fig. 2 elaborate all the ways in which these inclusions occur.

It is hypothesized that the tree structures presented in Fig. 2 correspond to potential folding pathways. The geometric relationships among compact units suggest assignments of specific roles in the folding process.

Unit growth. A unit may be wholly contained within the next consecutively larger unit. This situation occurs when successively larger compact n-tuples in the crystal structure are embedded within a given compact unit. In the underlying folding dynamics, this process might correspond to growth about a nucleation center.

Condensation. Compact units may merge. This situation occurs when two or more distinct compact units in the crystal structure are included within a common larger unit. In the folding process, it would correspond to a mutually stabilizing interaction between distinct structural intermediates.

A unit may be unlinked to other compact units on higher levels of the tree. This corresponds to a compact unit in the crystal structure that is not contained in any larger unit. In the proteins analyzed, such units contain smaller embedded units that fall into the above two cases. In the folding process, this situation corresponds to growth of a compact unit as an alternative to condensation. Such a unit is termed a "dead end."

As a final possibility, a unit may dissolve. In this case, the unit will not be in evidence in the crystal structure and would not be found by our procedure. In the folding process, such units would come about as nonproductive equilibrium intermediates that are ultimately unfolded as the whole set of low-stability equilibrium intermediates is pulled in the direction of successfully folding transition states (9).

Condensation and growth account for most of the paths that we observe.

The inclusion tree for a protein depicts all relationships between compact units consistent with the observed crystal structure. In these trees, the complete set of compact units is generated as described (Fig. 1), and all paths between related units are drawn, as shown in Fig. 2.

In Fig. 2, each compact unit is placed in a two-dimensional array with its sequence interval on the abscissa and its unit size on the ordinate; the central position is marked with a circle. Lines of connectivity are then drawn between the units.

We distinguish three types of unit connectivity in the inclusion trees shown in Fig. 2.

(*i*) Growth (heavy dashed line). A unit is connected to a larger unit by a heavy dashed line if the smaller is the only compact unit contained within the larger. The larger unit is thus composed of the smaller unit together with other less-compact material.

(*ii*) Condensation (solid line). Two or more units are connected to a larger unit by a heavy line if the larger one contains all the smaller ones, if no subunit of the larger unit also contains the smaller ones, and if the smaller ones are distinct. Condensation occurs by the merging of two or more separate compact units.

(*iii*) Condensation and growth (light dashed line). A given unit is connected to a larger unit by a light dashed line whenever the larger unit is a condensation of the unit in question and at least one other unit that is "similar." Units are deemed to be similar whenever they differ in length by no more than four

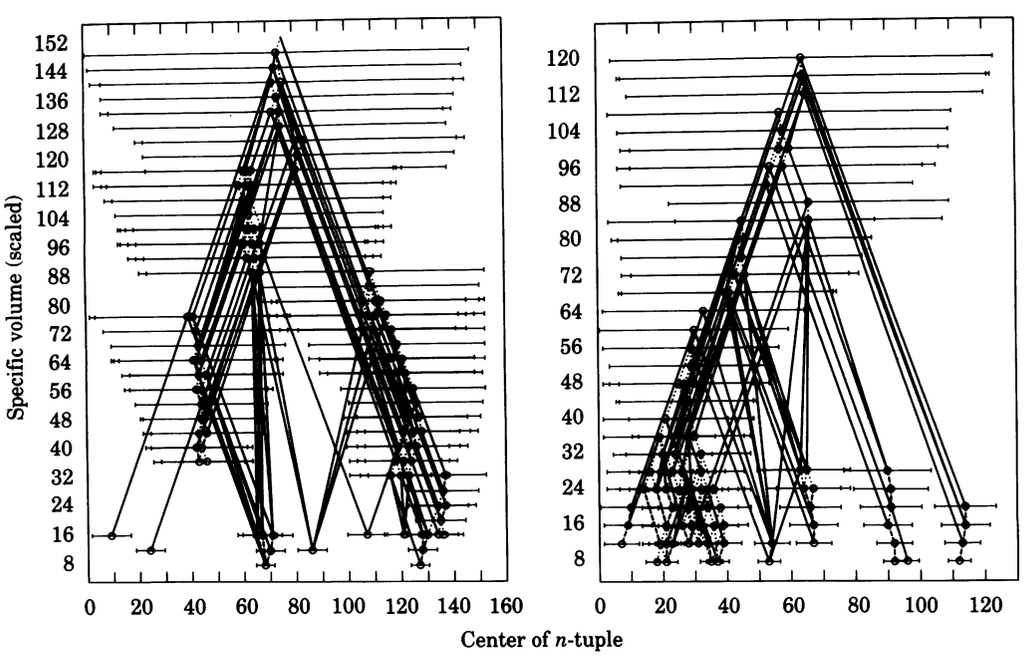

FIG. 2. Inclusion trees for Mb (*Left*) and RNase (*Right*). For each tree, units found to be compact are abstracted from Fig. 1 and interconnected as described in the text. The inclusion tree drawn in this way shows every possible hierarchic path between compact units.

amino acids and overlap by at least 50%. This process is more akin to growth than to condensation.

The procedure described here discovers all hierarchic paths of self-association between compact units. Inspection of the connecting paths in Fig. 2 reveals that many of the paths are similar in the general topology of their interactions and differ only in details. It is therefore possible to simplify the diagram by choosing only a single member to represent such a group.

In Fig. 3, we have abstracted representative nodes from the diagrams in Fig. 2 to form simplified inclusion trees. Each chosen representative is an actual node and is used in lieu of a cluster of similar nodes. These representative nodes are selected as follows. (*i*) In the case of a growth pathway (heavy dashed line), the largest unit that is not a dead end is taken as the representative. (*ii*) In the case of either a condensation pathway (solid line) or a growth and condensation pathway (light dashed line), the smallest node containing the same set of compact units is taken as the representative. These simplified trees are used for comparisons in the next section.

RESULTS

The method described in the preceding sections delineates the hierarchic folding paths between all compact, contiguous-chain units by using x-ray coordinates. With this method, we examined Mb and RNase. Each has been extensively studied both experimentally (21, 22) and in calculations (23–28).

Ptitsyn and Rashin (23) and later Richards and coworkers (24, 25) and Cohen *et al.* (27, 28) studied the folding of Mb helices. Ptitsyn and Rashin were specifically interested in hierarchic pathways.

An important aspect of the method presented here is that it does not require preidentification of the helical regions in Mb. Instead, the algorithm finds all compact units of minimum size, which, in the case of Mb, are the helices. The method is thus directly applicable to any x-ray protein structure.

The simplified inclusion tree for Mb (Fig. 3) can be compared with the pathways of self-organization described by Ptitsyn and Rashin (23) for this molecule. For example, it can be seen from Fig. 3 that the A helix cannot be merged with either B or BCD to form a compact unit; the smallest compact unit containing A is ABCDE. Additional differences are also apparent in Fig. 3.

Némethy and Scheraga (26) proposed a folding pathway for RNase by examination of the contact map. Based on their folding

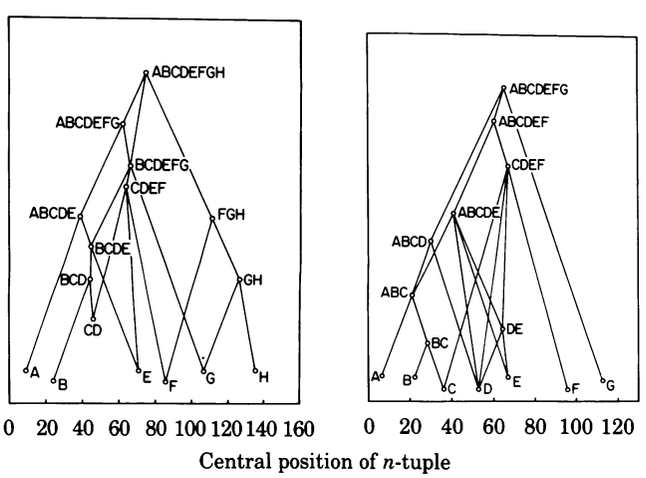

FIG. 3. Simplified inclusion trees for Mb (*Left*) RNase (*Right*). Each node in a simplified tree represents a group of similar nodes in the complete tree shown in Fig. 2.

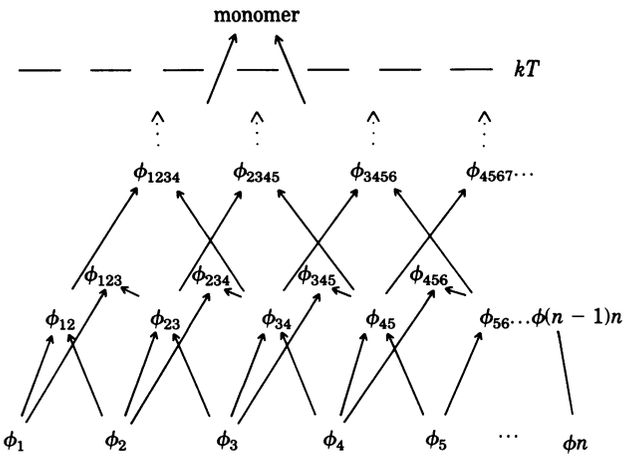

FIG. 4. Folding by hierarchic condensation. Local maxima in linear chain hydrophobicity (ϕ_1, ϕ_2, ϕ_3, . . . ϕ_n) serve as folding primitives (2, 3). These are brought together in stepwise fashion, forming successively larger intermediate structures and leading ultimately to the whole protein monomer. The smaller equilibrium intermediates are of low stability relative to kT and are readily interconvertible. Intermediates may terminate prematurely for steric reasons and will unfold as the population is pulled in the direction of successfully folding transtion states. Persisting intermediates will be visible in the native structure as compact units.

model (29), they suggested that the relative formation time of conformational intermediates can be read from the contact map by beginning at the diagonal and proceeding outward. Such a scheme is implicitly hierarchical.

We find the starting set of compact units for RNase in Fig. 3 to be approximately the same as that seen by Némethy and Scheraga, with the qualification that their region B corresponds to our regions B and C. However, we do find instances in which their postulated intermediates are not compact units by our criteria, possibly because a single intermediate may manifest itself as multiple sites of pairwise association on a contact map. We also find a larger number of possible pathways of self-association between units, all leading to the native structure.

DISCUSSION

In this paper, an algorithm to identify all compact structural units within an x-ray elucidated globular protein is introduced. The comprehensive set of hierarchic folding pathways for these units is then developed.

It was shown that the larger domains arise from combinations of smaller domains in an interative fashion. Every such unit is, by definition, a contiguous stretch of the linear polypeptide chain. This result both confirms earlier observations of hierarchic organization in proteins (4, 5) and identifies the full set of compact contiguous-chain units that serve as building blocks for the final protein structure.

The existence of hierarchic ordering in proteins implies that the structural interaction between domains is limited, in large part, to a superficial interdigitation of residues at the domain interface. The bringing together of domains in this way can minimize any change in backbone entropy upon association. In this view, the domains observed in the native protein are remnants of structural intermediates in the folding process, and they come about because the polymer chain organizes itself during folding so as to preserve the identity, although not necessarily the conformation, of these intermediates. Stated in terms of the folding process, residues that are near each other in space in the native structure will either be neighbors in the sequence or will have come together through a series of condensation

steps between intermediates composed of continuous stretches of the polypeptide chain.

The hierarchic organization of domains is structural evidence in favor of a model of folding by hierarchic condensation (9). In this model, neighboring hydrophobic chain elements interact to form folding clusters (2, 3), with further stepwise cluster association giving rise to a population of folding intermediates. The complete set of folding pathways leading to the native state can be described by a folding tree with nodes that correspond to one of these intermediates, as depicted in Fig. 4. Upon completion of folding, the intermediates would then be seen as structural domains in the native structure.

We thank Christian Sander, Shoshana Wodak, and Tom Creighton for useful discussion and Gene Davidson for his critical reading of the manuscript. Jack Seltzer of the University of Delaware Computer Center derived the formula for the surface area of an ellipsoid. X-ray coordinates were generously provided by S. Phillips (sperm whale Mb) and by the Protein Data Bank (30) (RNase, 2RSA, contributed by A. Wlodawer). This work was initiated during a 1979 summer workshop on protein structure sponsored by the Centre Européen de Calcul Atomique et Moléculaire in Orsay, France; and it was supported in part by Grants GM 29458 from the National Institutes of Health and PCM 8012007 from the National Science Foundation and by a U.S. Public Health Service Research Career Development Award to G.D.R.

1. Nozaki, T. & Tanford, C. (1971) *J. Biol. Chem.* **246,** 2211–2217.
2. Rose, G. D. (1978) *Nature (London)* **272,** 586–590.
3. Rose, G. D. & Roy, S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 4643–4647.
4. Rose, G. D. (1979) *J. Mol. Biol.* **134,** 447–470.
5. Crippen, G. M. (1978) *J. Mol. Biol.* **126,** 315–332.
6. Levitt, M. & Chothia, C. (1976) *Nature (London)* **261,** 552–558.
7. Richardson, J. S. (1977) *Nature (London)* **265,** 495–500.
8. Sternberg, M. J. E. & Thornton, J. (1977) *J. Mol. Biol.* **110,** 285–296.
9. Rose, G. D. (1980) *Biophys. J.* **32,** 419–422.
10. Kanehisa, N. I. & Tsong, T. Y. (1980) *Biopolymers* **19,** 1617–1628.
11. Creighton, T. E. (1978) *Prog. Biophys. Mol. Biol.* **33,** 231–297.
12. Creighton, T. E. (1980) in *Protein Folding,* ed. Jaenicke, R. (Elsevier/North-Holland, New York), pp. 427–441.
13. Lim, V. (1980) in *Protein Folding,* ed. Jaenicke, R. (Elsevier/North-Holland, New York), pp. 149–164.
14. Pattee, H. H., ed. (1973) *Hierarchy Theory* (Braziller, New York).
15. Goldstein, H. (1950) *Classical Mechanics* (Addison-Wesley, Reading, MA).
16. Tanford, C. (1961) *Physical Chemistry of Macromolecules* (Wiley, New York).
17. Sundaram, K. & Viswanadhan, V. N. (1980) *Physiol. Chem. Phys.* **12,** 187–191.
18. Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng.* **6,** 151–176.
19. Wodak, S. J. & Janin, J. (1980) *Proc. Natl. Acad. Sci.USA* **77,** 1736–1740.
20. Sander, C. (1981) *Structural Aspects of Recognition and Assembly in Biological Macromolecules* (Balaban, Philadelphia), pp. 183–196.
21. Baldwin, R. L. & Creighton, T. E. (1980) in *Protein Folding,* ed. Jaenicke, R. (Elsevier/North-Holland, New York), pp. 217–260.
22. Baldwin, R. L. (1980) in *Protein Folding,* ed. Jaenicke, R. (Elsevier/North-Holland, New York), pp.369–384.
23. Ptitsyn, O. B. & Rashin, A. A. (1975) *Biophys. Chem.* **3,** 1–20.
24. Richmond, R. J. & Richards, F. M. (1978) *J. Mol. Biol.* **119,** 537–555.
25. Cohen, F. E., Richmond, T. J. & Richards, F. M. (1979) *J. Mol. Biol.* **132,** 275–288.
26. Nemethy, G. & Scheraga, H. A. (1979) *Proc. Natl. Acad.Sci. USA* **76,** 6050–6054.
27. Cohen, F. E. & Sternberg, M. J. E. (1980) *J. Mol. Biol.* **137,** 9–22.
28. Cohen, F. E., Sternberg, M. J. E., Phillips, D. C., Kuntz, I. D. & Kollman, P. A. (1980) *Nature (London),* **286,** 632–634.
29. Tanaka, S. & Scheraga, H. A. (1977) *Macromolecules* **10,** 291–304.
30. Bernstein, F. C., Koetzle, T. G., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–42.