Proc. Nat. Acad. Sci. USA
Vol. 72, No. 10, pp. 3802–3806, October 1975
Chemistry

Model of protein folding: Inclusion of short-, medium-, and long-range interactions

(mechanism of folding/contact free energies/range of interactions/Monte Carlo)

SEIJI TANAKA* AND HAROLD A. SCHERAGA†

Department of Chemistry, Cornell University, Ithaca, New York 14853

Contributed by Harold A. Scheraga, July 30, 1975

ABSTRACT **A hypothesis for protein folding is proposed, in which the native structure is formed by a three-step mechanism: (A) formation of ordered backbone structures by short-range interactions, (B) formation of small contact regions by medium-range interactions, and (C) association of the small contact regions into the native structure by long-range interactions. Empirical interaction parameters (free energy of formation of a contact) between amino-acid residues were evaluated from the frequency of contacts in the x-ray structures of native proteins. On the basis of this mechanism, a Monte Carlo simulation of protein folding (with an accompanying decrease in the total contact free energy) was carried out for bovine pancreatic trypsin inhibitor. The predicted three-dimensional structure is in fairly good agreement with the experimental one.**

In order to predict the three-dimensional structure of a protein, it is necessary to circumvent the multiple-minimum problem and overcome the difficulties of treating the long-range interactions (1). For this purpose, a starting conformation was obtained from empirical prediction algorithms, and then its total conformational energy was minimized (2). As an alternative approach, the conformations of the unfolded protein can be obtained by statistical mechanical procedures involving only short-range interactions,‡ and altered (to a minimum-free-energy structure) by a Monte Carlo procedure involving short-, medium-, and long-range interactions. This paper describes a model for protein folding, and the results obtained for bovine pancreatic trypsin inhibitor (BPTI) by a Monte Carlo simulation of the model. A full description of the procedure and results will be published elsewhere.§

I. Hypothetical mechanism of protein folding

The folding of a polypeptide chain to the native structure of a protein in a given medium is assumed to occur in three steps (which *may* proceed simultaneously). (A) Because of short-range interactions, ordered¶ backbone structures, such as α-helical, extended, and chain-reversal conformations, are formed in a system at equilibrium under given physical conditions. (B) When these physical conditions are changed, so as to introduce medium-range interactions, the equilibrium is shifted, and small contact regions (defined in section II) are nucleated among the residues both in the ordered and in the unordered structures. In this step, the ordered backbone structures formed in step A may be rearranged to some extent to form more stable structures in these contact regions. (C) Finally, the small contact regions formed in step B associate, in response to long-range interactions, possibly involving further small rearrangements of the structures formed in steps A and B.

A statistical mechanical treatment in terms of a one-dimensional Ising model, based on short-range interactions, has been developed recently‡ to obtain the conformation of step A. Therefore, in this paper, we concentrate on the demonstration of the role of medium- and long-range interactions in steps B and C.

II. Empirical interaction parameters between amino acids of proteins from x-ray data

Contact regions (formed in steps B and C) arise from medium- and long-range interactions between residues that approach each other. The solvent (water) plays an important role in stabilizing such conformations. We describe specific local interactions as a contact between two groups, A and B, where the group may be the whole residue, or the side chain and the backbone, or each atom of the residue. A specific local interaction is said to exist between groups A and B in water if these groups are in contact or, more quantitatively, if the distance between them, r_{AB}, satisfies the relation

$$r_A^{(w)} + r_B^{(w)} \leq r_{AB} < r_A^{(w)} + r_B^{(w)} + 2r_{H_2O}^{(w)} \quad [1]$$

where $r_A^{(w)}$, $r_B^{(w)}$ and $r_{H_2O}^{(w)}$ are the van der Waals radii of groups A, B, and a water molecule, respectively [$r_{H_2O}^{(w)}$ is taken as 1.40 Å]. The local interaction between A and B is neglected when

$$r_{AB} \geq r_A^{(w)} + r_B^{(w)} + 2r_{H_2O}^{(w)}. \quad [2]$$

Infinite repulsion arises (excluded volume effect), when

$$r_{AB} < r_A^{(w)} + r_B^{-(w)}. \quad [3]$$

The three-dimensional structure of a protein can be represented symbolically by the presence or absence of contacts between the ith and jth residues [$1 \leq (i,j) \leq N$ and $i < j$, where N is the chainlength], when account is taken of the chain connectivity. Fig. 1A shows such a contact map for BPTI, based on its x-ray coordinates;‖ 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.

The same analysis, as in Fig. 1A, was made for 25 proteins whose x-ray structures are known. From these data, counts

Abbreviation: BPTI, bovine pancreatic trypsin inhibitor.

* From Kyoto University, 1972–1975.

† To whom requests for reprints should be addressed.

‡ S. Tanaka and H. A. Scheraga, manuscript submitted.

§ S. Tanaka and H. A. Scheraga, manuscripts to be submitted.

¶ The dihedral angles of, say, an α-helix are generally not the same in every residue of an α-helical segment of a protein, nor are they assumed to be in the Monte Carlo treatment used here. Therefore, we describe such structures as ordered rather than regular.

‖ Those of R. Huber, personal communication, already used in ref. 2.

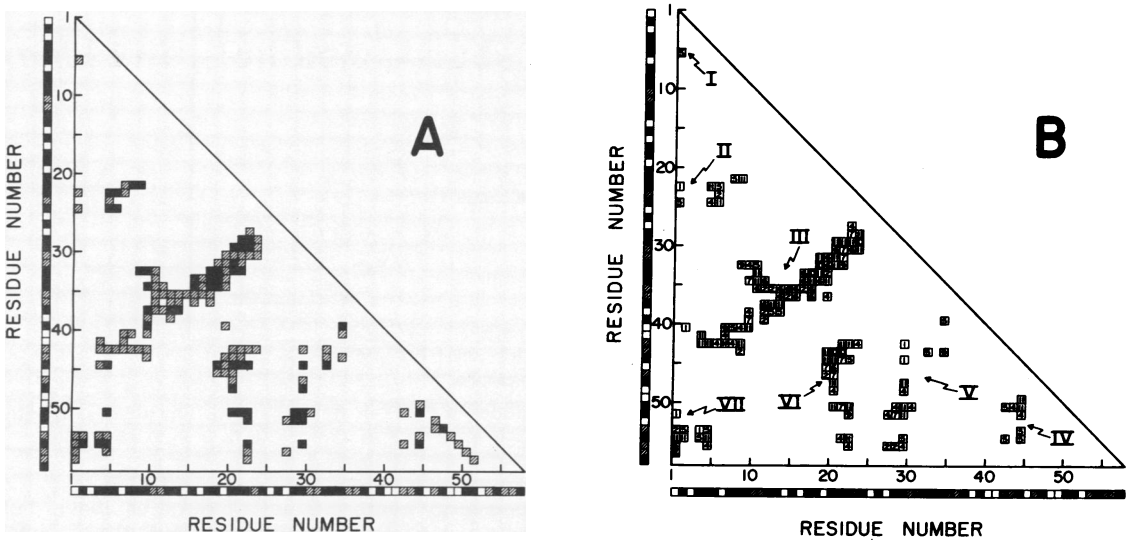

FIG. 1. Contact maps of BPTI for (A) consideration of all atoms, and (B) with side chains and backbone treated as spheres (see *text*). The solid, hatched, and open squares in the vertical and horizontal runs outside the triangle designate amino acids with highly nonpolar side chains (Cys, Ile, Leu, 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 (in A) designate contact between two highly nonpolar side chains, and between any other two side chains, respectively. The numerals in the squares (in B) show the sum for the various types of contact, where a side chain-side chain contact is designated by 1, a backbone–backbone contact by 2, and a side chain-backbone or backbone–side chain contact by 4 (a *complete* contact between two residues is indicated by 11, which is the sum of 1 + 2 + 4 + 4); every sum indicates a unique type of contact. Contacts from i to $i + 1$, $i + 2$, $i + 3$, or $i + 4$ were omitted from these diagrams in order to focus attention on the medium- and long-range interactions, and to obtain the long-range interaction parameters.

were made of the number of times an amino acid, A_k, of the type k had no contacts, and of the number of times, A_{kl}, that two residues A_k and A_l were in contact. We then defined an *empirical* equilibrium constant, K_{kl}, and its associated free energy, ΔG_{kl}, for formation of the complex A_{kl}, for contact between the *side chains* of all amino-acid residues except glycine. For glycine, we took the full residue, and evaluated its contact free energies with the other 19 side chains and glycine; glycine also served as a prototype of the backbone for non-glycine residues. A more detailed description, including the numerical values of ΔG_{kl}, will be reported elsewhere.[§] The values of ΔG_{kl} are used to compute the total contact free energy of any conformation of a protein that is generated by the Monte Carlo procedure (and does not have hard-sphere overlaps).

III. Prediction of three-dimensional structure by a Monte Carlo procedure

A. Simplified Model of Protein Chain. Backbone conformations were always generated from standard (3) bond lengths, bond angles, and planar *trans* peptide groups. However, hard-sphere overlaps were checked and empirical contact free energies were evaluated by representing the backbone NH-CH-CO group and the 20 side-chain R groups by spheres of given (van der Waals) radii. The effectiveness of such a spherical representation may be seen by comparing the contact maps of Fig. 1A and 1B.

B. Simulation of Hypothetical Mechanism of Protein Folding by a Monte Carlo Procedure. Since our interest in this paper is primarily the demonstration of the effectiveness of the Monte Carlo procedure for introducing medium- and long-range interactions, we will consider only steps B and C of the folding mechanism, and obtain an initial ordered conformation from the x-ray structure of BPTI, instead of from step A.[‡] It should be emphasized that, in contrast to our earlier procedure (2), the result of step A is not given in terms of values of ϕ and ψ but in terms of *symbols* designating the conformational states [helical (h), extended (ϵ), and other (c)] (see Fig. 2). Since conformations are selected randomly from these regions (see below), and a residue never moves out of the region assigned in step A, *regular* helical or extended sequences will occur rarely. Further, as demonstrated in ref. 2, while the starting conformation has correct short-range order, it lacks the proper medium- and long-range interactions, and hence does not resemble the native protein.

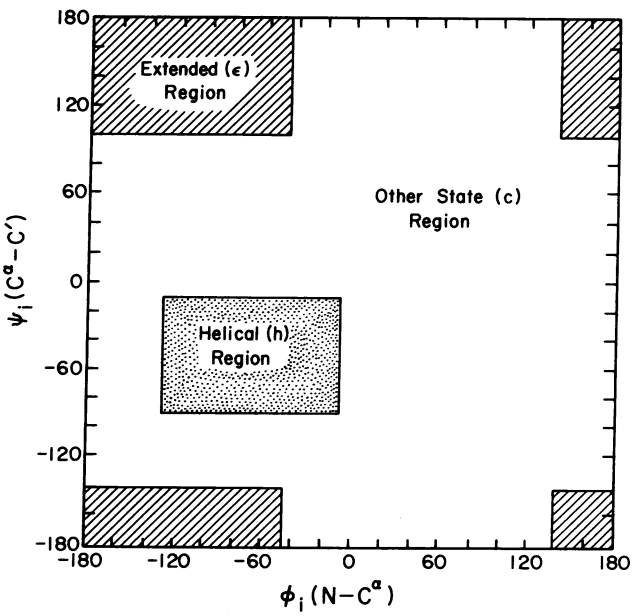

FIG. 2. Helical (h), extended (ϵ), and other (c) state regions, in terms of which the conformation of each residue is specified in step A.

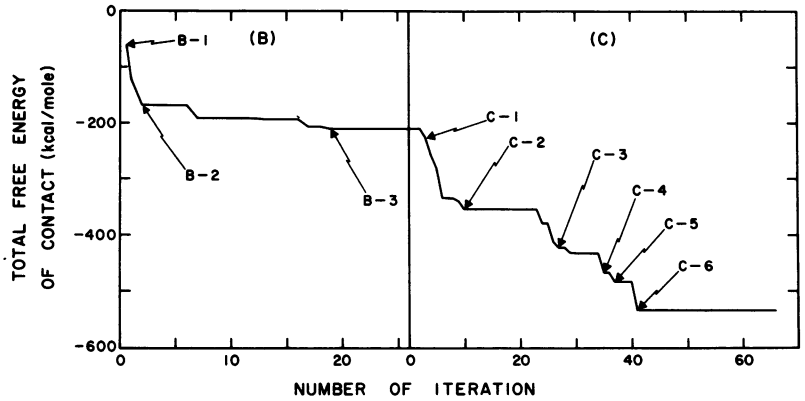

FIG. 3. Change in free energy due to formation of contacts between residues of BPTI in steps B and C. The abscissa should be multiplied by 1500 conformations in step B, and by 400 conformations in step C.

In step B, we make a random conformational change by randomly selecting the number of residues to vary, and their positions in the chain, and then altering the conformations of these residues randomly (*i*) within the *whole* of their own conformational regions; these conformational changes are repeated (1000 times in this study). The resulting lowest-free-energy conformation is altered as above, but now (*ii*) *restricted* to the region $|\phi,\psi| \leq 30°$; these conformational changes are repeated (500 times, here). Thus, one iteration of step B consisted of 1500 conformational deformations in this study.

Various types of random conformational deformations are

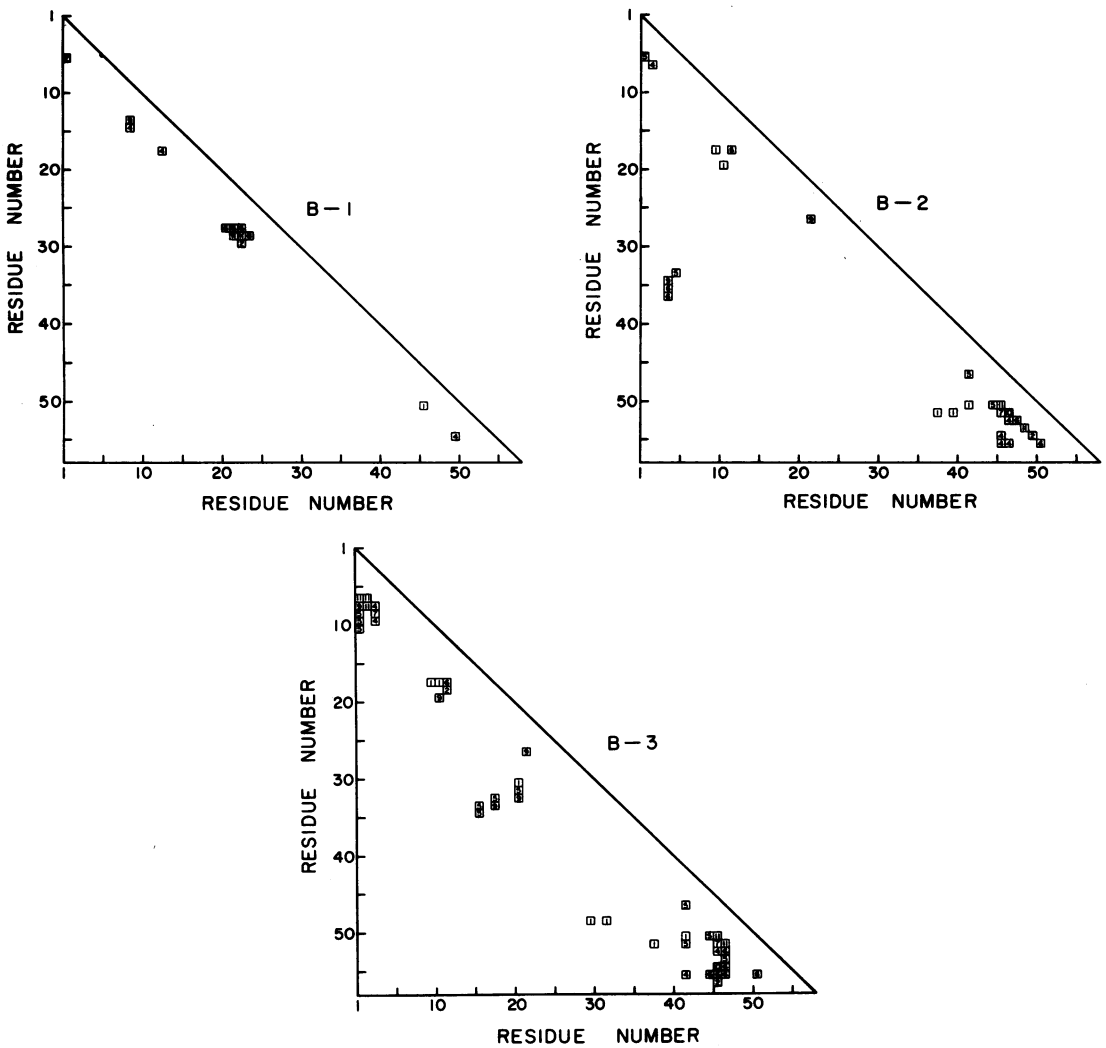

FIG. 4. Changes in the contact maps of BPTI during step B. The numbers in the squares have the same meaning as in Fig. 1B. The shorter-range contacts are omitted.

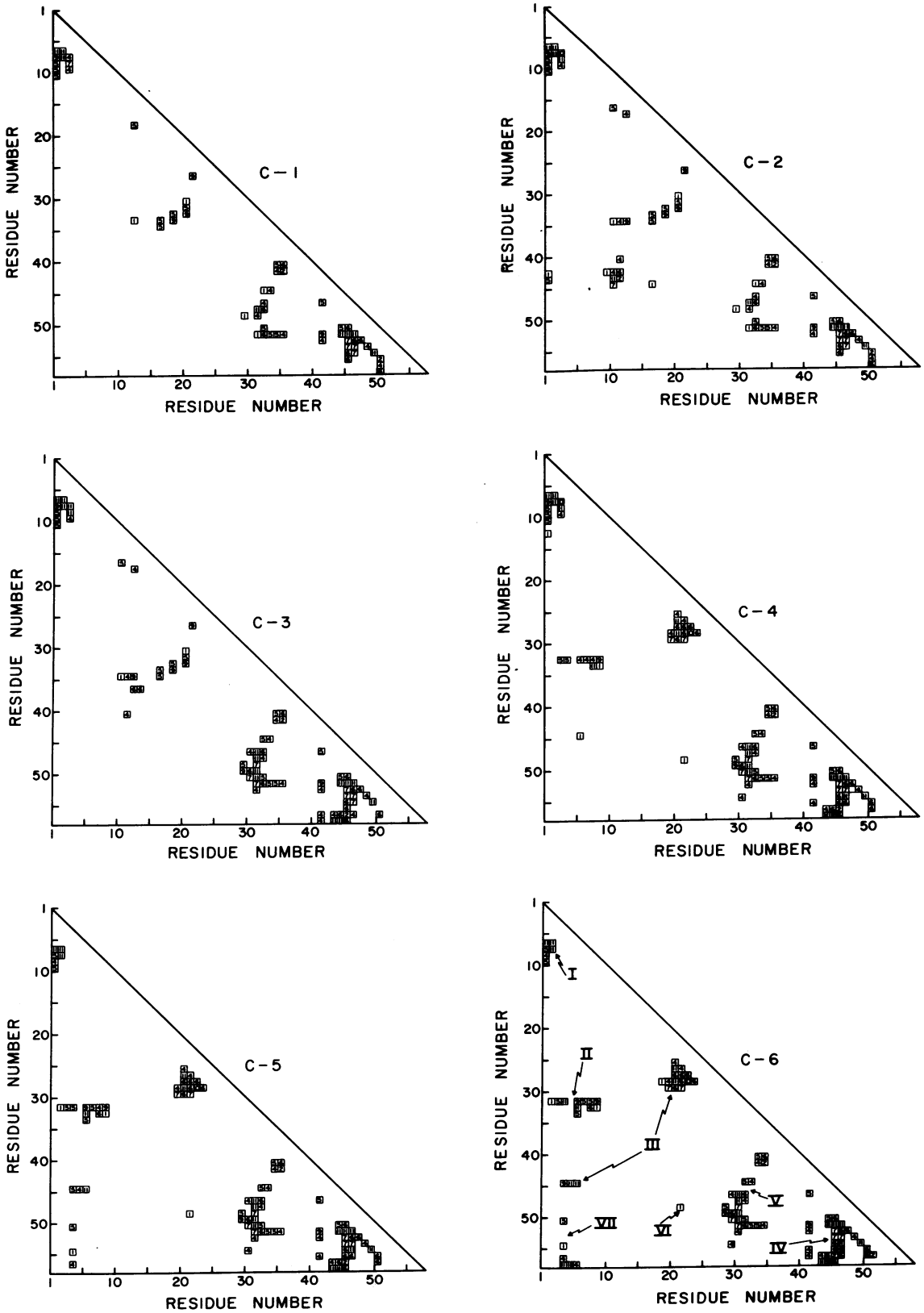

FIG. 5. Same as Fig. 4, but for step C.

produced in step B-*i*, typically (*a*) local deformations of the chain without a large change of shape of the whole molecule, and (*b*) drastic changes in the overall shape of the protein. However, only the deformations of type *a* effectively serve to form small local contact regions in the chain. The small conformational changes of step B-*ii* are made to stabi-

lize the contact regions formed in step B-*i*. Since large rings have a lower probability of formation than medium-size rings in a polymer chain, deformations of type *b* are not effective in bringing contact regions together (long-range order) in step B, even though such deformations are included in this step; i.e., the rare occurrences of contacts of long-range order are altered by the more stable contacts of medium-range order. At the end of step B (*i* and *ii*), local contact regions of medium-range order are formed, without yet achieving overall globularity.

In step C, drastic conformational changes are produced as follows. First, we randomly select a *region* (a given sequence of residues) to be varied, then randomly choose the number of residues to vary, and their positions, within the *region*, and finally we alter the conformations of these residues randomly within their own conformational domains (Fig. 2). Because of the *regional* conformational changes, some of the local contact regions formed in step B may rearrange, but the predominant effect is to bring these contact regions together (long-range order). By restricting these conformational changes to local regions, the contacts of long-range order are favored without an increase in free energy that would arise from the destruction of the local contact regions formed in step B.

In brief,[§] the chain was represented as a sequence of symbols h, ϵ, and c (the simulated result from step A). Then the following Monte Carlo simulation was carried out. Chain conformations were generated randomly, as described in step B-*i*, until one was found with no hard-sphere overlaps (Eq. 3). Then, the procedure described above for step B was followed, always checking for absence of hard-sphere overlaps. When a conformation with no overlaps was found, the free energy for all pair contacts was computed, using the parameters mentioned in section II. If this free energy was larger than that of the previous conformation, the latest conformational change was unfavorable (and was discarded), and the conformational changes of step B were started again from the previous conformation. If the new conformational changes gave a lower free energy than that of the previous conformation, the new one was retained as the starting conformation for further conformational changes in step B. This procedure was repeated until the free energy could not be lowered by an additional 10,000 applications of step B.

Then, the process was continued by application of step C, again until the free energy could not be lowered by an additional 10,000 applications of step C.

IV. Results and discussion

Fig. 3 shows the change in free energy of contact formation in steps B and C, starting with the x-ray (simulated step A) structure of BPTI. A horizontal line indicates that high-free-energy conformational changes were generated (and discarded) in these iterations. The conformational changes, corresponding to the free energies of Fig. 3, are illustrated in Figs. 4 and 5 for steps B and C, respectively.

To assess the predictive results, the final one (C-6 of Fig. 5) should be compared to Fig. 1 (the native structure). It can be seen that the main features of the native structure (indicated by I–VII in Fig. 1B) are reproduced in regions I–VII in C-6 of Fig. 5. This agreement is fairly good, especially if one takes into consideration the fact that we did not use a disulfide loop-closing function, as was used previously (2) to obtain globularity, and the fact that values of ϕ and ψ were not specified precisely initially, as previously (2).

The most important implication of these results is that the three-step mechanism of protein folding is necessary, because a random conformational change in the whole molecule can lead *only* to small contact regions *within about ten residues*, due to local medium-range interactions (in step B), as seen in Fig. 4. The introduction of random conformational changes, that are restricted to a limited number of residues (in step C), is required to produce large conformational changes and the globular form of the protein, in which the local contact regions of long-range order (*more than about 10 residues apart from each other*) are formed.

This work was supported by grants from the National Institutes of Health and from the National Science Foundation.

1. Scheraga, H. A. (1974) in *Peptides, Polypeptides and Proteins*, eds. Blout, E. R., Bovey, F. A., Goodman, M. & Lotan, N. (John Wiley, New York), pp. 49–70.
2. Burgess, A. W. & Scheraga, H. A. (1975) *Proc. Nat. Acad. Sci. USA* **72**, 1221–1225.
3. Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975) *J. Phys. Chem.*, in press.