Proc. Natl. Acad. Sci. USA
Vol. 74, No. 4, pp. 1320–1323, April 1977
Chemistry

Model of protein folding: Incorporation of a one-dimensional short-range (Ising) model into a three-dimensional model

(mechanism of folding/statistical mechanical averaging/Monte Carlo)

SEIJI TANAKA[*†‡] AND HAROLD A. SCHERAGA[*†§]

[*]Department of Chemistry, Cornell University, Ithaca, New York 14853; and [†] Biophysics Department, Weizmann Institute of Science, Rehovoth, Israel

Contributed by Harold A. Scheraga, February 2, 1977

ABSTRACT **In this paper, we have incorporated a one-dimensional short-range model into a three-dimensional model for protein folding. It has been applied, by extending the concept of the three-step mechanism for protein folding proposed in our previous paper, to simulate the folding of bovine pancreatic trypsin inhibitor, using a Monte Carlo procedure in all three steps, A, B, and C. The statistical mechanical ensemble treatment of the short-range model serves as a constraint on the Monte Carlo procedure, in which conformational transitions are introduced. The preliminary results of 10 independent Monte Carlo trials indicate that, while folding is achieved, improvements are required in order to account for the correct three-dimensional structure of a globular protein.**

This is an extension of our previous paper (1), in which protein folding was assumed to occur by a three-step mechanism (steps A, B, and C). Previously (1), step A was simulated by adopting the ordered (helical and extended) *segments* of the observed x-ray structure of bovine pancreatic trypsin inhibitor in order to demonstrate the role of medium- and long-range interactions[¶] in steps B and C, and the main features of the native structure were reproduced. However, in this paper, instead of simulating step A by adopting the observed ordered segments, we describe how a one-dimensional short-range[¶] model can be incorporated to simulate the conformation of the protein in step A and to serve as a constraint on the folding in steps B and C. It is recognized that short-range models are inadequate, by themselves, for describing a globular protein, and that they will provide less than the 100% accuracy (assumed in ref. 1) in assigning the conformation of each residue to one of three conformational regions in step A. However, it is hoped that the introduction of medium- and long-range interactions can compensate to some extent for the shortcomings of the short-range models. Our emphasis here is not so much on the results of such calculations, but on the introduction of the short-range interaction model into a three-dimensional protein-folding scheme. The latter is referred to as a three-dimensional scheme not only because it pertains to the three-dimensional structure of the protein, but also to distinguish it from the one-dimensional short-range models. Improvements of the three-dimensional model, and of the parameters used therein, are presently being investigated to determine whether these improvements will lead to a computed structure that resembles the native one. Thus, we present a three-step procedure, involving statistical mechanical averaging and a Monte Carlo technique, to fold a protein (bovine pancreatic trypsin inhibitor in this example).

The Monte Carlo procedure used here, and in our earlier paper (1), differs from the model-building approach of Ptitsyn and Rashin (2) and from the conformational energy minimization treatments of Levitt and Warshel (3) and of Burgess and Scheraga (4), all three of which were applied to the protein-folding problem. None of these treatments incorporates the ordered backbone conformations into a three-dimensional model with statistical mechanical procedures, as is done here.

Incorporation of short-range one-dimensional (Ising) model into a three-dimensional model

We use a statistical mechanical procedure to treat the one-dimensional cooperativity arising from short-range interactions, and incorporate it into a model for protein folding in which medium- and long-range interactions are also taken into account. The incorporation of a one-dimensional model into a three-dimensional model (for the helix-coil transition), using a Monte Carlo method, was done earlier by Tanaka and Nakajima (5). As shown in section VI of ref. 6, various molecular (or ensemble) averages can be obtained by statistical mechanical averaging over the whole protein molecule—within the framework of a short-range interaction model. Thus, the distribution of structures around the most probable one can be determined. The ensemble averages based on short-range interactions then serve as a constraint when medium- and long-range interactions are introduced, thereby restricting the folding pathway (in the Monte Carlo procedure) to conformations that satisfy the requirements of short-, medium-, and long-range interactions. While more states can be introduced (7–9), we confine ourselves here to a three-state short-range model, in which the three states, defined in Fig. 2 of ref. 1, are helical (h), extended (ϵ), and other (c).

We use the first-order *a priori* probability $F_{i;\eta_i}$, defined in Eq. 44 of ref. 6, that the ith residue is in conformational state η_i, where $\eta_i = h$, ϵ, or c, for $i = 1$ to N, where N is the number of residues in the chain. We emphasize that $F_{i;\eta_i}$ is obtained by statistical mechanical averaging over the whole molecule, using Eq. 44, the matrix of Eq. 26, and the partition function of Eq. 21 of ref. 6. The statistical weights used in Eq. 26 are given in ref. 9 (except that the values of $v_{h,j}{}^*$ were optimized to reproduce the mean helical and extended-state contents of the trypsin inhibitor). Using $F_{i;\eta_i}$, the ensemble of conformational states of the protein in step A is generated as follows.

The conformation of the ith residue of the chain is selected from h, ϵ, or c by a random number generator, satisfying the computed values of $F_{i;\eta_i}$. By repeating this for every residue, we obtain a specific conformational sequence $\{\rho\}$ for the backbone of the trypsin inhibitor; it is highly probable that this is among the nearly correct conformations—in step A. Repetition of this process produces an ensemble of conformational sequences $\{\rho\}$. Each specific conformational sequence $\{\rho\}$ is rep-

[‡] From Kyoto University; 1972–1975 at Cornell University, 1975–1976 at Kyoto University and Weizmann Institute of Science.

[§] To whom requests for reprints should be addressed at Cornell University.

[¶] 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 of a protein, then these terms pertain to interactions for which $|i - j| \leq 4$, 5 to 20, and >20, respectively.

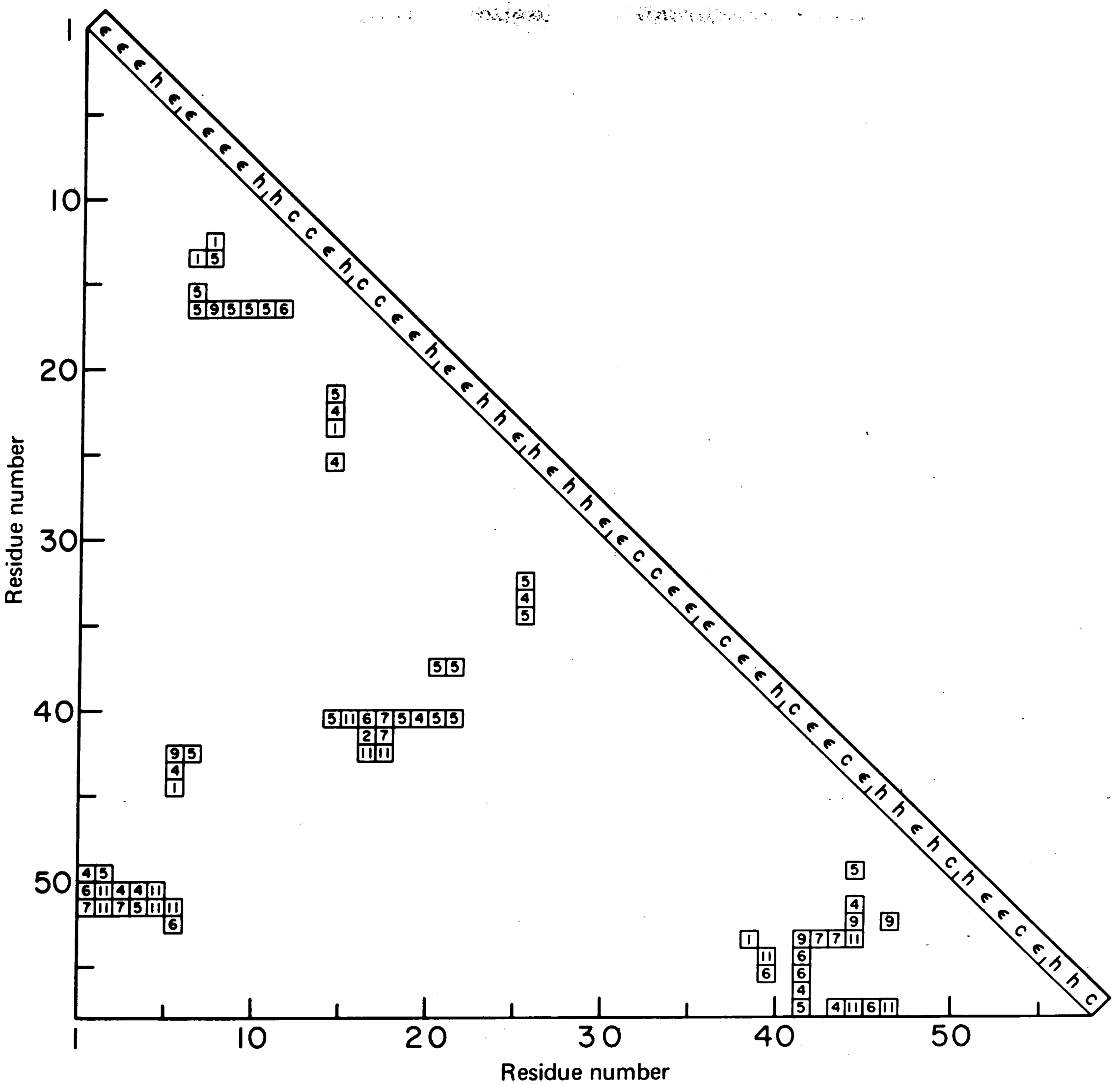

FIG. 1. The final structure obtained at the end of step C in one of the 10 independent Monte Carlo simulations. The numerals in the squares designate the type of contact between amino acids, as defined in the legend of Fig. 1 of ref. 1. The final backbone structure is shown on the diagonal of the map.

resented as a sequence of conformational symbols h, ϵ, and c. These symbols are converted to backbone dihedral angles ϕ and ψ by random selection in the h, ϵ, and c regions of Fig. 2 of ref. 1 (as described later in a discussion of steps B and C), thereby producing a three-dimensional structure of the chain for a specific $\{\rho\}$.

As described by Tanaka and Nakajima (5), the conformation generated by using the conformational probability, $F_{i;\eta_i}$, automatically includes the contribution of the short-range interactions; the ensemble of various conformational sequences $\{\rho\}$ thus generated pertains to a system at thermodynamic equilibrium (within the short-range interaction model). We assume, as an initial approximation, that the ensemble-average pertains to a specific conformation, and thus describes the contributions of short-range interactions to the free energy of a specific conformation; the medium- and long-range interactions are taken into account specifically, as described below.

Using spherical representations for the backbone and side-chain groups, respectively, as in ref. 1, the three-dimensional structure (for a specific conformational sequence $\{\rho\}$) is checked for overlaps between *all* spheres (excluded volume effect). This procedure (generation of $\{\rho\}$, conversion of $\{\rho\}$ to ϕ,ψ-space, and check for hard-sphere overlaps) is repeated until a specific sequence $\{\rho\}$ is found that has no overlaps. Then the conformational free energy of this structure is computed, using the standard free energies of formation of a contact between amino acids k and l, $\Delta G°_{k,l}$, that are reported in ref. 10. These values of $\Delta G°_{k,l}$ include not only medium- and long-range interactions, but also the effect of hydration. If this conformation has a lower free energy than the previous one, this one is selected over the previous one to obtain a conformational sequence $\{\rho\}$ that provides as low a free energy as can be obtained in step A (and which is used as the starting conformation in step B). But, if this conformation has a higher free energy than the previous one, it is discarded, and a new $\{\rho\}$ is selected from the ensemble. This procedure is repeated (10,000 times in this study). The conformation with the lowest free energy, i.e., the one obtained after these 10,000 trials, is taken as the starting conformation for step B.

In steps B and C, we incorporate the one-dimensional short-range model of step A into a three-dimensional model, and make alterations in the conformation to enable the medium- and long-range interactions to lower the free energy still further. [In step A, the short-range interaction model was used

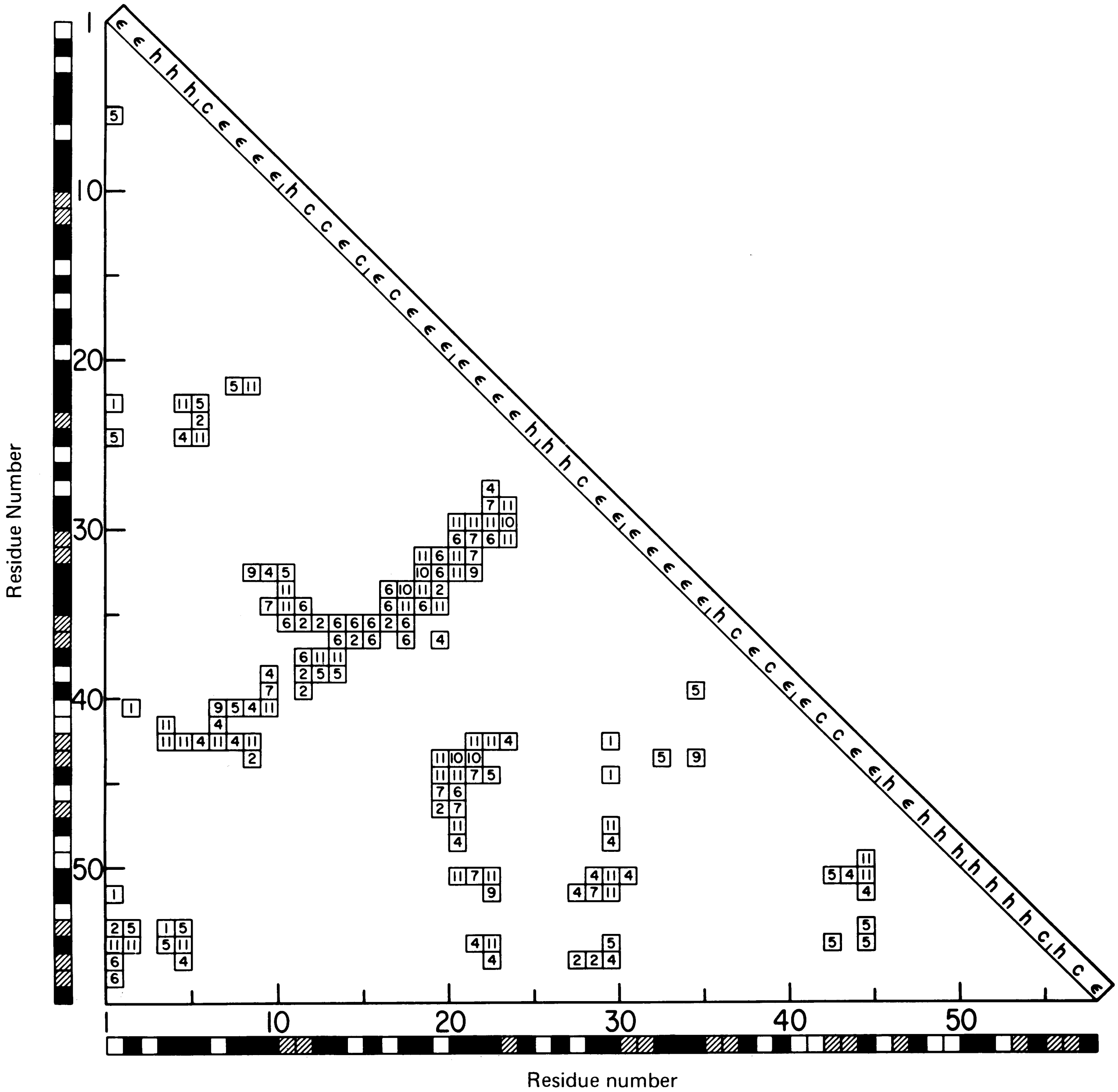

FIG. 2. Contact map for native structure of bovine pancreatic trypsin inhibitor with spherical representation of side chains and backbone. The backbone conformation of native inhibitor is indicated on the diagonal. See legend of Fig. 1 of ref. 1 for further details.

to define the conformational state, whose free energy was then computed; i.e., the free energy was used to select among the 10,000 conformations (with no overlaps) that had been generated by the short-range interaction model.]

In step B, random conformational changes are made in the structure obtained from step A by randomly selecting the number of residues to be varied, and their positions in the chain, and then altering the conformations of these residues. Such alterations of conformations are made, subject to the constraint (of the short-range interaction model) that the computed values of $F_{i;\eta_i}$ be satisfied. (*i*) In this alteration, each residue that is changed has the possibility of going back and forth among the h, ϵ, and c regions shown in Fig. 2 of ref. 1 (i.e., it can undergo a backbone conformational transition). [Whereas, in step A, $\{\rho\}$ was changed by allowing every residue in the chain to vary, in step B (and also in step C) the change is limited to the number of residues selected randomly, as described above (and in the later paragraph on step C).] The new conformational sequence $\{\rho\}$ is converted to a three-dimensional structure by random selection of ϕ–ψ values within the corresponding regions, η (see Fig. 2 of ref. 1). This procedure was repeated 500 times, each time starting with the final structure from step A (for the first iteration of step B) or from the previous structure obtained in (*ii*) of step B. (*ii*) Starting with the conformation of lowest free energy obtained in (*i*), only the values of ϕ and ψ were changed (in the manner described in the first sentence of this paragraph), *without* allowing the backbone structure to change among the h, ϵ, and c regions [i.e., the specific conformational sequence obtained in (*i*) was retained in (*ii*)]. The conformational changes introduced in (*ii*) were repeated 500 times to obtain a conformation of lower free energy. Thus, one iteration in step B consists of 1000 conformational deformations [500 in each of (*i*) and (*ii*)]. In contrast to ref. 1, where the deformation [(*i*) and (*ii*) of step B, as used in ref. 1] was restricted to $|\Delta\phi,\Delta\psi| \leq 30°$, here the whole range of ϕ,ψ in each region η was allowed in step B [(*i*) and (*ii*)].

The procedure of step B [(*i*) and (*ii*)] was iterated, always checking for the absence of hard-sphere overlaps after each conformational deformation. When a conformation with no overlaps was found, the free energy was computed, as in step A. If this free energy was larger than that of the previous conformation, the latest conformation was discarded (and the previous conformation was again considered). If the new conformational change led to a lower free energy than that of the previous conformation, the new conformation was retained as the starting one for further conformational changes in (*i*) and (*ii*) of step B. This procedure was repeated until the free energy could not be lowered by an additional 10,000 applications of step B.

In step C, the changes are restricted to randomly selected *local* segments of the chain, and the alterations within these *local* segments produce drastic conformational changes. In each segment selected, we randomly choose the number of residues to vary, and their positions within the segment, and alter the conformations of these residues randomly. Allowance is made for a conformational transition of the backbone structure among the h, ϵ, and c regions, but again subject to the constraint of the short-range interaction model that the computed values of $F_{i;\eta_i}$ be satisfied. This process was repeated until the free energy could not be lowered by an additional 10,000 applications of step C.

There are two points to be emphasized. First, in steps A, B, and C, conformational transitions in the backbone structure $\{\rho\}$, determined by the short-range one-dimensional model, are made to obtain a favorable backbone conformation, not only favorable as far as the short-range interactions are concerned, but also including the medium- and long-range interactions. Thus, a conformation formed in a given step may be rearranged, to some extent, to form a more stable structure in a later step.

Second, the conformational changes of step B can include the drastic ones of step C. The drastic conformational changes of step C are not effective in forming a compact globular structure of low free energy in the early and intermediate stages of step B (see also the third paragraph of section IIIB of ref. 1). However, the conformational changes of the type that occur in step B (which can also occur in step C) may sometimes become effective after the later stages of step B, where contacts of medium-range order have been established.

We have observed here, as previously (1), that few long-range contacts are formed in step B; i.e., step C is required to bring the contact regions (formed in step B) together. Thus, the computations carried out here again show that step B is not sufficient to produce long-range contacts (those far from the diagonal of a triangular contact map).

Results and discussion

Ten independent Monte Carlo simulations (i.e., 10 independent steps A, B, and C) of the folding of the trypsin inhibitor were carried out. In each one, the total free energy decreased in a manner similar to that of Fig. 3 of ref. 1, and the contact regions formed in a manner similar to those shown in Fig. 4 (step B) and Fig. 5 (step C) of ref. 1. The final stage of step C for one of the 10 independent Monte Carlo simulations is shown in Fig. 1. By comparison with the native structure (shown in Fig. 2), it can be seen that, while some features of the native structure appear, the agreement is not as good as was obtained in our previous study (1), in which a specific backbone conformational sequence $\{\rho\}$ (taken from the x-ray structure) was maintained throughout steps B and C; in this study, the backbone was allowed to change conformational regions (subject to the constraints of the short-range one-dimensional model). Also, there are differences in the appearances of the triangular maps among the 10 independent Monte Carlo simulations.

Some improvement in the appearance of the triangular maps was achieved by the following modifications of the procedure: helical and extended segments were introduced initially, using the short-range model of ref. 9; a free energy of formation of a hydrogen bond was included in these ordered regions to reduce the likelihood of their melting once they are formed and to make it more difficult to break a helical sequence in its interior than at its ends; an initial application of the short-range model was then made only to the nonordered segments (followed by the normal step A, B, and C procedures—even in the previously assigned ordered regions). However, the maps were still not as good as those obtained previously (1) by assigning the conformational regions of each residue with 100% accuracy in step A.

Further improvements are under consideration in this simple and reasonable Ising-Monte Carlo model. These involve a finer subdivision of the conformational space of each residue than is used in the three-state short-range model in order to localize the conformation more precisely [with improved statistical weights obtained from observed x-ray structures, with an improved treatment of the energy of the short-range interaction model, and with improved long-range (hydration) parameters].

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

1. Tanaka, S. & Scheraga, H. A. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 3802–3806.
2. Ptitsyn, O. B. & Rashin, A. A. (1975) *Biophys. Chem.* **3,** 1–20.
3. Levitt, M. & Warshel, A. (1975) *Nature* **253,** 694–698.
4. Burgess, A. W. & Scheraga, H. A. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 1221–1225.
5. Tanaka, S. & Nakajima, A. (1972) *Macromolecules* **5,** 714–720.
6. Tanaka, S. & Scheraga, H. A. (1976) *Macromolecules* **9,** 159–167.
7. Tanaka, S. & Scheraga, H. A. (1976) *Macromolecules* **9,** 812–833.
8. Tanaka, S. & Scheraga, H. A. (1977) *Macromolecules* **10,** 9–20.
9. Tanaka, S. & Scheraga, H. A. (1977) *Macromolecules,* in press.
10. Tanaka, S. & Scheraga, H. A. (1976) *Macromolecules* **9,** 945–950.