(Reprinted Nature, Vol. 253, No. 5494, pp. 6 94-6 98, February 2 7, I9 75)

# Computer simulation of protein folding

**Michael Levitt* & Arieh WarshelI***

Department of Chemical Physics, Weizmann Institute of Science, Rehovoth, Israel

*A new and very simple representation of protein conformations has been used together with energy minimisation and thermalisation to simulate protein folding. Under certain conditions, the method succeeds in 'renaturing' bovine pancreatic trypsin inhibitor from an open-chain conformation into a folded conformation close to that of the native molecule.*

PROTEIN molecules owe their enormous functional versatility to the fact that they spontaneously fold into complicated and unique conformations determined by the particular amino-acid sequence[1]. Discovering the relationship between protein sequence and conformation is a fascinating theoretical problem of fundamental importance, Most previous theoretical work has used the concept of 'local structure', in which the conformation of a short segment of polypeptide chain is supposed to depend almost entirely on the sequence of that segment. Although this approach has helped understand local secondary structure[2,3], it has not shown how residues distant along the chain can come together to form the overall conformation. The only promising attempt to study the tertiary folding of a protein, in this case myoglobin, was based on the packing of cylinders supposed to represent a helices[4]. The method was not implemented on a computer and cannot be applied more generally to other proteins not built entirely from helices.

Here we tackle the problem differently. First, we simplify the representation of a protein by averaging over the fine details. This is done both to make the calculations much more efficient and also to avoid having to distinguish between many conformations that differ only in these finer details. Second, we simulate the folding of this simple structure by the combined use of convergent energy minimisation and normal mode thermalisation, which accelerate the process by avoiding the many non-productive random fluctuations that occur in nature. Tests of the procedure on bovine pancreatic trypsin inhibitor (PTI), show that under certain conditions it can rapidly reproduce the correct overall folding of this small protein molecule.

## Simple representation of protein structure

Even the smallest protein (say 50 residues) is extremely complicated, with about 750 atoms and 200 degrees of freedom (single-bond torsion angles). Calculating its free energy presents severe computational difficulties, in particular when considering interactions with the rapidly moving solvent molecules and the thermal motion of parts of the protein itself. Our method is designed to overcome these problems and is based on two assumptions: (1) that much of the protein's fine structure can be eliminated by averaging, and (2) that the overall chain folding can be obtained by considering only the most effective variables (those that vary most slowly yet cause the greatest changes in conformation).

Averaging over groups of atoms in the full structure gives a simplified structure with each residue represented by only two centres, the $C^{\alpha}$ atom, and the centroid of the side chain. Interactions are assumed to occur only between side chains, while the $C^{\alpha}$ positions define the chain path (Fig. 1). Each amino-acid residue only has one degree of freedom, the torsion angle about the line joining two adjacent $C^{\alpha}$s (known here as a). Although a simple representation based on virtual bonds has been used before to study polypeptide random coils[5], it has never been applied to ordered globular proteins. This simplification reduces the degrees of freedom by a factor of four and the number of interaction centres by a factor of fifteen. One might also hope that the reduced space used here to describe different conformations would have many fewer energy minima. The space is of lower dimension and the side chains are smooth spheres without all the minor bumps of the all-atom structures.

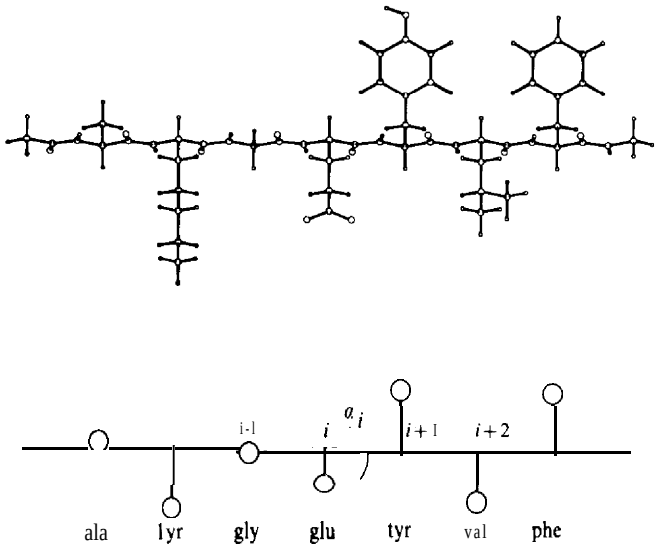


Fig. 1 Relationship between the simplified model of protein structure introduced here and the real alf-atom structure of proteins. The two reference points for each residue in the simplified model correspond to the centroid of the side chain and the $C^{\alpha}$. Each residue is only allowed one degree of freedom: the torsion angle a between the 4 successive $C^{\alpha}$s of residues (i-1, i, i+ 1, i+2). All the side chains of a given type have the same simplified geometry. The bond lengths, bond angles, and torsion angles used to define the geometry of the simplified molecule were taken as the average values found in eight protein conformations, though they could just as well have been taken from amino-acid model compounds.

The effect of the fine details and more rapidly changing variables is included in the effective time-averaged potential functions used. (By the ergodic theorem[6], this time averaging is equivalent to Boltzmann weighted spatial averaging over conformations generated by changing the fast variables.) For rotations about the torsion angfe a, the effective potential is obtained by averaging the energy over all those conformations of a dipeptide that have a particular value of a. As it was impossible to study all 400 different dipeptides, calculations were done on six considered most representative: ala-ala, ala-gly, ala-pro, gly-gly, gly-ala and pro-ala. This showed that

*Present address; MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK.

the effective potential only depended on the nature of the second amino acid, giving different potentials for the a preceding ala, gly, and pro. The alanine potential had a deep minimum at $\alpha = 210°$ (twisted $\beta$ chain) and a more shallow minimum at $\alpha = 45°$ ($\alpha$ helix); the glycine potential had a broad minimum at $\alpha = 0°$ (reverse turn); and the proline potential had two sharp minima at a = 60° and $\alpha = 210°$. Because aspartic acid and asparagine were found to occur as frequently in the reverse turns of known protein conformations as glycine, the same potential was used for all three. The alanine-type potential was used for all other amino acids except proline. All the atoms were included in these dipeptide calculations, which used energy parameters derived from crystals of amino acids, amides, and hydrocarbons.

The interaction potential between a pair of identical amino-acid side chains was also calculated by spatial averaging. Each side chain was assumed to be spherically symmetrical with a radius equal to the average radius of gyration of that side chain. The effective potential was calculated at various distances apart as a sum of the interaction energy of all atoms anywhere in one sphere with all atoms anywhere in the other sphere. This effective potential between identical side chains was approximated by a Lennard-Jones type function, and potentials between pairs of different side chains were obtained by a geometric mean combining law. Because proteins fold in water not a vacuum, interactions with the solvent are included by assigning to each side chain a hydrophobic energy taken from the solubilities of amino acids in water and in ethanol'. In the calculations, the energy of transfer between these two solvents was taken as the difference in energy of the side chain when isolated in water and when completely surrounded by other residues. When surrounded by an intermediate number of neighbours, the hydrophobic energy was varied according to a sigmoid function. More complicated models that include hydrogen bonds and S-S bridges will be described elsewhere, together with full details of the standard geometry and all energy parameters used.

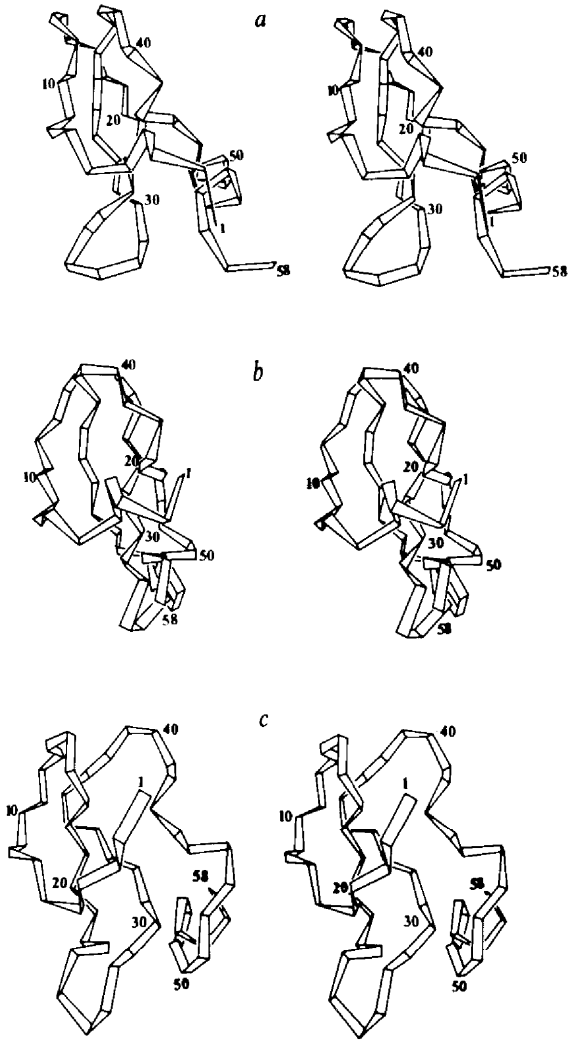


Fig. 2 Stereo ribbon drawings of PTI in: a, the idealised native molecule, *b*, the minimum energy conformation generated starting at the idealised native conformation and c, the best conformation generated by folding from an extended chain with a terminal helix (the final conformation in Fig. 4). (The programs used to rotate the molecules into the same orientation and then draw the ribbon between $C^{\alpha}$s were provided by Dr A. D. McLachlan.)

The folding of this idealised protein can be simulated by solving the equations of molecular dynamics at sufficiently small time intervals. In a viscous medium like water, these equations of motion can be approximated by Langevin equations, where the change in the variables is directed down the energy gradient with a random deflection due to Brownian motions. For greater computational efficiency we neglect these thermal fluctuations while the chain folds, and the end point of the trajectory is the potential energy minimum accessible from the starting conformation. We minimise the energy of the idealised protein chain with respect to all the a angles using a powerful quadratically convergent method (VA09A, by R. Fletcher and taken from the Harwell Subroutine Library). After reaching a minimum, thermal fluctuations are reintroduced and the conformation is considered to be vibrating about the minimum so that each normal mode has average kinetic energy $kT/2$ (where $k$ is the Boltzmann constant and $T$ the absolute temperature). A new starting conformation for the next pass of energy minimisation is chosen by suddenly stopping the thermal vibration. At this time each normal-mode coordinate will be displaced randomly from the minimum by $(R(n)kT/\lambda)^{\frac{1}{2}}$, so that the associated energy becomes $R(n)kT/2$. Here $\lambda$ is the eigenvalue of the energy second derivative matrix corresponding to the particular normal mode, and R(n) is a random number uniformly distributed between 0 and 1. (An exponential distribution of random numbers between 0 and $\infty$ would be more realistic.) Normal-mode thermalisation avoids non-productive changes in conformation for it knows which combinations of angle changes should cause the greatest change in conformation for a given energy increase.

## Testing the simplified representation

The drastic simplifications used in the present representation of a protein conformation were tested by minimisation from near the native folded conformation. Bovine pancreatic trypsin inhibitor was chosen for this test as it is the only small protein (less than 100 residues) of known conformation that has a single polypeptide chain and no additional prosthetic group. As a first step, a simplified native PTI conformation was obtained by taking the $C^{\alpha}$ positions and side-chain centroids from the X-ray coordinates[9] (kindly supplied by Drs Huber and Steigemann). Next an idealised chain, based on the PTI sequence and having the same geometry for all side chains of the same type, was made to fit the simplified native co-ordinates by adjusting the *a* torsion angles. This conformation, known as the idealised native structure, deviates by 1.1 Å r.m.s. from the simplified native structure. The r.m.s. deviation is

$$\left\{\frac{1}{N}\sum_{i,j}^{N}(\Delta r_{ij})^2\right\}^{1/2}$$

where $\Delta r_{ij}$ is the difference, in the two structures, of the distance between side-chain centroids (i and j). Energy minimisation

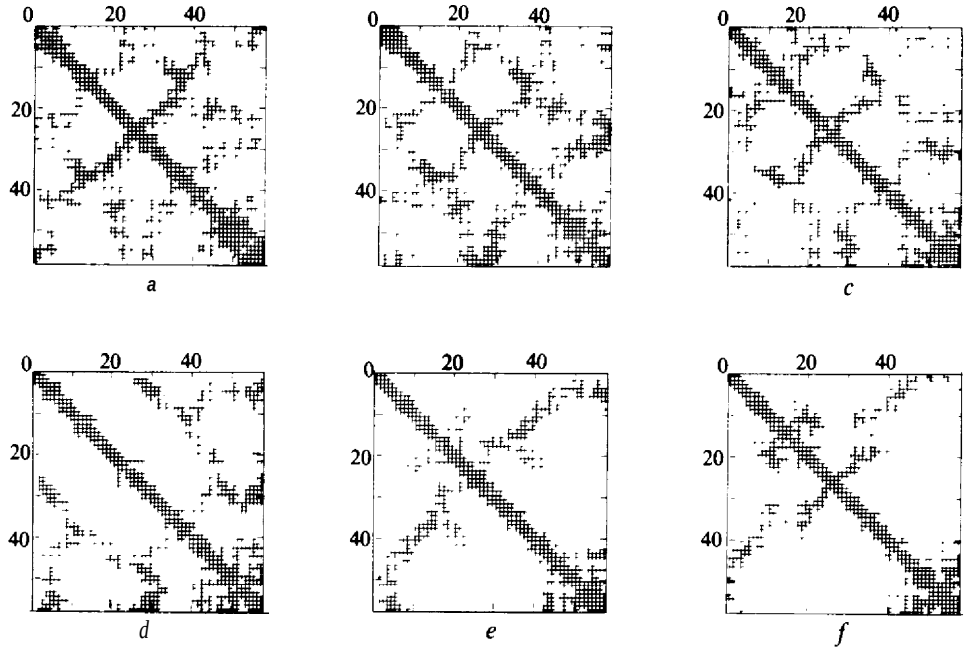


Fig.3 Contact maps[11] of the following conformations: a, idealised native; *b*, folded from idealised native; c, folded from extended chain with terminal helix; *d*, e and *f*, folded from the same starting angles and with the same parameters as c, but using different sequences of random numbers for the thermalisation. The six structures shown here have energies of *-332.0, -52.0, -48.9, -44.9, -32.5, -28.7* kcalorie $mol^{-1}$, respectively and r.m.s. deviations from the idealised native conformation of 1.1, 3.4, 5.3, 6.3, 11.7 and 12.4 Å respectively. (The energy of the idealised native conformation is so high because it has not been minimised.) A cross at the intersection of row *i* and column j indicates that residue *i* is within 10 Å of residue j. In these maps, helices feature as a broadening of the diagonal (down from top left to bottom right), antiparallel β sheet as a band perpendicular to the diagonal, and parallel β sheet as a band running parallel to the diagonal.

from this starting conformation was then carried out to reproduce the stability of native PTI.

After 558 cycles a perfect minimum is reached at an energy of -52.0 kcalorie $mol^{-1}$ and a r.m.s. deviation of only 3.37 Å from the simplified native conformation. Thermal randomisation about this minimum does not lead to further movement from the native molecule on subsequent minimisation. Randomly disturbing the initial best-fit angles (with a disturbance between – 15° and + 15°) has little effect on the conformation obtained by subsequent minimisation. Figure 2 compares the minimum energy and native chain folding in stereo, and Figure 3 compares the contact maps. Because of a general twisting of the molecule, the comparison of the two structures should be done in stereo. Because main-chain hydrogen bonds have been omitted, the terminaf helix becomes distorted and consequently packs too tightly against the β hairpin centred at residue 27.

## Simulation of PTI folding

Having shown that so simple a model can represent the stable conformation of a folded protein, we tried to simulate the actual process of folding. Most tests were done with two **open** starting conformations, one fully extended (all $\alpha = 180°$), and one extended apart from the C-terminal helix ($\alpha = 180°$, except for residues 48 to 58 where a = 45"). Retaining the terminal helix from the native structure is justified here as we are more concerned with the process of folding than with prediction of the native conformation of an unknown protein. In the latter case a statistical rule (see ref. 2) could be used to guess the position of the a helices in the starting conformation. Figure 4 shows the iteration history of minimisation from the second of these starting points, which was the most successful run of those obtained to date. Thermal randomisation about the first minimum, an irregular but extended conformation, raises the energy and in this case causes the chain to bend back on to itself decreasing the r.m.s. deviation. From this point, minimisation first opens the molecule again, but then reaches a new minimum where the chain now has kinks that could become the bends of β hairpins. After a second randomisation, minimisation rapidly folds the molecule bringing the terminal helix close to the β hairpin centred on residue 27. More minimisation and thermalisation first brings together the two top loops (near residues 15 and 40), and then brings the N-terminal tail on to the rest of the molecule. The final folded conformation of Fig. 4 is remarkably like the native molecule (Figs 2 and 3). In both conformations the chain bends back on itself near residues 14, 27, and 40. In both conformations, the pairs of half-cystine residues that are experimentally known to form S-S bridges, are close together (< 10 Å). It is interesting that the C-terminal helix is the part of the native molecule reproduced feast well even though these residues had been set to a perfect helix in the starting conformation; this is due to the omission of peptide-peptide hydrogen bonds which stabilise the helix and could now be introduced.

Repeating the folding simulation from the fully extended starting conformation also lead to a compact structure with many of the features of native PTI, although after the same number of cycles the r.m.s. fit was a little worse (7.7 Å instead of 6.5 Å) and the energy was higher (-27.7 kcaforie $mol^{-1}$ instead of -44.9 kcalorie $mol^{-1}$). Almost all the differences in conformation of these two folded structures involved the last 10 residues which remained extended if not pre-set to a helix and consequently failed to pack against the rest of the molecule.

## Variation of folding conditions

Changing either set of starting torsion angles by a random value between 15" and 15" had little effect on the final conformation. Folding at a lower initial temperature ($T$ = 300 K, rather than $T$ = 1,000 K) failed to reach a compact conformation, as the thermal disturbances were too small to get out of the local minima corresponding to an extended chain.

Four additional runs of 600 cycles, under the same conditions as those used in Fig. 4 but based on different sequences of random numbers, gave rise to different folded shapes. In one of these, the folded molecule was close to the native structure (r.m.s. deviation of 6.3 Å), although the β sheet was formed between parallel rather than antiparallel chains (Fig. 3*f*). In two others, the antiparallel β sheet centred near residue 27 was formed, but this hairpin did not subsequently fold on to itself to give a compact shape (Fig. 3*d* and a). Conformations that deviated more from the native structure always had higher energies, which gives an independent criterion for choosing the best conformation and suggests that more passes of thermalisation and minirnisation lead to a conformation closer to the native one. Of the five runs using different random numbers for the thermalisation step, two succeeded in getting to within 6.5 Å of the simplified native structure in less than 600 cycles. That certain folding pathways are less successful is consistent with the experimental results of Creighton[10] who

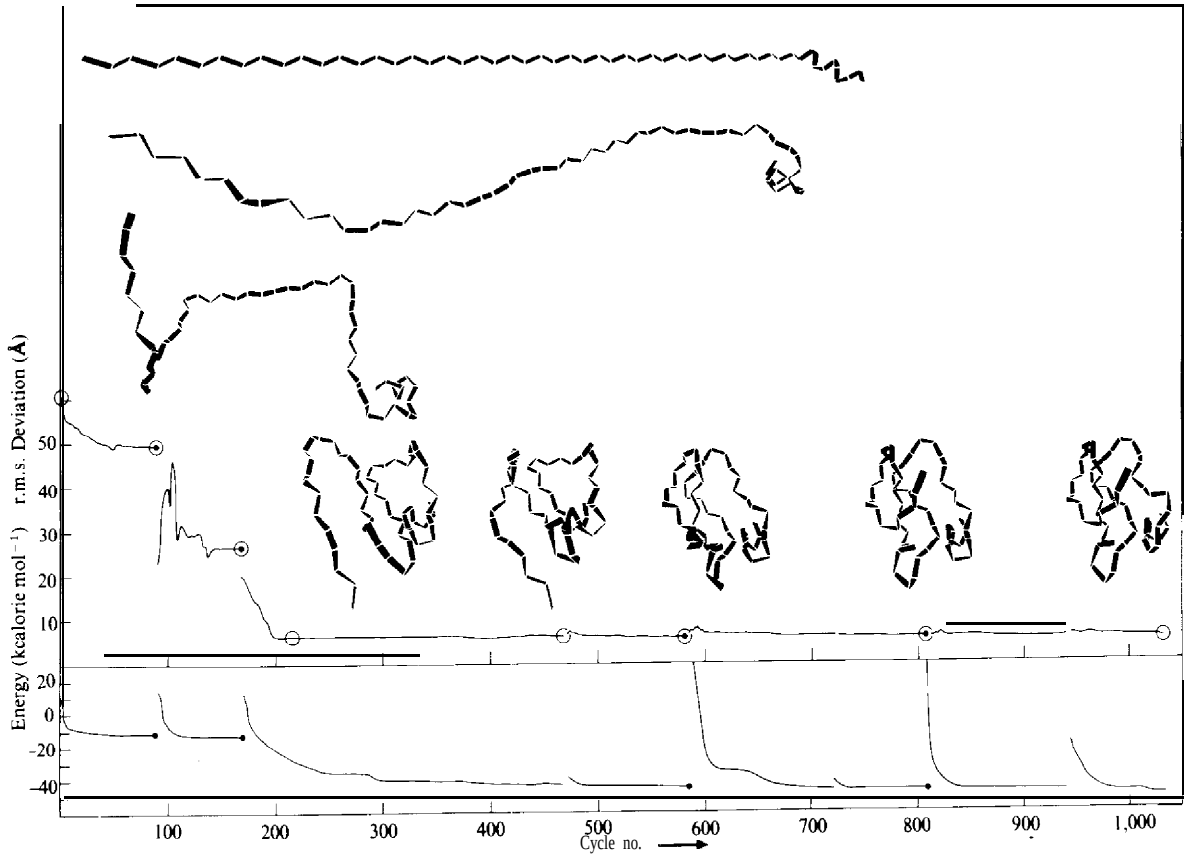


Fig. 4 Simulation of PTI folding from an extended starting conformation with the terminal helix (α = 180° for all except 48 to 58 where a = 45°). No knowledge whatsozver about native PTI is used during this simulation (apart from setting the terminal helix). The conformation was thermalised at the end of each minimisation except near cycles 490 and 730 when the energy rises slightly because the minimisation was restarted after rounding the torsion angles to one d cgree. In the first two thermalisations, each normal mode was perturbed in the **plus** direction to raise the associated energy to $R(n)kT/2$ with $T$ = 1,000 K. In the other three thermalisations, the perturbations were randomly in the plus and minus directions but always such as to raise the energy by $kT/2$ with $T$ = 300 K. (Because the random numbers are distributed uniformly rather than exponentially, these temperatures do not correspond to the macroscopic temperature.) The 8 ribbon diagrams, which show the $C^{\alpha}$ chain path, refer from left to right to the 8 conformations at the circled points on the r.m.s. deviation curve, respectively. The last five conformations have progressively lower energies and are each a little closer to the native structure (E = -43.3, -45.7, -46.0, -46.9 and -48.9 kcalorie $mol^{-1}$, respectively; r.m.s. deviation = 6.08, 5.7, 5.6, 5.4 and 5.3 Å, respectively). The solid dots at the end of a minimisation indicate that a perfect minimum was reached (r.m.s. gradient less than $10^{-6}$ kcalorie $mol^{-1}$-$rad^{-1}$). One cycle takes about 0.6 s on an IBM 370/165 computer.

has analysed the predominant kinetic intermediates present at different times after starting PTI renaturation and found several with the wrong tertiary fold.

## General model for protein folding?

It seems remarkable that so simple a model based on time-averaged forces can account for the stability and folding of a molecule as complicated as a protein. Looking at known protein conformations closely, one is struck by the precise geometry of the interatomic contacts that stabilise the molecule: all possible interior hydrogen bonds are well formed, and many of the nonpolar side chains interlock to form a close-packed interior. As the forces responsible for this precise geometry fall off rapidly with distance and improper orientation, it would seem that folding must depend on a very rare random fluctuation that happened to bring the right residues close together with sufficient precision for the short-range forces to take effect. It therefore seems unlikely that these short-range forces could 'direct' the folding from an open disordered structure. In view of the present results, however, the time average of these short-range forces may play an important role in directing protein folding. These effective forces, which are weak, fairly long range, and not *too* dependent on orientation, restrict the number of low energy conformations severely; they cause the chain to fold into the approximate shape rapidly and without having to pass through many local minima. Because this approximately folded molecule corresponds to a large region in the space of possible protein conformations, folding would not be so rare an event.

As a general model for protein folding we propose that initially, when the chain has a flexible open structure, the effective time-averaged forces between the residues play a central role, folding the chain into a compact shape with most groups close to their final positions (say within 5 Å). Once the chain becomes compact with less freedom of movement, the specific short-range interatomic forces become important; they form a precise conformation provided that the resulting gain in enthalpy overcomes the loss in entropy. The process would be rather like crystallisation, with the atoms simply falling into place from their nearby positions in the approximate folded conformation.

To simulate this second step one switches over to progressively more detailed models gradually incorporating more atoms and ending with the all-atom structures considered in earlier work[12,13]. Although calculating the energy of the all-atom molecule would be time consuming, one would have the great advantage of starting close to the right conformation and could minimise successive overlapping zones of a few residues at a time without having to search through many local minima.

The general concept of using a simple model based on effective time-averaged forces when the detailed forces are too

complicated has many potential applications; for example, the formation of protein quarternary structure and multienzyme complexes, virus assembly and so on. At each level of complexity, forces would be time averaged over those substructures that are relatively fixed or seem to play a less important role in the assembly. Such a hierarchical approach might eventually lead to an understanding and simulation of very complicated biological assembly processes.

We thank the Weizmann Institute Computer Center for facilities and the European Molecular Biology Organisation for supporting one of us (M.L.).

Received October 29, 1974; revised January 15, 1975.

1 Anfinsen, C. B., Science, 181,223-230 (1973).
2 Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, 0. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, B., and Nagano, K., Nature, 250, 140-142 (1974).
3 Scheraga, H. A., in Current *topics* in biochemistry (edit. by Anfinsen, C. B., and Schechter, A. N.), 1-42 (Academic, New York, 1974).
4 Ptitsyn, 0. B., *Vestnik Akad. Nauk S.S.S.R.*, 5, *57-68 (1973).*
5 Flory, P. J., in *Statistical Mechanics of Chain Molecutes, 248-306* (Wiley, New York, 1969).
6 Hill, T. L., in *Statistical Mechanics*, 15–17 (McGraw-Hill, New York, 1956).
7 Nozaki, Y., and Tanford, C., *J. biol. Chem., 246,* 2211-2217 (1971).
8 Simon, E. M., *Biopolymers, IO, 973-989 (1971).*
9 Huber, R., Kukla, D., Ruhlmann, A., and Steigemann, W., *Cold Spring Harbor Symp. quant. Biol.*, 36, 141–148 (1971).
10 Creighton, T. E., *J. molec. Biol., 87, 603-624 (1974).*
11 Phillips, D. C., in *British Biochemistry, Past and Present* (edit. by Goodwin, T. W.), 11-28 (Academic, London, 1970).
12 Levitt, M., and Lifson, S., *J. molec. Biol.*, 46, 269-279 (1969).
13 Warme, P. K., and Scheraga, H. A., *Biochemistry,* 13, 757-767 (1974).