J. Mol. Biol. (1976) 104, 59-107

A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding

MICHAEL LEVITT

Weixmann Institute of Science, Rehovoth, Israel
and
Medical Research Council Laboratory of Molecular Biology
Hills Road, Cambridge, England†

(Received 12 December 1975, and in revised form 19 February 1976)

This report is one of a series of papers that introduce and use a new and highly simplified treatment of protein conformations. The first paper (Levitt & Warshel, 1975) outlined fhe approach and showed how it could be used to simulate the "renaturation" of a small protein. The present paper describes fhe representation in some detail and tests the methods extensively under a variety of different conditions. The third paper (Warshel & Levitt, 1976) is devoted to a study of the folding pathway and stability of a mainly a-helical protein.

In this work, I show how the concept of time-averaged forces, introduced previously (Levitt & Warshel, 1975), can be used to simplify conformafional energy calculations on globular proteins. A detailed description is given of the simplified molecular geometry, the parameterization of suitable force fields, the best energy minimization procedure, and fhe techniques for escaping from local minima. Extensive tests of the method on the native conformation of pancreatic trypsin inhibitor show that the simplifications work well in representing the stable native conformation of this globular protein. Further tests show that simulated folding of pancreatic trypsin inhibitor from open chain conformations gives compact calculated conformations that have many features in common with the actual native conformation. Folding simulafions are done under a variety of conditions, and fhe relevance of such calculations to the actual *in* vitro folding process is discussed at some length. These same techniques have many potential applications including enzyme-substrate binding, changes in protein tertiary and quaternary structure, and protein-protein inferactiona.

1. Introduction

Protein molecules fulfil almost all the catalytic and structural roles in the living cell; they are both the machine tools and building blocks of the cell's factories. Such functional and structural versatility is entirely due to the folding of different amino acid sequences into different three-dimensional conformations. In each of these folded conformations, the position of every atom is precise and depends uniquely on the particular amino acid sequence (Anfinsen *et al.*, **1961;** Anfinsen, **1973).** This relationship between a protein's sequence and its three-dimensional structure constitutes the second part of the translation of genetic information into functional

† Present address.

protein molecules. The first part, the synthesis of the correct sequence of amino acids from the DNA sequence, differs in almost all respects from the second part: protein synthesis is a complicated biochemical process depending on hundreds of enzymes, transfer RNAs, and the ribosome; protein folding is a simple physical process depending on the same interatomic forces that stabilize the simplest molecules. Besides the attractiveness to theoreticians of the protein folding problem, the ability to calculate the conformation of any protein from its sequence would be an invaluable aid to molecular biology in general.

In principle, one should be able to use the methods of conformational analysis developed for small molecules on proteins. With these methods, one finds stable conformations of the molecule by changing selected variables to minimize the total energy, which is expressed as an analytical function of the atomic positions, the chemical connectivity, and the interatomic forces. In practice, severe problems arise when extending techniques that work well for small systems to much larger systems. (a) Proteins have too many atoms. As an important contribution to the energy is summed over all pairs of atoms, calculating the total energy of a protein is time-consuming. (b) Proteins are stable in water at room temperature. While the properties of small molecules in *vacuo* at low temperatures can be computed fairly easily, much less is known about the effect of the solvent and atomic thermal motion on the interatomic forces. (c) Protein structures need to be described by too many variables. Even if one considers only the torsion angles about single bonds, a small protein still has several hundred degrees of freedom? making energy minimization much less efficient.

Previous theoretical studies on proteins illustrate these difficulties clearly. Much attention has been given to the allowed conformations of single amino acids, essentially a small-molecule problem. The earliest work was based on the simple idea of forbidding conformations that have very close non-bonded contacts (Ramachandran *et al.*, 1963), whereas the most recent studies include solvent and entropic effects (Lewis *et al.*, 1973). Studies of bigger systems with up to, say, 20 residues have led to the introduction of more powerful techniques (Gibson & Scheraga, 1967, 1969), but the results were disappointing as in calculations on gramicidin S (Vanderkooi *et al.*, 1966; Liquori *et al.*, 1966; Momany *et al.*, 1969; De Santis & Liquori, 1971). In recent years, following the first energy calculations on known protein conformations (Levitt & Lifson, 1969), more attention has been given both to the energy refinement of X-ray co-ordinates of proteins (Levitt, 1974; Warme & Scheraga, 1974; Hermans & McQueen, 1974; Gelin & Karplus, 1975), and to the binding of a substrate to a protein (Platzer *et al.*, 1972; Levitt, 1972, 1974*b*). In all these cases the calculated changes in conformation are small (< 1 8).

Recently, Burgess & Scheraga (1975) applied the methods used before, in the refinement of protein X-ray co-ordinates, to a conformation of pancreatic trypsin inhibitor that had the correct local structure (all (ϕ, ψ) angles had been set to within 30° of the native values), but did not have the correct native tertiary structure. Although this calculation required considerable computer time, their results were disappointing: it was not possible to refine the (ϕ, ψ) angles to get back the native tertiary structure, even if the S-S bonds were artificially brought together. These difficulties arise from the intrinsic complexity of protein structure in the conventional all-atom representation, the time-consuming evaluation of the energy with so many atoms, and the large number of variables that must be considered.

of interactions with more distant groups. As this compression also amounts to about 10%, both these factors cancel giving $r^\circ = 2r_v$.

Electrostatic interactions (i.e. hydrogen bonds) between polar groups of side chains were not considered here. These interactions are of a very specific nature and cannot be included in a simple one-centre model of side chains. In any case, such interactions are very rare in the native conformations of globular proteins.

Another type of interaction between side chains is also too specific to be treated by spherical averaging, namely the disulphide bonds between half-cystine residues (Cys). This special interaction is treated in two ways. In one, the van der Waals' forces are simply not calculated between 2 Cys residues; there is no attractive S-S bond but also no non-bonded repulsion to prevent close approach of these 2 groups. In the other, a harmonic potential is introduced to constrain the centroid separation of selected pairs of Cys residues to 4·2 A, a value found in S-S bridges in proteins. To use this constraint one must specify which pairs of half-cystines are to be linked.

(ii) *Side chain-solvent interactions*

Interactions with the solvent are clearly an important factor, as proteins are stable in water for which different side chains have very different affinities. Water molecules carry a large dipole moment, are in continual thermal motion, and tend to favour specific orientations around the protein due to hydrogen bonds; for these reasons the full explicit treatment of protein-water interactions is not feasible in most calculations. Instead it is assumed here that the interaction energy of a side chain with water is proportional to the amount of water in contact with the group. As the exact calculation of surface area in contact with water is too complicated (Lee & Richards, 197 **1),** the fraction of water lost from atom i due to the approach of atom j is approximated by the simple sigmoid function

$$g(r_{ij}) = 1 - \tfrac{1}{2}\{7x^2 - 9x^4 + 5x^6 - x^8\}, \quad x < 1$$
$$= 0, \quad x \geq 1$$

for $x = r_{ij}/r_{max}$,

where r_{ij} is the separation of side chains i and j, and r_{max} has the fixed value of 9 A. The form of $g(r_{ij})$ is to some extent arbitrary; it was chosen for its simplicity and the fact that the function and its first and second derivatives are zero at $x = 1$, making the change to $g(r_{ij}) = 0$ continuous. For calibration, a side chain completely surrounded by other side chains is assumed to have **10** near-neighbours each at $r_{ij} = 6{\cdot}3$ Å. Such a side chain loses all its hydration shell since

$$\sum_{j=1}^{10} g(r_{ij}) = 10\, g(0{\cdot}7) \approx 1.$$

Each type of side chain is assigned a solvent interaction energy estimated from solubilities of amino acids in water and ethanol (Nozaki & Tanford, **1971).** The measured energy of transfer between these 2 solvents was set to the energy difference of a side chain when isolated in water ($\sum_j g(r_{ij}) = 0$) and when completely surrounded by other side chains ($\sum_j g(r_{ij}) = 1$). The total loss in solvent interaction energy due to the close approach of residues i and j is $(s_i + s_j)g(r_{ij})$, for in such an encounter water is displaced from both i and j. Table 3 gives the values of s_i used here, and Fig. 3 shows the variation of the solvent energy with group separation.

Estimating the accessible surface area from the number of near-neighbour contacts should work better in the simplified model than for conventional representations as the groups are all about equal size and there are no close-bonded neighbours. Nevertheless, the present model is deficient in that the value of $\sum_j g(r_{ij})$ can exceed 1 if there are several close contacts; this corresponds to the physically unreasonable removal of more than all of the hydration shell.

(iii) *Peptide-hydrogen bonds*

The main backbone interactions in proteins are the hydrogen bonds between pairs of

peptide groups. The geometry of the simplified peptide group is based on 2 atoms (0' and N') instead of the 4 atoms used conventionally (see Fig. 1.) The corresponding energy parameters were chosen. so that the simple peptide group would reproduce the angular and distance dependence of the all-atom peptide hydrogen bond (Fig. 4). For some calculations a "swollen" peptide with an O'–N' separation of 2 Å (instead of 1 Å) was used to reproduce the weakening of the hydrogen bond in water. In the latter case, the shape of the energy contour (Fig. 4) is unaltered except for a shift of the minimum from an interpeptide distance of 4·5 Å to 5·5 A.

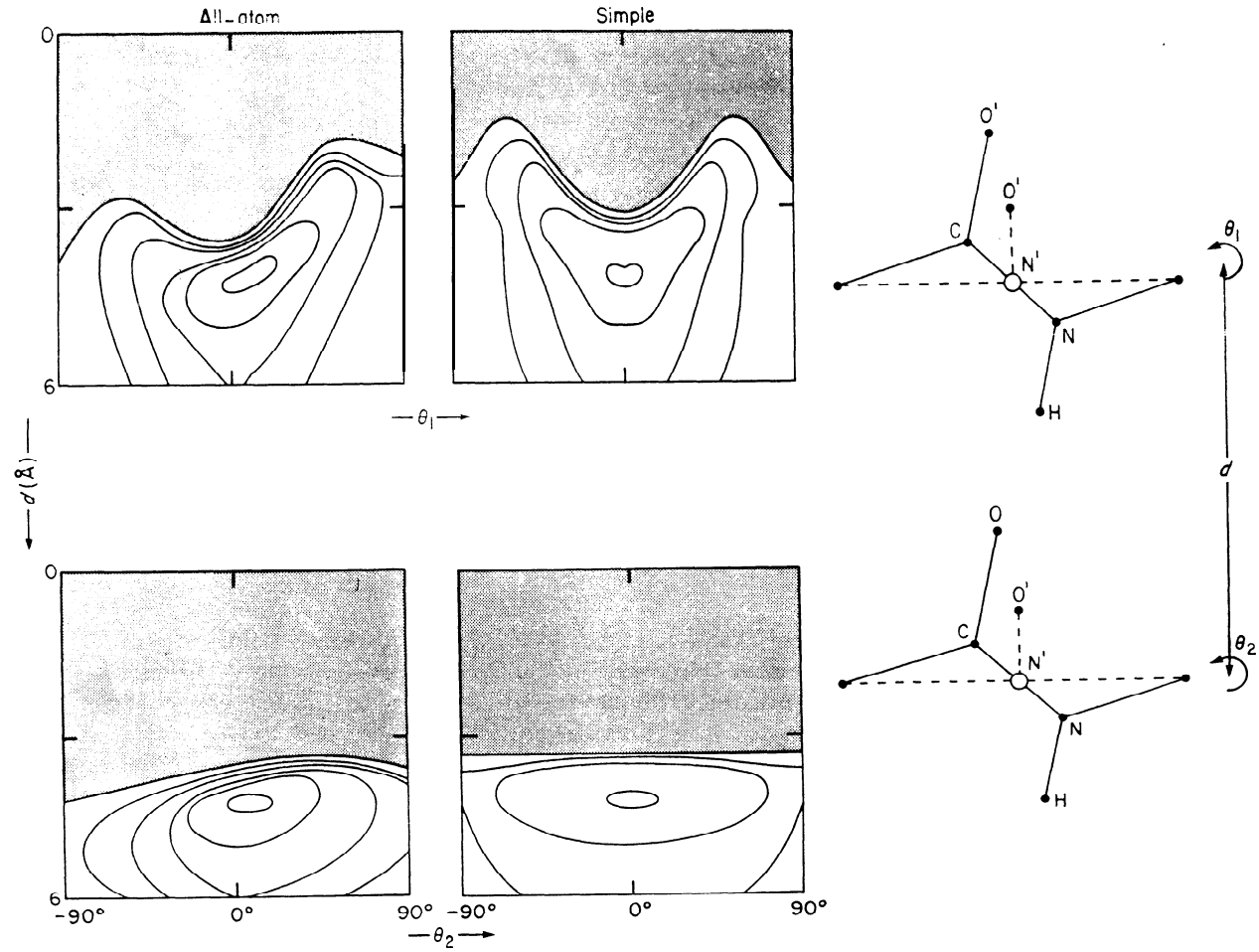


FIG. 4. Showing how the simplified peptide hydrogen bonds reproduce the properties of the conventional all-atom, peptide hydrogen bonds. The simplified peptide unit consists of only 2 effective atoms : N_i' midway between C^a_{i-1} and C^a_i; and O_i' displaced 1 Å from N_i' in a direction perpendicular to the plane formed by $C^a_{i-1}, C^a_i, C^a_{i+1}$, i.e.

$$\mathbf{r}(C_i') = \mathbf{r}(N_i') + \mathbf{U}_i/|\mathbf{U}_i|,\ \mathbf{U}_i = [\mathbf{r}(C^a_{i+1}) - \mathbf{r}(C^a_i)] \times [\mathbf{r}(C^a_{i-1}) \quad \mathbf{r}(C^a_i)]$$

The two left-hand contour maps show the variation of the all-atom peptide interaction energy as a function of d and θ_1, and d and θ_2, respectively. d is the separation of peptide centres, θ_1 is the rotation of the upper peptide about its centre, and θ_2 is the corresponding rotation of the lower peptide. Energy parameters used in the all-atom calculation were those that best fit the packing in amide crystals (Hagler et *al.*, 1974). The 2 right-hand contour maps show the corresponding variation of the energy of the simple peptide using energy parameters that give a good fit to the left-hand contours. These parameters are: (a) partial charges of 0·74 electrons on N' and — 0·74 on 0' (reproducing the magnitude and direction of the observed peptide dipole moment), and (b) van der Waals' interactions between the 2 pairs of N' and 0' atoms with $E = \epsilon_p \{(r_p/r)^{12} - 2(r/r)^6\}$ for $\epsilon_p = 0{\cdot}2$ kcal/mol and $r_p = 4{\cdot}6$ Å.

(iv) Near-neighbour non-bonded interactions

Near-neighbour non-bonded interactions between atoms that are separated by a few covalent bonds along the chain play an important role in restricting the allowed conformations of the polypeptide backbone. This is clearly illustrated by the good agreement between the allowed regions of the (ϕ, ψ) contact map of alanine dipeptide (Ramachandran *et al.*, **1963)** and the (ϕ, ψ) values found in known proteins. In the present representation, where the conventional backbone torsion angles ϕ and ψ are replaced by the α angle, energetically favourable chain conformations depend on the particular values of a.

The smallest possible peptide with the 4 adjacent C^{α} atoms needed to define a single α angle consists of a pair of linked amino acids (see Fig. 5). The non-bonded energy of this assembly of 3 peptide groups and 2 side chains will depend on the 4 backbone torsion angles ϕ_1, ψ_1, ϕ_2, and ψ_2 and any side chain torsion angles χ_1, χ_2, etc. Following the general formulation, α is taken as the most effective degree of freedom and the energy is averaged over all other degrees of freedom **at** a particular α value. In this way, the influence of near-neighbour side chain and backbone interactions is treated by the time-averaged potential

$$V_{\text{eff}}(\alpha) = \frac{\sum V(\theta) \exp\{-V(\theta)/kT\}\, J(\theta)d\theta}{\sum \exp\{-V(\theta)/kT\}\, J(\theta)d\theta},$$

where θ is the vector of single-bond torsion angles i.e. ϕ_1, ψ_1, ϕ_2, ψ_2, χ_1, χ_2, etc. $V(\theta)$ is the total non-bonded energy of the dipeptide at a particular θ, k is the Boltzman constant, T the absolute temperature, and $J(\theta)d\theta$ is the Cartesian space volume element corresponding to the torsion space volume element $d\theta$. ($J(0)$ is the Jacobian of x with respect to θ.) Many peptide conformations that have a particular α value are generated using all-atom standard geometry and selected values of the torsion angles θ. At each conformation the non-bonded energy is calculated with van der Waals' and atomic partial charge parameters derived from packing studies on 25 crystals (Levitt & Lifson, unpublished work).

As this calculation was time-consuming, it could not be done for all 400 combinations of the pair of side chains. Instead 6 peptides thought to be most representative were studied : Ala-Ala, Gly-Ala, Pro-Ala, Gly-Gly, Ala-Gly, and Ala-Pro. Fig. 5 shows the variation of $V_{\text{eff}}(\alpha)$ with α in each *case*. The curve for Gly-Ala is most like that of Ala-Ala, whereas the curve for Ala-Gly is most like that for Gly-Gly indicating how the shape of the potential depends most on the nature of the second side chain. A qualitative explanation for this dependence is as follows : a $\alpha \approx \phi_2 + \psi_1 + 180°$, with ψ_1 depending on the nature of the first side chain, and ϕ_2 on that of the second side chain; but the side chain has more influence on ϕ than on ψ due to steric hinderance with the oxygen of the peptide preceding the side chain.

TABLE 5

Fourier coefficients of torsional potential

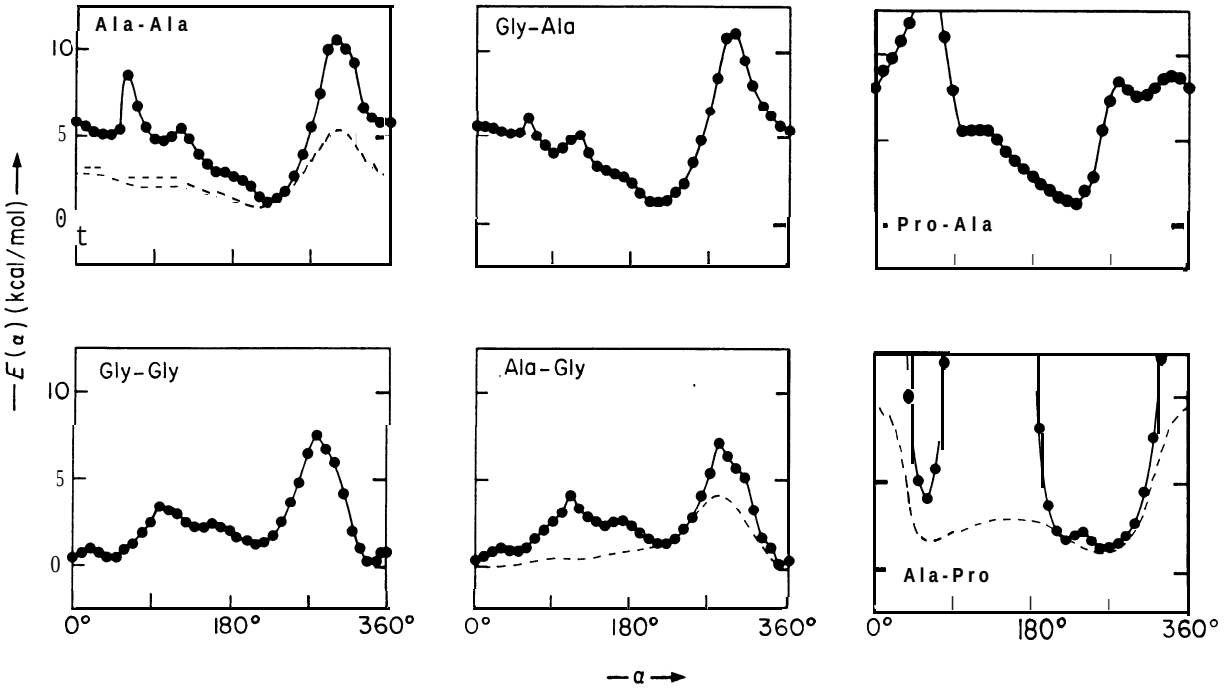
k	Class of potential? Ala		Gly		Pro	
	A_k	B_k	A_k	B_k	A_k	B_k
1	0·505	0·0	0·548	0-0	1·276	0-0
2	0·403	−0·367	0·022	−0·637	1·518	−0·260
3	−0·178	−0·429	−0·440	−0·227	1·066	−0·728
4	−0·225	−0·014	−0·238	0·222	0·234	−0·208
5	−0·067	0·140	0·083	0·144	−0·054	0·076
6	0·018	O-090	0·049	−0·046	-09082	0·166

The α torsion potential in kcal/mol is expressed as a Fourier series given by

$$V(\alpha) = K_a \sum_{k=1}^{6} \{A, \cos(k\alpha - a) + B_k \sin(k\alpha - a)\}$$

where K_a is a scale factor normally taken as 2.

† The 3 classes of torsion potential are referred to as Ala-like, Gly-like, and Pro-like.



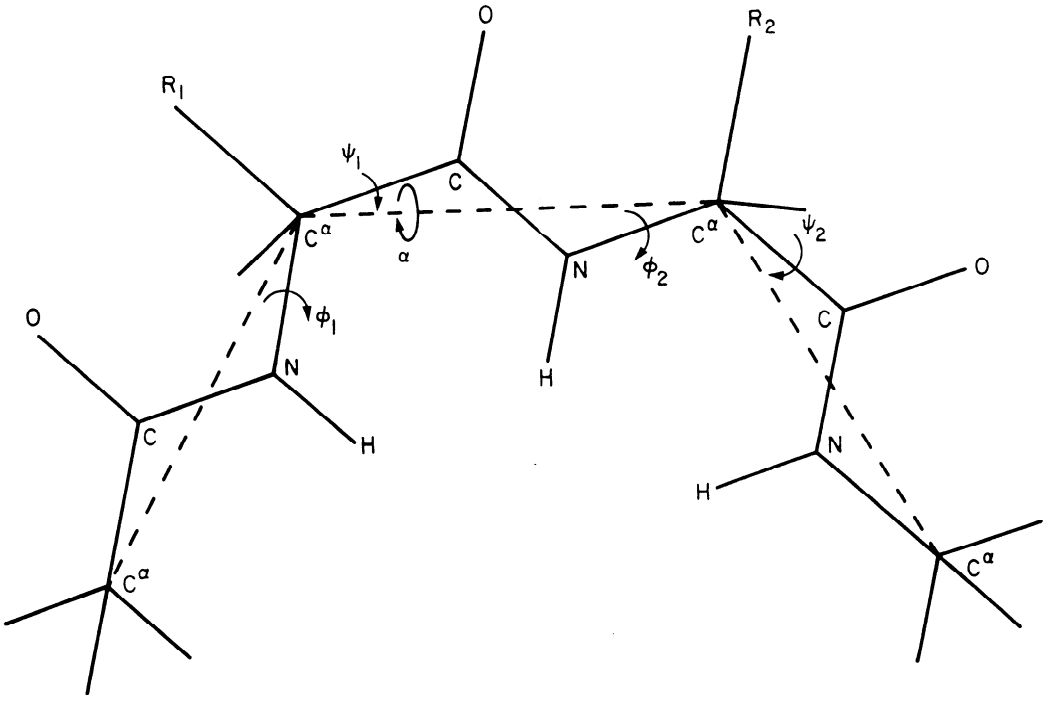


FIG. 5. Showing the variation of the all-atom non-bonded energy with torsion angle α for the following pairs of linked amino acids: Ala-Ala, Ala-Gly, Gly-Ala, Gly-Gly, Ala-Pro, Pro-Ala. Each point on the curve is the Boltzmann weighted average of the non-bonded energy for all those conformations that have an α angle value within 5° of the value shown. The sharp peaks on these plots would be smoothed out if more conformations had been chosen at each α interval, and the higher energy **values would be reduced if bond** angles had been allowed to relax at each conformation. The actual torsional potentials derived from these data take account of these corrections and are drawn as dashed lines on the plots for Ala-Ala, Ala-Gly, and Ala-Pro.

Three types of $V_{eff}(\alpha)$ potential were used in the actual simplified energy calculations. Analytical forms for the Ala-like, Gly-like, and Pro-like potentials were based on the corresponding curves in Fig. 4, after some of the sharp peaks had been smoothed out (see Table 5). The Gly-like potential differs from the other two in that the region around $a = 0°$ is of low energy. Because aspartic acid and asparagine were found to occur almost as frequently in reverse turns as glycine for the proteins analysed, the same potential was used for all three†. The Ala-like potential was used for all other amino acids except proline. In all cases the type of potential used depended only on the nature of the residue at position $i + 1$. Venkatachalam (1968) considered a similar system of three linked peptide units but confined his attention to those allowed conformations that had at least one hydrogen bond.

(v) The complete energy function

Adding together the different energy contributions described above gives the complete energy function of the simplified protein. For convenience, this function is presented in full together with some of the finer details of the energy calculation:

$$\begin{aligned} V_{\text{tot}}(\alpha) = & \sum_{i,j} \epsilon_{ij} \{3(r^{\circ}_{ij}/r_{ij})^8 - 4(r^{\circ}_{ij}/r_{ij})^6\} \\ & + \sum_{\substack{i,j \\ r_{ij}<9\text{Å}}} (s_i + s_j)\, g(r_{ij}) + \sum_{\text{SS bonds}} K_{\text{ss}}(r^{\text{ss}}_{ij} - r^{\text{ss}}_{\circ})^2 \\ & + \sum_{i,j} \epsilon_{\text{p}} \{r^{\circ}_{\text{p}}/r_{\text{NO}})^{12} - 2(r^{\circ}_{\text{p}}/r_{\text{NO}})^6 + (r^{\circ}_{\text{p}}/r_{\text{ON}})^{12} - 2(r^{\circ}_{\text{p}}/r_{\text{ON}})^6\} \\ & + 332 \sum_{i,j} q^2_{\text{p}} \{1/r_{\text{NN}} + 1/r_{\text{NO}} - 1/r_{\text{NO}} - 1/r_{\text{ON}}\} \\ & + \sum_i (2 \sum_{k=1}^{6} A^i_k \cos[(k-1)\alpha_i] + B^i_k \sin[(k-1)\alpha_i]\} \end{aligned}$$

The first term, which is the side chain van der Waals' energy, is summed between all pairs of side chain centroids i and j ($j > i + 1$), except for pairs of half-cystines. The centroid separation is r_{ij}, and the energy parameters ϵ and r° are taken from Tables 3 or 4, depending on the set of van der Waals' parameters used.

The second term, which is the side chain solvent interaction energy, is summed between all side chain centroids i and j ($j > i$) less than 9 Å apart. The parameters s_i and s_j for the different types of side chain are given in Table 2; the sigmoid function g(x) (where $x = r_{ij}/9$) is

$$1 - \tfrac{1}{2}\{7x^2 - 9x^3 + 5x^6 - x^8\}.$$

The third term, which is the S-S bond energy, is summed over pairs of half-cystines i and j that are considered to be bonded together. These side chains, r^{ss}_{ij} apart, are forced to be $r^{\text{ss}}_{\circ}$ apart ($r^{\text{ss}}_{\circ} = 4{\cdot}2$ A) through the force constant K_{ss} (taken as 10 kcal/mol per Å^2). This interaction is often omitted as prior knowledge of the correct S-S is needed.

The fourth and fifth terms, which together give the peptide hydrogen bonds, are summed over all peptide groups i and j ($j > i + 2$). The distances $r_{\text{NN}}, r_{\text{OO}}, r_{\text{NO}}$, and r_{ON} are between N . . N, O . . O, N . . O, and O . . N, respectively. Suitable values for the parameters ϵ_{p}, r_{p} and q_{p} are given in the legend to Fig. 4. No hydrogen bond is allowed for the peptide preceding proline.

The sixth and final term, which is the effective short range non-bonded energy expressed as a function of a, is summed over all α_i. The Fourier coefficients A^i_k and B^i_k are given in Table 5.

(c) Changing the conformation

A rigorous way to simulate the conformational changes of the simplified protein model is to solve the equations of molecular dynamics. In this approach the conformation of the molecule at time $t + \Delta t$ depends on its conformation, velocity, and the operative forces at a slightly earlier time, t. With a medium as viscous as water, inertial forces are damped out verv rapidly (in less than 10^{-13} s) and the frictional forces, which depend on the velocity'; always tend to balance the applied forces giving

$$c\frac{da}{dt}(t) = F(\alpha(t)) + R(t),$$

† Note added in proof.

When Ptitsyn & Rashin (1973, 1974) studied the refolding of the mainly a-helical protein myoglobin they abandoned computerized energy calculations in favour of a physical model consisting of nine preformed a-helices connected by flexible pieces of chain. They manually searched for possible packings of the helices after making some assumptions about neighbouring helices along the chain coming together first, and estimated the helix interaction energy by counting which residues were taken out of water at the helix contact surface. This approach is exciting in its originality and simplicity but suffers from the drawbacks inherent in manual searches and rough estimations of energy.

In this paper I describe a highly simplified representation of protein conformation that reduces both the number of effective atoms and number of degrees of freedom. The basis of this simplification (Levitt & Warshel, **1975)** is to average over the finer details and to consider only those degrees of freedom that have the greatest effect on the conformation. Averaging like this saves having to distinguish between conformations that differ only in the finer details. It also simplifies the inclusion of solvent and atomic thermal motion effects, and makes the energy calculation less sensitive to the exact energy parameters. The methods and energy parameters needed to define the new representation are presented in some detail. Test calculations are done on a small protein, bovine pancreatic trypsin inhibitor? starting at both the native folded conformation and at the much more open denatured conformations. With this protein, the calculation simulates the folding rather well and arrives at native-like structures under a variety of conditions. In such a computer simulation one can study each step in the refolding process to find the roles played by local interactions, different energy contributions, and intermediate partially folded conformations.

2. Materials and Methods

(a) *Simplified geometry*

As a first step towards the simplification of the protein structure, groups of atoms are combined into single effective atoms (interaction centres). The peptide group, which forms the backbone repeating unit, is simplified by combining the C, N, and H atoms into an effective N' atom and replacing the 0 by an effective 0' atom (see Fig. 1). The side chain is simplified by treating it as a single effective atom located at the centroid of the side chain atom positions (the C^α is treated as part of the side chain).

In conventional representations of polypeptides, the backbone degrees of freedom are the ϕ and ψ torsion angles (dihedral angles) about the N—C^α and C^α—C bonds, respectively. As the peptide group is planar and the two bonds C^α—C and N—C^α are parallel to within 10°, the path of the adjacent C^α atoms depends mainly on $\phi_{i+1} + \psi_i$, while the orientation of the peptide group between C^α atoms depends mainly on $\phi_{i+1} - \psi_i$ (Diamond, 1965). One can, therefore, replace these 2 degrees of freedom by one variable related to $\phi_{i+1} + \psi_i$ and still retain most of the chain's conformational flexibility. Here a new backbone degree of freedom is used: the torsion angle a defined by the positions of the 4 adjacent C^α atoms of residues $i-1$, i, $i+1$, and $i+2$. A simple formula giving the approximate relationship between α and the conventional ϕ and ψ angles is

$$\alpha_i = 180° + \phi_{i+1} + \psi_i + 20°(\sin\phi_i + \sin\psi_{i+1}).$$

As *a* depends on the positions of 4 successive C^α atoms and consequently on 2 pairs of ϕ and ψ values, there is no unique *a* value corresponding to a particular (ϕ, ψ) pair, even for repeating structures with $\phi_{i+1} = \phi_i$, $\psi_{i+1} = \psi_i$. Nevertheless, the most populated regions of the (ϕ, ψ) map, the right-handed a-helix and the β-sheet, correspond to well separated *a* values of 45° and 210°, respectively **(see Fig. 2(a)). The more irregular chain**

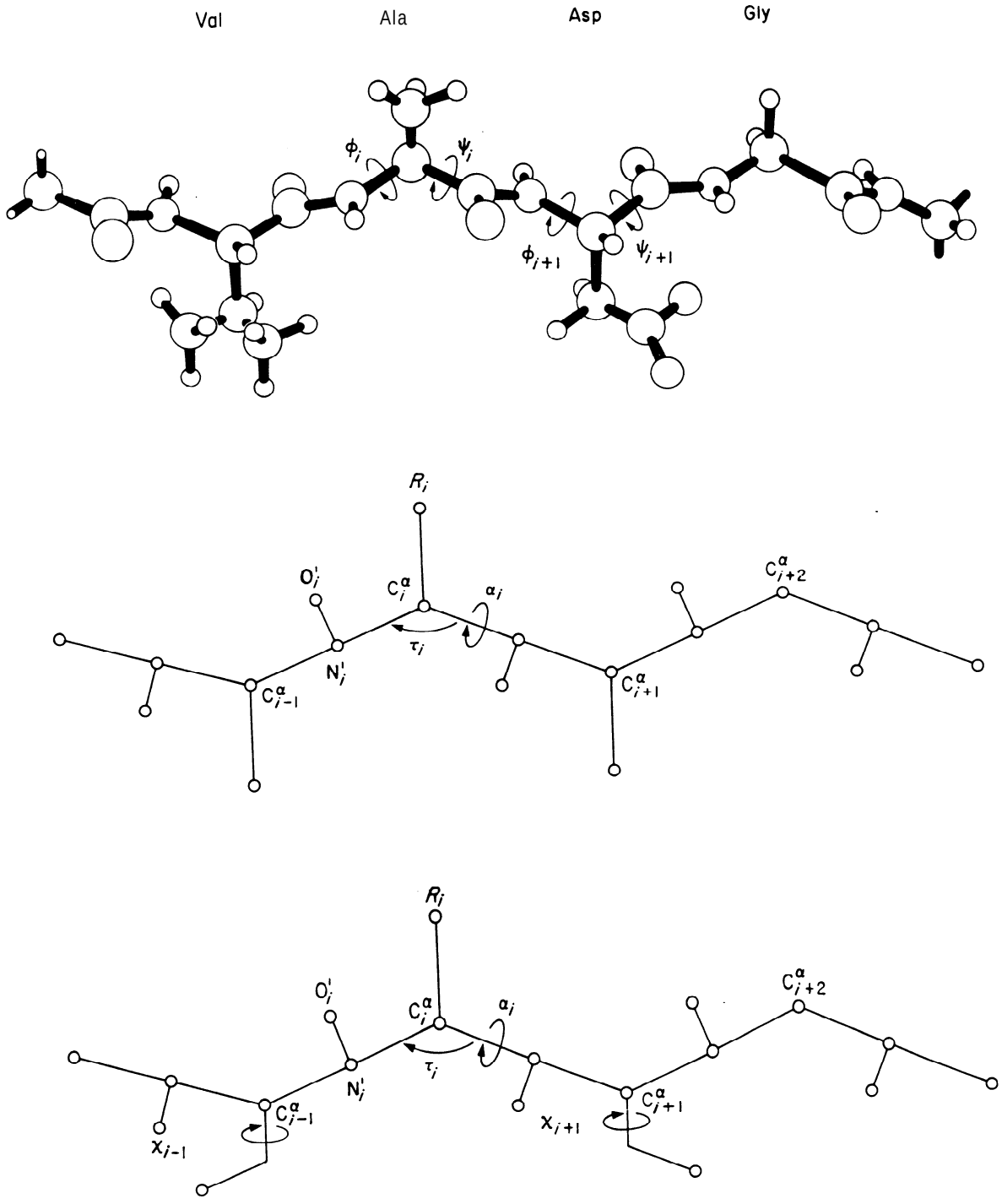


FIG. 1. The relationship between the simplified models used here and the conventional all-atom polypeptide chain. The effective atoms or interaction centres of the simplified model include: (a) the 2 atoms (0' and N') of the peptide, and (b) the centroid of the side chain. In the first model, each residue has only one degree of freedom, the torsion angle α defined by the positions of 4 adjacent C^α atoms belonging to residues $i - 1$, i, $i + 1$, and $i + 2$. (For the first residue α is defined with respect to N' instead of C^α, and there is no a for the last 2 residues.) a is taken as zero for the *cis* conformation, and it increases with a clockwise rotation of the bond furthest from the observer relative to the bond nearest to him, i.e. in accordance with the usual convention. In the second more complicated representation, certain side chains have an additional degree of freedcm, the torsion angle χ defined by C^α_{i-1}, C^α_i, C^β_i and the side chain centroid, R_i. The second model has no more interaction centres than the first as both the C^α and C^β atoms are used only to define axes of rotation.

conformations, in which successive residues have very different (ϕ, ψ) values, will be represented less well, but the important reverse turn conformation for which α is close to 0° will still be clearly defined. Although a simple representation based on virtual bonds has been used before to study polypeptide random-coil conformations (Flory, 1969), it has never been applied to ordered globular proteins.

As the rotation axes of ϕ_{i+1} and ψ_i are parallel but not co-linear, the bond angle between C^{α}_{i-1}, C^{α}_{i} and C^{α}_{i+1}, known here as τ_i, changes as α_i changes (see Fig. 2(b)). Rather than treat the τ_i as extra independent variables, each τ_i is made to depend functionally on α_i using

$$\tau_i = 106° + 13° \cos(\alpha_i - 45°).$$

By this device, the representation of backbone geometry is greatly improved, especially for the a-helix, while still retaining the simplicity of the model.

For the majority of residues in globular proteins, ϕ is within 30° of $-90°$ (Fig. 2(a)), so that the plane of the peptide group may be taken as fixed to a first approximation (which is very closely correct for a-helices and β-sheets). In the present work this is achieved by making the N′—0′ bond of the simplified peptide perpendicular to the C^{α}_{i-1}—C^{α}_{i}—C^{α}_{i+1} plane.

In conventional treatments, the side chain degrees of freedom are the torsion angles about single bonds (known collectively as χ angles). In the present simplified representation 2 different models of side chains are introduced. In the first model, the side chain is completely rigid and sticks out from the backbone. In the second model, certain side chains have a single degree of freedom equivalent to the conventional χ_1 about the C^{α}—C^{β} bond. With the second representation. many side chains are almost as flexible as in conventional treatments.

Tables 1 and 2 give the "standard" bond lengths, bond angles, and torsion angles used to construct the simplified protein chain in these 2 representations. The values given are averages taken from the atomic co-ordinates of 13 proteins determined by X-ray crystallography. The Cartesian co-ordinates of the simplified structure of any sequence of amino acids at specified α values are calculated from these standard internal co-ordinates using the well known matrix transformations (Eyring, 1932).

(b) *The molecular force field*

In the previous section the geometry of the simplified model was defined, here a suitable molecular energy function must be found. In conventional conformational analysis the energy is calculated as a sum of pairwise interactions between all non-bonded atoms with additional terms to allow for bond stretching, bond -angle bending, and bond twisting energy. In the simplified models used here, bonds and bond angles are kept fixed at their initial standard values, and the non-bonded interaction between whole groups of atoms is treated by an effective potential. As a basic formalism, atoms in the particular groups under consideration are assumed to be in constant thermal motion, and the effective potential is taken as the sum of all the individual pairwise interactions between the groups averaged over a long time. Rather than simulate the thermal motion of the atoms in each group and calculate an effective time-averaged potential, the potential is averaged over all positions in space occupied by the moving atoms. By the egodic theorem (Hill, 1956), a spatial average weighted by the exponential Boltzmann factor is equivalent to a time average. When this formalism is applied to the derivation of various energy parameters further simplifications are introduced whenever necessary for computational efficiency.

(i) *van der Waals' interactions between side chains*

In this section the average van der Waals' forces between side chains are considered; the backbone van der Waals' forces will be considered in the discussion of peptide interactions. The most effective degrees of freedom of a side chain are the single bond torsion angles χ. Were one to consider side chain thermal motion to be rotation about these torsion angles, the effective time-averaged potential would depend on both the orientation and separation of the interacting side chains. So complicated a potential is not justified

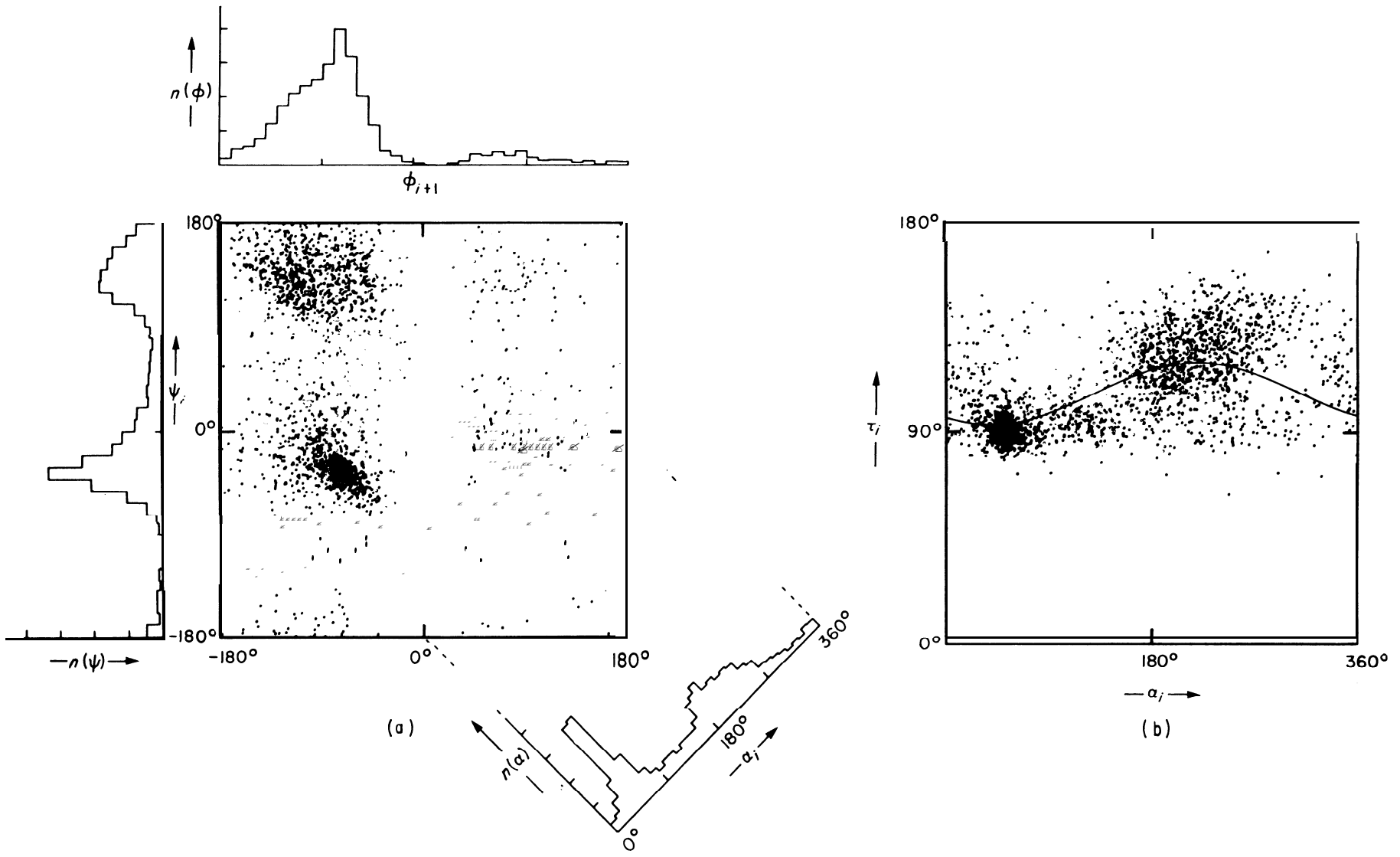


FIG. 2. (a) Showing the distribution of ϕ and ψ values found in the 13 proteins analysed here. Cases with glycine as residue $i + 1$ have been left out for greater clarity. As the new torsion angle α is given approximately by $\alpha_i = \phi_i + {}_1 + \psi_i + 180°$, it is almost constant along lines running from top left to bottom right on the plot, and it increases fastest along lines running from bottom left to top right. The 3 main regions of the (ϕ,ψ) map, the right-handed α-helix, the extended chain, and the left-handed or-helix, overlap on the ϕ and ψ histograms but have separated α values at $\alpha = 45°$, $\alpha = 210°$, and $\alpha = 3\ 1\ 5°$, respectively.

(b) The distribution of τ_i and α_i values from the proteins analysed (τ_i is the bond angle between 3 adjacent C^α atoms). One would expect τ_i to depend on both ϕ_i and ψ_i but as ϕ is relatively constant being centred about -- 90°, τ_i depends mainly on ψ_i and hence α_i. The solid curve, judged by eye to give this dependence, is given by $\tau_i = 106"\ \ 13" \cos(\alpha_i - 45°)$.

TABLE 1

Standard geometry† for rigid side chains

Amino acid	b_{AR} (Å)	θ_{AAR} (deg.)	ϕ_{AAAR} (deg.)	Radius of gyration (Å)
Ala	0·77	121·9	243·2	0·77
Val	1·49	121·7	220·3	1·29
Leu	2·08	118·1	205·6	1·54
Ile	1·83	118·9	217·9	1·56
Cys	1·38	113·7	209·4	1·22
Met	2·34	113·1	204·0	1·80
Pro	1·42	81·9	237·4	1·25
Phe	2·97	118·2	203·7	1·90
Tyr	3·36	110·0	195·6	2·13
Trp	3·58	118·4	203·7	2·21
Asp	1·99	121·2	215·0	1·43
Asn	1·98	11795	207·1	1·45
Glu	2·63	118·2	213·6	1·77
Gln	2·58	118·0	205·4	1·75
His	2·76	118·2	219·9	1·78
Ser	1·28	117·9	232·0	1·08
Thr	1·43	117·1	226·7	1·24
Arg	3·72	121·4	206·6	2·38
Lys	2·94	122·0	210·9	2·08
Gly	0·0	—	—	—

† Two additional parameters, the mean inter-C^α bond length and mean C^α bond angle, are needed to build the Cartesian co-ordinates of any protein from the standard geometry: $b_{AA} = 3{\cdot}808$ Å and $\tau_{AAA} = 106{\cdot}3°$ were used. The subscripts A and R refer to the C^α and side chain centroid, respectively. τ_{AAR} is defined as the angle between the 3 atoms C^α_{i-1}, C^α_i and R_i. ϕ_{AAAR} is defined as the torsion angle between the 4 atoms C^α_{i+1}, C^α_{i-1}, C^α_i and R_i. The values given are averages taken from 13 well-refined protein conformations : carboxypeptidase A (Quiocho & Lipscomb, 1971), concanavalin A (Edelman et *al.*, 1972; Hardman & Ainsworth, 1972), α-chymotrypsin (Birktoft & Blow, 1972), high potential iron protein (Carter et *al.;* 1974), insulin (Adams et al., 1969), lysozyme (Blake et *al.*, 1967), myogen (Kretsinger & Nuckolds, 1973), papain (Drenth et al., 1971), ribonuclease S (Wykoff et *al.*, 1970), rubredoxin (Watenpaugh et *al.*, 1973), staphylococcal nuclease (Arnone et *al.*, 1971), subtilisin BPN' (Wright et *al.*, 1969), and thermolysin (Colman et *al.* 1972).

in the first attempt at protein energy function simplification. Instead all the atoms of a side chain were assumed to be spread out uniformly by thermal motion into a sphere centred at the side chain centroid and of radius equal to the average radius of gyration of the particular group (see Table 1).

The effective potential (V_{eff}) between 2 identical side chains was then calculated at various distances apart using

$$V_{eff}(R) = \sum_{i,j} \int_{\substack{\text{Spheres} \\ 1 \text{ and } 2}} \{A/r_{ij}^9 - B/r_{ij}^6\} \, dv_1 dv_2,$$

where the pairwise atomic van der Waals' potential is integrated over distances r_{ij} of atom i anywhere in one sphere and atom j anywhere in the other sphere, when the inter-sphere distance is R. The energy parameters A and B were derived from 25 hydrocarbon, amide, and amino acid crystals (Levitt & Lifson, unpublished work).

Such van der Waals' potentials were calculated for interactions of the 9 pairs of non-polar side chains. A disturbing feature of these potentials was the steep variation of the repulsive part, approximately like $1/R^{16}$. So steep a potential is due to the uniform spherical distribution of all the atoms of the group. In reality, those atoms that make

TABLE 2

Standard geometry? for flexible side chains

Amino acid‡	b_{BR} (Å)	ϕ_{ABR} (deg.)	χ_{AABR} (deg.)
Val	0·33	74	250
Leu	1·01	103	285
Ile	0·59	115	285
Cys	0·60	66	310
Met	1·31	116	290
Phe	2·12	108	290
Tyr	2·53	109	310
Trp	2·70	112	290
Asp	1·09	95	300
Asn	1·09	95	300
Glu	1·47	122	285
Gln	1·50	122	300
His	1·87	106	300
Ser	0·57	52	70
Thr	0·36	70	240
Arg	2·57	130	270
Lys	1·84	123	250

† b_{BR} is the distance between the C^{β} atom (B) and the side chain centroid (R); ϕ_{ABR} is the bond angle between atoms C_i^{α}, C_i^{β} and R_i; and χ_{AABR} is the torsion angle between atoms C_{i-1}^{α}, C_i^{α}, C_i and R_i. Other necessary parameters, which are the same for all amino acids, are b_{AB} = 1·54 A, (between C_i^{α} and C_i^{β}), ϕ_{AAB} = 120" (between C_{i-1}^{α}, C_i^{α} and C_i^{β}), ϕ_{AAAB} = 240" (between C_{i+1}^{α} C_{i-1}^{α}, C_i^{α} and C_i^{β}). The main chain geometry is as defined in Table 1.

‡ Rigid side chain geometry (Table 1) is used for amino acids not included here.

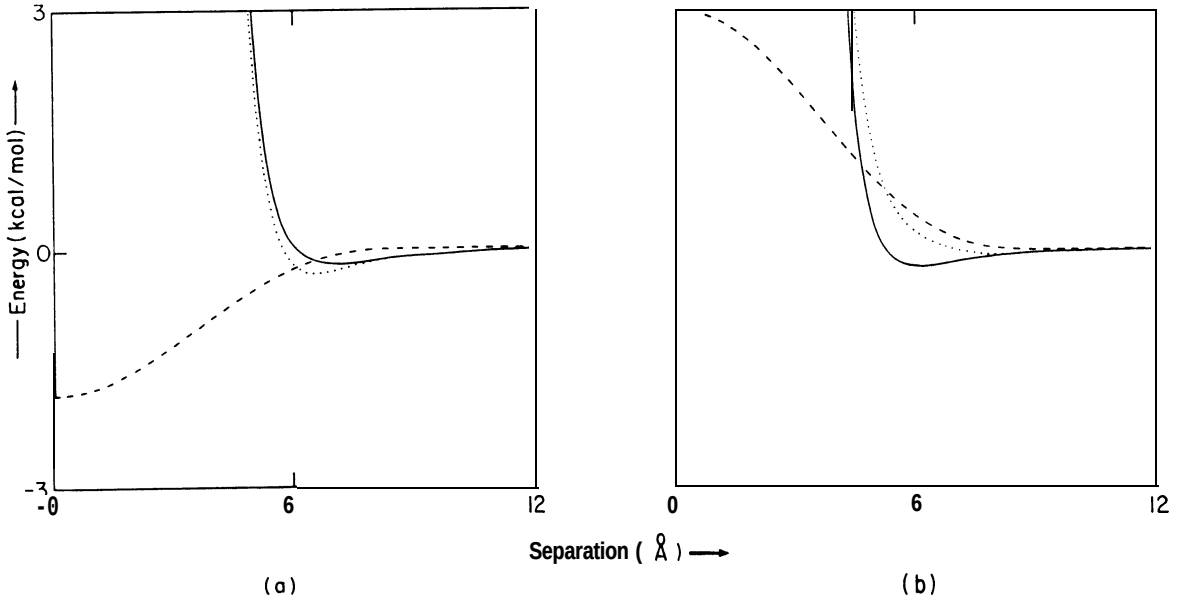


FIG. 3. The shape of typical side chain interaction potentials between (a) a pair of leucine side chains, and (b) a pair of lysine side chains. The dotted curve is the total energy as a function of separation, the solid curve is the van der Waals' energy, and the broken curve is the solvent energy. The analytical functions are given in the text, and the parameters are taken from Table 3.

the closest contacts when the 2 groups come together would move so as to reduce the strong repulsive potential. To allow for this an 8–6 Lennard–Jones type potential was used (see Fig. 3), which is less steeply repulsive than the more conventional 12–6 potential and has the following form

$$\epsilon\{3(r^0/r)^8 - 4(r^0/r)^6\}.$$

r^0 and ϵ are the position and depth on the minimum of the plot of $V_{nn}(R)$ against R. Table 3 gives the r^0 and ϵ values derived in this way for the 9 non-polar amino acids. Parameters for the polar amino acids were roughly estimated allowing for the more extended shapes and greater **flexibility of the side chains, which together weaken the van der** Waals' **interactions. The energy between a pair of non-identical side chains of types i and j is calculated using $\epsilon_{ij} = (\epsilon_i\epsilon_j)^{\frac{1}{2}}$ and $r^\circ_{ij} = (r^\circ_i r^\circ_j)^{\frac{1}{2}}$.**

Sometime **after deriving and using these energy parameters known here as set A, two**

TABLE *3*

van der Waals' (set A) and solvent parameters

Amino acid side chain	r° (Å)	ϵ (kcal/mol)	s† (kcal/mol)	‡ Number of atoms	§ Degrees of freedom	Atoms per /I degree of freedom
Ala	5·2	0·025	−0·5	2	0	–
Val	6·4	0·15	−1·5	4	½	8
Leu	7·0	0·19	1·8	5	1	5
Ile	7·0	0·19	1·8	5	1	5
Cys	6·1	0·10	−1·0	3	½	6
Met	6·8	0·19	−1·3	5	3½	1½
Pro	6·2	0·17	−1·4	4	0	–
Phe	7·1	0·39	−2·5	8	1½	5
Tyr	7·1	0·39	−2·3	9	1½	6
Trp	7·6	0·56	−3·4	11	1½	7
Asp	5·0	0·10	2·5	5	1½	3
Asn	5·0	0·10	0·2	5	1½	3
Glu	6·0	0·10	2·5	6	2½	2
Gln	6·0	0·10	0·2	6	2½	2
His	6·0	0·10	−0·5	7	1½	5
Ser	4·9	0·025	0·3	3	0	–
Thr	5·0	0·10	−0·4	4	½	8
Arg	6·0	0·20	3·0	8	3½	2
Lys	6·0	0·20	3·0	6	3½	2
Gly	4·2	0·025	0·0	1	0	–

† The **values of the hydrophobic parameters** s_i are taken as the measured free energy of transfer from water to ethanol (Nozaki & Tanford, 197 1). When experimental values were not available they were roughly estimated from the relationship between accessible surface area (Lee & Richards, 1971) and hydrophobicity (Chothia, 1974). The affinity of the different polar groups for water by virtue of the hydrogen bonding groups (ΔF_{polar}) were : - O H = 1 kcal/mol; – $CONH_2$ = 2 kcal/mol ; – COO = 4·5 kcal/mol; $-NHC(NH_2)_2^+$ = 6 kcal/mol; – NH_3^+ = 5. In fact $s_i = -0{\cdot}025 \times A_{acc} + F_{polar}$ kcal/mol, where A_{acc} is the accessible area of the residue in Å².

‡ Only count non-hydrogen atoms but include C^α as part of the side chain.

§ Counting single bond torsion angles except those that merely rotate a methyl group at the end of the side chain. The torsion angle about a bond connected to a tetrahedral carbon with 2 non-hydrogen substituents e.g. χ_1 in Val, is given ½ degree of freedom as rotation about this bond is more restricted than normal.

‖ The number of non-hydrogen atoms in the side chain per degree of freedom gives some idea of the effectiveness of non-bonded interactions to that side chain. With more degrees of freedom, the chain will move more under thermal influences and be able to avoid good binding interactions. All the non-polar side chains except Met have a ratio of 5 or more, whereas all the polar side chains apart from His and Thr have a ratio of 3 or less.

other sets were derived in a more consistent way for all side chains (see Table 4). One set (B) depended on the number of non-hydrogen atoms (n) in the side chain alone, whereas the other set (C) depended on the number of such atoms in the residue as a whole. Values of ϵ were calculated from the relationship $\epsilon = 0{\cdot}6n - 0{\cdot}08$ kcal/mol (obtained from a straight line fit to a plot of the non-polar ϵ values given in Table 3 against the particular n values).

TABLE 4

Sets B and C van der Waals' parameters

Amino acid	Set B Whole residue r^{o}† (A)	Set B Whole residue ϵ‡ (kcal/mol)	set c Side chain only r^{o}† (Å)	set c Side chain only ϵ‡ (kcal/mol)
Ala	5·5	0·21	4·6	0·05
Val	6·5	0·33	5·8	0·16
Leu	6·9	0·39	6·3	0·21
Ile	6·9	0·39	6·2	0·21
Cys	5·9	0·27	5·0	0·10
Met	6·8	0·39	6·2	0·21
Pro	6·3	0·33	5·6	0·16
Phe	7·35	0·56	6·8	0·39
Tyr	7·4	0·62	6·9	0·45
Trp	7·7	0·73	7·2	0·56
Asp	6·3	0·39	5·6	0·21
Asn	6·4	0·39	5·7	0·21
Glu	6·7	0·45	6·1	0·27
Gln	6·7	0·45	6·1	0·27
His	6·8	0·51	6·2	0·33
Ser	5·7	0·27	4·8	0·10
Thr	6·3	0·33	5·6	0·16
Arg§	7·35	0·56	6·8	0·39
Lys	6·9	0·45	6·3	0·27
Gly	5·0	0·16	3·8	0·025

† r^{o} is calculated as $(6V/\pi)^{\frac{1}{3}}$, where V is the residue or side chain volume found in known protein conformations (Chothia, 1975). The side chain volumes were taken as 38 Å^3 less than the whole residue volumes, allowing for the peptide volume.

‡ ϵ is derived from the number of heavy atoms (non-hydrogen) in the side chain or residue using the equation given in the text.

§ Chothia (1975) gives no Arg volume, so its r^{o} and ϵ values are taken to be like those of Phe which has the same number of heavy atoms as Arg.

Values of r° could have been found in the same way, Instead they were taken from the average amino acid packing volumes found in globular proteins (Richards, 1974; Chothia, 1975). Each r° was taken as twice the radius of a sphere (r_v) whose volume equals that of the side chain in proteins. This can be justified as follows. Were hard spheres representing each amino acid to pack into the volume of a protein, the volume of each sphere would be about 25% less than the side chain volume calculated by Richard's methods due to the spaces between the spheres (the radius would be about 10% less). In the present calculation, the side chains are soft spheres and when they interact inside a globular protein, the final space-filling equilibrium separation is less than r° due to the compressive effect

where F(a(t)) **is** the vector of forces acting on the system defined by the variables $a(t)$, c is a constant relating the frictional forces to the velocity, and $R(t)$ is a random perturbing force due to interactions with the thermal environment. Now if $a(t)$ changes to $a(t) + \Delta\alpha(t)$ in a small time interval Δt, $d\alpha(t)/dt = da(t)/\&$, **and**

$$da(t) = -\frac{\Delta t}{c}\frac{\partial V}{\partial \alpha}(\alpha) + \frac{\Delta t}{c}R(t),$$

where $V(\alpha)$ is the potential energy of the molecule. This equation has been used by Simon (1971), who describes how suitable values of ***At,*** c, and ***R(t)*** can be estimated. For the present discussion it is sufficient to notice that the changes in co-ordinates suggested above ($\Delta\alpha(t)$) are directed down the energy gradient with additional random perturbations due to thermal motion. At sufficiently low temperature, ***R(t)*** is negligible, and $\Delta\alpha(t)$ becomes simply the change of variables that would be obtained when minimizing the energy by the well known method of steepest descent. As this method is one of the least efficient minimization methods, which never really converges to the minimum, molecular dynamics based on the previous equation will not work very rapidly. At room temperature the continual random perturbations will only make matters worse.

As rapid movement towards a stable conformation is essential if one is to study protein folding, it was decided to simulate the dynamics of the molecule more efficiently. First it is assumed, as above, that all momentum was indeed completely damped by the viscous drag of the water. Next it is assumed that the random perturbation could be omitted during the motion towards a stable conformation and then re-introduced at **that con**formation. With these assumptions, the molecule will simply move down the potential energy surface until it reaches the nearest accessible energy minimum.

Here the very sophisticated minimization routine VA09D (Fletcher, 1970), taken from the Harwell Subroutine Library, is used to minimize the energy as rapidly as possible. This method is the most powerful of the variable metric minimization methods, which use analytical first derivatives to build up gradually an approximation to the energy second derivative matrix. Some of the advantages of this particular routine include: (a) efficiency, in that each cycle requires less than 2 energy and derivative evaluations, often averaging as few as 1·2 evaluations per cycle ; and (b) stability, in that the metric **matrix** is factorized to ensure that it remains positive definite. In some tests VA09D succeeded **in minimizing** the energy with respect to 56 variables to a precision of 10^{-6} radians after less than 80 function evaluations. On the other hand, when there was no easily accessible minimum, the method did not stop in the nearest valley but always continued until it found a, true minimum where all the first derivatives were zero. (In practice, minimization stopped when the root-mean-square gradient was less than 10^{-6} kcal/mol per radian root-mean-square.)

Such **efficient minimization** depends on having the first derivative of the energy in analytical form. As most terms in the complete energy function depend on the separation cf 2 atoms, it is relatively easy to collect the vector of Cartesian co-ordinate derivatives $\partial V/\partial \mathbf{r}k$ during the energy calculation. Once all contributions to this derivative have been summed, the required torsion angle derivatives $\partial V/\partial\alpha_i$ are obtained as

$$\frac{\partial V}{\partial \alpha_i} = \sum_k \frac{\partial V}{\partial r_k}[\mathbf{n}_i \times (\mathbf{r}_k - \mathbf{r}_i)],$$

where $\mathbf{n}_i$ is the unit vector directed along the rotation axis of angle α_i that begins at the Cartesian position $\mathbf{r}_i$. The derivative of those energy contributions that depend explicitly on ***a*** are obtained by direct differentiation. Special attention must be given to the derivatives with respect to the dependent variables: the backbone bond angles τ_i are functionally related to the backbone torsion angles α_i. In this case, there are additional contributions to $\partial V/\partial\alpha_i$ given by

$$\frac{\partial V}{\partial \tau_i}\cdot\frac{\partial \tau_i}{\partial \alpha_i},$$

where $\partial V/\partial\tau_i$ is obtained from $\partial V/\partial \mathbf{r}_k$ in the same ways as $\partial V/\partial\alpha_i$.

(i) *Normal-mode thermalization*

After reaching a stable conformation at an energy minimum, the interaction with the thermal environment is re-introduced (see Fig. 6). Although there is a single minimum energy conformation **at** 0°K (ignoring the zero point energy), the conformation vibrates about the minimum at higher temperatures. Each mode of vibration has on average $\frac{1}{2}(kT)$ of potential energy above that at the minimum. To a first approximation, the energy near the minimum varies quadratically

$$V(\alpha) = V_0 + \tfrac{1}{2} \sum_{i,j} \delta^2 V/\delta\alpha_i \delta\alpha. A\ \alpha_i\ \Delta\alpha_j,$$

where V_0 is the energy value at the minimum, $\delta^2 V/\delta\alpha_i\delta\alpha_j$ is an element of the energy second derivative matrix, **M,** at the minimum, and $\Delta\alpha = a - a$, *is* the change in co-ordinates from the minimum at a,. This equation is written more simply in terms of the normal-mode co-ordinates q

$$T'(a) = V_0 + \tfrac{1}{2} \sum \lambda_i (\Delta q_i)^2.$$

$\Delta q = \mathbf{U}(\Delta\alpha)$, where the transformation matrix U diagonalizes M to Λ, so that $\mathbf{U}^T$ M U $= \Lambda$. The above expression for the energy in terms of the λ_i and Δq_i contains no terms in Aq, Δq_j

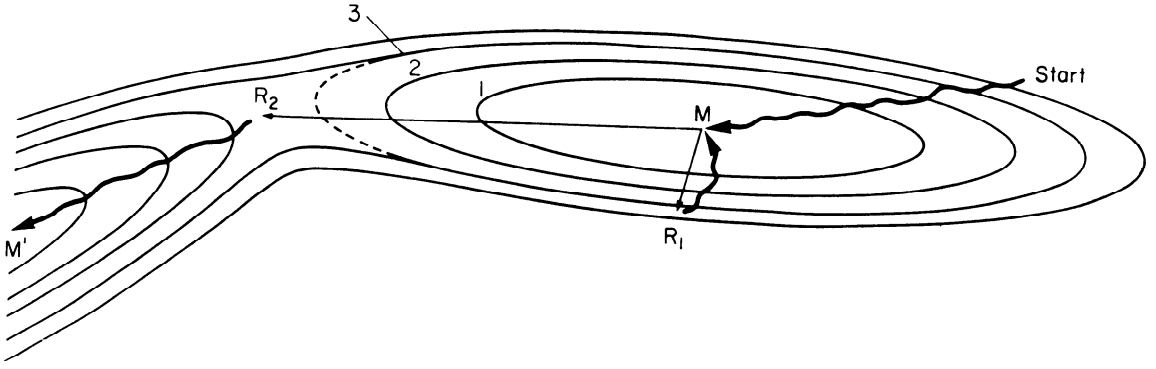


FIG. 6. Showing energy contours of a hypothetical 2-dimensional energy function to illustrate how normal-mode thermalization can escape from a minimum. The energy is contoured at intervals of kT (i.e. $\frac{1}{2}(kT)$ for each degree of freedom) above the energy of the present minimum at point M. Contours further away from the minimum become increasingly less quadratic (not such perfect! ellipses). The thermalization technique jumps to a random point uniformly distributed within the 2nth contour, where *n* is large enough to get over low barriers around the minimum. If this random point is in the right direction, subsequent minimization can move the conformation away from the old minimum towards a new minimum, M'. Points chosen at random without considering the normal modes would be uniformly distributed within a circle around the minimum so that there would be much less chance of choosing a point like R_2 that is a long way from M yet still at an energy of only $3(kT)$ above it.

and the co-ordinates q are therefore independent degrees of freedom of the conformation about the minimum at a,. By the equipartition of energy theorem, each degree of freedom will have on average kinetic energy and potential energy above the minimum of $\frac{1}{2}(kT)$. As a consequence of this kinetic energy, the value of the particular normal-mode co-ordinate will oscillate about the equilibrium value. Due to random fluctuations, the potential energy associated with the particular mode at time *t* will differ from the average value of $\frac{1}{2}(kT)$ and it is taken here as $R_n(t)(kT)/2$, where $R,(t)$ is a suitably distributed random number. The perturbation of the conformation that results from this energy fluctuation is given by $\frac{1}{2}(\lambda_i(\Delta q_i)^2) = R_n(t)(kT)/2$ or $\Delta q_i = (R_n(t)(kT)/\lambda_i)^{\frac{1}{2}}$. After applying these shifts to the torsion angle co-ordinates at the present minimum, minimization can be restarted.

The random number $R,(t)$ was chosen in two ways : (a) uniformly distributed between 0 and $2n$, and (b) exponentially distributed ($P(R,) = \exp(-R,)$) between 0 and $n \times \log(I_{\max})$ (where $I_{,,,} = 2^{31} - 1$, and depends on the word-length of the computer). **In** both distributions, the mean of $R,(t)$ is n, which was adjusted so that the disturbance at

van der Waals' parameters and other energy contributions. Without torsional energy terms, the chain becomes very much more twisted and kinked than in either the native conformation or the conformation obtained with the full force field. The C-terminal a-helix is badly distorted, and generally there is better packing of the side chains at the expense of a distorted backbone conformation. Without van der Waals' forces one gets a collapsed conformation with much inter-penetration of side chains; almost all side chains are closer than 10 A as indicated by the uniformly shaded contact map. Without the solvent energy term the molecule is less well packed; the α-helix has moved away from the p-hairpin. With additional S-S bonds both the a-helix and β-hairpin become distorted as the strong bonding forces and the strong van der Waals' repulsion associated with set C parameters oppose one another.

As a final test of the effect of energy terms on the near-native minima, the results obtained with set A van der Waals' parameters are given in Table 7. These parameters, the earliest set used, are like set B for the non-polar side chains, but use smaller radii and less strong attraction for interactions between the polar side chains (to allow for the more extended shape and greater conformational flexibility of these side chains). The best near-native minimum has a r.m.s. deviation of 2·91 A without hydrogen bonds but with S-S bonds. Without S-S bonds the fit worsens to 3·04 A r.m.s. ; with hydrogen bonds it worsens to 3·23 A. Under the same conditions and with all the energy terms including hydrogen bonds, the set B van der Waals' parameters give a significantly better near-native minimum than do the set A parameters (r.m.s. deviations of 2·57 Å and 3·23 A, respectively). The set B conformation is more compact and has lower solvent and van der Waals' energy contributions; the set A conformation is better stabilized by the peptide hydrogen bond contribution. Because the set A side chains have smaller van der Waals' radii, the additional constraint of S-S bonds can be tolerated better than with the set C parameters.

(iii) *Minimization with the additional χ variables*

Here the second representation of protein structure is investigated in which many residues have an additional degree of freedom allowing the side-chain centroid to move relative to the backbone. Thirty seven out of the 58 residues in PTI were assigned variable χ angles (see Table 2), and the others had the same rigid geometry used above (Table 1). Now the conformation of the molecule was defined by *a* total of 93 variables rather than the 56 *a* angle variables used in the more simple representation above. With the set A van der Waals', hydrogen bond, torsional, and solvent energy terms the near-native minimum energy conformation has a much lower energy than with rigid side chains (−97·6 kcal/mol instead of −59·6 kcal/mol) but the r.m.s. deviation is no better (3·14 Å instead of 3·23 A) (see Table 7). With the extra variables the conformation becomes considerably more compact (R_g is 14·7 Å, instead of 16·5 A), which is much more compact than native PTI and could explain

FIG. 9. Showing the effect of omitting various energy contributions on conformations generated by minimization from the idealized native starting conformation. Set C van der Waals' parameters were used, and the following contributions were omitted from the different runs. (a) None omitted (except the peptide hydrogen bonds that cannot be used with the set C parameters as the radii of the side chain spheres have been specially enlarged to include the volume of the peptide group). (b) No torsion energy, i.e. the backbone lacks any stiffness. (c) No van der Waals' energy. (d) No solvent interaction energy. (e) None omitted, but there is an additional S-S bonding potential that holds together the native PTI disulphide bonds.

TABLE 7

Minimization using set A van der Waals' parameters

Conformation and conditions	Energy (kcal/mol) Total	Torsion	H-bond	van der Waals'	Solvent	R_g (Å)	r.m.s. deviation $\Delta\alpha_i$ (deg.)	Ab_{SS} (Å)	Δr_{ij} (Å)
Initial									
Idealized native PTI	301·2	38·1	14·9	282·5	−34·3	16·1	0·0	0·3	1·3
Minimized									
A, .	−59·6	20·8	−30·6	−15·0	−34·8	16·5	38·2	3·1	3·23
A, no H-bonds	−44·4	20·1		−17·4	−47·1	15·9	40·7	2·9	3·04
A, no H-bonds, no torsion	−74·8	—		−16·4	−58·5	15·5	49·9	4·1	3·47
A, no H-bonds, no van der Waals'	−355·0	43·0			−398·0	9·3	97·8	9·3	9·46
A, no H-bonds, no solvent	−8·2	15·4		−23·6	—	15·9	42·7	4·8	3·45
A, no H-bonds, and S-S bonds	−31·1	20·4		−13·3	−38·2	15·8	38·6	0·0	2·91
Flexible side chains									
Initial idealized	−38·3	26·6	−17·4	17·0	−30·5	16·0	—	0·8	1·00
Minimized	−97·6	23·4	−36·7	−23·0	−61·5	14·7	34·8	2·1	3914

part of the r.m.s. deviation. Minimization was repeated, therefore, with the bigger side chain radii of the set C parameters. again including the peptide hydrogen bonds and other energy terms. The r.m.s. deviation still shows no improvement over the rigid side chain model and the conformation is still too compact (see Table 6). More attention will have to be given to finding the most suitable energy parameters for the flexible side chain model of protein structure; no further results with this model are reported here.

(iv) ***Minimization*** *of the simplified native co-ordinates*

In the results reported above, the polypeptide chain was constructed using idealized standard geometry taken from averages found in known protein conformations. Each side chain of a given type had the same shape and no allowance was made for the adjustments of side chain conformation necessary to pack better. Here the actual simplified native co-ordinates are used and the minimization is started at a **con-**formation where each side chain is at the centroid of the side chain in the **actual** X-ray co-ordinates of PTI. With this geometry and with rigid side chains, the closest near-native energy minima deviate by 2·46 A and 2·72 Å r.m.s. with the set B and C van der Waals' parameters? respectively (see Table 6). The energy values at these minima are lower than when using the idealized side chain geometry, especially the solvent contributions and the conformations have a radius of gyration closer to that of the native co-ordinates. It is surprising that the r.m.s. deviations are so similar to those obtained with the idealized geometry. Clearly the side chains do **occupy** better positions in the simplified native conformation than in the idealized native conformation as evidenced by the lower energy in the former conformation (Table 6).

That there is no energy minimum closer than 2·5 Å r.m.s. deviation from **the** simplified native co-ordinates seems to be a property of the simple spherical shape of the side chain potentials rather than due to side' chain rigidity or to side **chains** occupying unfavourable positions. This limitation could be expected from a **model** that averages the proper shape of side chains into a structureless sphere, **moving** individual atoms by about 1·5 Å in the process.

(v) *Properties of the idealized near-native minima*

Although the present model greatly simplifies the structure of a protein chain, it has been shown above that there are stable equilibrium conformations very close **to** the native structure (r.m.s. deviation ≈2·5 A). Here the best near-native minimum energy conformations are studied in more detail and compared to the actual **native** conformation of PTI.

The best near-native minimum energy conformation, obtained with set B van der Waals' parameters. hydrogen bonds, etc, deviates by 254 Å r.m.s. (see Fig. 10). The individual deviations of each residue vary between 1·83 Å and 4·25 A. The average deviation of the non-polar residues is 2·48 Å, while for the polar residues it is 2·80 Å. Those residues that are in the hydrophobic core of PTI (4, 5, 21, 22, 23, 33, 35, 45. 51 and 55) deviate on average by 2·28 A, while those in secondary structure (2 to 6. 18 to 24, 29 to 35 and 47 to 55) deviate on average by 2·25 Å r.m.s.

The conservation of the native side chain interactions was also checked. **These** side chain interactions were defined as between a pair of side chain centroids closer than the sum of the set C van der Waals' radii (Table 4) plus 2·8 Å and having the side chains pointing towards each other. (The line joining the side chains must not

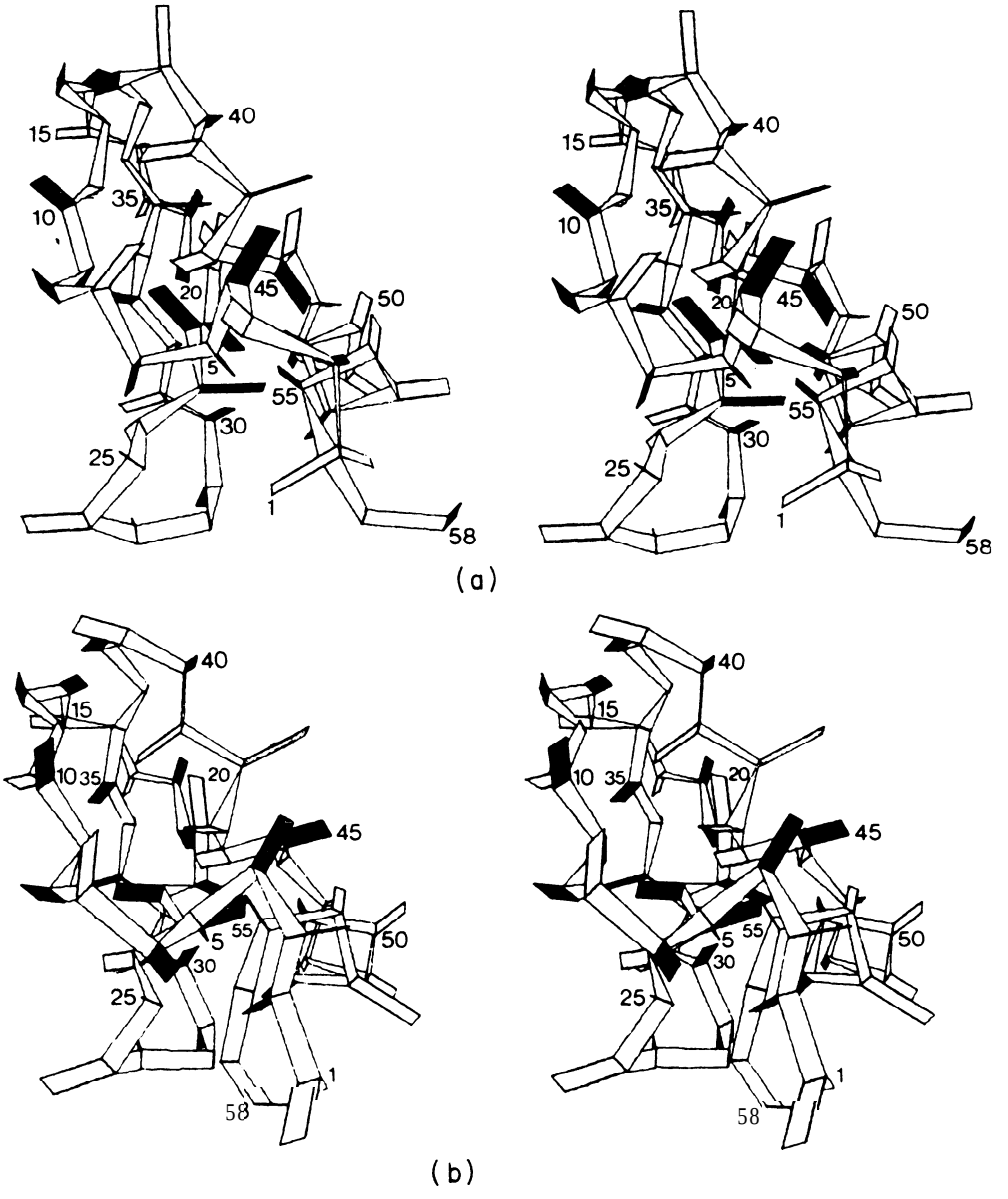
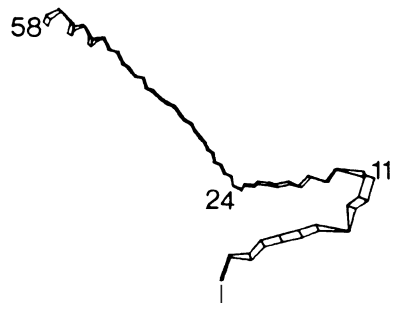


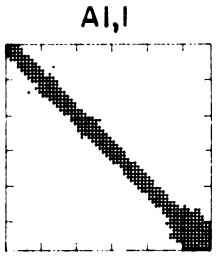
FIG. 10. Showing (a) the idealized native conformation and (b) the B parameter near-native minimum, which deviates by 2·57 Å r.m.s. from the native PTI co-ordinates. Here the ribbon traces both the C^{α} backbone and the side chain centroids. The ribbon segment between the C^{α} and the side chain centroid is shaded solid if that particular amino acid is non-polar (hydrophobic). The Figure shows the great similarity of the native PTI conformation and the near-native minimum.

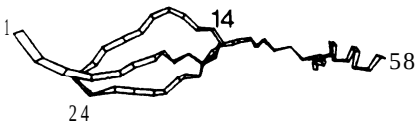
intersect with the planes through each C^{α} that are perpendicular to the particular C^{α}-to-side chain centroid vector.) In the simplified native PTI structure there are 92 such contacts, 35 between a pair of non-polar side chains: 9 between a pair of polar side chains, and 48 between one non-polar and one polar side chain (mixed). In the set B near-native minimum energy conformation there are 82 contacts (30 non-polar, 9 polar, and 43 mixed) of which 40 are identical to contacts in native **PTI.** Similar results are also found for the set A and set C near-native conformations.

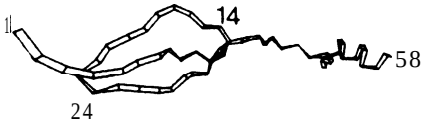
The values of the r.m.s. deviation obtained here for the near-native minima! indicate a high degree of conformational similarity. C^{α} r.m.s. deviations of 3·2 Å and 2·2 Å were obtained between the 31 residue structural repeats of mvogen (McLachlan. 1972). and between the 70 framework residues of the V and C domains of the Bence-Jones dimer (Padlan & Davis, 1975). For the B near-native minimum. the corresponding C^{α} r.m.s. deviation i x 1·9 Å. (The r.m.s. deviation usuall v quot cd in the

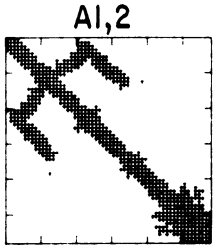

58
11
24
1
L

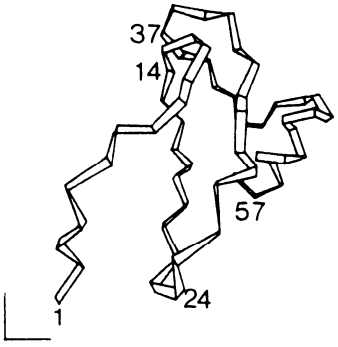

58
11
24
1
L

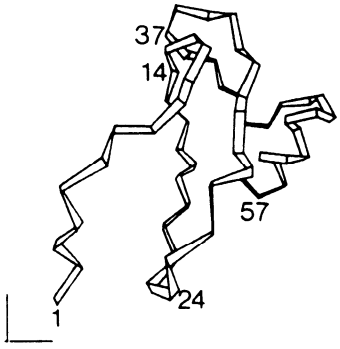

A1,1

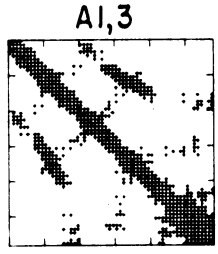

1
14
58
24

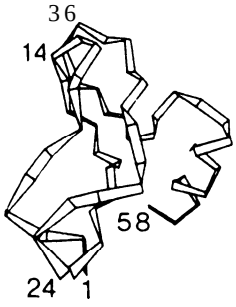

1
14
58
24

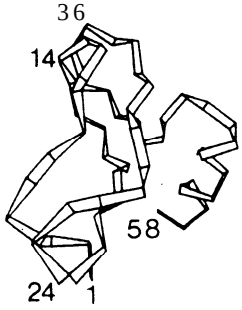

A1,2

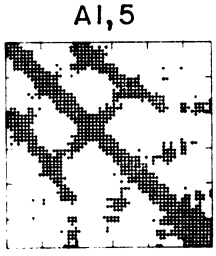

37
14
57
24
1

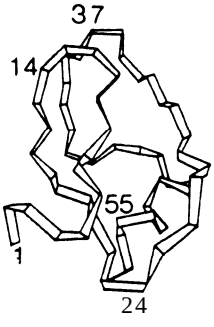

37
14
57
24
1

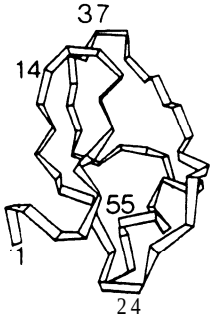

A1,3


36
14
58
24
1


36
14
58
24
1


A1,5


37
14
55
1
24


37
14
55
1
24


A2,1

present work is for the side chain only, and it is always higher than the deviation of the C^{α} backbone.)

(c) Folding from open chain conformations

In this section the simple model of protein conformation, used above to reproduce the geometry and stability of native PTI, is used to simulate folding from "denatured" open chain conformations. One aim of this approach was to evaluate the power of the minimization and thermalization techniques, and to examine the nature of the energy surface. For this, I tried to estimate: (a) the rate of convergence to an energy minimum : (b) the difficulty of escaping from a local minimum ; (c) the frequency of occurrence of local minima or the distance between minima on the energy surface ; and (d) the difficulty of folding to a compact conformation. The second aim was to simulate realistically the actual folding of a small protein as a test of the usefulness of the simple model. For this, I tried to estimate: (a) the speed and pathways of folding from an open conformation; (b) the conformations and stability of intermediates ; (c) the chances of getting to native-like conformations ; (d) the significance of pre-formed secondary structure; and (e) the role of the different energy terms in the simulated folding. As the results of some simulations without hydrogen bonds have already been reported (Levitt & Warshel, 1975), more attention is now given to folding simulations with hydrogen bonds.

(i) *Set A folding with a pre-formed helix*

Figure **11** and Table 8 show the chain fold, side chain contact maps, and energies of conformations generated by alternating minimization and normal-mode thermalization using set A van der Waals', hydrogen bond, torsional, and solvent energy terms. The starting conformation was an open chain with $a = 180°$ for all residues except the last ten **(17** % of the molecule), which were set to an a-helix as in native PTI. In the first minimum energy conformation (Al ,1), which is still very open, there are bends before Gly12 and Asn24, but there is little side chain interaction as shown by the lack of off-diagonal shading on the contact map (see Fig. **11).** The energy has dropped from $-3{\cdot}0$ kcal/mol to $-18{\cdot}3$ kcal/mol in 44 cycles, reaching a perfect minimum. The second minimum energy conformation (Al ,2) shows some secondary structure with a p-sheet hairpin centred on residues 12 and **13;** there is also a third β-strand running parallel to the first (the three strands consist of residues 2 to **10,** 14 to 21, and *24* to *34).* All the energy contributions show a significant fall in value, though the conformation is still rather open ($R_g = 29{\cdot}8$ A). The hydrogen bonding between the three β-strands is good with peptide-to-peptide distances less than 6 Å for the following pairs of peptide groups : *6* : **21, 7** : 20, 8 : 19, 10 : **16,** 9 : **15** for strands **1** and 2, and 5:25, 6:26, 8:27, 9:29, 10:30, 11:31, 12:32 for strands **1** and 3. The

FIG. 11. Conformations generated when folding from an open chain conformation of PTI. In this starting conformation all α_i = 180°, except for residues 48 to 58, where $\alpha_i = 45°$, i.e. all fully extended chain with a preformed C-terminal α-helix like that in native PTI. Set A van der Waals' parameters were used together with peptide hydrogen bonds, torsional forces, and solvent interactions. The thermalization done between the minimization runs was with exponentially distributed random numbers at a temperature of 330°K and for $n = 1{\cdot}5$. Before the first pass of minimization the *a* torsion angles were uniformly randomized to a maximum perturbation of 10" (using the IBM supplied subroutine RANDU with IX = 55555 initially). Conformations A1,1, A1,2, A1,3 and A1,5 are the first 4 successive minima (A1,1 and A1,2 are drawn at half-scale). Conformation A2,1 was obtained by minimization from conformation A1,5 with an additional S-S bonding constraint to force the native S-S bridges to be formed.

TABLE 8

Open chain folding with set A van der Waals' parameters

Conformation	Energy (kcal/mol) Total	Torsion	H-bond	van der Waals'	Solvent	R_g (Å)	r.m.s. Δb_{ss} (Å)	r.m.s. Δr_{ij} (Å)	n†
Set helix									
Near-native	59·7	20·8	30·6	15·0	−34·8	16·5	3·1	3·2	368
Initial	− 3·0	18·1	−15·4	− 1·4	− 4·4	71·4	97·6	59·2	0
A1,1	−18·3	7·1	−15·1	−2·4	−7·9	49·2	56·6	36·6	44
A1,2	−37·9	14·9	−22·6	−10·4	−19·8	29·8	31·5	18·5	75
A1,3	−30·9	16·7	−15·1	−12·0	−20·5	18·5	15·9	7·9	38
A1,4	−59·8	16·7	−27·6	−16·0	−32·9	16·4	12·6	6·3	134
A1,5	−65·4	17·5	−28·3	−15·3	−39·3	16·3	12·4	6·6	134
A2,1 (SS bonds)	−24·2	20·5	−18·9	2·5	−28·3	16·3	1·8	5·8	119
A3,1 (fix helix)	−63·6	17·2	−29·6	−16·0	−35·2	16·5	12·6	6·4	134
Long H-bonds									
Near-native	51·1	23·6	−22·3	18·0	34·4	16·4	3·6	2·9	376
Initial	3797	19·3	24·5	1·9	4·2	72·2	98·3	59·9	0
A6,1	20·1	8·9	−8·6	−3·2	17-2	39·4	46·2	26·9	102
A6,2	31·2	15·6	−12·4	10·9	23·5	24·2	22·5	12·4	106
A7,1 (push 5 Å)	44·8	15·7	14·6	14·4	31·5	18·2	8·2	7·3	288
A8,1 (weak torsion)	−56·3	11·1	−16·7	−16·0	−34·7	17·1	6·8	6·4	132
A9,1 (SS bonds)	−38·7	21·3	−13·8	14-8	−31·4	16·6	0·2	5-9	153
No helix set									
Initial	9·9	17·2	3·3	−1·0	−3·0	77·1	106·2	64·7	0
A11,1	−4·8	5·3	−2·6	−1·7	−5·8	55·0	70·7	42·4	33
A11,2	−19·4	9·8	−7·0	−8·6	−13·6	28·5	24·1	15·4	99
A11,3	−32·0	11·2	−16·4	−9·2	−17·6	31·6	32·4	19·4	128
A12,1 (push 5 Å)	−28·1	9·1	−13·2	−9·4	−14·6	28·6	22·9	16·3	119
A13,1 (push 15 Å)	−41·3	11·6	−11·7	−13·6	−27·6	20·3	11·5	9·7	98
A14,1 (push 5 Å)	−46·0	14·0	15·5	−10·8	−33·7	19·0	7·8	8·5	138

† n is the number of cycles of minimization required to generate the conformation from the previous stable conformation.

next conformation found by minimization (A1,3) is more compact than before (R_g = 18·5 A), but its energy is high as the minimization was stopped after only 38 cycles. The thermalization that occurred at the beginning of this pass of minimization has opened the p-hairpin, moving the second strand (17 to 21) away from the third strand (24 to 36). The α-helix has folded against this distorted β-sheet with bends immediately before Gly36, Arg39, and Asn43. Continuing the minimization gives a fourth conformation (A1,4) after 134 cycles, and then a fifth conformation (A1,5), a true energy minimum, after another 134 cycles. This final minimum energy conformation is as compact as native PTI, the energy is low (−65·4 kcal/mol), and the r.m.s. deviation is down to 6·6 A. In fact,, the energy of conformation A1,5 is lower than that of the corresponding near-native minimum, which has an energy of −59·7 kcal/mol and a r.m .s. deviation of 3·23 A. Only one native PTI S-S bond can be formed easily in conformation A1,5, that between Cys14 and Cys38 (they are 1·8 Å apart). When an additional S- S bonding potential is used to connect the native S-S bridges (between 5 : 55, 14 : 38 and 30 : 51), the central p-hairpin (residues 14 to 32) is forced to come apart. The energy of the resulting conformation (A2,1) is higher at −24·2 kcal/mol and the hydrogen bond energy is much less favourable than in conformation A1,5. These distortions are caused by the formation of the 5 :55 S-S bond through the middle of the p-hairpin. Were a weaker S-S bond force-constant than 10 kcal/mol per Å^2 to be used, the N and C termini of the molecule could possibly come together by going around the p-hairpin not through it.

(ii) *Set A folding with long hydrogen bonds*

The next folding simulation (see Fig. 12 and Table 8) was done using the same conditions as above except for a modification to the hydrogen bond potential. The geometry of the simplified peptide was changed by increasing the distance between 0' and N' from 1 A to 2 A. Now the minimum of the interaction energy between a pair of peptides occurs at a separation of 5·5 Å instead of 4·5 A. The first minimum energy conformation (A6,1) was again very open. With the long hydrogen bonds, the α-helix is not stable and has become very distorted. Although the hydrogen bond energy is now much higher than before (in conformation A1,1), the solvent energy is lower so that both conformations have similar energies. The next conformation (A6,2) has a similar arrangement of β-strands as before (conformation A1,2), but now the anti-parallel chains forming the hairpin (14 to 23 and 27 to 36) interact more strongly than the parallel pair (3 to 11 and 27 to 36). Thermalization and energy minimization from conformation A6,2 did not readily lead to a compact structure, and a pushing potential *was* used that dropped from 100 kcal/mol at conformation A6,2 to 0 kcal/mol at any conformation that was at least 5 A r.m.s. deviation from conformation A6,2. This gave a third minimum energy conformation (A7,1), in which what is left of the a-helix packs against the p-hairpin. The β-hairpin is well formed and shows the characteristic right-handed twist found in the actual native PTI conformation. On the next pass of minimization, the torsional energy was halved in the hope that with a less stiff backbone the conformation would become more compact. Indeed? the resulting conformation (A8,1) is more compact (R_g = 17·1 A) and deviates less from the native co-ordinates (6·4 A, r.m.s.). All the native S-S bonds can be introduced easily into conformations A7,1 and A8,1; minimization with the S-S bonding energy gives conformation A9,1, in which the β-hairpin is still intact (see Fig. 12).

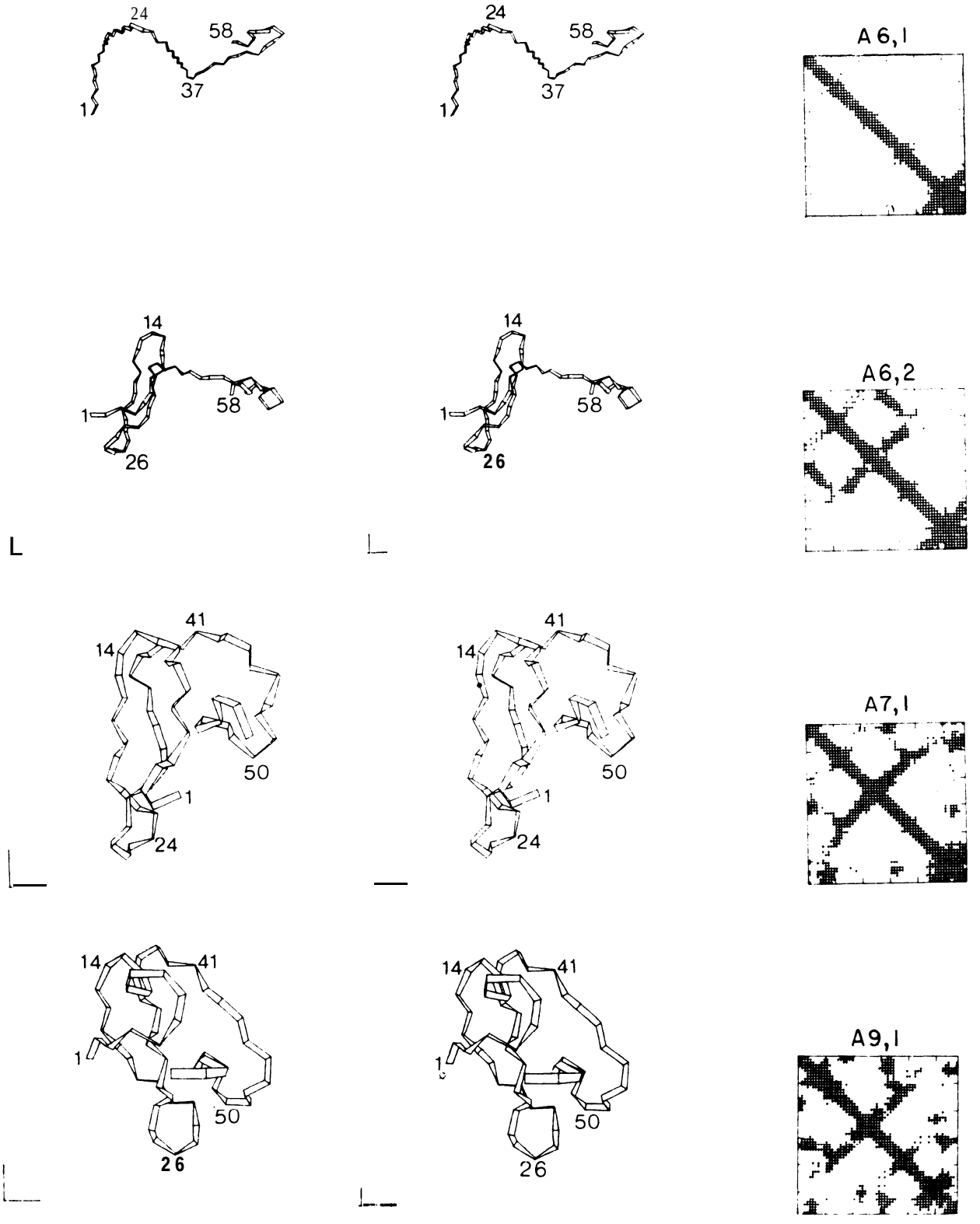


FIG. 12. **The conformations generated with the same potentials and other conditions as for Fig. 11, except that the peptide groups were swollen to have a 0'-N' separation of 2 Å instead of 1 Å (long peptide hydrogen bonds), and the random number generation started with IX = 11111. Conformations A6,1 and A6,2 are the first 2 successive energy minima (drawn half-scale). Conformation A7,1 was generated from conformation A6,2 by minimization with a pushing potential that dropped from 100 kcal/mol to 0 as the r.m.s. deviation from A6,2 increased to 5 Å. Conformation A9,1 was generated from A7,1 by minimization with the S-S bond constraint.**

(iii) Set ***B folding with pre-formed helix***

Two separate simulations were done next using the same starting conformation as above but with set B van der Waals' parameters. In the first run, the van der Waals' energy was full-strength, whereas in the second run it was halved. The first simulation (see Table 9) gave results very like those with the set A parameters (see Fig. **11).** In the fourth and final minimum energy conformation (B2,1) there is a three-stranded β-sheet that has the strongest association between strands **1** and 2 (residues 3 to **11** and **15** to 23). The p-hairpin so formed is centred on Gly12, unlike the b-hairpin in native PTI, which is centred around residue 25. The a-helix of conformation B2,1 remains intact and packs against the p-sheet . The r.m.s. deviation of this conformation from the native co-ordinates is 7·10 A, but none of the three native Cys pairs are closer together than **10 A.**

The second simulation with set B van der Waals' parameters at half-strength was

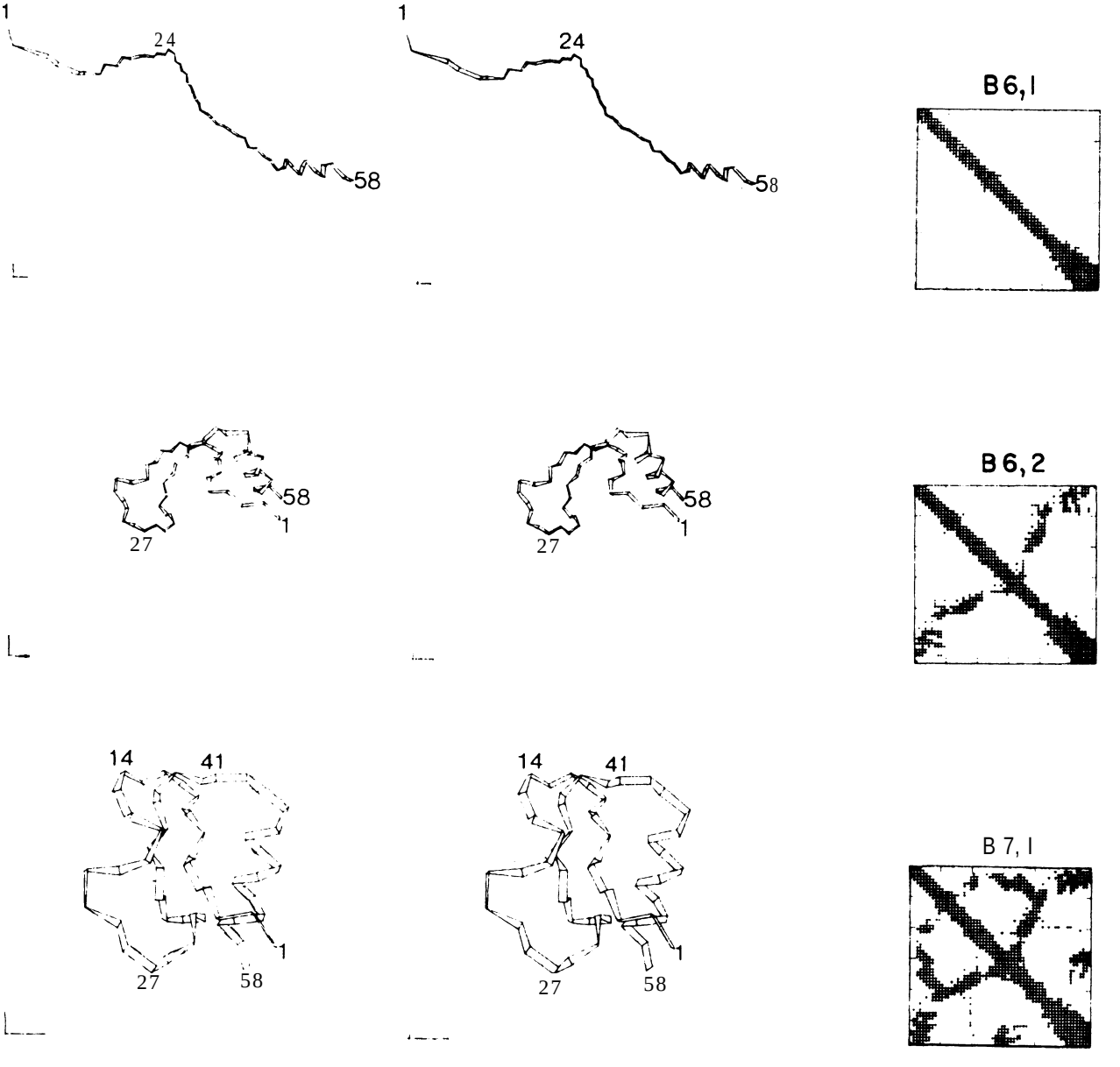


FIG. 13. Conformations generated with half-strength set B van der Waals' parameters, starting from the fully extended chain and pre-set a-helix, without any randomization of the initial a angles. Conformations B6,1, B6,2 and B7,1 are the first 3 minima obtained by alternating minimization with thermalization (conformation B 1,1 is half-scale).

the minimum was big enough to get over the barrier around the minimum; for $T = 330°K$, $n = 1{\cdot}5$ worked well. The exponential distribution is more realistic as it allows for much bigger fluctuations although with a correspondingly low probability; it is used for most of the calculations reported here.

In effect, the energy is minimized without any thermal disturbances until a minimum is reached. At this point, the conformation is given a random displacement to get it out of the minimum so that minimization can continue. As these perturbations are in the directions of the normal-mode vibrations, and as the amplitude of each perturbation is realistically related to the energy associated with each mode, this simple procedure usually works well. Arbitrary random perturbations of the conformation work much less well as most of the perturbations cause the energy to rise very steeply without changing the overall conformation very much.

(ii) *Holding potentials*

Experience with normal-mode thermalization showed that it worked best when the protein chain had an open conformation, but less well once the chain had folded. The quadratic approximation to the energy at the minimum, which forms the basis for normal-mode thermalization, is much less valid for the compact conformations, and perturbations of the conformation can cause unexpectedly large increases in the energy. Subsequent minimization can then move the conformation unnecessarily far from the previous minimum due to the very large initial forces. This problem was combatted by holding the current conformation close to the starting conformation until it had relaxed sufficiently to eliminate any large forces. The energy function was modified by an additional term giving:

$$V'(a) = V(a) + K(V_{\text{init}}) \text{ RMS } (r_{ij} - r_{ij}^{\text{init}}).$$

The right-hand factor is the root-mean-square value of the difference in the distance between groups i and j in the present conformation (r_{ij}) and the same distance in the starting conformation (r_{ij}^{init}). $K(V_{\text{init}})$ is a constraining force constant that depends on the energy of the starting conformation for the particular minimization run. It is chosen to be strong enough to hold the conformation against the repulsive forces operative at the initial conformation.

$$K(V_{\text{init}}) = V_{\text{init}}/\sigma_{\text{hold}}^2 \text{ for } V_{\text{init}} > V_{\text{low}}$$
$$= 0 \qquad \text{for } V_{\text{init}} \leq V_{\text{low}}.$$

σ_{hold} is twice the root-mean-square distance that the conformation could move if the molecular potential varied quadratically; it was taken as 3 A, which works well. V_{low} is a threshold energy below which the conformation is considered to have relaxed sufficiently; it was usually taken as 0 kcal/mol. This technique is useful when starting minimization at any high energy conformation, e.g. the actual native conformation in the simplified model.

(iii) *Pushing potentials*

In some cases the, walls around the minimum were so steep and non-quadratic that normal-mode thermalization could not get over them; subsequent minimization then always led back to the previous minimum. This problem was combatted by modifying the energy to include a term that pushed the conformation away from the previous minimum.

$$V'(a) = V(\mathbf{a}) + K_{\text{push}}\{1 - \tfrac{1}{2}(7y^2 - 9y^4 + 5y^6 - y^8)\} \text{ for } 0 < y < 1,$$

where

$$y = \left\{\frac{1}{N}\sum_{i,j}^{N}(r_{ij} - r_{ij}^{\text{min}})^2\right\}^{\frac{1}{2}} \Big/ \sigma_{\text{push}}$$

is the scaled root-mean-square deviation of the current conformation (r) from the previous minimum energy conformation $(\mathbf{r}^{\text{min}})$. This sigmoid-shaped pushing potential function has a maximum value at the previous minimum energy conformation, and gradually drops to zero as the conformation changes and the root-mean-square deviation increases to σ_{push}. The particular functional form is somewhat arbitrary; it was chosen for its simplicity and the fact that the pushing potential and its first and second derivatives are zero at y = 1.

more interesting (see Fig. 13 and Table 9). Now the folding from open to compact conformations follows a different pathway from those described above. The first conformation (B6,1), a minimum reached after 51 cycles, is extended, but has a 90° bend near Asn24. The second conformation (B6,2) is a long imperfect hairpin centred at 24 and running the length of the molecule. Each strand of the hairpin bends by about 90" before Gly12 and Gly36. The third minimization, which now used full-strength van der Waals' energy terms, gave a compact conformation, B7,1 (R_g = 16·4 A) that deviates by 6·2 Å r.m.s. from native PTI. In this conformation the β-hairpin is more irregular at the centre of the bend than in conformations A1,5 and A7,1 discussed above. Another feature is that the first four residues of the molecule come between the β-hairpin and the C-terminal α-helix. In conformation B7,1 all three pairs of Cys residues that form S- S bonds in native PTI are closer together than 10 Å, with the closest association between Cys5 and Cys55 (5·5 Å). The energy of this final conformation is lower than that of the other minimum energy conformation obtained with the set B van der Waals' parameters (B2,1) (energy values of −72·8 kcal/mol and −60·7 kcal/mol, respectively). The energy of B7,1 is in fact lower than that of the B parameter near-native minimum energy conformation, which has an energy of −67·8 kcal/mol.

(iv) Set *A* folding *from a fully extended chain*

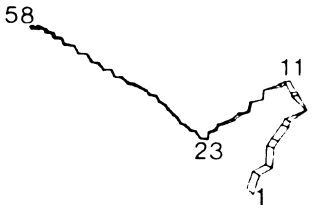
In this simulated folding (see Fig. 14 and Table 8) we use exactly the same condition as for the first set A parameter simulation (section (i) above)? except that the conformation of the last ten residues is not set to an a-helix in the starting conformation. The first minimum energy conformation reached after only 33 cycles looks just like the first minimum obtained with the pre-formed α-helix (Al ,1). The bends occur in the same places along the chain showing how the starting conformation of the last ten residues initially has little effect, on the conformation of the rest of the chain. The second minimum energy conformation (Al 1,2) is reached after normal-mode thermalization at the first minimum followed by a further 99 cycles of minimization. Now the chain has formed a long p-hairpin centred at residue 25 and involving the first 45 residues of the molecule. The 90° bends of the hairpin before Glv12 and Gly37 are like those found before in conformation B6,2 under verv different conditions. The conformation at this point looks very much like two separate anti-parallel β-sheets, each with the usual right-handed twist, at right-angles to one another. Repeated thermalization of conformation All.2 followed by minimization always gave a conformation that was just like A11,2. Clearly, the long β-hairpin conformation is in a deep minimum. A pushing potential was therefore used to force the molecule away from this verv stable but rather open conformation (R_g = 28·5 Å). On the first attempt the pushing potential was set to drop from a maximum value of 100 kcal/mol at conformation A11,2 to 0 kcal/mol once the r.m.s. deviation from conformation All.2 exceeded 5 A. The resulting conformation (A12,1) is about 6 Å r.m.s. away from conformation Al 1.2. but still looks similar to that conformation with the same stable central p-hairpin. Note how the hairpin has actually become more uniform than before; the hydrogen bond energv term has dropped from −7·0 kcal/mol to −13·2 kcal/mol, and the clear patches on the contact map near residues 10 and 36 have disappeared (see Fig. 14).

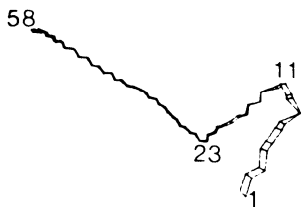
The minimization from conformation All.2 was repeated using a pushing potential that fell to zero only after the conformation had moved 15 Å r.m.s. away from Al 1.2.

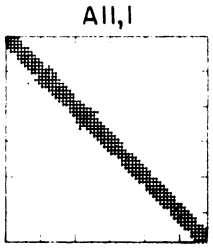
TABLE 9

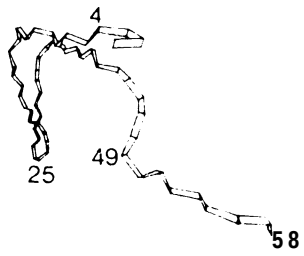
Open chain folding with set B van der Waals' parameters and pre-formed α-helix

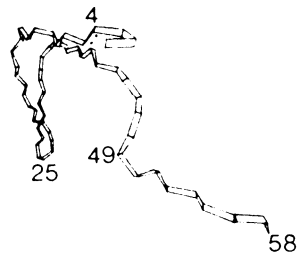
Conformation	Energy (kcal/mol) Total	Torsion	H-bond	van der Waals'	Solvent	R_g (Å)	r.m.s. Δb_{ss} (Å)	r.m.s. Δr_{ij} (Å)	n
Near-native	−67·6	21·9	27·7	21·1	40·6	16·0	2·0	2·6	393
First run									
Initial	−3·6	18·1	15·4	1·9	−4·4	71·4	97·6	59·2	0
B1,1	−37·5	11·3	−19·8	−11·9	17·1	27·7	27·2	15·5	100
B1,2	−44·9	13·5	24·2	15·3	18·9	28·6	28·6	16·9	129
B1,3	−58·1	16·6	26·3	21·5	26·9	17·6	12·3	7·8	86
B2,1 (push 5 Å)	60·7	16·2	26·1	21·8	29·0	17·3	11·2	7·1	124
Second run									
Initial (weak van der Waals')	5·4	18·8	−8·0	−2·1	−3·3	72·5	100·1	60·3	0
B6,1 (weak van der Waals')	−12·7	8·4	−9·4	-492	−7·5	51·2	63·3	38·4	51
B6,2 (weak van der Waals')	−28·2	16·1	−11·0	14·9	−18·4	21·8	16·1	10·4	90
B7,1	−72·8	21·8	−37·8	20·1	36·6	16·4	4·7	6·2	131

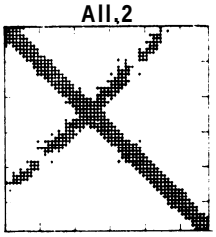

58
11
23
1


58
11
23
1

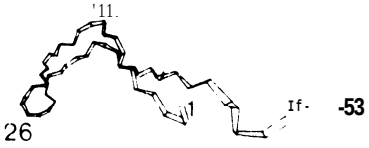

AII,I

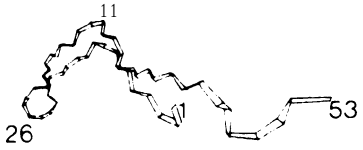

4
49
25
58

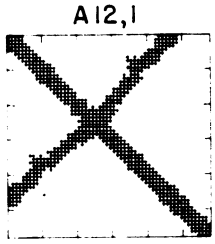

4
49
25
58

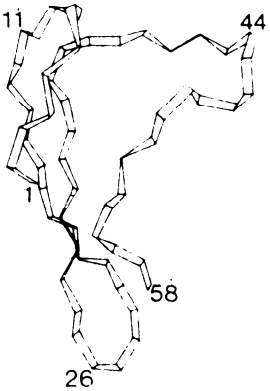

AII,2

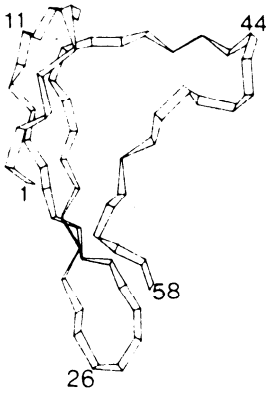
L


11
26
-53

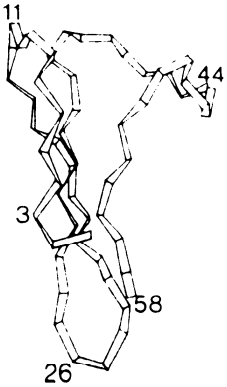

11
26
53

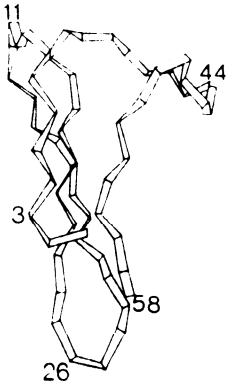

A12,I

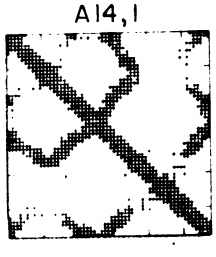

11
44
1
58
26


11
44
1
58
26


A 13,I


11
44
3
58
26


11
44
3
58
26


A14,I

The resulting conformation (A13,1) is more compact ($R_g = 20{\cdot}3$ A) and has a much lower energy (−41·3 kcal/mol). The original p-hairpin has folded in two, the two ends have come apart (near residues 1 to 5 and 38 to 45), and the C-terminal part of the chain (residues 51 to 58) has packed onto the central p-hairpin. A final conformation (A14,1) was then generated by minimization starting at conformation A13,1 and using a weak pushing potential that fell to zero after 5 Å r.m.s. shift. This conformation has a lower energy than A13,1 (−46·0 kcal/mol), is more compact ($R_g = 19{\cdot}0$ A), and has bends before Gly12, Lys26, Gly37, Phe45, and Asp50. Although the last 15 residues pack rather loosely against the p-hairpin, they could conceivably detach themselves, form into an a-helix, and then pack more strongly against the p-hairpin to give a conformation closer to native PTI. The energy of the final conformation folded without a pre-formed a-helix (A14,1) is 19·4 kcal/mol less stable than the corresponding conformation folded with the a-helix (A1,5).

(d) *Folding without hydrogen bonds under various conditions*

Some results of folding simulations using set A van der Waals' parameters without hydrogen bonds have been presented previously (Levitt & Warshel, 1975). The results presented here (see Fig. 15 and Table 10) are for folding simulations of *an* extended chain with a pre-formed a-helix using set C van der Waals' parameters without any hydrogen bonds. With these parameters the van der Waals' interactions are more strongly attractive than for sets A or B. Collapse to a compact conformation occurs more rapidly with no intermediate open chain energy minima. As the randomization associated with normal-mode thermalization is not needed, the relationship between the initial open and final compact conformation is more direct. It becomes possible to investigate the effects of changes in the energy function on the folded conformation. The first conformation C1,1 was generated using the complete force field ; it is compact ($R_g = 17{\cdot}6$ A), quite close to the native PTI co-ordinates (r.m.s. deviation of 7·34 A), has a low energy (−62·4 kcal/mol), and was reached after only *300* cycles. In this conformation the p-hairpin centred at residue 25 is too open near the centre (residues 20 to 30), and the ends of the chain (residues 1 to 8 and *45* to 58) pack on opposite sides of the hairpin. It would be difficult to join Cys5 and Cys55 to form the native S-S bridge without distorting the b-hairpin even more. The other native S-S bridges, 14: 38 and 30: 51, are closer together than 5 Å and could be connected without distorting the conformation. Although there are no hydrogen bonds, the a-helix remains intact; with the set A parameters without hydrogen bonds the a-helix breaks up (Levitt & Warshel, 1975).

The next conformation (C2,1) was generated using the same conditions as above except for slightly altered solvent interaction parameters for Pro and Cys residues. Many features of the conformation' are like those of C1,1. The β-hairpin is less well formed (it is open near residues 18 and 35), but the a-helix now packs on the same side of the sheet as the N-terminal chain segment (residues 1 to 8). Now all three

---

FIG. 14. Conformations generated with the set A parameters (as in Fig. 11), but with a fully extended starting conformation (no pre-formed C-terminal oc-helix). Conformations A1 1,1 and A1 1,2 are the first 2 minima (half-scale). Conformation A1 2,1 (also half-scale) was generated using an additional pushing potential which dropped from 100 to 0 kcal/mol as the r.m.s. deviation from conformation A 11,2 increased to 5 A. Conformation A13,1 was generated with a pushing potential that dropped from 100 to 0 kcal/mol as the r.m.s. deviation from A11,2 increased to 15 Å. Conformation A14,1 was generated from A13,1 using a 5 Å pushing potential,

TABLE 10

Set C parameter open ***chain folding*** *omitting certain energy contributions*

Conformation	Energy (kcal/mol) Total	Torsion	H-bond	vander Waals'	Solvent	R_g (Å)	r.m.s. Ab_{ss} (Å)	r.m.s. Δr_{ij} (Å)	n
Near-native	−61·2	22·3	—	−58·8	−24·7	17·2	4·4	2·8	319
Set helix									
Initial	18·2	18·8	—	2·7	− 33	72·5	100·1	60·3	0
C1,1	−62·4	12·5	—	−57·5	−19·2	17·6	8·8	7·4	205
C2,1 (modified solvent)	−59·7	14·8	—	− 55-9	−18·6	17·5	5·6	6·6	280
C3,1 (no solvent)	−40·2	12·7	—	−52·9	—	19·2	18·2	9·1	279
C4,1 (no torsion)	−87·6	—	—	−59·9	−27·7	16·9	14·2	8·0	289
C5,1 (no bends)	−18·9	9·2		−15·0	−13·1	61·1	58·6	49-o	92
C5,2 (no bends)	−33·2	15·1		−33·9	−15·1	29·4	23·7	17·9	192

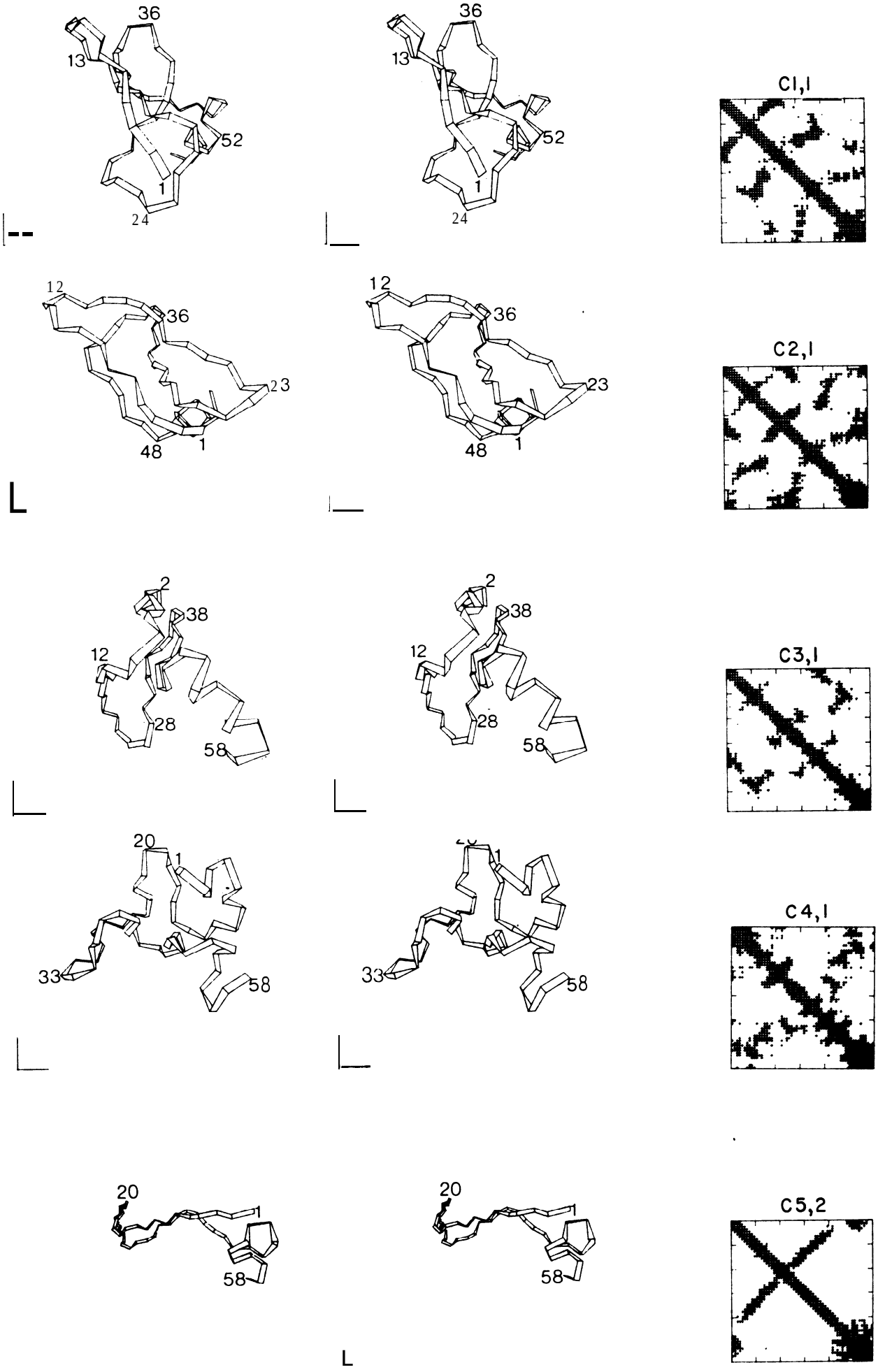


FIG. 15 (caption on page 100)

native S-S bonds can be formed readily. The r.m.s. deviation from the native co-ordinates is 6·4 Å for conformation C2,1.

The solvent interaction term was then omitted completely giving conformation C3,1. Although the a-helix is still well preserved the molecule shows no signs of any β-sheet secondary structure. The conformation is not as compact as before (R_g = 19·2 A), the r.m.s. deviation is 9·1 A, and there is much less resemblance to native PTI than for the two preceeding folding simulations (conformations Cl ,1 and C2,1).

In the next minimization the torsional energy term was omitted giving conformation C4,1. Although the molecule is now compact (R_g = 16·9 A) and well stabilized (energy of −87·6 kcal/mol), the α-helix has broken up and the chain is so kinked as to bear no resemblance to real protein structures. Note the patchy shading on the contact map.

Finally, all the energy terms were included in the energy function but the same torsional potential was used for Gly, Asp, and Asn as for all the other amino acids (except Pro). The minimization no longer reached a compact conformation without passing through stable intermediates. The first minimum energy conformation (C5,1) is very open (R_g = 61·1 A), and even the second minimum energy conformation obtained after thermalization and minimization (C5,2) is still not compact (R_g = 29·4 A). As conformations generated after the same total number of minimization cycles are compared here, the minimization was not continued further than conformation C5,2. In this conformation, the β-hairpin, centred at residue 21, is like that obtained before under very different conditions (B6,2 and **Al** 1 ,1). Further minimization, possibly combined with a pushing potential, could well cause the β-hairpin to bend, giving a more native-like conformation. Note that this p-hairpin is formed without any explicit peptide hvdrogen bond function. The α-helix is much more distorted in conformation C5,2 than in conformations **Cl** ,1 and C2,1, and this cannot be attributed to any differences in the energy functions used. In fact, the disruption of the α-helix occurred when the first minimum energv conformation (C5,1) was perturbed by normal-mode thermalization, and without hydrogen bonds the helix could not be formed again.

(e) *Computing requirements*

Even though the present calculations are about three orders of magnitude faster than folding simulations of PTI using conventional all-atom geometries and energy functions. the demands on computer time are still quite high. For the IBM 370/165 computer' the calculation of the energy and analytical first derivative at a particular conformation of PTI (58 residues) takes 0·6 seconds c.p.u. time without peptide hydrogen bonds and 1·1 seconds with hvdrogen bonds. **A** typical simulation requires about 400 energy evaluations (about 4 min without hydrogen bonds). The additional

FIG. 15. The effect of different energy contributions on folded conformations obtained with the set C van der Waals' energy parameters. With these stronger van der Waals' forces folding is very rapid without the usual open intermediate conformations. Conformation C1,1 was obtained using the normal solvent and torsional energy terms. Conformation C2,1 was generated using modified solvent energy parameters for Pro and Cys: s_i = −1·4 instead of 0·0 kcal/mol, and s_i = −1·0 instead of −1·5 kcal/mol, respectively. Conformation C3,1 was generated without using any solvent energy at all. Conformation C4,1 was generated without using any torsional energy terms. Conformation C5,2 is the second minimum generated using the same torsional potentials for all residues except Pro, i.e. there were no regions of the chain susceptible to bending by virtue of local interactions (conformation C5,2 is half-scale).

time required to construct the chain from standard geometry, diagonalize the energy second derivative matrix, apply the normal-mode perturbations, and output the contact maps is less than ten seconds. For molecules with up to 110 residues, the programs, which are written in FORTRAN and extensively overlayed, require 200K bytes of core store and almost no disc back-up storage space. For this particular computer a typical multiplication and core access takes 2 μs.

4. Discussion

(a) *Validity of the simple approach*

(i) *The chain geometry and energy*

In purely geometrical terms the simplifications used here work well in that the simplified native co-ordinates can be fitted to within 1·0 A r.m.s. The energy function also works well, in that there are true minimum energy conformations less than 3 A r.m.s. deviation from the native co-ordinates. This deviation from the native co-ordinates is worse when torsional, solvent, or van der Waals' energy terms are omitted, but it is relatively insensitive to the presence or absence of peptide hydrogen bonds and to which of the three different sets of van der Waals' parameters is used. Preliminary results show that using the more sophisticated side chain geometry with an additional degree of freedom χ does not lead to near-native energy minima with a much lower r.m.s. deviation. Using rigid side chains that have the actual geometry of the native co-ordinates (not idealized standard geometry) also makes no significant improvement to the r.m.s. deviation. The inability to find near-native minima closer than 2·5 Å r.m.s. deviation from the native co-ordinates stems neither from the use of incorrect energy parameters nor from the use of over-simplified chain geometry. This limitation may be due to the intrinsic lack of side-chain detail: when the side chains are so perfectly spherical, they are not able to interlock and form a stable conformation very close to the native co-ordinates. Were each side chain to contain several interaction centres as in the conventional all-atom treatments, there would probably be minima closer to the native PTI conformation.

Another limitation of the simple model is that the nearest near-native energy minimum does not always have the lowest energy, and that many different minima have rather similar energies (see Fig. 8). Again this seems to be an intrinsic limitation of the simple model; it could only be avoided using a more detailed representation of the side chain interactions. Work on such a representation is in progress.

In spite of these limitations, the calculations done on the native PTI conformations indicate several important advantages of the present model over conventional approaches : (a) all energy calculations are faster by at least two orders of magnitude ; (b) it is much easier to include the effects of solvent, group vibration, and thermal energy perturbations ; (c) the energy surface is much less convoluted with far fewer energy minima ; and (d) with fewer variables and analytical first derivatives minimization is faster by an order of magnitude.

(ii) *Validity of folding simulation*

Many of the energy minimization runs reported here started from very open, generally structureless, chain conformations. In most runs with a preset terminal a-helix the chain rapidly folded into a compact conformation having the size, shape, β-sheet secondary structure? hairpin turns, and a-helix on to β-sheet packings so

characteristic of native protein conformations. The other unsuccessful runs (about 30%) did not reach a sufficiently compact shape after about 400 cycles of minimization, and they were abandoned.

More than half of the successful simulation runs ended in conformations that were within 6·5 A r.m.s. deviation of the native PTI co-ordinates as determined by X-ray crystallography. When viewed in stereo from the same orientation, the folded and actual native conformations have many structural features in common. The folding simulations succeed equally well with the three different sets of van der Waals' parameters used, and are also not sensitive to the details of the starting conformation. Nevertheless, with the stronger set C van der Waals' energy term, folding occurs about twice as rapidly, and without passing through any stable non-compact intermediates. Omitting the hydrogen bond energy has little effect on the folding from an open conformation. Omitting the solvent or torsional energy terms is more serious; the resulting conformations have little resemblance to PTI or proteins in general in that there are no straight chain segments packing parallel to each other. Other factors that seem to make the correct folding more rapid include: pre-formation of secondary structure (e.g. the C-terminal a-helix of PTI); the general stiffness of the polypeptide chain; and the presence of residues that favour a bend in the chain by virtue of local interactions. The combination of general chain stiffness and increased flexibility at special turn-promoting residues makes the torsion angles at the bends highly effective : they allow the energy to drop by the formation of *a* p-hairpin. Lewis et *al.* (1971) considered β-bends to play an important role in bringing together distant chain segments. Walton (1973), on the other hand, considered that these bends formed when the chain was forced to fold due to the favourable interactions between the chain segments. In the present calculations it seems that the chain is forced to bend by long-range interactions but that these bends do occur more readily at specially flexible regions of the chain.

Assuming that the energy drops 30 kcal/mol during the simulated folding in which each side chain sphere moves about 100 A and, applying Stokes' law to the movement against the viscous drag of water, gives a folding time of 3×10^{-9} seconds. (The validity of the use of this law at the molecular level is confirmed by the similarity of the measured bulk viscosity of water and the value calculated from the coefficient of diffusion using Stokes' law.) While such rapid folding is very convenient computationally, it may be too fast to give a true picture of the real folding process. With slower folding, some stable secondary structure could possibly form before the collapse to a compact conformation.

During the folding from an open chain conformation, stable, but not very compact, intermediate conformations were often generated. Most of these intermediate minimum energy conformations had some well defined secondary structure in the form of p-hairpins and other p-sheets that were not interacting strongly with the preformed ***a-helix.*** The p-hairpins, which formed spontaneously, had the same right-handed twist found in real p-hairpins. Often there was a third strand extending the hairpin into a sheet by either a parallel or anti-parallel association.

Although the net forces on these intermediate conformations were zero, the minima were not very deep; it was relatively easy to escape from these minima using normal-mode thermalization. Normal-mode thermalization works well in the denaturation of these intermediates as it uses the low frequency "breathing" vibrations to cause a large deformation of the conformation for a small increase in energy.

When the conformation becomes more compact, the quadratic approximation used to calculate the normal-modes is less valid. In these cases the best way to escape from a minimum is to add an artificial pushing potential to the energy of the molecule. Although pushing the conformation away from the previous minimum is less physically realistic than normal-mode thermalization, it can be likened to giving the conformation at the previous minimum a certain kinetic energy which is gradually used up in the movement of the side chains from their Cartesian positions at that minimum.

Technically, the combined use of energy minimization, normal-mode thermalization, and pushing potentials works well in that the chain is made to fold rapidly from an open to compact conformation. Often the conformation is made to change by 100 A r.m.s. in Cartesian co-ordinates and 100° in torsion angle co-ordinates after less than 400 evaluations of the energy function and its first derivatives.

(b) *General implications of the folding simulations*

That such an over-simplified model works at all well in the folding simulations of PTI suggests that certain features of the model may be relevant to the real *in vivo* protein folding process. The native conformations of proteins as studied by X-ray diffraction are very highly ordered structures stabilized by interactions of very precise geometry: all possible interior hydrogen bonds are well formed and the non-polar side chains form a close-packed hydrophobic core. In fact, the atomic geometry inside native proteins resembles that found inside crystals of the individual amino acids (Chothia, 1975). It is almost *as* if the right numbers of each amino acid had co-crystallized into a globular protein micro-crystal. The forces responsible for such precise arrangements fall off rapidly with distance and improper orientation. The simplified protein conformations obtained by the present simulated folding calculations are stabilized by very different types of forces, which are spherically symmetrical and fall off slowly with distance. Such forces would be like the average forces between side chains in rapid thermal motion at high temperatures as then the fine details of the atomic structure would become blurred.

One way to reconcile the success of the simple forces used here with the precision of actual native protein conformation is to imagine a two-stage folding process. Initially, when the chain conformation is open, the side chains would move freely and the forces would be like those used here. The present calculations show that such averaged forces are able to cause the chain to fold rapidly from an open to compact conformation in which the residues are quite close to their final spatial positions (say to within 6 A r.m.s.). As the molecule becomes more compact, the thermal motion of the side chains would become more restricted due to interactions with the other side chains. At this stage, the second in the hypothetical process, detailed atom-atom interactions would come into effect. The conformation would become much more ordered as the favourable enthalpy gained from precise hydrogen bonds and van der Waals' contacts compensated for the energetically unfavourable decrease of the chain entropy. The latter process would be akin to crystallization with each residue falling into place precisely from its nearby position in the approximately folded conformation. In the present calculations these initial approximately folded conformations are labile and can readily move between the various neighbouring minima. There are many minima of low energy near the native conformation (within 6 Å

(The same functional form is used in the calculation of solvent interaction energy.) Typically the maximum value, K_{push}, was taken as 100 kcal/mol, and σ_{push} was set to a value between 5 and 15 Å that was big enough to generate a new minimum energy conformation sufficiently different from the previous one. Such a pushing constraint works much better on Cartesian co-ordinates than on the torsion angles (i.e. using RMS (α_i)), as each torsion angle affects the conformation differently, with some angles causing almost no change in conformation and others causing very big changes. Gibson & Scheraga (1969) also considered modifications of the energy function as a way to escape from local minima but they worked in torsion angle co-ordinates and used a pushing function that never dropped to zero.

3. Results

(a) Idealized native conformation

The simplified representation of protein geometry and the corresponding potential energy function were first tested on the native conformation of a small protein. Bovine pancreatic trypsin inhibitor was chosen as it is the only protein with less than 100 amino acids that does not have any prosthetic group. The native co-ordinates of PTI† as determined by X-ray crystallography (Huber et *al.*, 1970,1971; Deisenhofer & Steigemann, 1974) were averaged to give only three centres per residue: the peptide group centroid, the side chain centroid, and the C^{α} atom. These simplified native co-ordinates, also referred to as the native co-ordinates, are used throughout this work to assess the adequacy of the methods.

The rigid side chain model was tested first. The co-ordinates of a simplified polypeptide chain having the sequence of PTI **and built up** from the standard **bond** lengths, bond angles, and torsion angles (Table 1) were calculated using the actual *a* torsion angles of the simplified native structure. As the simplified standard chain geometry differs from that of the native structure, these calculated co-ordinates were not close to the native structure (see Fig. 7). The crystallographically determined co-ordinates of PTI had been extensively refined (Deisenhofer & Steigemann, 1974), and although all bond lengths and most bond angles were kept at standard values, both the peptide planarity and the bond angle at the C^{α} ($\tau(C^{\alpha})$) had been changed in the refinement. While this may explain part of the deviation from the native structure of the co-ordinates calculated above with the native *a* angles, it is well known that molecular Cartesian co-ordinates depend very sensitively on the exact values of the backbone torsion angles. In order to bring the structure with standard geometry closer to the simplified native structure, the *a* torsion angles were changed to minimize the r.m.s. deviation:

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

where r_{ij} is the distance between side chain centres i and j in the calculated idealized co-ordinates and r_{ij}^{obs} is the corresponding distance in the simplified native co-ordinates (Nishikawa *et al.*, 1972). This meaure of deviation does not depend on the relative orientation of the two sets of co-ordinates; it is related to the conventional r.m.s. deviation of atomic positions between two optimally oriented sets of co-ordinates by $\mathrm{RMS}\,(r_{ij}-r_{ij}^{obs}) = \sqrt{(\frac{2}{3})}\,\mathrm{RMS}\,(r_i - r_i^{obs})$.

The conformation with standard geometry that best fitted the simplified native conformation deviated by only 1·2 A r.m.s. This best fit conformation, which is used

† **Abbreviations used: PTI, pancreatic trypsin inhibitor; r.m.s., root-mean-square.**

r.m.s.) all with similar energies. Because the different amino acid side chains are more similar to one another in the simple model than in the detailed all-atom structures, many different sequences could give the same approximate fold. Nevertheless, only a small sub-set of these sequences could make enough detailed interactions to overcome the high chain entropy and so form a precise native conformation.

The question of whether secondary structure is formed before the tertiary structure of proteins has been considered often. In many hypothetical schemes for protein folding the secondary structure forms first and these rigid chain segments then pack together to give the final tertiary structure (Levinthal, 1966 ; Anfinsen, 1973; Ptitsyn, 1973). Most attention has focussed on the early formation of α-helices, but as many proteins have little or no a-helix, p-hairpins must also be considered as secondary structure units that form first (Nagano, 1974). In many of the PTI folding simulations presented here there are stable intermediate conformations with some well-defined β-hairpin secondary structure. The a-helix, which was set in the starting conformations, was also stable in the open intermediate conformations. The two-stage folding scheme outlined above, in which the chain first folds approximately and then in more detail, can also be applied to the early formation of secondary structure. The transition between the use of a simple force-field and the use of a detailed force-field need not occur at the same time for all residues. Once any local region of the chain has become more compact forming either an a-helix or p-hairpin, the detailed forces should become operative and stabilize this pre-formed secondary structure. It is very difficult to simulate this gradual transition from a simple to a detailed force-field, but pre-formed secondary structures can be incorporated by freezing regions of the chain that fold first. This is part of the rationale behind setting the C-terminal a-helix in some of the folding simulations of PTI. Pre-formed a-helices are used more extensively in folding simulations of carp myogen (Warshel & Levitt, 1976).

Although β-sheets form readily in the present simulations, a-helices have never formed spontaneously. One cause of this may be that the a-helix is greatly stabilized by detailed interactions omitted in the simple model. Ways of correcting the force-field to enable a-helices to form spontaneously are currently under consideration.

(c) *Other applications of the model*

With the present simple protein representation it becomes computationally much easier to study various properties of native protein conformations. In such studies, it is best to use the actual simplified X-ray co-ordinates of the protein, with each simple side chain sphere at the centroid of the real all-atom side chain. Questions that could be answered include: (a) what is the stability of the native protein conformation, and of its secondary structure sub-assemblies ; (b) what is the pathway of most rapid denaturation ; (c) can one predict the way in which two protein molecules will interact; (d) what is the initial mode of binding of a freely rotating substrate and is there a "funnel" directing the substrate into its precise orientation in the enzyme active site. In each of these problems the computed results would be subject to the 2 A limit of accuracy associated with the simple spherical side chains, but the energy surface would be simple enough to avoid the many false minima of more detailed all-atom treatments.

An exciting application of the model would be the predication of the conformation of an unknown protein. One approach would be to repeatedly simulate the folding

until one obtained a compact low energy conformation that had reasonable properties. Another, possibly more productive, approach would be to use the available experimental data as follows: (a) set the secondary structure from one or a combination of the various prediction rules (cf. Schulz et *al.*, 1974) ; (b) require the native S-S bonds to be formed ; (c) require that a particular S—S bond be formed first as happens in PTI (see Creighton, 1974,1975); (d) use the orientation of rigid parts of the molecule such as α-helices as determined from low-resolution X-ray or electron microscope studies (as in tobacco mosaic virus, see Champness *et al.*, 1976; or in purple membrane protein, see Henderson & Unwin, 1975); (e) use known group reactivities and the proximity of certain pairs of side chains obtained from chemical studies. The computer can then be used to give a model of the protein which satisfies any or all of the above constraints as well as the geometric and energetic requirements.

Computing facilities were generously provided by the Weizmann Institute and Cambridge University Computer Centres. During the course of this work I was supported by the European Molecular Biology Organization and the **Medical Research Council.**

REFERENCES

Adams, M. J., Baker, E. N., Blundell, T. L., Harding, M. M., Dodson, E. J., Hodgkin, D. C., **Dodson, G. G., Rimmer, B.,** Vijayan, M. & Sheats, S. **(1969).** *Nature* (London), *224,* **491-495.**

Anfinsen, **C. B. (1973).** Science, 181, **223-230.**

Anfinsen, C. B., Haber, E., Sela, M. & White, F. H. (1961). *Proc. Nat. Acad. Sci., U.S.A.* 47, **1309-1314.**

Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. **E. Jr, Richardson, D.** C., **Richardson, J. S. & Yonath, A. (1971).** *J. Biol. Chem. 246,* **2302-2316.**

Birktoft, J. J. & Blow, D. M. (1972). *J. Mol. Biol.* 68, 187-240.

Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *Proc. Roy. Soc. ser. B,* 167, 365-385.

Burgess, A. W. & Scheraga, **H. A. (1975).** *Proc. Nat. Acad. Sci., U.S.A.* **72, 1221-1225.**

Carter, C. W. Jr, Kraut, J., Freer, S. Y ., **Xuong, Ng.H., Alden, R. A.** & Bartsch, **R.** G. **(1974).** *J. Biol. Chem. 249,* **4212-4225.**

Champness, J. N., Bloomer, A. C., Bricogne, G., Butler, P. J. G. & Klug, **A. (1976).** *Nature (London), 259, 20-24.*

Chothia, C. (1974). *Nature (London), 248, 338-339.*

Chothia, *C.* **(1975).** *Nature (London), 254, 304-308.*

Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* 70, 701–724.

Creighton, T. E. (1974). *J. Mol. Biol.* ***87, 603-624.***

Creighton, T. E. (1975). *J. Mol. Biol. 95,* 167–199.

Deisenhofer, J. & Steigemann, W. (1974). *Acta Crystallogr. sect. B,* 31, 238-250.

De Santis, P. & Liquori, **A. M. (1971).** *Biopolymers,* 10, **699-710.**

Diamond, R. (1965). *Acta Crystallogr.* 19, 774-789.

Drenth, J., Jansonius, J. M., Koekoek, R. & Wolthers, B. *G.* **(1971).** *Advan. Protein Chem. 25,* **79-115.**

Edelman, G. M., Cunningham, B. A., Reeke, G. N. Jr, **Becker, J. W., Waxdal, M.** J. & **Wang, J. L. (1972).** *Proc. Nat. Acad. Sci., U.S.A. 69,* 2580-2584.

Eyring, H. (1932). *Phys. Rev. 39,* **746-751.**

Fletcher, R. (1970). *Computer J.* 13, 317-322.

Flory, P. J. (1969). In *Statistical Mechanics* ***of*** *Chain Molecules,* **pp. 248-306, Wiley, New York.**

Gelin, **B. R. & Karplus, M.** *(* **1975).** *Proc. Nat. Acad. Sci., U.S.A.* 72, 2002-2006.

Gibson, K. D. & Scheraga, H. A. (1967). *Proc. Nat. Acad. Sci., U.S.A. 58,* 420-427.

Gibson, K. D. & **Scheraga, H. A. (1969).** *Proc. Nat. Acad. Sci., U.S.A.* 63, **9-15,**

Hagler, A. T., Huler, E. & Lifson, S. (1974). *J. Amer. Chem. Soc. 96,* **5319-5326.**
Hardman, **K. D. & Ainsworth, C. F. (1972).** *Biochemistry,* 11, 4910-4919.
Henderson, R. & Unwin, P. N. T. (1975). Nature (London), 257, 28-32.
Hermans, **J. Jr & McQueen, J. E. Jr (1974).** *Acta Crystallogr. sect. A,* **30,** 730–739.
Hill, T. L. (1956). In *Statistical Mechanics,* **pp. 15-17, McGraw-Hill, New York.**
Huber, R., Kukla, D., Ruhlmann, A. & Steigemann, W. (1970). In Proceedings of the *International Research Conference on Proteinase Inhibitors,* **Munich** 1970 **(Fritz, H. & Tschesche, H., eds), pp. 56-64, Walter de Gruyter, Berlin.**
Huber, R., Kukla, D., Ruhlmann, A. & Steigemann, W. (1971). *Cold Spring Harbor Symp. Quant. Biol. 36,* **141-150.**
Kretsinger, R. H. & Nuckolds, **C. E. (1973).** *J. Biol. Chem.* **248, 3313-3326.**
Lee, B. & Richards, F. M. (197 1). *J. Mol. Biol.* **55,** 379–400.
Levinthal, *C. (1966). Scientific Amer. 214,* **no. 6, 42-52.**
Levitt, M. (1972). Ph.D. Thesis, Cambridge University.
Levitt, M. *(1974a). J. Mol. Biol. 82,* **393–420.**
Levitt, M. (1974b). In *Peptides, Polypeptides and Proteins* **(Blout, E. R., Bovey,** F. **A., Goodman, M. &** Lotan, **N., eds), pp. 99–113, Wiley, New York.**
Levitt, M. & Lifson, S. (1969). *J. Mol. Biol. 46, 269-279.*
Levitt, M. & Warshel, A. (1975). *Nature (London), 253, 694-698.*
Lewis, P. N., Momany, **F. A. & Scheraga, H. A. (1971).** *Proc. Nat. Acad. Sci., U.S.A.* **68, 2293-2297.**
Lewis, P. N., Momany, **F. A. & Scheraga, H. A. (1973).** *Isr. J. Chem.* 11, **121-138.**
Liquori, **A. M., De Santis, P., Kovacs, A. L. & Mazzarella, L. (1966).** *Nature (London),* **211,** 1039–1041.
McLachlan, **A. D. (1972).** *Nature New Biol.* **240, 83-85.**
Momany, **F. A., Vanderkooi, G., Tuttle, R. W. & Scheraga, H. A.** *(1969). Biochemistry,* **8, 744-746.**
Nagano, K. (1974). *J. Mol. Biol. 84, 337-372.*
Nishikawa, K., Ooi, T., Isogai, **Y. & Saito,** *N.* **(1972).** *J. Phys. Soc. Japan,* **32, 1331-1337.**
Nozaki, Y. & Tanford, **C. (1971).** *J. Biol. Chem.* **246, 2211-2217.**
Padlan, E. A. & Davies, D. R. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72, 819-823.**
Phillips, D. *C. (1970).* **In** *British Biochemistry, Past and Present* **(Goodwin, T. W., ed.), pp. 11-28, Academic Press, London.**
Platzer, K. E. B., Momany, **F. A. & Scheraga, H. A. (1972).** *Int. J. Peptide Protein Res.* **4,** 201–219.
Ptitsyn, *0.* **B. (1973).** *Vestnik Akad. Nauk SSSR, 5, 57-68.*
Ptitsyn, 0. B. & Rashin, **A. A. (1973).** *Dokl. Akad. Nauk SSSR,* 213, 473-475.
Ptitsyn, 0. B. & Rashin, **A. A. (1974).** *Biophys. Chem. 3,* **l-20.**
Quiocho, F. A. & Lipscomb, W. N. (1971). *Advan. Protein Chem.* **25, l-78.**
Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol. 7, 95-99.*
Richards, F. M. (1974). *J. Mol. Biol. 82,* **1-14.**
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. & Nagano, K. (1974).** *Nature (London),* **250,** 140–142.
Simon, E. M. (1971). *Biopolymers,* 10, 973-989.
Vanderkooi, G., Leach, S. J., Nemethy, G., Scott, R. A. & Scheraga, H. A. (1966). Bio-*chemistry, 9,* **2991-2997.**
Venkatachalam, *C.* **M. (1968).** *Biopolymers, 6,* **1425-1436.**
Walton, A. *G. (1973). Croatica Chemica Acta,* **45, 59-66.**
Warme, **P. K. & Scheraga, H. A. (1974).** *Biochemistry,* 13, 757-767.
Warshel, A. & Levitt, M. (1976). *J. Mol. Biol.* **In the press.**
Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1973). *Acta Crystallogr. sect. B, 29, 943-956.*
Wright, C. S., Alden, R. A. & Kraut, *J.* **(1969).** *Nature (London),* 221, 235-242.
Wyckoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B. & Richards, F. M. (1970). *J. Biol. Chem.* **245,** 305–328.

Note added in proof: The assignment of the Gly-like torsional potential to both Asp and Asn (Materials and Method, section (b) (iv)) was made by counting how often each type of amino acid occurred in a turn. Residue $i+1$ was defined as in a turn if the torsion angle α_i (defined by the C^α values of residues $i-1$, i, $i+1$, and $i+2$) was between $-90°$ and $90°$ and the residue was not part of α-helix or β-sheet secondary structure. Originally this frequency in turns was counted from 11 known protein conformations, but the frequencies recently found in 28 independent unrelated proteins (Levitt & Greer, 1976, manuscript in preparation) confirm that Gly, Asp and Asn occur most frequently in turns. The normalized turn frequencies of the 20 amino acids (random $=1{\cdot}0$) are as follows: Ala, 0·67; Val, 0·49; Leu, 0·69; Ile, 0·53; Cys, 1·0; Met, 0·52; Pro, 1·12; Phe, 0·97; Tyr, 1·20; Trp, 0·51; Asp, 1·60; Asn, 1·32; Glu, 1·12; Gln, 0·81; His, 0·82; Ser, 1·21; Thr, 0·91; Arg, O-73; Lys, 0·97; Gly, 1·82.

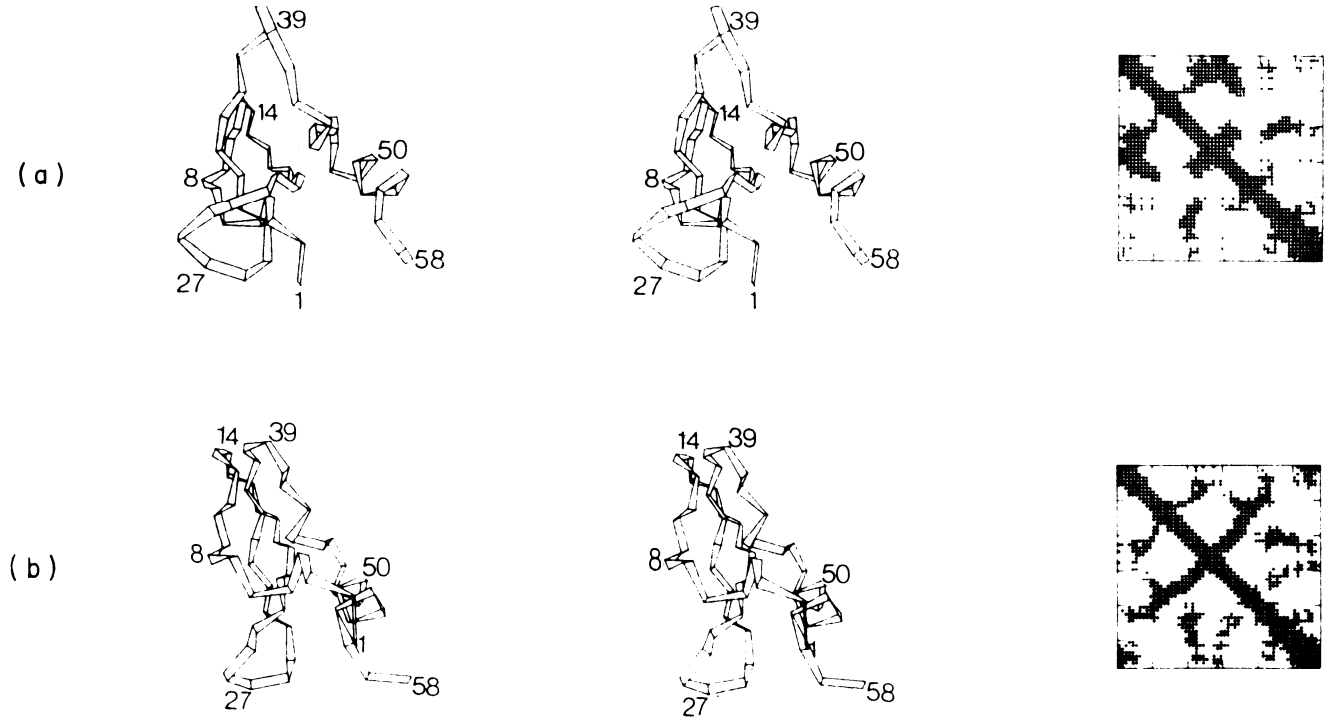


FIG. 7. Showing how the conformation generated with standard geometry and the native α torsion angles (a) differs from the idealized native conformation (b), which has other α angles chosen to get as close as possible to the simplified native conformation of PTI. The r.m.s. deviations (see text for definition) from the simplified native conformation are 4·3 Å and 1·2 Å for (a) and (b) respectively. While the local structure of conformation (a) is similar to that of (b), the fit is much worse for residues further apart along the sequence. The stereo ribbon drawings showing the path of the C^{α} backbone were drawn using a program kindly supplied by Dr A. D. McLachlan and Mr P. Barber. Each conformation has been rotated to superimpose on the native PTI conformation, and, unless otherwise stated, each is drawn to the same scale. The contact maps (Phillips, 1970; Nishikawa *et al.*, 1972) show those pairs of residues that are closer together than 10 Å in each conformation. A plus sign at the (i, j)th position of the map indicates that residues i and j satisfy this criterion (i increases from top to bottom, and j increases from left to right). In these maps secondary and tertiary structure features clearly. α-helix (e.g. at the C-terminal of PTI) shows up as a broad diagonal band; anti-parallel β-sheet (e.g. residues 16 to 35 of PTI) shows up as a band running perpendicular to the diagonal; and parallel β-sheet (not present in PTI) shows up as a band parallel to but displaced from the diagonal of the map. The similarity of the local structure in conformations (a) and (b) is also evident in the contact maps : the pattern off the diagonal is much less well preserved than that on the diagonal.

as the starting point of some of the energy minimization runs to be described later, is known as the idealized native conformation. The C^{α} ribbon drawing of this conformation is almost indistinguishable from that of the actual native co-ordinates. Attempts to fit the chain without any dependence of τ_i on α_i gave a worse deviation of 1·6 Å r.m.s. The fit of only the C^{α} positions with the τ_i/α_i dependence was very good at 0·4 Å r.m.s.

The flexible side chain model also was fitted to the simplified native co-ordinates in this way. Standard geometry from Table 2 was used to construct the chain, which had as additional variables the side-chain torsion angles χ. In this case the best fit deviation was 0·9 Å r.m.s., showing only a slight improvement over the model with rigid side chains.

(b) *Energy minimization from the native structure*

In the previous section it was shown that a simplified chain geometry can fit the native conformation well; here the energy of the two idealized native structures is

minimized? one with rigid side chains, and the other with some flexible side chains. The aims of this approach were: (a) to find a stable calculated conformation as near the experimental conformation as possible. If one cannot find a near-native minimum when starting from the native conformation, there is little point in trying to do so from an open "denatured" conformation. (b) To assess the effects of the different energy contributions on the deviation of the calculated near-native minimum from the native co-ordinates. This should indicate what energy terms are most important, whether certain contributions can be safely omitted and whether a better fit to the native co-ordinates could be obtained with different values of the parameters.

(i) *Holding potentials*

Energy minima close to the native co-ordinates had to be found using a holding potential that constrained the conformation to the native co-ordinates until the energy had relaxed sufficiently. Without these potentials, minimization starting from the idealized native conformation behaved unpredictably with the different sets of van der Waals' energy parameters. In some cases the minimum found differed by 3 Å r.m.s. from the native conformation, while in other cases the deviation was as much as 7 Å. It was felt that such extreme differences did not truly reflect the validity of the energy parameters, and noticed that the r.m.s. deviation from the native co-ordinates was greater when the minimization started from a higher energy. The minimization method used here seems to behave in some ways like molecular dynamics in a vacuum: those runs (or trajectories) that start with a high potential energy move away from the initial conformation so rapidly that they skip over the closest minima. Pure energy minimization should set the momentum to zero at each step and should never have enough kinetic energy to skip over minima. It is not clear why the method used, VA09D, does not work like this.

When a holding potential is used: the conformation is constrained to stay close to the native co-ordinates. By repeating such constrained minimization several times with a constraint that is progressively less strong, the conformation can relax yet still stay close to the native co-ordinates (r.m.s. deviation less than 1·5 Å). Once the energy of the conformation before the start of a minimization run falls below a threshold value, the holding potential is removed and the final pass of minimization is completely unconstrained. The conformations shown in Figure 8 demonstrate the importance of this constrained pre-relaxation if one is to regard the r.m.s. deviation from the native conformation as a measure of the usefulness of the energy parameters. When the conformation is allowed to relax more before removing the holding potential, the final r.m.s. deviation is lower. The existence of several different, low energy conformations all within 6 Å r.m.s. deviation of the native structure will be discussed later in the context of limitations of the simple representation of the side chain interactions.

(ii) *Variation of the energy function*

In the previous section a consistent way to find a near-native energy minimum as close as possible to the native co-ordinates for any starting conformation and energy functions was established; here three sets of van der Waals' parameters are tested. omitting various energy contributions? and monitoring the effect on the resulting near-native minimum energy conformations. Most of the work was done with the more simple rigid side chain model,

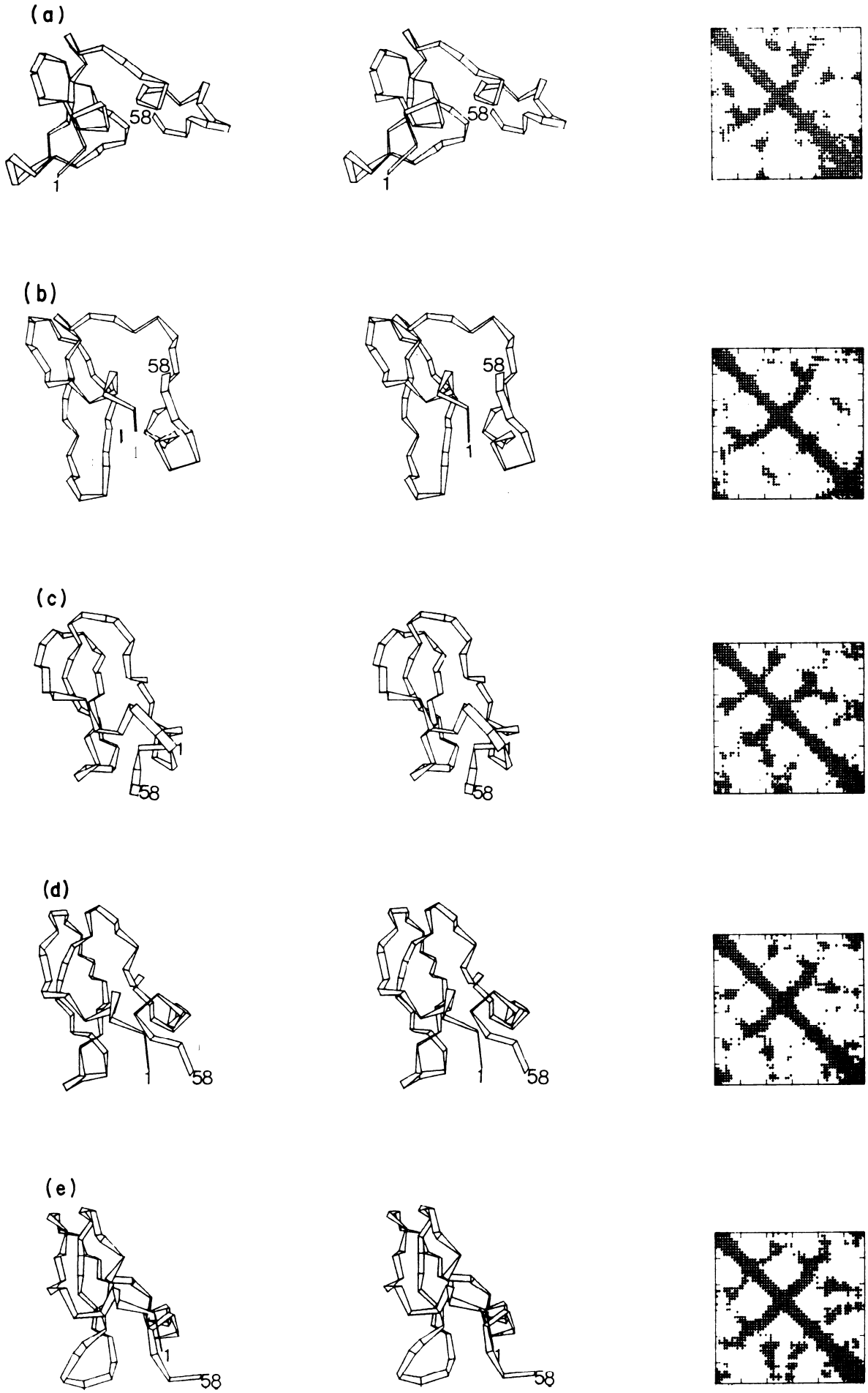

(a)
58
1
58
1
(b)
58
1
58
1
(c)
1
58
1
58
(d)
1
58
1
58
(e)
1
58
1
58

Table 6 shows the energy contributions and r.m.s. deviations of conformations obtained using the set B and set C van der Waals' parameters sometimes omitting torsional, solvent, and S-S bond energy terms. With the set C van der Waals' parameters, the smallest deviation from the native co-ordinates (2·75 Å r.m.s.) was obtained with all the energy terms apart from the S-S bond energy. With the S-S bond energy, the deviation is worse (3·51 A r.m.s.), and the van der Waals' energy contribution becomes less favourable. Without either the torsional or solvent energy, the r.m.s. deviation is also worse (3·81 Å and 3·43 A, respectively). Omitting the van der Waals' energy term completely has a more drastic effect (r.m.s. deviation of 9·46 A). The near-native minimum obtained with the full C parameter energy function is a little more swollen than the native conformation (radii of gyration (R_g) = 17·1 Å and 16·2 A, respectively). Without torsional forces the molecule becomes more tightly packed ($R_g = 15{\cdot}9$ A) ; while without van der Waals' forces, there is a total collapse due to the strong unopposed solvent interaction forces ($R_g = 9{\cdot}3$ A).

The energy values in Table 6 show the detailed balance expected when using convergent minimization of a multi-component energy function. For example, when either the torsional or solvent energy is omitted, the van der Waals' energy is able to reach a lower value; when there are extra energy terms like S-S bonds or helix rigidity, the van der Waals' energy is forced to a higher value. Each term in the energy function has a different conformational requirement: the solvent energy is very low when the molecule is completely collapsed, the torsional energy is low when the chain is extended and the conformation very open, and the van der Waals' energy is most favourable somewhere in between these extremes. The true native conformation is maintained by a delicate balance of these opposing tendencies.

With the set B van der Waals' parameters, the near-native minimum deviates by only 2·57 A r.m.s. from the simplified native co-ordinates. As the set B side chains are smaller than for set C, the van der Waals' energy is higher and the solvent energy lower than in the corresponding set C near-native minimum. The set B near-native minimum is also less swollen with a radius of gyration close to that of the native co-ordinates ($R_g = 16{\cdot}0$ Å).

Because the simplified native and idealized native conformations have not been allowed to relax, their energy values are generally higher than those of the near-native minima. The solvent energy is an exception with values of -34 kcal/mol in the native structure and −27·7 kcal/mol in the C set near-native minimum. The backbone becomes much less strained after minimization as the torsional energy drops from 38 kcal/mol to 22 kcal/mol.

Figure 9 shows some of the minimum energy conformations obtained with set C

FIG. 8. Showing how the conformation generated by minimization from the idealized native conformation is affected by the starting energy and bad contacts in the initial conformation. Set **C van der Waals' parameters, torsional forces, and solvent interactions are used in each run.** The conformation was constrained to be close to the simplified native co-ordinates until the energy had dropped below a selected threshold value. For a lower selected energy value, more **complete passes of the constrained pre-relaxation** were **needed. Once the energy dropped below** the threshold value, minimization was restarted without any constraint whatsoever. The conformations shown here were generated as follows. (a) Without any pre-relaxation; the starting energy was over 1000 kcal/mol. (b) Unconstrained minimization from a starting energy of 400 kcal/mol. (c) Starting energy of 67 kcal/mol. (d) Starting energy of — 37 **kcal/mol.** (e) The idealized **native conformation shown for** comparison. The **r.m.s.** deviations from the native **PTI co-ordinates** are **6·35 Å, 4·80 Å, 3·10 Å, 2·54 Å,** and **1·2 Å,** respectively. The final energies are -63, -65, -68, — 62, and **1111** kcal/mol, respectively.

TABLE 6

Minimization using sets B and C van der Waals' parameters

Conformation and conditions	Energy contribution† (kcal/mol) Total	Torsion	H-bond	van der Waals'	Solvent	R_g‡ (Å)	r.m.s. deviation $\Delta\alpha_i$ (deg.)	Δb_{ss} (Å)	Δr_{ij} (Å)
Rigid side chains									
Initial									
Simplified native PTI, B	18·9	30·7	−16·5	30·2	34·0	16·2		0·5	0·0
Simplified native PTI, C	259·9	30·7		363·2	−34·0	16·2		0·5	0·0
Idealized native PTI, B	230·8	38·1	14·9	212·1	−34·3	16·1	0·0	0·3	1·3
Idealized native PTT, C	1111·5	38·1	--	1107·7	−34·3	16·1	0·0	0·3	1·3
Minimized §									
B (with H-bonds)	−67·6	21·9	−27·7	−21·1	−40·6	16·0	33·2	2·0	2·57
C	61·2	22·3		−58·8	−24·7	17·2	33·4	4·4	2·75
C, no torsion	−99·2	—		−63·5	−35·7	15·9	59·7	6·4	3·81
C, no van der Waals'	−355·0	43·0			-398-O	9·3	97·8	9·3	9·46
C, no solvent	−45·4	19·9		−65·3		16·9	39·5	4·2	3·43
C, S-S bonds	−54·3	24·3		−54·0	−24·6	16·8	51·2	0·0	3·51
C, rigid helix	−52·6	27·8		−55·5	−24·9	16·5	28·8	4·9	2·89
Flexible side chains									
Initial, idealized	47·5	26·6	−17·4	68·8	−30·5	16·0	0·0	0·8	1·00
RI inimized, and H-bonds	−145·1	28·3	−29·8	-92-o	−51·7	14·8	35·5	2·4	3·06
Non-idealized starting‖									
B	−87·6	22·5	−30·4	−20·0	−59·6	16·1		2·0	2·46
C	−80·5	22·1		−67·3	-- 33·3	16·2		3·9	2·72

-f-Force field BC and C use different van der Waals' parameters. Hydrogen bonds are not included in the C field, which has correspondingly bigger side chains.

‡ The radius of gyration, R_g, is calculated as $\left\{\frac{1}{N}\sum_{i,j}^{N}(r_{ij})^2\right\}^{\frac{1}{2}}$ where r_{ij} is the separation of side chain centres i and j.

§Minimization was started at the idealized native conformation.

‖ Jn this case minimization was started from the actual simplified PTI co-ordinates which had non-idealized bond lengths, bond angles and torsion angles.

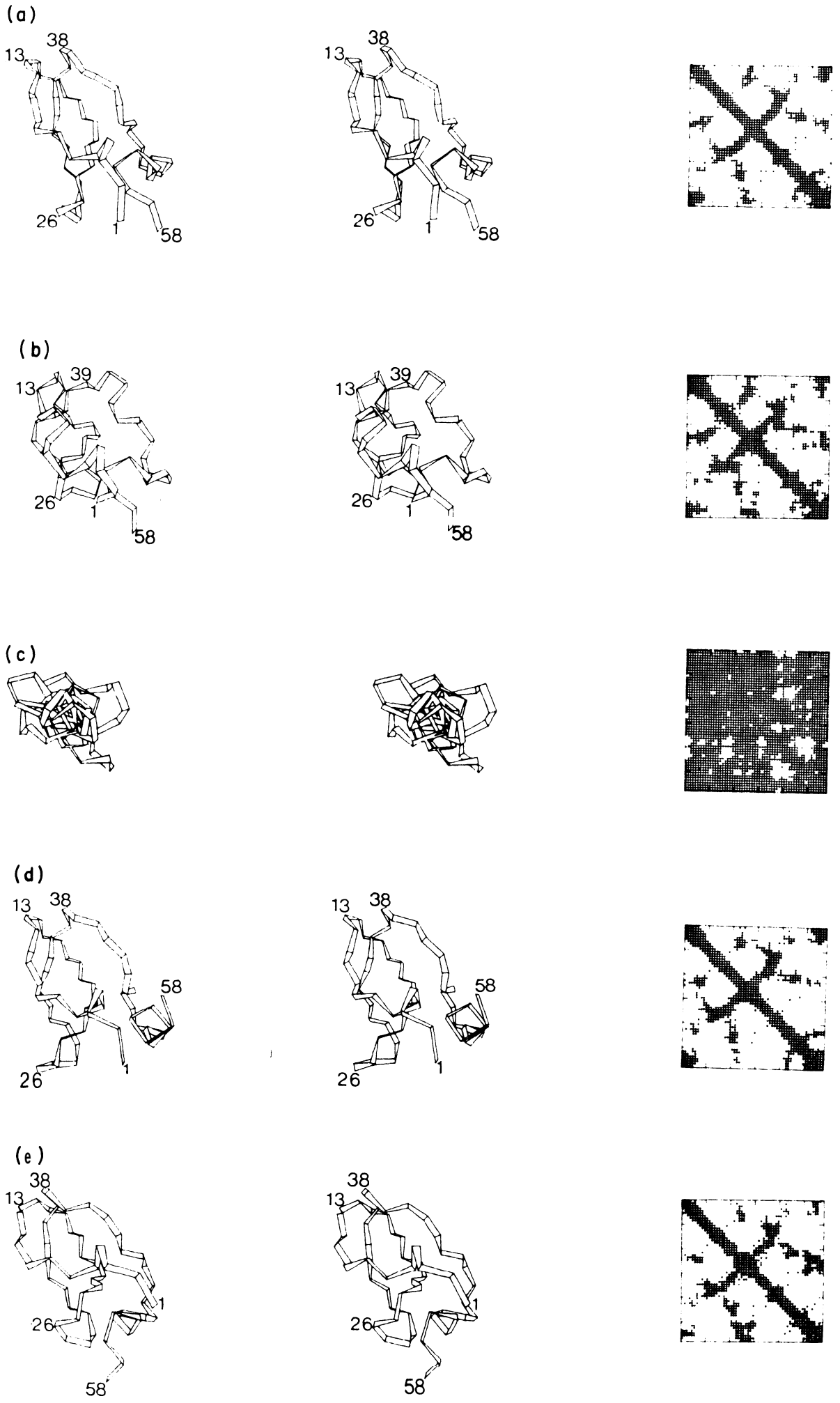


FIG. 9 (caption on page 84)