Table V
Polyisoprenes of M_n = 50 000 Prepared Using Diplit 5.1

Initiator age, days	Polyisoprene characteristics		
	$M_n \times 10^{-3}$	M_w/M_n	M_n of top 10% $\times 10^{-3}$
1	39	1.38	90–506
2	41	1.42	97–510
3	50	1.30	106–533
6	49	1.31	106–514
7	46	1.29	96–350

Table VI
Polymerization Time vs. Molecular Weight for Polyisoprenes of M_n = 25 000 Prepared with Diplit 5.1

Polymerization time at 60 °C, h	Polyisoprene characteristics		
	$M_n \times 10^{-3}$	M_w/M_n	M_n of top 10% $\times 10^{-3}$
1	25	1.27	52–166
2	27	1.27	56–243
3	28	1.29	56–238
4	27	1.31	56–238

The relationship found to hold between initiator makeup and characteristics of polyisoprene of M_n = 25 000 also holds for higher molecular weight polymers. Data on the use of diplit 5.1 as the initiator in polymerizations calculated to give polyisoprenes of 50 000 molecular weight are given in Table V. Correspondence between calculated and actual M_n was within usual experimental variation. Polydispersity in all cases was well below 1.5.

In all these experiments, a polymerization time of 2 h was arbitrarily used. Though 2 h at 60 °C is known to result in high conversion of isoprene to polyisoprene, it was of interest to examine various polymerization times and relate the effect of polymerization time to differences in the product. We found very little difference in products prepared using shorter or longer polymerization periods. This is illustrated in Table VI which shows results obtained by GPC on polymerizations run at 60 °C for 1 to 4 h. Molecular weight is quite constant varying only from 25 000 to 27 000, which is within experimental error and dispersities ranged from 1.27 to 1.31.

Acknowledgment. The authors gratefully acknowledge the assistance of Drs. D. D. Bly, J. R. Downing, and the late T. D. Swartz who developed the GPC procedures and molecular weight data supporting this work. We are also indebted to Dr. H. N. Cripps for developing the procedure for preparing *m*-diisopropenylbenzene.

References and Notes

(1) Contribution No. 2334.
(2) L. J. Fetters and M. Morton, *Macromolecules*, **2**, 435 (1969).
(3) G. Karoly, *Polym. Prepr., Am. Chem. Soc., Div. Polym. Chem.*, **10**, 837 (1969).
(4) Lithium Corporation of America, Product Bulletin No. 192.
(5) W. H. Sharkey, R. P. Foss, H. W. Jacobson, and H. N. Cripps, *Polym. Prepr., Am. Chem. Soc., Div. Polym. Chem.*, **17**, 211 (1976).
(6) M. Morton, L. J. fetters, J. Inomata, D. C. Rubio, and R. N. Young. *Rubber Chem. Technol.*, **49**, 303 (1976).
(7) M. Morton and L. J. Fetters, *Rubber Chem. Technol., Chem. Rev.*, **48**, 359 (1975).
(8) H. Gilman, W. Langham, and F. W. Moore, *J. Am. Chem. Soc.*, **92**, 2327 (1970).
(9) H. Gilman, F. W. Moore, and O. Baine, *J. Am. Chem. Soc.*, **63**, 2479 (1941).
(10) C. W. Kamienski and R. C. Morrison, U.S. Patent 3 954 894, May 4, 1976, to Lithium Corporation of America.

Hypothesis about the Mechanism of Protein Folding[1]

Seiji Tanaka and Harold A. Scheraga*[2]

Department of Chemistry, Cornell University, Ithaca, New York 14853.
Received June 1, 1976

ABSTRACT: A three-step mechanism of protein folding, proposed in our previous paper, is applied here to postulate the nature of the intermediates in the folding of rubredoxin, ferricytochrome *c*, and lysozyme. Contact maps are calculated for these three proteins, and it is shown that they contain much information (such as the polarity of residues in contact regions) about the structure of the native protein. Elementary processes are described for the formation of contact regions. Based on these concepts, details of the pathways of folding these three proteins from the unfolded to the native structure are postulated, focusing on the formation of ordered backbone structures (such as α-helical, extended, and chain-reversal conformations) in step A of the three-step mechanism and on the formation of contact regions in response to medium- and long-range interactions in steps B and C. It was found that chain reversals can often play an important role in forming contact regions in step A (short-range interactions) and in step B (medium-range interactions) but not in step C.

In a previous paper,[3] a hypothesis was proposed for protein folding, wherein the globular structure of a native protein is thought to form by a three-step mechanism. It was also demonstrated[3] that such a three-step mechanism is necessary if the globular structure is to result from the folding process.

This proposed mechanism for the folding of a polypeptide chain to the globular structure of a native protein in a given medium involves the following three steps (which *may* proceed simultaneously): (A) Because of short-range interactions,[4] ordered backbone structures,[7] such as α-helical, extended, and chain-reversal conformations, are formed in a system at equilibrium under given conditions (e.g., above the denaturation temperature). (B) When these physical conditions are changed (e.g., by changing the temperature and/or solvent composition), so as to introduce medium-range interactions, the equilibrium is shifted, and small contact regions (defined in Figure 1 of ref 3 and in section III of ref 3), involving medium-range interactions, are nucleated among the amino acid residues both in the ordered and in the unor-

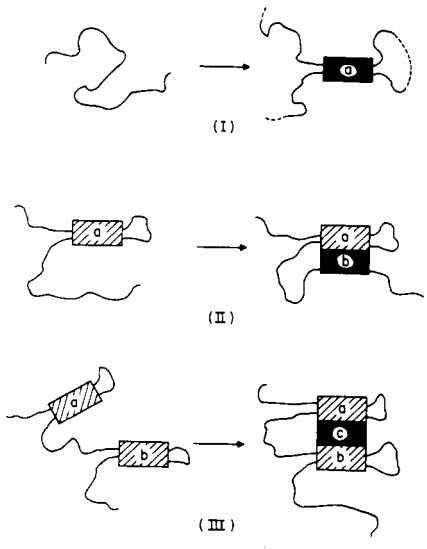

Figure 1. Schematic description of the elementary processes by which contact regions can form in steps B and C. The solid rectangles show the newly formed contact regions in each process, and the hatched rectangles show the contact regions that had already formed in the previous process. See section I of the text for a more detailed description of elementary processes (I)–(III).

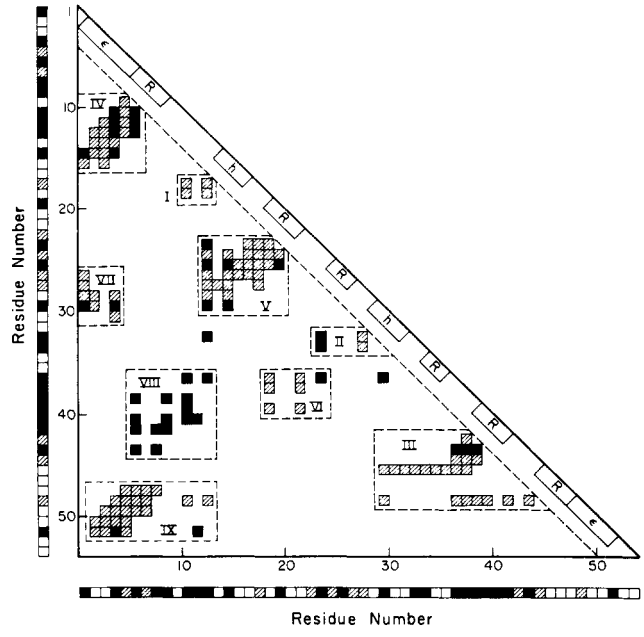

Figure 2. Contact map of native rubredoxin. The solid, hatched, and open rectangles in the vertical and horizontal runs outside the triangle designate amino acids with highly nonpolar side chains (Cys, Ile, Leu, Met, Phe, Trp, Tyr, Val, Ala, Pro), those with weakly nonpolar or weakly polar side chains (Asn, Gln, Gly, Ser, Thr), and those with highly polar side chains (Arg, Asp, Glu, His, Lys), respectively. The solid and hatched squares *within* the triangle designate contact between two amino acids with highly nonpolar side chains and any other two amino acids, respectively. Contacts from i to $i+1, i+2, i+3$, and $i+4$ (i.e., $|i-j| \leq 4$) were omitted from the diagram in order to focus attention on the medium- and long-range interactions in steps B and C. Instead of showing the short-range contacts, the *locations* of the ordered (helical, extended, and chain-reversal) backbone conformations formed in response to short-range interactions are designated by helical (h), extended (ϵ), and chain-reversal (R) on the diagonal (actually, between the diagonal and the dashed line parallel to it); the diagonal is the locus of single residues and does *not* represent *contacts* between pairs of residues. (As far as the chain-reversal conformations are concerned, the locations of those formed not only in step A [types (i) and (iii)] but also those formed in step B [type (ii)] are shown on the diagonal). Duplication between chain-reversal and helical, extended, or other chain-reversal conformations are shown by dotted lines in the diagonal region. The Roman numerals in the domains bounded by dashed lines designate areas discussed in the text.

dered structures.[7] In this step, the ordered backbone structures formed in step A may be rearranged to some extent to form stable intermediate structures in these contact regions. (C) Finally, the small contact regions formed in the intermediate structures in step B associate, in response to long-range interactions, with possibly further small rearrangements of the intermediate structures formed in steps A and B.

Statistical mechanical treatments in terms of one-dimensional short-range interaction models were developed to describe the conformations formed in step A;[8-13] i.e., in this model, ordered and other characteristic backbone conformations found in native proteins form in response to short-range interactions. By introducing medium- and long-range interactions (steps B and C) and using a Monte Carlo simulation of protein folding (applied to bovine pancreatic trypsin inhibitor),[3] it was demonstrated that the three-step mechanism is required to obtain the globular structure of the native protein.

In this paper, we show how this three-step mechanism can provide a picture as to how various other proteins fold from the unfolded to the globular form of the native structure. Since we have already described the details of step A in terms of statistical mechanical treatments,[8-13] we will concentrate in this paper on steps B and C. Instead of using a Monte Carlo simulation, as we did previously,[3] we will simply examine the contact maps[3] of native proteins and, from them, postulate how the contact maps should look during intermediate stages of folding; i.e., we will describe how the contact regions might form in steps B and C, according to the three-step mechanism. The specific role of chain-reversal conformations during folding will also be discussed.

In using a contact map of the native structure to postulate which contact regions exist at intermediate stages of folding, we cannot make any statement about structures that might exist at intermediate stages of folding but do not persist toward the end of the folding process. In essence, we are assuming that *most* intermediate structures *do* persist through step C, even though there may be small rearrangements of such structures throughout the folding process. This assumption seems reasonable because of the following two facts: (i) the ordered structures of proteins that are predicted with the one-dimensional, short-range interaction models[8-13] (that simulate the conformations in step A) correlate fairly well with those found in the native state, i.e., at the end of step C; (ii) it was observed,[3] in the Monte Carlo simulation of protein folding, that most of the contact regions formed in step B (see B-1 to B-3 of Figure 4 of ref 3) also appear during step C (see C-1 to C-6 of Figure 5 of ref 3), with small rearrangements and the formation of new contact regions (but with the maintenance or rearrangement of old contact regions) except in cases where little free energy is lost in the disruption of a contact region (e.g., an isolated contact region between two amino acid residues can be destroyed easily). Because of the small rearrangements that may take place in contact regions during folding, the intermediate contact regions (deduced here from the contact maps of native structures) should be regarded as "approximate" ones. Since the only intermediates postulated here also appear in the contact maps of the native protein, there is no question about being able to maintain these intermediate contact regions in a chain whose amino acid residues are joined together in the proper sequence.

I. Formation of Contact Regions

It was already demonstrated in our previous paper[3] that the three-dimensional structure of a protein can be represented

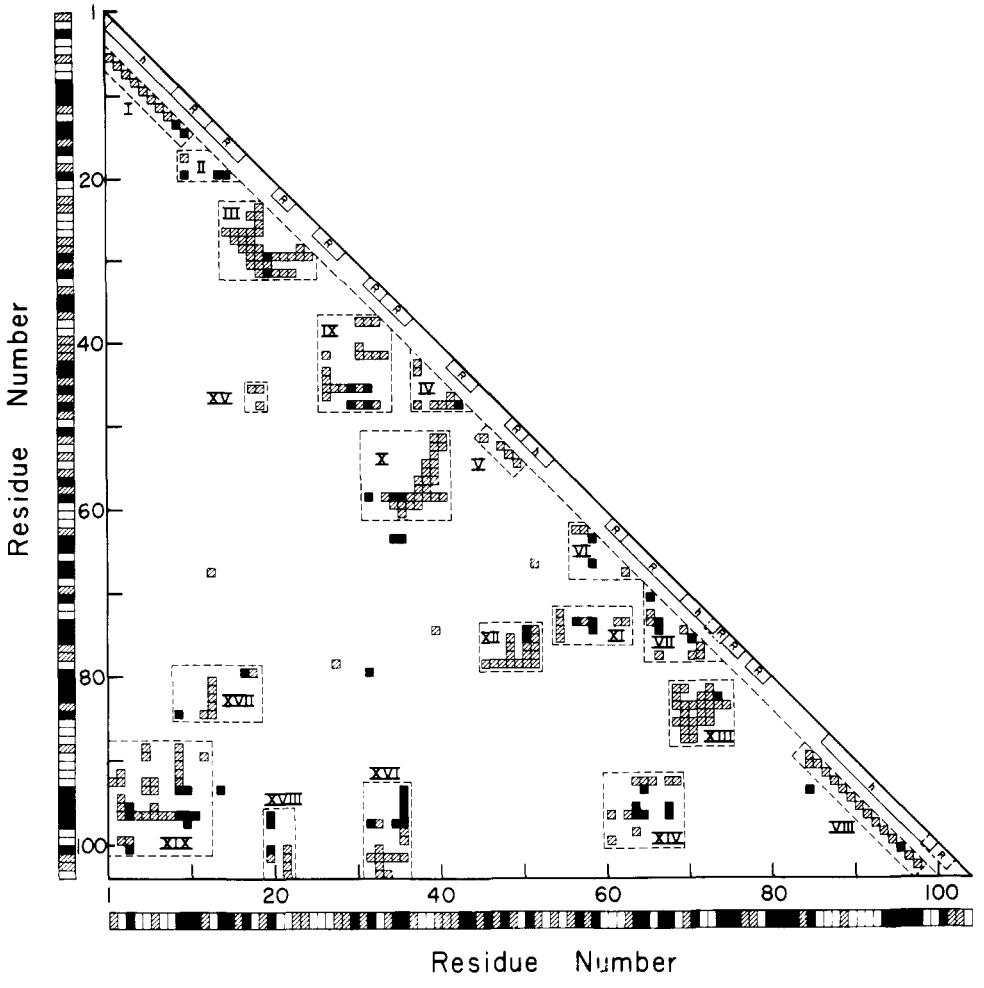

Figure 3. Same as Figure 2, but for native ferricytochrome *c*.

symbolically by the presence or absence of contacts between the amino acid residues of which the protein is constituted. The definition of a contact between two or more amino acids (as used here) was also presented in ref 3 (see, especially, eq 1 of ref 3). The folding mechanism can then be described by specifying how the contacts between amino acids are formed as the polypeptide chain passes from the unfolded to the globular form of the native protein.

Some of the contact regions found in native proteins, such as those arising simultaneously with the formation of helical, chain-reversal, and unordered conformations, can actually form in step A,[14] in response to short-range interactions. However, most of the contacts among the amino acids are formed in steps B and C; more specifically, as described in ref 3, the medium-range contacts are formed in step B, while the long-range ones are formed in step C. The formation of these contact regions is brought about by various kinds of chain movements, simulated by a Monte Carlo procedure in ref 3. As a result of these various chain movements, the contact regions form by simple elementary processes or combinations thereof. Three elementary processes, by which a contact region can form, are the following:

(I) Two segments (those that have not yet formed a contact region) are brought together to form a new contact region, *a*, as seen in (I) of Figure 1. Such a region can be formed by short-, medium-, or long-range interactions, i.e., in steps A, B, or C.

(II) A contact region, *a*, that had already been formed in process (I) forms a new contact region, *b*, with another segment that had not been involved in a contact region, as seen in (II) of Figure 1. Region *b* is formed by medium- or long-range interactions, i.e., in steps B or C, because the contact regions that formed previously participate in this process.

(III) Two previously formed contact regions, *a* and *b*, are brought together to form a new additional contact region, as seen in (III) of Figure 1. Type (III) contact regions are formed predominantly at the end of step B or in step C.

The criterion for deciding when a particular contact region is formed is given in section III.

As folding proceeds, combinations of these elementary processes result. For example, three contact regions, *a*, *b*, and *c*, may form an additional contact region, *d*, or two additional contact regions, *d* and *e*. We will designate all such combinations of elementary processes (I)–(III) as type (IV). Most contact regions formed in the final stages of step C are of type (IV); hereafter, if a contact region in step C is not specified as types (I)–(III), it should be regarded as a type (IV).

II. Calculation of Contact Maps for Globular Proteins

We will use contact maps[3] to represent the contact regions between amino acids and thereby describe the globular form of the native protein (and the partially folded forms of postulated intermediate structures). To represent the native protein, we use the x-ray coordinates and evaluate the contacts between the ith and jth residues [$1 \leq (i, j) \leq N$ and $i > j$, where N is the chain length]; a contact is said to exist if at least one pair of atoms (one atom in the ith and one in the jth residue) satisfies eq 1 of ref 3. This method of calculating a contact map is the same as that used to obtain Figure 1A of ref 3. The contact maps thus obtained are shown in Figures 2–4 for rubredoxin, [15]ferricytochrome *c*,[16] and lysozyme.[17]

Table I
Structures Nucleated in Step A

Proteins	No. of amino acid residues	Ordered backbone structure[a]	Residues involved in ordered backbone structure[b]	Contacts formed by ordered backbone structure[c]	Contact region on triangle map[d]	e
Rubredoxin[f]	54	ϵ	2–6			
		h	15–17	18···13, 19···13	I	2/4
		h	30–32	33···28, 34···28	II	2/4
		ϵ	49–52			
Ferricytochrome c[g]	104	h, R[h]	2–9 (h)	6···1, 7···2, 8···3, 9···4, 10···5, 11···6, 12···7	I	10/10
			9–13 (R)[i]	(11···6),[j] (12···7),[j] 13···8, 14···9, 15···10		
		R	14–17 (R)[i]	20···14, 20···15	II	2/4
		R	43–45 (R)[i]	47···42, 48···40, 48···41, 48···42, 48···43	IV	5/8
		R[h], h	50–51 (R)	53···48, 54···49,	V	4/4
			52–54 (h)	54···49, 55···50		
		R[h], h, R[h]	64–70 (R)[i]	71···66, 73···66,	VII	9/11
			71–74 (h)	75···70, 76···71, 77···72,		
			74–75 (R)	76···71, 77···72, 78···71, 78···72		
		h, R[h]	88–100 (h)	90···85, 91···85, 91···86, 92···87, 93···88, 94···85, 94···89, 95···90, 96···91, 97···92, 98···93, 99···94, 100···95, 101···96, 102···97, 103···98	VIII	16/16
			100–102 (R)[i]	(102···97),[k] (103···98)[k]		
Lysozyme[l]	129	ϵ	1–4			
		h, R[h]	5–14 (h)	7···2, 8···3, 9···4, 10···5, 11···3, 11···6, 12···7, 13···8 14···9, 15···10, 17···12	I	13/14
			14–15 (R)	(17···12),[m] 18···12, 18···13		
		h, R[h]	25–35 (h)	28···23, 30···25, 31···26, 32···27, 33···28, 34···29, 35···30, 36···31, 37···32, 38···29, 38···30, 38···32, 38···33	II	13/13
			35–36 (R)	(37···32),[n] (38···32),[n] (38···33)[n]		
		ϵ	43–46			
		h, R[h], h	80–84 (h)	87···82	IV	7/7
			86–87 (R)	88···83, 90···85		
			89–99 (h)	93···88, 94···89, 95···90, 96···91		
		R,[h] h, R[h]	104–108 (R)[i]	111···103, 111···104, 111···105, 111···106	VI	11/13
			109–113 (h)	(111···106),[o] 112···106, 112···107, 113···108, 114···109, 115···110, 116···106, 116···111		
			113–115 (R)[i]	(115···110),[o] (116···111)[o]		
		R,[h] R,[h] h, R[h]	116–117 (R)	120···115	VII	4/6
			120–123 (h)	123···118, 125···119, 125···120		
			123–124 (R)	(125···120)[p]		

[a] The symbols h, ϵ, and R designate helical, extended, and chain-reversal conformations. This use of the symbol R differs from that in ref 11–13. [b] These residues are the ones involved in the ordered backbone conformations given in column 3 and are given on the diagonal in Figures 2 to 4. Some chain reversals [the category (ii) ones, listed in column 3 of Table V] are not listed here [because they are not formed in step A, i.e., they are not category (i) or (iii) chain reversals]; however, they are shown on the diagonal in Figures 2 to 4 to indicate the location of all chain reversals. The observed h, ϵ, and R regions are quoted from Table I of ref 11. [c] Contacts formed by the ordered backbone structures of columns 3 and 4. The first and second numbers are the ith and jth residues, respectively (see also column 2 of Tables II–IV). [d] The contact regions corresponding to the Roman numerals in this column can be found in Figures 2 to 4. [e] (Number of contacts formed by ordered backbone structure)/(total number of contacts in region). The number of contacts formed by the ordered backbone structure is the sum of the contacts given in column 5, and the total number of contacts in the region is the sum of the contacts listed in column 2 of Tables II–IV. [f] See ref 15 for the x-ray coordinates. [g] See ref 16 for the x-ray coordinates. [h] The contact region in column 6 consists of two or more ordered backbone structures out of h, ϵ, and R and is designated, for example, as h, R in column 3. [i] A multiple chain-reversal region.[21] [j] Contacts 11···6 and 12···7 may also be considered to be formed by helical sequence 2–9 [since they involve the residues in 2–9 (h)] and are not counted twice in the number of contacts given in the numerator in column 7. [k] Contacts 102···97 and 103···98 were already counted above as those formed by helical sequence 88–100 and are not included in the number of contacts given in the numerator in column 7. [l] See ref 17 for the x-ray coordinates. [m] Contact 17···12 was already counted above as that formed by helical sequence 5–14 and is not included in the number of contacts given in the numerator in column 7. [n] Contacts 37···32, 38···32, and 38···33 were already counted above as those formed by helical sequence 25–35; see footnote *m*. [o] Contacts 111···106 and 115···110 and 116···111 were already counted above as those formed by 104–108 (R) and by 109–113 (h), respectively; see footnote *m*. [p] Contact 125···120 was already counted above as that formed by 120–123 (h); see footnote *m*.

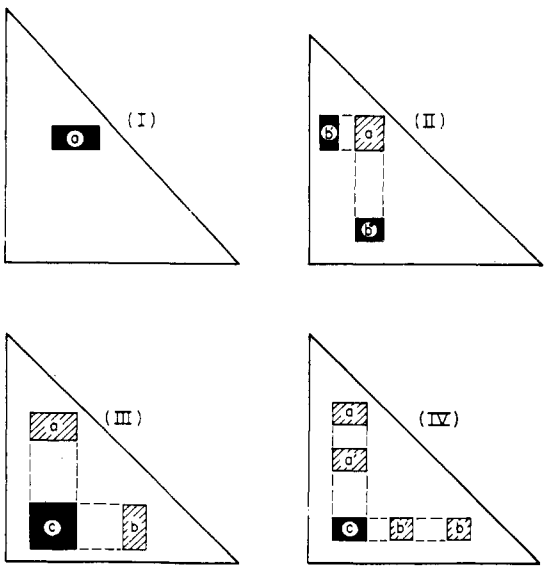

Figure 5. Schematic contact map corresponding to the elementary processes, (I)–(III) of Figure 1 [and also process (IV)], for formation of contact regions. See text for details.

of Figure 5. If a, a', b, and b' form at an earlier stage than c, then the next stage of folding [type (IV)] can be represented by the following reaction:

$$a + a' + b + b' \rightarrow a + a' + b + b' + c$$

where the relative positions of these regions are as shown in (IV) of Figure 5; i.e., the contact regions in the range bounded by the vertical and horizontal dashed lines in (IV) of Figure 5 contribute to the formation of the new contact region, c.

In summary, the following criteria are used to deduce the structures formed in each step: (i) In step A, ordered backbone structures are formed, i.e., α-helical, extended, and *some* chain-reversal conformations [those classified as category (i) and (iii) chain reversals in section IV]. These ordered structures (except for extended sequences[14]) may also serve to bring other nearby residues together to form contact regions in step A; these nearby residues can be in the range of $5 \leq \text{Min}|i - j| \leq 20$ [e.g., see the role of category (i) chain reversals in section IV] if the following criteria are satisfied:

$$k + 1 - l \leq [\text{the values of both } i \text{ and } j \text{ in contact}] \leq k + m + l \quad (1)$$

where i, j, and k indicate positions in the amino acid sequence, and the *ordered* backbone conformation $\{\rho\}$ consists of the m residues from $k + 1$ to $k + m$, i.e.,

$$\{\rho\} = \eta_{k+1} \cdots \eta_{k+m} \quad (2)$$

where η designates a particular conformational state of a residue, e.g.,

$$\{\rho\} = \text{hh} \ldots \text{hh} \quad (3)$$

for a helical sequence,[14] or

$$\{\rho\} = \text{RS or RDD} \ldots \text{DDS} \quad (4)$$

for chain-reversal or multiple chain-reversal conformations,[20,21] respectively. The value of l designates the range of the possible contact formed by the ordered backbone structure(s) in step A; in this paper, we arbitrarily take $l = 3$ as a criterion to assign such (otherwise medium range) contacts formed by ordered (i.e., α-helical or chain reversal)[14] backbone structures to step A. The residues i and j that are in contact do not necessarily lie in the ordered sequence $\{\rho\}$. If more than 50% of the contacts, in a contact region formed by the ordered backbone structure, satisfy eq 1, then the contact region is said to arise by formation of the ordered backbone structure ($\{\rho\}$ of eq 2) in step A (see column 7 of Table I).

(ii) In step B, contact regions that satisfy the criterion[3,4] $5 \leq \text{Min}|i - j| \leq 20$ are formed because of local medium-range interactions. As described in section IV [category (ii) chain reversals], the chain-reversal conformation plays a role in the formation of a contact region in this step; in order to identify this category of chain reversals, we use the following criteria: a chain reversal (indicated by eq 4) will bring about a contact between residues i and j if it occurs between i and j ($i > j$) in the amino acid sequence, i.e.,

$$j < \text{position of chain reversal} < i \quad (5)$$

and satisfies the relations

$$k + 1 - l' \leq [\text{the values of both } i \text{ and } j \text{ in contact}] \leq k + m + l'' \quad (6)$$

and

$$5 \leq l' + m + l'' \leq 20 \quad (7)$$

The values of l' and l'' are varied, from case to case, from 1 to a maximum value (depending on the value of m) to satisfy eq 7. Expression 7 is the criterion for designating interactions of medium-range[4] order. If more than 50% of the contacts in a contact region satisfy eq 6 and 7 (i.e., are formed by a chain reversal), then the chain reversal is regarded as contributing to the formation of the contact region (see column 6 of Table V).

Thus, contact regions with $5 \leq \text{Min}|i - j| \leq 20$ can form in either step A or step B. They are assigned to step A if the formation of such a region is promoted by formation of an ordered backbone conformation of short-range order, satisfying eq 1–4. They are assigned to step B if these contacts of medium-range order are brought about by a category (ii) chain reversal and satisfy eq 5–7.

(iii) In step C, contact regions (separated by more than about 20 residues from each other,[3] i.e., those with $\text{Min}|i - j| > 20$) are formed in response to long-range interactions.

The actual appearances of the contact maps, postulated to arise during the three-step folding of proteins, will be shown below for the three proteins used as illustrations here.

(A) Rubredoxin. The contact map of rubredoxin is depicted in Figure 2. The nucleation of the regions of short-range ordered backbone structures[4] (in the range of $|i - j| \leq 4$) that are postulated to form in step A are shown (but only in terms of the symbols h, ϵ, and R) along the diagonal (and not *within* the contact map) in Figure 2. The residues that are in contact in each contact region are given in column 2 of Table II. In column 3 of the same table, the smallest value of $|i - j|$ in the contact region ($\text{Min}|i - j|$) is given, where i and j are the residues that are in contact; this is a measure of the order of formation of the contact region. As described in the second and fourth paragraphs of section III, contact regions with $5 \leq \text{Min}|i - j| \leq 20$ are formed in step A or step B, depending on whether the contact is due to the formation of an ordered backbone conformation (step A) or not (step B). Contact regions with $\text{Min}|i - j| > 20$ are formed in step C. For example, regions I–VI (see column 3 of Table II) are formed in either step A or step B, while regions VII–IX are formed in step C. To obtain a unique folding sequence, we assume that contact regions form in the order given by $\text{Min}|i - j|$ in each of steps A, B, and C; probably, the contact regions form simultaneously in any one of these three steps.

The ordered backbone structures that are formed in step A are the helical sequences at residues 15–17 and 30–32 and the extended sequences at residues 2–6 and 49–52 (see Table I). Contact region I is formed not only by nucleation of the

Table I
Structures Nucleated in Step A

Proteins	No. of amino acid residues	Ordered backbone structure[a]	Residues involved in ordered backbone structure[b]	Contacts formed by ordered backbone structure[c]	Contact region on triangle map[d]	e
Rubredoxin[f]	54	ϵ	2–6			
		h	15–17	18···13, 19···13	I	2/4
		h	30–32	33···28, 34···28	II	2/4
		ϵ	49–52			
Ferricytochrome c[g]	104	h, R[h]	2–9 (h)	6···1, 7···2, 8···3, 9···4, 10···5, 11···6, 12···7	I	10/10
			9–13 (R)[i]	(11···6),[j] (12···7),[j] 13···8, 14···9, 15···10		
		R	14–17 (R)[i]	20···14, 20···15	II	2/4
		R	43–45 (R)[i]	47···42, 48···40, 48···41, 48···42, 48···43	IV	5/8
		R[h], h	50–51 (R)	53···48, 54···49,	V	4/4
			52–54 (h)	54···49, 55···50		
		R[h], h, R[h]	64–70 (R)[i]	71···66, 73···66,	VII	9/11
			71–74 (h)	75···70, 76···71, 77···72,		
			74–75 (R)	76···71, 77···72, 78···71, 78···72		
		h, R[h]	88–100 (h)	90···85, 91···85, 91···86, 92···87, 93···88, 94···85, 94···89, 95···90, 96···91, 97···92, 98···93, 99···94, 100···95, 101···96, 102···97, 103···98	VIII	16/16
			100–102 (R)[i]	(102···97),[k] (103···98)[k]		
Lysozyme[l]	129	ϵ	1–4			
		h, R[h]	5–14 (h)	7···2, 8···3, 9···4, 10···5, 11···3, 11···6, 12···7, 13···8 14···9, 15···10, 17···12	I	13/14
			14–15 (R)	(17···12),[m] 18···12, 18···13		
		h, R[h]	25–35 (h)	28···23, 30···25, 31···26, 32···27, 33···28, 34···29, 35···30, 36···31, 37···32, 38···29, 38···30, 38···32, 38···33	II	13/13
			35–36 (R)	(37···32),[n] (38···32),[n] (38···33)[n]		
		ϵ	43–46			
		h, R[h], h	80–84 (h)	87···82	IV	7/7
			86–87 (R)	88···83, 90···85		
			89–99 (h)	93···88, 94···89, 95···90, 96···91		
		R,[h] h, R[h]	104–108 (R)[i]	111···103, 111···104, 111···105, 111···106	VI	11/13
			109–113 (h)	(111···106),[o] 112···106, 112···107, 113···108, 114···109, 115···110, 116···106, 116···111		
			113–115 (R)[i]	(115···110),[o] (116···111)[o]		
		R,[h] R,[h] h, R[h]	116–117 (R)	120···115	VII	4/6
			120–123 (h)	123···118, 125···119, 125···120		
			123–124 (R)	(125···120)[p]		

[a] The symbols h, ϵ, and R designate helical, extended, and chain-reversal conformations. This use of the symbol R differs from that in ref 11–13. [b] These residues are the ones involved in the ordered backbone conformations given in column 3 and are given on the diagonal in Figures 2 to 4. Some chain reversals [the category (ii) ones, listed in column 3 of Table V] are not listed here [because they are not formed in step A, i.e., they are not category (i) or (iii) chain reversals]; however, they are shown on the diagonal in Figures 2 to 4 to indicate the location of all chain reversals. The observed h, ϵ, and R regions are quoted from Table I of ref 11. [c] Contacts formed by the ordered backbone structures of columns 3 and 4. The first and second numbers are the ith and jth residues, respectively (see also column 2 of Tables II–IV). [d] The contact regions corresponding to the Roman numerals in this column can be found in Figures 2 to 4. [e] (Number of contacts formed by ordered backbone structure)/(total number of contacts in region). The number of contacts formed by the ordered backbone structure is the sum of the contacts given in column 5, and the total number of contacts in the region is the sum of the contacts listed in column 2 of Tables II–IV. [f] See ref 15 for the x-ray coordinates. [g] See ref 16 for the x-ray coordinates. [h] The contact region in column 6 consists of two or more ordered backbone structures out of h, ϵ, and R and is designated, for example, as h, R in column 3. [i] A multiple chain-reversal region.[21] [j] Contacts 11···6 and 12···7 may also be considered to be formed by helical sequence 2–9 [since they involve the residues in 2–9 (h)] and are not counted twice in the number of contacts given in the numerator in column 7. [k] Contacts 102···97 and 103···98 were already counted above as those formed by helical sequence 88–100 and are not included in the number of contacts given in the numerator in column 7. [l] See ref 17 for the x-ray coordinates. [m] Contact 17···12 was already counted above as that formed by helical sequence 5–14 and is not included in the number of contacts given in the numerator in column 7. [n] Contacts 37···32, 38···32, and 38···33 were already counted above as those formed by helical sequence 25–35; see footnote *m*. [o] Contacts 111···106 and 115···110 and 116···111 were already counted above as those formed by 104–108 (R) and by 109–113 (h), respectively; see footnote *m*. [p] Contact 125···120 was already counted above as that formed by 120–123 (h); see footnote *m*.

Table II
Contact Regions Formed during the Folding of Rubredoxin

Contact region[a]	Residues involved in contact region[b]	Minimum value of $\|i-j\|$[c]	Contribution of contact regions formed previously[d]	
I	18 18 19 19 11 13 11 13	5		
II	33 33 34 34 24 28 24 28	5		
III	43 44 44 44 45 45 45 46 46 46 46 46 46 46 38 37 38 39 37 38 39 30 31 32 33 34 35 36 46 46 49 49 49 49 49 49 49 37 38 30 37 38 39 40 42 44	5	II (33, 34)	
IV	10 11 11 11 12 12 12 12 13 13 13 13 13 14 5 4 5 6 3 4 5 6 2 3 4 5 6 2 14 14 15 15 15 15 16 16 3 4 1 2 3 4 1 3	5	I (11, 13)	
V	24 24 24 24 25 25 25 25 25 25 26 26 26 26 26 13 17 18 19 13 15 17 18 19 20 13 15 16 17 18 26 26 27 27 27 27 27 28 28 28 28 29 29 30 30 19 20 13 15 16 17 18 13 14 15 18 13 15 13 15	5	I ([13],[e] 18, 19) II (24, 28) III (30) IV (13, 14, 15, 16)	
VI	37 37 37 38 38 40 40 19 22 24 19 22 19 22	13	III (37, 38, 40) V (19, 24)	[I (19)],[f] [II (24)]
VIII	37 37 39 39 39 40 41 41 41 41 42 42 42 44 44 11 13 6 9 11 11 6 9 11 12 6 8 9 6 8	24	III (37, 39, 40, 42, 44) IV (6, 11, 12, 13)	[I (11, 13)], [V (13)] [VI (37, 40)]
VII	27 28 29 29 29 30 30 30 31 1 1 1 2 4 1 2 4 4	25	IV (1, 2, 4) V (27, 28, 29, 30) III ([30], 31)	[II (28)]
IX	48 48 48 48 49 49 49 49 49 49 50 50 50 50 50 5 6 7 8 4 5 6 7 11 13 3 4 5 6 7 51 51 51 51 52 52 52 52 52 53 2 3 4 5 2 3 4 5 12 2	36	III (49) IV (2, 3, 4, 5, 6, 11, 12, 13) VIII ([6], 8, [11, 12, 13])	[I (11, 13)] [V (13)] [VII (2, 4)]

[a] See the contact map of Figure 2. The contact regions in this column are listed in the order of their formation during folding (see text). [b] These are the residues i and j that are in contact, with the value of j appearing directly under the corresponding value of i. [c] This is the minimum number of residues intervening between two residues i and j that are in contact in a contact region. [d] The contact region in column 1 is formed by contributions (i.e., associations by elementary processes such as those of Figure 1) from the contact regions in this column. The arabic numerals in parentheses denote the residues of the previously formed contact regions (given by Roman numerals) that contribute to the formation of the new contact region (i.e., the one given in column 1). The contact regions or residues in square brackets appeared earlier in other contact regions in this same column, e.g., see footnotes *e* and *f*. [e] Residue 13 appears in both regions I and IV. However, such a residue is enclosed in brackets (to indicate its appearance in *two* regions) in the region (i.e., region I) that contains the *smaller* number of residues in contact, there being three residues in region I and four residues in region IV. [f] Whereas footnote *e* deals with the appearance of the same *residue* in two or more contact regions, here we deal with two or more contact *regions*, where all residues in contact in one of them are the same as some of those in another region. For example, in forming region VI, region I contains only one residue in the relevant contact (i.e., residue 19), but region V also contains residue 19 in contact. Thus, a bracket is placed around the whole region, i.e., [I (19)]. The bracket is placed around the region with the *smaller* number of residues in contact, i.e., around I (19) and *not* around V (19, 24).

helical segment at residues 15–17 but also by contacts between residues 18···13 and 19···13 which incorporate this helical segment in the contact region (we designate a contact by specifying i before j; e.g., $i = 18$ and $j = 13$ in the 18···13 contact; see footnote *c* of Table I and footnote *b* of Tables II–IV); this region (see column 5 of Table I) meets the condition that more than 50% of the contacts in this region that includes this helical sequence satisfy eq 1 (i.e., two such contacts out of a total of four contacts in region I, as indicated in column 7 of Table I and column 2 of Table II). In a similar manner, contact region II is formed by nucleation of the helical segment at residues 30–32 and contacts between residues 33···28 and 34···28. Thus, contact regions I and II, as well as the extended backbone structures shown in Table I, are nucleated (due to formation of helical sequences) in step A; regions I and II are shown as A-1 and A-2 in Figure 6.

Following the formation of regions I and II, the ordered backbone structures and contact regions formed in step A associate (in step B) to form new contact regions of medium-range order ($5 \leq \text{Min}|i-j| \leq 20$), but without formation of other ordered backbone structures. Elementary processes (I) to (IV) are involved in step B; these are shown as B-1, B-2, B-3, and B-4 in Figure 6.

The following stages of step B can be identified.

(1) Contact region II (residues 33 and 34, and helical segment 30–32) associates [by a type (II) elementary process] with the unordered sequence at residues 43–49 (but not including 47 and 48) with the aid of chain-reversal formations at residues 35–36, 40–42, and 46–48 (see Table V) to form a new contact region III (B-2 of Figure 6); the contacts in region III are summarized in Table II. As seen in column 6 of Table V, all contacts in region III are formed by chain reversals 35–36, 40–42, and 46–48, since all contacts satisfy eq 6 and 7.

(2) The extended sequence at residues 2–6 (including the neighboring residue 1) associates [by a type (II) elementary

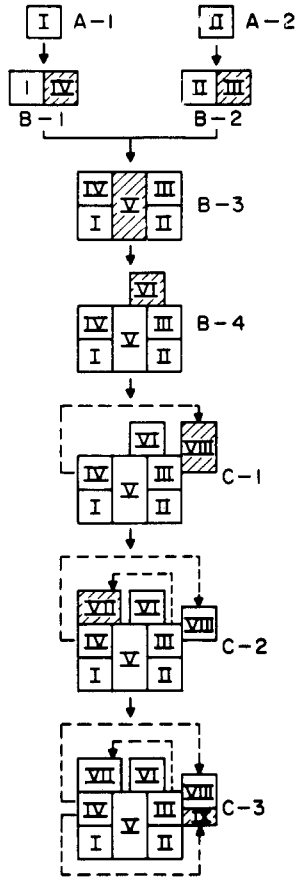

Figure 6. Schematic illustration of the mechanism of folding of rubredoxin. The Roman numerals correspond to those of Figure 2, and the folding is postulated to take place (by association of contact regions) from top to bottom of the figure. The hatched region is the new one formed at each stage. The dashed lines indicate contacts (in three dimensions) that cannot be represented easily in this two-dimensional diagram. This diagram is obtained by following the order of formation of the contact regions given in column 1 of Table II, with the participation of the contact regions given in column 4 of Table II; for simplicity, the duplicate regions given in column 5 of Table II are omitted from the diagram[22] (see, also, footnote *f* of Table II).

process] with region I (at residues 11 and 13) to form region IV (B-1 of Figure 6), with the aid of a chain reversal at residues 7–9 (see Table V). As seen in column 6 of Table V, all contacts in region IV are formed by the chain reversal at residues 7–9. The contacts in region IV are summarized in column 2 of Table II, and the contribution of region I to the formation of region IV is indicated in column 4 of Table II.

(3) Regions I (at residues [13], 18 and 19), II (at residues 24 and 28), III (at residue 30), and IV (at residues 13, 14, 15, and 16) associate by formation of chain reversals at residues 20–22 and 26–27 (see Table V) to form region V (B-3 of Figure 6). (See footnote *e* of Table II for the reason that residue 13 is placed in brackets in region I.)

(4) Since $\text{Min}|i-j| = 13$ for region VI, which is greater than $\text{Min}|i-j| = 5$ for regions III–V, region VI is formed later (in the final stage of step B) than regions III–V. Contact region III (at residues 37, 38, and 40) and region V (at residues 19 and 24) [or region I (at residue 19) and region II (at residue 24)] associate to form region VI (B-4 of Figure 6).[22]

In Step C, further association takes place to form contact regions involving interactions of long-range order ($\text{Min}|i-j| > 20$) (see Figure 2 and column 3 of Table II).

(1) Region III (at residues 37, 39, 40, 42, and 44) and region IV (at residues 6, 11, 12, and 13) associate to form region VIII[22] (C-1 of Figure 6), for which $\text{Min}|i-j| = 24$.

(2) Regions IV (at residues 1, 2, and 4), V (at residues 27–30), and III (at residues [30] and 31) associate to form region VII[22] (C-2 of Figure 6), for which $\text{Min}|i-j| = 25$.

(3) Finally, region IX (for which $\text{Min}|i-j| = 36$) is formed by association of regions III (at residue 49), IV (at residues 2–6 and 11–13), and VIII (at residues [6], 8, [11–13]) (C-3 of Figure 6).[22]

Thus, the folding of rubredoxin (the order of formation of contact regions and the specific regions that interact to form new contact regions) is represented schematically in Figure 6 and summarized in Table II. The roles of the chain-reversal conformations during the folding process are summarized in Table V and discussed in section IV.

(B) Ferricytochrome *c*. The contact map of ferricytochrome *c* is depicted in Figure 3, with the regions of short-range ordered backbone structures shown along the diagonal. The residues that are in contact in each contact region are given in column 2 of Table III in the order of formation, i.e., of increasing values of $\text{Min}|i-j|$.

In step A, short-range interactions would produce the ordered backbone structures listed in Table I, and these would lead to the formation of contact regions I, II, IV, V, VII, and VIII, for all of which $\text{Min}|i-j| = 5$ (see Table III). (The numbering system does not indicate the order of formation of these contact regions in step A) see below for the omission of regions III, VI, and IX from step A.

(1) The helical sequence at residues 2–9 leads to formation of contacts between residues 6···1, 7···2, 8···3, 9···4, 10···5, 11···6, and 12···7. A chain reversal at residues 9–13 (see column 4 of Table I) leads to contacts between residues (11···6), (12···7), 13···8, 14···9, and 15···10, where the parentheses indicate that the contacts 11···6 and 12···7 may be considered to be formed by either the helical sequence at residues 2–9 [since 11···6 and 12···7 appear in contacts formed by h(2–9)] or the chain reversal at residues 9–13. (All of the contacts mentioned above satisfy eq 1.) There are 10 such contacts formed by h(2–9) and R(9–13) which are all found in region I (see Tables I and III). Thus, region I is formed in step A by a combination of the helical sequence at residues 2–9 with the chain reversal at residues 9–13.

(2) The formation of the chain reversal at residues 14–17 leads to contacts between residues 20···14 and 20···15 (see Table I) that satisfy eq 1; these amount to 50% of the total number of contacts in region II, i.e., a total of four contacts at residues 18···10, 20···10, 20···14, and 20···15, as seen in Table III. Thus, contact region II (see Figure 3) can be considered as arising from the formation of a chain reversal at residues 14–17.

(3) The formation of a chain reversal at residues 43–45 introduces contacts between residues 47···42, 48···40, 48···41, 48···42, and 48···43, i.e., five contacts (see Table I) out of a total of eight in region IV (see Table III). Thus, region IV is formed by a chain reversal at residues 43–45.

(4) Combination of the helical sequence at residues 52–54 with the chain reversal at residues 50–51 forms contact region V, since contacts 54···49 and 55···50 are formed by h(52–54) and contacts 53···48 and 54···49 are formed by R(50–51) (all four contacts out of a total of four are involved in region V), as seen in column 7 of Table I and column 2 of Table III.

(5) Combination of the helical sequence at residues 71–74 with the chain reversals at residues 64–70 and 74–75 yields contact region VII, since R(64–70) forms contacts 71···66 and 73···66, h(71–74) forms contacts 75···70, 76···71, and 77···72, and R(74–75) forms contacts 76···71, 77···72, 78···71, and 78···72 (see column 5 of Table I); this amounts to 9 contacts out of a total of 11 in region VII (see column 2 of Table III).

(6) The helical sequence at residues 88–100 and the chain reversal at residues 100–102 (see column 5 of Table I for the contacts formed by these ordered backbone structures) form

16 contacts, which is the total that is found in region VIII (see Table III).

It should be noted from column 3 of Table III that, even though $\text{Min}|i-j| = 5$ for contact regions III, VI, and IX, they are not included in step A because they do not arise by formation of ordered backbone structures. Therefore, the formation of these three regions is assigned to step B.

The chain reversals that play a role in the formation of contact regions in step A (mentioned above) are listed in Table V.

The progress of step B is as follows:

(1) At an early stage of step B (since $\text{Min}|i-j| = 5$ in region III), the chain reversals at residues 22–23 and 27–29 bring contact region III together with residues 15–32 [as seen in Table V, the formation of contacts in region III does not satisfy the condition of eq 1 for a short-range contact formed by a chain reversal but does satisfy the conditions of eq 5–7 (for formation of a chain reversal in a medium-range contact)]. Therefore, region III is not attributed to step A but to step B. Region III involves 32 contacts, 31 of which are brought about by chain reversals at residues 22–23 and 27–29, as listed in Table V. The formation of region III, as indicated above, occurs by elementary process type (I).

(2) Region VI arises by formation of a chain reversal at residues 62–63 that brings two segments, 57–63 (except for residues 60–62) and 63–68 (except for residues 65 and 66) together by a type (I) elementary process. To be more specific, contacts 63···57, 63···58, 64···59, and 67···59 (see column 4 of Table V) are formed by the chain reversal at residues 62–63; that amounts to 4 contacts out of a total of 5, in region VI. [As indicated in stage (5) of step A, the chain reversal at residues 64–70 helps form contact region VII; although the formation of region VI in step B involves residues 64, 67, and 68, which are part of the chain reversal at residues 64–70, the chain-reversal character of resdues 64, 67, and 68 does not play a role in forming contact region VI.]

(3) Regions III (at residues 26–32), IV (at residues 38, 41, 42, 44, 47, and 48), and V (at residues 46 and [48]) are brought together by formation of chain reversals at residues 33–34 and 35–37. This forms 20 contacts (as listed in column 4 of Table V) out of a total of 22 (see Table III) in region IX [by a type (IV) elementary process].

(4) Then, region XIII, with $\text{Min}|i-j| = 9$, is formed by association of region VII (at residues 70–73) with region VIII (at residues 85–86), and with the additional residues 82–84 and 69 (as seen in columns 2 and 4 of Table III), with the aid of chain reversals at residues 76–77 and 79–80. This involves a type (III) elementary process.

(5) Regions IV (at residues 38, 40 and 41), V (at residues 52–55), VI (at residues 57–59), and IX (at residues [38], 32, and [41]) associate to form region X.

(6) Regions VI (at residues 57–59 and 63), VII (at residues 73–76), and X (at 55 and [57–59]) associate to form region XI.

Step C involves the formation of contact regions for which $\text{Min}|i-j| > 20$ and proceeds as follows:

(1) Regions V (at residues 46, 48, 49, 50, and 52), VII (at residues 75–78), and IX (at residues [46], 47, and [48]) associate to form region XII (see Table III).

(2) Region XIV is formed by association of regions VI (at residues 63, 64, and 68), VIII (at residues 94, 95, 98, and 99), and X (at residue 61), as seen in Table III.

(3) Region XV is formed by association of regions III (at residues 17 and 18) and V (at residues 46 and 48).

(4) Region XVI is then formed by association of regions VIII (at residues 94–103), IX (at residues [32] and 33), and X (at residues 32, 34, 35, and 36).

(5) Region XVII is then formed by association of regions I (at residues 9, 12, and 13), III (at residues 17 and 18), and XIII (at residues 82–85).

(6) Region XVIII is formed by association of region III (at residues 20 and 22) with region XVI (at residues 97, 98, and 101–104).

(7) Finally, regions I (at residues 1–12) and VIII (at residues 89–101, with the exception of 99) associate to form region XIX, completing the native structure of ferricytochrome *c*.

In summary, regions I, II, IV, V, VII, and VIII are nucleated in step A, as summarized in Table I. Then, in step B, contact regions III, VI, IX, XIII, X, and XI are formed. Finally, regions XII and XIV–XIX are formed in step C. The contacts involved in each region are listed in Table III. In Table V, the roles of the chain reversals are summarized. A schematic diagram of the folding of ferricytochrome *c*, such as that of Figure 6 for rubredoxin, is not shown here, but such a schematic diagram can be obtained easily by following the description of the formation of the contact regions in columns 1 and 4 of Table III (simply by connecting the new region given in column 1 of Table III with the regions, previously formed, given in column 4 of Table III).

(C) Lysozyme. The contact map of lysozyme is depicted in Figure 4, and the residues involved in the contact regions are listed in Table IV in the order of formation of the contact regions (except for those formed in step A). The folding mechanism proposed here assumes that the formation of disulfide bonds is an independent process which does not influence the folding pathway.

In step A, the ordered backbone structures listed in Table I are formed; they are the extended sequences at residues 1–4 and 43–46, the helical sequences at residues 5–14, 25–35, 80–84, 89–99, 109–113, and 120–123, and the category (i) chain-reversal conformations at residues 14–15, 35–36, 86–87, 104–108, 113–115, and 116–117 (see Table V). These ordered backbone conformations can introduce contacts between amino acids (hence, produce contact regions) in step A as follows:

(1) Combination of the helical sequence at residues 5–14 and the chain reversal at residues 14–15 forms contacts 7···2, 8···3, 9···4, 10···5, 11···3, 11···6, 12···7, 13···8, 14···9, 15···10, and 17···12 (involving the helix) and (17···12), 18···12, and 18···13 (involving the chain reversal) (see Table I); these constitute 13 out of the total of 14 contacts that form region I (see column 2 of Table IV for the total number of contacts in each region).

(2) Similarly, contact region II is formed by combination of the helical sequence at residues 25–35 and the chain reversal at residues 35–36; 13 contacts (listed in column 5 of Table I) are involved, and these constitute all of the contacts of region II.

(3) The combination of the helical sequences at residues 80–84 and 89–99 and the chain reversal at residues 86–87 forms region IV, since these ordered backbone conformations promote the formation of contacts 87···82 [involving h(80–84)], 88···83, and 90···85 [involving R(86–87)], and 93···88, 94···89, 95···90, and 96···91 [involving h(89–99)]; this amounts to 7 out of a total of 7 contacts in region IV (as seen in Tables I and IV).

(4) The combination of the helical sequence at residues 109–113 and the two chain reversals at residues 104–108 and 113–115 forms region VI. The contacts formed by each ordered backbone conformation are listed in column 5 of Table I and amount to 11 contacts out of the total of 13 found in region VI.

(5) The combination of the helical sequence at residues 120–123 and the chain reversals at residues 116–117 and 123–124 forms region VII, since the helical sequence (120–123) involves the contacts 123···118, 125···119, and 125···120, and the chain reversals (116–117 and 123–124) involve the contacts 120···115 and 125···120, respectively (see Table I). Thus, 4

contacts out of a total of 6 in region VII are formed by these ordered backbone conformations.

Thus, five contact regions (I, II, IV, VI, and VII) are formed in step A, as indicated in Table I.

Step B then proceeds as follows:

(1) The amino acid residues involved in the sequence 48–70 are brought together to form region III by formation of chain reversals at residues 55–56, 60–62, and 67–68 [by a type (I) elementary process]. The contacts formed by these three chain reversals are listed in Table V and amount to 35 out of a total of 38 contacts in region III. Although contact region III arises by formation of ordered backbone conformations (i.e., the chain reversals at residues 55–56, 60–62, and 67–68, as seen in Table V), the contacts (listed in Table V) formed by the chain reversals do not satisfy the condition of eq 1 (short-range contact) but do satisfy eq 6 and 7 (medium-range contact). This is the reason why region III is assigned to step B and not to step A.

(2) Contact regions IV (at residues 95 and 96) and VI (at residues 103–105, 107, and 108) (see column 4 of Table IV) are brought together by formation of a chain reversal at residues 99–101 [type (III) elementary process]. The contacts formed by this chain reversal (99–101) are listed in Table V; these constitute 9 of the 10 contacts found in region V.

(3) Regions I (at residues 5, 6, 8, 9, 12, 13, 17, and 18) and II (at residues 23, 25, 28, 29, 32, and 33) associate by formation of chain reversals at residues 18–19 and 20–22 by a type (III) elementary process. The contacts formed by these two chain reversals are listed in Table V; these constitute 18 contacts out of the total of 27 found in region VIII, as seen in column 4 of Table V.

(4) Contact regions II (at residues 31, 32, 35–38), III (at residues 50–53 and 57), and VIII (at residues [32] and 33) associate to form the contacts of region IX that are listed in Table IV.

(5) Contact regions III (at residues 51, 53, 58, 59, 60, 61, 63–66), IV (at residue 83), and IX (at residues [51, 53], 54, and 55) are brought together with the aid of chain reversals at residues 70–71 and 75–76 to form region X. The contacts formed by these two chain reversals are listed in column 4 of Table V and amount to 40 out of the 55 contacts found in region X.

(6) Contact regions IV (at residues 82, 83, 90, 91, 93–95), V (at residues [95] and 98), and X (at residues 75, 76, 78, [83 and 84]) associate to form region XI, as seen in Table IV, in the final stage of step B (since $\mathrm{Min}|i-j| = 7$).

Thus, regions III, V, VIII, IX, X, and XI are formed in step B (since $\mathrm{Min}|i-j|$ in these contact regions are in the range of $5 \leq |i-j| \leq 20$, as seen in column 3 of Table IV).

Then, step C proceeds as follows:

(1) Regions IV (at residues 88, 90, 91, 94, and 95), IX (at residues [53], 54, [55], and 56), X (at residues 53, 55, 58, 63, and 64), and XI (at residues [90, 91, 94, 95], 97 and [98]) associate[22] to form contact region XII by a type (IV) elementary process.

(2) Regions I (at residues 1–6 and 8) and IX (at residues 37–41) associate to form region XIII by a type (III) elementary process, as seen in Table IV.

(3) Then, region XIV is formed by association of regions IV (at residues 83, 85, and 87), IX (at residues 40–43), and XII (at residue 88).

(4) Region XV is then formed by association of regions I (at residues 11, 12, 14, 15, and 17), IV (at residues 87, 88, 93, 95 and 96), V (at residues [95, 96], 97, and 100), VII (at residues [12, 17], 20, 21), and XII (at residues [88], 92, [95], and 97).

(5) Regions II (at residues 23, 26–28, 30, 31, and 34), VI (at residues 103–106, 108–111, and 114–116), VIII (at residues [23], 24, [28]), and IX (at residues [31] and 35) associate to form region XVI.

(6) Region XVIII is formed by association of region I (at residues 1 and 3) with region XIV (at residues 84 and 86–88).

(7) Region XVII is formed by association of regions II (at residues 25–27, 29, 30, 33, and 34), VII (at residues 120 and 123), and XVI (at residues 24, 26, 27, 30, and 34).

(8) Finally, region XIX is formed by association of region I (at residues 5–7, 9, 10, and 13, located at the N terminus of the chain) with regions XVII (at residues 121, 123, and 124) and VII (at residues [123] and 125, located at the C terminus of the chain) to complete the native structure.

We have not presented the schematic diagram of the folding process described above, which can be depicted in a similar fashion to Figure 6, because it can be represented simply by following the above description and connecting the regions given in column 4 of Table IV. The chain reversals described above are classified and summarized in Table V, and will be discussed in section IV.

The folding of lysozyme (see Table IV) can be reexamined to demonstrate how it folds into two separate "wings", which then associate with each other. As seen in Figure 5 of ref 17a, one wing of lysozyme consists of residues 1–39 and 104–129 and the other wing consists of residues 43–100. The two wings are connected at residues 40–42 and 101–103. More specifically, as demonstrated below, contact regions involving residues 1–39, 104–129, and 43–100 are formed separately in the early stages of folding. The contact regions between these three parts (including those between 1–39 and 104–129 in one wing, as well as those between the two wings) are established in the later stages of folding.

One wing (involving residues 1–39 and 104–129) is formed in the early stages of folding as follows:

(1) Region I (with contacts among residues 1 to 18) and region II (with contacts among residues 23 to 38) are formed in step A; these constitute a part of this wing (residues 1–39).

(2) Region VI (with contacts among residues 103–116) and region VII (with contacts among residues 111–125) are also formed in step A; these constitute another part of this wing (residues 104–129).

(3) Region VIII (with contacts among residues 5–33) is formed in the early stages of step B).

(4) Region XIII (with contacts among residues 1 to 41) is formed in the early stages of step C and constitutes a part of this wing (residues 1–39).

Thus, up to now, one wing (involving residues 1–39 and 104–129) is formed without yet achieving contacts between these two parts of this wing (i.e., residues 1–39 and 104–129) and between the two wings.

The other wing (involving residues 43–100) is also formed in the early stages of folding as follows:

(1) Region IV (with contacts among residues 82–96) is formed in step A.

(2) Region III (with contacts among residues 48 to 70), region X (with contacts among residues 51 to 84), and region XI (with contacts among residues 75 to 98) are formed in step B.

(3) Region XII (with contacts among residues 53 to 98) and region XIV (with contacts among residues 40 to 88) are formed in the early stages of step C.

Thus, the other wing (involving residues 43–100) has been formed. The contacts among the three parts (residues 1–39, 104–129, and 43–100) are formed at the later stage of step C (except for region V and region IX which are formed in step B).

(1) The contact region V (with contacts among residues 95 to 108), near the hinge at residues 101–103, is formed in step B.

(2) Region IX (with contacts among residues 31 to 57) that

is located near the other hinge (at residues 40–42) is formed in step B.

(3) Region XV is formed by association of residues 11 to 21 (that are part of residues 1–39 of one wing) with residues 87 to 100 (that are part of residues 43–100 of the other wing) in step C.

(4) Region XVI is formed by association of residues 23–35 (that are part of residues 1–39) with residues 103–116 (that are part of residues 104–129 of the same wing) in step C.

(5) Region XVIII is formed by association of residues 1 to 3 (that are part of one wing) with residues 84 to 88 (that are part of residues 43–100 of the other wing).

(6) Region XVII is formed by association of residues 24 to 30 (that are part of residues 1–39 of one wing) with residues 120 to 124 (that are part of residues 104–129 of the same wing) in step C.

(7) Region XIX is formed by association of residues 5 to 13 at the amino end of the chain (that are part of residues 1–39) with residues 121 to 129 at the carboxyl end of the chain (that constitute another part of the same wing) in step C.

IV. Role of Chain-Reversal Conformations during Protein Folding

Up to now, we have focused attention on the formation of contact regions as folding proceeds and have found that chain-reversal conformations sometimes play an important role in forming contacts among the amino acid residues of a protein. In examining the folding of these three proteins, we found that the role of the chain-reversal conformation in the folding process may be divided into three categories:

(i) Chain-reversal (or multiple chain-reversal) regions themselves can form contact regions (sometimes, by involving neighboring amino acids of unordered segments) due to short-range interactions (step A) because of their compact form.

(ii) The chain-reversal conformation plays an important role in bringing sequences of ordered backbone structure (such as α-helical or extended segments), and also contact regions, together to form a new contact region in which medium- or long-range interactions come into play (in step B or step C). In its formation, this category of chain reversals promotes the formation of contact regions by elementary process (I)–(IV), described in section I.

(iii) The chain-reversal conformation can exist as a specific structure (formed by short-range interactions in step A) without contributing to the formation of contact regions as in category (i) (which leads to a contact of short-range order, satisfying eq 1) and category (ii) (which leads to a contact of medium-range order, satisfying eq 5–7). Category (iii) differs from category (i) in that the latter involves the formation of a contact region (by association of the chain reversal with neighboring residues), whereas the former consists only of the chain reversal (with no association with other residues and, hence, no formation of a contact region), serving to form a "kink" structure.

These roles of chain-reversal conformations (and the criteria described above for assigning the three categories) were demonstrated in discussing the folding of rubredoxin, ferricytochrome *c*, and lysozyme in section III. In Table V, we have specified explicitly the role played by each chain reversal in the folding of these three proteins and the type of elementary process [(I), (II), (III), or (IV)] involved in category (ii) chain reversals.

As summarized in Table V, the predominant role of the chain reversal appears to be of the type of category (ii), i.e., to bring about contact regions in which medium-range interactions come into play (step B), and, secondarily, of the type of category (i); chain reversals rarely bring together more than two contact regions to form new contact regions in which long-range interactions come into play (step C). Among these three proteins, chain reversals of category (iii) are found only in lysozyme (see Table V). From the Monte Carlo simulation of chain folding,[3] it appears that the major conformational changes of step C, that induce globularity, do not involve the formation of chain reversals; rather, large segments of the chain are brought together in step C by *moderate* bending over a limited region (but larger than that of a chain reversal), thereby preserving the contact regions that were formed in earlier stages of folding (although there may be small rearrangements of the contact regions).

As seen in Table V, a considerable number of chain reversals fall into category (ii). While chain reversals in categories (i) and (iii) arise from short-range interactions in step A, we cannot, at first sight, argue that chain reversals of category (ii) also arise from short-range interactions. However, we cite two facts which support the conclusion that short-range interactions also play a role in category (ii) chain reversals in proteins. First, we recently analyzed the native structures of 26 proteins and found the relative tendencies of each of the 20 naturally occurring amino acids to form chain-reversal conformations in proteins.[11] As described in the first paragraph of section VI of ref 11, Ala, Gln, Glu, Met, and Pro have a relatively strong preference to adopt an R conformation[20] at residue $(i - 1)$, whereas Asn, Asp, Cys, Gly, His, Phe, Trp and Tyr have a strong preference to adopt an S conformation[20] at residue i of an RS chain-reversal conformation at two successive residues, $(i - 1)$ and i. The other amino acid residues (Arg, Ile, Leu, Lys, Ser, Thr, and Val) do not have such special preferences. Furthermore, we found[11] that Asn, Asp, Gly, and Pro have a relatively stronger tendency to form a chain-reversal conformation, compared to helical, extended, or other conformations (see Table III of ref 11). Second, we found[11] a good correlation between the observed chain reversals in proteins and positions predicted from the probabilities (based on a short-range interaction model) of finding chain reversals. These two facts imply that certain amino acids have a specific tendency to form chain-reversal conformations and that this tendency is accounted for within the short-range interaction model. Thus, it appears that, even in step B, chain-reversal conformations form *predominantly* because of short-range interactions, even though they play an additional role in forming contact regions in step B. In other words, the positions of most chain reversals can be predicted *primarily* from the amino acid sequence of the protein.[11] However, they are stabilized by medium-range interactions as contact regions form in step B. Also, the medium-range interactions, i.e., the formation of contact regions, may serve to some extent to promote the formation of chain reversals. We may conclude that, aside from whatever role they play in step B [category (ii)], the possible locations of chain-reversal conformations (formed by short-range interactions) can be predicted with a high degree of reliability by a short-range interaction model, as in step A. The introduction of medium- and long-range interactions can lead to rearrangements of the chain-reversal (and other ordered structural) conformations formed in step A.

V. Discussion

Besides being useful for the investigation of protein folding (as in section III), the contact maps of Figures 2–4 provide important information about the structure of native proteins, specifically about the role of the solvent (water in most cases). It is well known that the solvent plays an important role in stabilizing the structure of native proteins. A typical effect of solvent is the hydrophobic interaction or, in the present context, the interactions involved in the formation of contact regions involving nonpolar amino acids. In the contact maps

Table IV. Contact Regions Formed during the Folding of Lysozyme

3

Contact region[a]	Residues involved in contact region[b]	Minimum value of $\|i-j\|$[c]	Contributions of contact regions formed previously[d]
I	7 7 8 9 10 11 11 12 13 14 15 17 18 18 1 7 3 4 5 3 6 7 8 9 10 12 12 13	5	
II	28 30 31 32 33 34 35 36 37 38 38 38 38 23 25 26 27 28 29 30 31 32 29 30 32 33	5	
III	57 58 58 58 59 59 59 59 59 60 60 60 60 60 61 61 63 64 64 52 51 52 53 48 50 51 52 53 49 50 51 52 53 48 50 58 51 53 64 64 65 66 66 66 66 68 68 69 69 69 69 69 69 69 69 70 70 58 59 60 50 51 53 60 49 53 48 49 50 51 53 59 60 61 48 49	5	
IV	87 88 90 93 94 95 96 82 83 85 88 89 90 91	5	
V	100 103 104 104 104 105 107 108 108 108 108 95 98 98 99 99 98 95 96 98 99	5	IV (95, 96) VI (103, 104, 105, 107, 108)
VI	111 111 111 111 111 112 112 112 112 113 114 115 116 116 103 104 105 106 103 105 106 107 108 109 110 106 111	5	
VII	120 120 123 123 125 125 111 115 115 118 119 120	5	
VIII	23 23 24 24 24 25 25 25 25 25 25 28 28 28 28 28 28 17 18 17 18 19 8 9 12 13 17 18 19 8 12 17 18 19 20 21 29 29 29 29 29 32 32 33 5 6 8 9 12 8 12 8	5	I (5, 6, 8, 9, 12, 13, 17, 18) II (23, 25, 28, 29, 32, 33)
IX	50 50 51 51 51 51 52 52 52 52 52 53 53 53 53 53 54 54 54 44 45 43 44 45 46 42 43 44 45 46 40 41 42 43 44 36 39 40 54 54 54 55 55 55 55 55 55 55 55 55 55 56 56 56 56 57 41 42 43 31 32 33 35 36 37 38 39 40 41 42 31 32 35 36 35 57 57 57 57 57 36 39 42 43 44	5	II (31, 32, 35, 36, 37, 38) III (50, 51, 52, 53, 57) VIII ([32], 33)
X	71 72 72 72 72 72 72 73 73 73 73 74 74 74 74 74 75 75 75 61 60 61 62 64 65 66 60 61 62 63 64 62 63 64 65 61 62 63 75 76 76 76 76 77 77 78 78 78 78 79 79 79 80 80 80 80 80 64 58 62 63 64 63 64 63 64 65 66 64 65 66 51 53 58 59 60 80 80 80 80 81 81 81 81 83 83 84 83 83 83 84 84 84 63 64 65 66 53 64 65 66 53 54 55 58 63 64 53 54 55	6	III (51, 53, 58, 59, 60, 61, 63, 64, 65, 66) IV (83) IX ([51, 53], 54, 55)

Table IV (continued)

4

Contact region[a]	Residues involved in contact region[b]	Minimum value of $\|i-j\|$[c]	Contribution of contact regions formed previously[d]	
XI	90 90 90 90 91 91 91 93 94 94 94 95 97 97 98 76 78 82 83 82 83 84 76 76 78 83 76 75 76 76	7	IV (82, 83, 90, 91, 93, 94, 95) V ([95], 98) X (75, 76, 78, [83, 84])	
XII	88 90 91 91 91 91 91 92 92 92 94 94 94 95 95 97 98 98 55 58 53 54 55 56 58 58 55 56 58 63 64 56 58 63 58 63	30	IV (88, 90, 91, 94, 95) IX ([53], 54, [55], 56) X (53, 55, 58, 63, 64) XI ([90, 91, 94, 95], 97, [98])	[III (53, 58, 63, 64)] [V (95, 98)]
XIII	37 37 38 38 38 38 38 38 38 38 39 39 39 40 40 40 41 2 3 1 2 3 4 5 6 8 1 2 3 1 2 3 1	30	I (1, 2, 3, 4, 5, 6, 8) IX (37, 38, 39, 40, 41)	[VII (5, 6, 8)] [II (37, 38)]
XIV	83 83 84 84 84 84 85 86 87 88 40 41 40 41 42 43 40 40 40 60	41	IV (83, 85, 87) IX (40, 41, 42, 43) XII (88)	[X (83, 84)] [XI (83, 84)] [XIII (40, 41)]
XV	87 88 88 88 88 89 89 92 92 92 92 93 95 95 96 96 96 96 96 15 11 17 15 17 14 15 12 15 16 17 15 17 20 15 16 17 20 21 97 99 99 99 100 100 20 17 20 21 20 21	71	I (11, 12, 14, 15, 17) IV (87, 88, 93, 95, 96) V ([95, 96], 97, 100) VII ([12, 17], 20, 21) XII ([88], 92, [95], 97)	[XI (93, 95, 97)] [XIV (87, 88)]
XVI	103 104 105 105 105 105 105 105 106 108 108 108 108 109 110 110 110 23 24 23 24 27 28 31 23 28 31 35 35 31 34 35 111 111 111 111 111 111 111 114 114 115 115 115 115 115 116 23 27 28 30 31 35 30 34 26 27 30 31 34 27	73	II (23, 26, 27, 28, 30, 31, 34) VI (103, 104, 105, 106, 108, 109, 110, 111, 114, 115, 116) VIII ([23], 24, [28]) IX ([31], 35)	[V (103, 104, 105, 108)] [VII (111, 115)]
XVIII	84 86 86 87 88 1 1 3 3 3	83	I (1, 3) XIV (84, 86, 87, 88)	[IV (87, 88)] [X (84)], [XI (84)] [XII (88)], [XIII (1, 3)]
XVII	120 120 120 210 121 121 123 123 123 123 123 123 124 124 24 26 27 30 24 26 25 26 29 30 33 34 24 25 124 124 26 29	89	II (25, 26, 27, 29, 30, 33, 34) VII (120, 123) XVI (24, 26, 27, 30, 34)	[VIII (24, 25, 29, 33)]
XIX	121 122 123 123 123 123 126 124 124 125 125 126 126 127 127 5 5 5 6 9 5 6 9 5 6 5 6 5 6 127 127 128 128 129 129 129 129 7 9 6 10 6 9 10 13	114	I (5, 6, 7, 9, 10, 13) XVII (123, 123, 124) VII ([123], 125)	[VIII (5, 6, 9, 13)]

a, b, c, d - See corresponding footnotes of Table II.

Table III. Contact Regions Formed during the Folding of Ferricytochrome c

Contact region[a]	Residues involved in contact region[b]	Minimum value of $\|i-j\|$[c]	Contribution of contact regions formed previously[d]	
I	6 7 8 9 10 11 12 13 14 15 1 2 3 4 5 6 7 8 9 10	5		
II	18 20 20 20 10 10 14 15	5		
III	24 25 25 26 27 27 27 27 27 28 28 28 29 29 29 30 30 19 18 19 19 15 16 17 18 19 16 17 18 17 18 19 17 18 30 30 30 31 31 31 31 31 31 31 31 31 32 32 32 19 18 25 18 19 20 21 22 23 24 25 26 28 19 20	5		
IV	43 44 47 48 48 48 48 48 38 38 42 38 40 41 42 43	5		
V	52 53 54 55 46 48 49 50	5		
VI	63 63 64 67 68 57 58 59 59 63	5		
VII	71 73 74 74 75 75 76 77 78 78 78 66 66 66 67 67 70 71 72 67 71 72	5		
VIII	90 91 91 92 93 94 94 95 96 97 98 99 100 101 102 103 85 85 86 87 88 85 89 90 91 92 93 94 95 96 97 98	5		
IX	38 38 38 41 42 42 42 42 42 44 45 46 46 46 46 46 46 31 32 33 30 26 30 31 32 33 26 26 26 27 28 29 30 31 47 48 48 48 48 26 29 30 31 32	5	III (26, 27, 28, 29, 30, 31, 32) IV (38, 41, 42, 44, 47, 48) V (46, [48])	
XIII	82 82 82 82 82 83 83 84 84 84 84 85 86 69 70 71 72 73 70 71 69 70 71 69 69	9	VII (70, 71, 72, 73) VIII (85, 86)	
X	52 52 53 53 54 55 55 56 56 57 57 57 58 58 59 59 59 40 41 40 41 40 39 40 39 40 38 39 40 38 39 32 34 35 59 59 59 59 59 59 60 60 60 60 61 36 37 38 39 40 41 35 36 37 38 36	11	IV (38, 40, 41) V (52, 53, 54, 55) VI (57, 58, 59) IX ([38], 32, [41])	[III (32)]
XI	73 74 74 74 74 74 74 75 75 75 76 55 55 57 58 59 62 63 55 59 55	11	VI (57, 58, 59, 63) VII (73, 74, 75, 76) X (55, [57, 58, 59])	[XII (73)] [V (55)]
XII	73 75 76 76 76 77 77 77 78 78 78 79 79 79 79 79 79 79 51 52 49 51 52 49 51 52 49 51 52 46 47 48 49 50 51 52	23	V (46, 48, 49, 50, 52) VII (75, 76, 77, 78) IX ([46], 47, [48])	[IV (47, 48)] [X (52)] [XI (75, 76)]

Table III (continued)

2

Contact region[a]	Residues involved in contact region[b]	Minimum value of $\|i-j\|$[c]	Contribution of contact regions formed previously[d]	
XIV	94 94 95 95 95 95 95 98 99 64 68 61 63 64 65 68 64 61	26	VI (63, 64, 68) VIII (94, 95, 98, 99) X (61)	[XI (63)]
XV	46 46 48 17 18 18	28	III (17, 18) V (46, 48)	[II (18)] [IV (48)] [IX (49, 48)] [XII (46, 48)]
XVI	94 95 96 97 98 98 98 98 99 100 101 102 102 102 102 36 36 36 36 32 33 35 36 36 36 33 32 33 34 35 102 103 104 104 36 33 33 34	58	VIII (94, 95, 96, 97, 98, 99, 100, 101, 102, 103) IX ([32], 33) X (32, 34, 35, 36)	[III (32)] [XIV (94, 95, 98, 99)]
XVII	80 80 81 82 83 84 85 85 85 17 18 13 13 13 13 9 12 13	62	I (9, 12, 13) III (17, 18) XIII (82, 83, 84, 85)	[VIII (85)] [XIII (17, 18)]
XVIII	97 98 101 101 102 102 103 104 20 20 20 22 20 22 22 22	77	III (20, 22) XVI (97, 98, 101, 102, 103, 104)	[II (20)] [VIII (97, 98, 101, 102, 103)] [XIV (98)]
XIX	89 89 90 90 90 91 92 93 93 93 93 93 94 94 94 94 95 5 9 5 9 12 9 2 1 2 5 6 9 5 6 9 10 2 96 96 96 97 97 97 97 97 97 97 97 97 98 100 100 101 2 3 6 2 3 4 5 6 7 8 9 10 11 10 2 3 3	78	I (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) VIII (89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 100, 101)	[XIV (94, 95, 98)] [XVI (94, 95, 96, 97, 98, 100, 101)] [XVII (9, 12)] [XVIII (97, 98, 101)]

a, b, c, d - see corresponding footnotes a-d of Table II.

Table V. Role of Chain Reversals and Contacts Formed by Chain Reversals

Protein[a]	Role of chain reversal[b] (i)[c]	(ii)[d]	Contacts formed by category (ii) chain reversal in step B[e]	Contact region[f]	No. of contacts[g] formed by category (ii) chain reversal / Total No. of contacts in the region	(iii)[b]
Rubredoxin		35-36 (II)	46···30, 46···31, 46···32, 46···33, 46···34, 46···35, 49···30			
		40-42 (II)	43···38, 44···37, 44···38, 44···39, 45···37, 45···38, 46···39, (46···30), (46···31), (46···32), (46···33), (46···34), (46···35), 46···36, 46···37, 46···38, (49···30), 49···37, 49···38, 49···39, 49···40	III	23/23	
		46-48 (II)	(49···30), (49···37), (49···38), (49···39), (49···40), 49···42, 49···44			
		7-9 (II)	10···5, 11···4, 11···5, 11···6, 12···3, 12···4, 12···5, 12···6, 13···2, 13···3, 13···4, 13···5, 13···6, 14···2, 14···3, 14···6, 15···1, 15···2, 15···3, 15···4, 16···1, 16···3	IV	22/22	
		20-22 (IV)	24···13, 24···17, 24···18, 24···19, 25···13, 25···15, 25···17, 25···18, 25···19, 25···20, 26···13, 26···15, 26···16, 26···17, 26···18, 26···19, 26···20, 27···13, 27···15, 27···16, 27···17, 27···18, 28···13, 28···14, 28···15, 28···18, 29···13, 29···15, 30···13, 30···15	V	30/30	
		26-27 (IV)	(27···13), (27···15), (27···16), (27···17), (27···18), (28···13), (28···14), (28···15), (28···18), (29···13), (29···15), (30···13), (30···15)			
Ferricytochrome c	9-13 14-17 43-45 50-51 64-70 74-75 100-102	22-23 (I)	24···19, 25···18, 25···19, 26···19, 27···15, 27···16, 27···17, 27···18, 27···19, 28···16, 28···17, 28···18, 29···17, 29···18, 29···19, 30···17, 30···18, 30···19, 31···18, 31···19, 31···20, 31···21, 31···22, 32···18, 32···19, 32···20	III	31/32	
		27-29 (I)	(29···17), (29···18), (29···19), (30···17), (30···18), (30···19), 30···25, (31···18), (31···19), (31···20), (31···21), (31···22), 31···23, 31···24, 31···25, 31···26 (32···18), (32···19), (32···20)			
		62-63 (I)	63···57, 63···58, 64···59, 67···59	VI	4/5	
		33-34 (IV)	38···31, 38···32, 38···33, 41···30, 42···28, 30···42, 42···31, 42···32, 42···33, 44···26, 45···26, 46···27, 46···28, 46···29, 46···30, 46···31, 48···29, 48···30, 48···31, 48···32	IX	26/22	
		35-37 (IV)	(38···31), (38···32), (38···33), (41···30), (42···26), (42···30), (42···31), (42···32), (42···33), (44···26), (45···26), (46···27), (46···28), (46···29), (46···30), (46···31), (48···29), (48···30), (48···31), (48···32)			

Table V (continued)

Protein[a]	Role of chain reversal[b] (i)[c]	(ii)[d]	Contacts formed by category (ii) chain reversal in step B[e]	Contact region[f]	No. of contacts[g] formed by category (ii) chain reversal / Total No. of contacts in the region	(iii)[b]
		76-77 (III)	82···69, 82···70, 82···71, 82···72, 82···73, 83···70, 71···83, 84···69, 84···70, 84···71, 85···69, 86···69	XIII	12/12	
		79-80 (III)	(82···69), (82···70), (82···71), (82···72), (82···73), (83···70), (83···71), (84···69), (84···70), (84···71), (85···69), (86···69)			
Lysozyme	14-15 35-36 86-87 104-108 113-115 116-117 123-124	55-56 (I)	57···52, 58···51, 58···52, 58···53, 59···48, 59···50, 59···51, 59···52, 59···53, 60···49, 60···50, 60···51, 60···52, 60···53, 61···48, 61···50, 64···51, 64···53, 66···50, 66···51, 66···53, 68···49, 68···51, 68···53, 69···50, 69···51, 69···53			37-38 40-41 125-126
		60-62 (I)	63···38, (64···51), (64···53), 64···58, 64···59, 65···60, (66···50), (68···51), (66···53), 66···60, (68···49), (68···51), (68···53), (69···50), (69···51), (69···53), 69···59, 69···60	III	35/38	
		67-68 (I)	(68···49), (68···51), (68···53), (69···50), (69···51), (69···53), (69···59), (69···60), 69···61			
		99-101 (III)	103···98, 104···98, 104···99, 105···99, 107···98, 108···95, 108···96, 108···98, 108···99	V	9/10	
		18-19 (III)	23···17, 23···18, 24···17, 24···18, 25···8, 25···9, 25···12, 25···13, 25···17, 25···18, 28···12, 28···17, 28···18, 29···12	VIII	14/27	
		20-22 (III)	(23···17), (23···18), (24···17), (24···18), 24···19, (25···8), (25···9), (25···12), (25···13), (25···17), (25···18), 25···19			
		70-71 (IV)	72···60, 72···61, 72···62, 72···64, 72···65, 72···66, 73···60, 73···61, 73···62, 73···63, 73···64, 74···62, 74···63, 74···64, 74···65, 73···61, 75···62, 75···63, 75···64, 76···58, 76···62, 76···63, 76···64, 77···63, 77···64, 78···63, 78···64, 78···65, 78···66, 79···64, 79···65, 79···68, 80···63, 80···64, 80···65, 80···68, 81···64, 81···65, 81···66, 83···64	X	40/55	
		75-76 (IV)	(76···58), (76···62), (76···63), (76···64), (77···63), (77···64), (78···63), (78···64), (78···65), (78···66), (79···64), (79···65), (79···66), (80···63), (80···64), (80···65), (80···66), (81···64), (81···65), (81···66), (83···64)			

Footnotes to Table V

(a) See footnotes f, g and ℓ of Table I for the X-ray coordinates of these proteins.

(b) The locations of chain-reversal conformations were determined in section III of ref. 11, and these regions can be found in column 6 of Table I of ref. 11. See section IV of the text for the meaning of the chain-reversal category symbols (i)-(iii).

(c) The contacts formed by chain reversals of category (i) (in step A) are listed in column 5 of Table I.

(d) The capital Roman numerals in parentheses denote the types of elementary processes for formation of contact regions (see section I of the text).

(e) The total number of contacts in each region (designated in column 5) is given in column 2 of Tables II-IV. The contacts in parentheses in this column are ones that already appeared earlier in this column, within the same contact region. Such contacts are formed by two chain reversals; e.g., contact 46···31 of region III is formed by both chain reversals 35-36 and 40-42 [and the parenthesis is placed arbitrarily around the 46···31 contact formed by the 40-42 chain reversal, to avoid duplication in counting the number of contacts formed by the category (ii) chain reversals (given in the numerator in column 6)].

(f) The contact regions are shown in the triangular maps of Figs. 2-4.

(g) The total number of contacts in the region given in column 5 is given in column 2 of Tables II-IV. The number of contacts formed by a category (ii) chain reversal is given in column 4. If more than 50% of the contacts formed in this region is due to the chain reversal (and satisfies eqs. 6 and 7), then the chain reversal is regarded as contributing to the formation of the contact region (see section III of the text).

of Figures 2–4 (and in Figure 1A of ref 3 for bovine pancreatic trypsin inhibitor), the contacts between nonpolar and nonpolar amino acids are indicated by solid squares, while those between any other pair of nonpolar, polar, or neutral amino acids are indicated by hatched squares.[23] Thus, these contact maps reveal the composition of the contact regions, which conveys structural information about the protein. Contrary to some current views of protein folding, as seen in Figures 2–4, most of the contact regions are *not* stabilized only by contacts between nonpolar and nonpolar amino acids. Only the contact region VIII of rubredoxin (see Figure 2) is seen to consist of all nonpolar contacts. Most contact regions are stabilized by various types of contacts other than hydrophobic interactions.

We recently analyzed the three-dimensional structures of numerous native proteins to obtain the interaction parameters between amino acids.[24] In that analysis,[24] we found many cases in which two ionized side chains of the same sign of charge were in the same contact region. Also, we found cases in which polar and nonpolar residues were in the same contact region in the interior of a native protein; the reason for this appears to be that, because of the chain connectivity, it sometimes is impossible for polar residues next to nonpolar ones to avoid being in a contact region formed by interactions among nonpolar residues. Thus, chain connectivity, as well as the arrangement of the amino acids in protein sequences, is important in determining the contact regions (three-dimensional structure), as it is in determining the ordered backbone structures, as discussed at the end of section VB of our previous paper.[10] Thus, simple generalizations such as "the interior of a native protein is nonpolar and its exterior is polar" do not describe the structures of native proteins.

Recently, Kuntz[25] (like Phillips,[26] Nishikawa et al.,[27] and Rossman and Liljas[28]) attempted to illustrate the characteristic structures of native proteins, using distance maps, where the distances between α-carbons are plotted. Such distance maps are convenient for representing geometrical information about the backbone structures of proteins. However, contact maps provide a simpler visualization of the structural features of a native protein. In addition, contact maps provide information about the types of interactions between residues in, and the (polar or nonpolar) composition of, the contact regions.

In summary, in this paper we have described possible pathways for the folding of three proteins, based on the three-step mechanism for protein folding proposed in our

previous paper.[3] In other papers,[8-13] we developed statistical mechanical treatments of protein conformation, using a short-range interaction model, to carry out step A. We then simulated steps B and C of protein folding by a Monte Carlo procedure in which medium- and long-range interactions are introduced.[3,24] We are currently[29] trying to simulate the complete protein folding process (from step A to step C), by incorporating the statistical mechanical treatment of protein conformation[8-13] (based on short-range interactions) into our protein folding model[3] (in which medium- and long-range interactions are taken into account), in order to detect the intermediates and final structure in steps A, B, and C.

Acknowledgment. We are indebted to Dr. George Némethy for helpful discussions of this manuscript.

Miniprint Material Available: Full size copies of Tables III–V showing contact regions formed during protein folding (14 pages). Ordering information is given on any current masthead page.

References and Notes

(1) This work was supported by research grants from the National Institute of General Medical Sciences, U.S. Public Health Service (GM-14312), and from the National Science Foundation (BMS75-08691).
(2) To whom correspondence should be addressed.
(3) S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **72,** 3802 (1975).
(4) In this paper, as well as in ref 3 and 5, the terms short-, medium-, and long-range interactions are defined as follows: if i and j refer to the positions of two residues in the amino acid sequence, then these terms pertain to interactions for which $|i - j| \leq 4$, 5 to 20, and >20, respectively. This differs from the nomenclature of ref 6, where medium range pertained to $1 \leq |i - j| \leq 4$.
(5) S. Tanaka and H. A. Scheraga, *Macromolecules,* **9,** 945 (1976).
(6) P. K. Ponnuswamy, P. K. Warme, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **70,** 830 (1973).
(7) The dihedral angles of, say, an α-helix are generally not the same in every residue of an α-helical segment of a protein. Therefore, we describe such structures as ordered rather than regular. We use the term "ordered backbone structure" to designate α-helical, extended, and chain-reversal conformations, and the term "unordered backbone structure" to designate all other conformations.
(8) S. Tanaka and H. A. Scheraga, *Macromolecules,* **9,** 142 (1976).
(9) S. Tanaka and H. A. Scheraga, *Macromolecules,* **9,** 159 (1976).
(10) S. Tanaka and H. A. Scheraga, *Macromolecules,* **9,** 168 (1976).
(11) S. Tanaka and H. A. Scheraga, *Macromolecules,* **9,** 812 (1976).
(12) S. Tanaka and H. A. Scheraga, *Macromolecules,* **10,** 9, (1977).
(13) S. Tanaka and H. A. Scheraga, *Macromolecules,* in press.
(14) As will be shown in Table I, extended structures rarely form contact regions in step A. However, the extended structures are stabilized additionally in steps B or C by hydrogen bonds which, in this case, constitute contacts of medium- or long-range order.
(15) K. D. Watenpaugh, L. C. Sieker, J. R. Herriott, and L. H. Jensen, *Acta Crystallogr., Sect. B,* **29,** 943 (1973).
(16) R. E. Dickerson, T. Takano, D. Eisenberg, O. B. Kallai, L. Samson, A. Cooper, and E. Margoliash, *J. Biol. Chem.*, **246,** 1511 (1971).
(17) (a) C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma, *Nature (London),* **206,** 757 (1965); (b) D. C. Phillips, *Proc. Natl. Acad. Sci. U.S.A.*, **57,** 484 (1967).
(18) Undoubtedly, prosthetic groups (such as the heme of ferricytochrome *c*) play some role in folding and in stabilizing the native structure of a protein. In this paper, we regard the introduction of a prosthetic group as equivalent to the alteration of the physical conditions (e.g., like the alteration of solvent composition) around the amino acids *near* the prosthetic group. Since we are concentrating here only on the formation of contact regions between various amino acid residues and *not* between the prosthetic group and its nearby residues, we do not consider here the role played by the prosthetic group in the folding process.
(19) The contacts are grouped into contact *regions* on a triangle contact map (such as Figures 2–4) primarily by visual observation, wherein contiguous contacts are assigned to the same region; e.g., see regions IV and V of Figure 2. However, sometimes the contacts are not contiguous, but nevertheless are close together; these are grouped together in the same contact region for the sake of simplicity and convenience in describing the folding mechanism, as is done in section III.
(20) See ref 11 for definition of R and S conformations.
(21) See section IIIB of ref 11 for the definition of a multiple chain-reversal region.
(22) We designate the duplicate appearance of residues in two or more *regions,* such as regions I (19) and II (24), by placing brackets around them, i.e., [I (19)] and [II (24)], as seen in the last column of Table II (see footnote *f* of Table II for the criterion for selecting the region around which to place the brackets). While this is done for the sake of simplicity, residues 19 and 24 may be considered to be involved in regions I and II, respectively, as well as in region V. Hereafter, in describing contact regions in the text, we will not mention the duplicate (bracketed) ones, for the sake of simplicity; however, all such bracketed regions are given in the last columns of Tables II–IV.
(23) Parenthetically, it should be noted that one can differentiate between pairs of different types of amino acids (polar, nonpolar, or neutral) by using different symbols in the contact map. However, even though we do not discriminate between pairs other than between nonpolar and nonpolar amino acids in Figures 2–4, it is possible to discriminate between the types of contacts (represented by hatched squares in these contact maps) simply by looking at both the vertical and horizontal runs attached to the outside of the triangles. In addition, one can include much more information about the contacts between amino acids. For example, one can distinguish between side chain–side chain, side chain–backbone, and backbone–backbone contacts, as shown in Figure 1B of ref 3.
(24) See ref 5. The interaction parameters (i.e., the free energy of formation of a contact between a pair of amino acids) were used in ref 3.
(25) I. D. Kuntz, *J. Am. Chem. Soc.*, **97,** 4362 (1975).
(26) D. C. Phillips, "British Biochemistry, Past and Present", T. W. Goodwin, Ed., Academic Press, London, 1970, p 11.
(27) K. Nishikawa, T. Ooi, Y. Isogai, and N. Saito, *J. Phys. Soc. Jpn.*, **32,** 1331 (1972).
(28) M. G. Rossman and A. Liljas, *J. Mol. Biol.*, **85,** 177 (1974).
(29) S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, in press.