Protein Science (1992), *1*, 625–640. Cambridge University Press. Printed in the USA.

Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin β_4

MANFRED J. SIPPL, MANFRED HENDLICH, AND PETER LACKNER

Institute for General Biology, Biochemistry & Biophysics, Department of Biochemistry, University of Salzburg, Hellbrunnerstraße 34, A-5020 Salzburg, Austria

(RECEIVED October 25, 1991; ACCEPTED December 20, 1991)

Abstract

Recently we developed methods for the construction of knowledge-based mean fields from a data base of known protein structures. As shown previously, this approach can be used to calculate ensembles of probable conformations for short fragments of polypeptide chains. Here we develop procedures for the assembly of short fragments to complete three-dimensional models of polypeptide chains.

The amino acid sequence of a given protein is decomposed into all possible overlapping fragments of a given length, and an ensemble of probable conformations is calculated for each fragment. The fragments are assembled to a complete model by choosing appropriate conformations from the individual ensembles and by averaging over equivalent angles. Finally a consistent model is obtained by rebuilding the conformation from the average angles. From the average angles the local variability of the structure can be calculated, which is a useful criterion for the reliability of the model.

The procedure is applied to the calculation of the local backbone conformations of myoglobin and lysozyme whose structures have been solved by X-ray analysis and thymosin β_4, a polypeptide of 43 amino acid residues whose structure was recently investigated by NMR spectroscopy. We demonstrate that substantial fractions of the calculated local backbone conformations are similar to the experimentally determined structures.

Keywords: knowledge-based prediction; molecular force field; protein folding; protein modeling; statistical mechanics

The calculation of polypeptide and protein conformations from amino acid sequences belongs to the most interesting problems in molecular biology. A wealth of information on protein systems has accumulated from experimental and theoretical studies but in spite of enormous efforts the problem is still unsolved.

Folding and unfolding of many protein chains are reversible processes that can be induced by changing the physical parameters of the environment. This has been demonstrated 30 years ago on ribonuclease (Anfinsen, 1973) and has been subsequently confirmed on a number of different proteins. From these studies the conclusion has been drawn that the folding of protein chains depends solely on the amino acid sequence and the surrounding solvent and that the native state corresponds to the minimum of the free energy of the protein-solvent system, which is accessible in the approach to equilibrium.

In principle it should be possible to simulate the protein-solvent system by suitable energy functions or molecular force fields. Over the last two decades a number of attempts have been made to calculate native three-dimensional protein folds from amino acid sequences. Many of these approaches rely on a seemingly simple recipe involving two problems: (1) construction of a suitable

Reprint requests to: Manfred J. Sippl, Institute for General Biology, Biochemistry & Biophysics, Department of Biochemistry, University of Salzburg, Hellbrunnerstraße 34, A-5020 Salzburg, Austria.

molecular force field for the protein-solvent system and (2) search for the global minimum on the molecular energy surface.

Several semiempirical force fields have been developed that are used in macromolecular modeling studies (e.g., Weiner & Kollman, 1981; Burkert & Allinger, 1982; Brooks et al., 1983; van Gunsteren et al., 1983; Carson & Hermans, 1985). Most of these force fields are based on Coulomb's law for electrostatic forces, the Lennard-Jones potential as a model of core repulsion, and Van der Waals interactions and harmonic and periodic terms for covalent interactions and rotatable bonds, respectively. These force fields are important tools for the study of the stability and motions of macromolecular systems in vacuo but the simulation of solvent effects still poses an enormous problem.

Recently we presented a novel approach for the construction of a knowledge-based molecular force field based on the compilation of potentials of mean force from a data base of known protein structures (Sippl, 1990). A major strength of these potentials is that they contain all forces that stabilize the native states of soluble proteins including solvent effects.

We demonstrated that this force field can be used to model the complex conformational behavior of peptide fragments (Sippl, 1990), and we have shown that the force field is able to identify the native fold of a large number of globular proteins among several thousand alternatives (Hendlich et al., 1990; Sippl & Weitckus, 1991). The results indicate that this approach provides a reasonable model for protein-solvent systems and therefore could be used for the calculation of protein folds from amino acid sequences.

To reach this goal we have to concentrate on several theoretical and technical problems. A major technical problem is the search for the global minimum on the molecular energy surface. An exhaustive search of the conformational space of a polypeptide is computationally prohibitive, and therefore it is necessary to develop alternative procedures.

A very promising route for the development of efficient algorithms is to search the data base of known protein structures for possible models of an unknown fold. We have shown, that the knowledge-based mean field is able to correctly identify globin folds for the most distantly related globin sequences even if the force field is constructed from a data base devoid of globins (Sippl & Weitckus, 1991). Recently we have generalized this approach by allowing gaps in the sequence and/or structure (unpubl.).

It is clear, however, that in spite of the growing data base of known structures the unknown fold of a given sequence will, in many cases, have no counterpart in the data base. In such cases the unknown fold has to be constructed from scratch. In the present study we develop and apply procedures that can be used to construct backbone conformations from the conformational preferences of short fragments.

Our strategy is to calculate ensembles of probable conformations for short overlapping fragments along the entire protein sequence. The fragments are assembled to a complete three-dimensional model by choosing appropriate conformations from the ensembles. Because the preferred conformation of isolated short fragments are determined from interactions within the fragment, the resulting model is optimized with respect to the local interactions only.

We apply the procedures to the calculation of the local backbone conformations of myoglobin and lysozyme whose structures are known from X-ray analysis and to thymosin β_4, a polypeptide of 43 amino acids whose conformational preferences have been determined by NMR methods, and we show that the results of our calculations agree with many of the local structural features of these molecules.

Mean force potentials

We briefly review the concepts used to compile the potentials of mean force from a data base of known protein structures and the calculation of ensembles for short peptides. A detailed account of the concepts involved is found in Sippl (1990) and Hendlich et al. (1990).

The probability density $f(r)$ of a particular interaction is approximated by the relative frequencies $g(r)$ of the corresponding interatomic distances. The distances are sampled in intervals r from a data base of known protein structures (Bernstein et al., 1977). The potentials are obtained from the inverse Boltzmann law by

$$E_k^{ac,bd}(r) = -k_B T \ln[g_k^{ac,bd}(r)] - k_B T \ln [Z_k^{ac,bd}], \quad (1)$$

where a and b denote amino acids, c and d correspond to specific atoms of a and b (e.g., a = Ala, $c = N$ of Ala, and b = Val and $d = C^\beta$ of Val), k is the separation of a and b along the amino acid sequence (or structural level), and k_B and T are Boltzmann's constant and absolute temperature, respectively. Z_k^{ab} is the Boltzmann sum or partition function of the potential.

To use the potentials consistently we have to introduce an appropriate reference state. We are interested in the relative energies of polypeptide conformations with respect to one particular sequence. In this case the appropriate reference state is the probability density averaged over all amino acid pairs. Thus, the potential of mean force of the reference state is

$$E_k^{c,d}(r) = -k_B T \ln[g_k^{c,d}(r)] - k_B T \ln[Z_k^{c,d}], \quad (2)$$

where $g_k^{c,d}(r)$ is relative frequency of atoms c and d as a function of distance r, and $Z^{c,d}$ is the corresponding partition function. Note that the reference state is defined

as an average over all types of amino acid pairs but not over atom types or different values of k.

We obtain the specific interactions between two atoms c and d of amino acids a and b of sequential separation k by subtracting the reference state energy from the potential of mean force

$$\Delta E_k^{ac,bd}(r) = E_k^{ac,bd}(r) - E_k^{c,d}(r) = -k_B T \ln\left[\frac{g_k^{ac,bd}(r)}{g_k^{c,d}(r)}\right] - k_B T \ln\left[\frac{Z_k^{ac,bd}}{Z_k^{c,d}}\right]. \quad (3)$$

This is the net potential of mean force with respect to the average conformation in the data base (Sippl, 1990).

The problem of sparse data is treated as reported previously (Sippl, 1990).

$$\Delta E_k^{ac,bd}(r) = -k_B T \ln\left[\frac{1}{1+\sigma m_{ab}} + \frac{\sigma m_{ab}}{1+\sigma m_{ab}} \frac{g_k^{ac,bd}(r)}{g_k^{c,d}(r)}\right] + C_k^{ac,bd}, \quad (4)$$

$m_{a,b}$ is the absolute frequency of the amino acid pair (a,b) with sequential separation k in the data base, and σ controls the relative weight of one measurement with respect to the reference state. $C_k^{ac,bd}$, which is independent of the conformational variable s, contains the partition functions $Z_k^{ac,bd}$ and $Z_k^{c,d}$, which are constants for a given temperature.

The calculation of the total net energy $\Delta E(S,C)$ of an amino acid sequence S in a particular conformation C is straightforward once the net potentials of mean force have been compiled from the data base. The distance intervals r for all atom-atom pairs are calculated from the distances of conformation C, the net energy is obtained from the associate potentials, and the sum over all these individual interactions yields the total net energy:

$$\Delta E(S,C) = \sum_{ij} \Delta E_k^{ac,bd}(r_{ij}), \quad (5)$$

where a, b, c, d, and k are functions of the atom indices i and j.

Calculation of probable conformations for short fragments

The identification of probable conformations for a particular amino acid sequence requires a search for low conformational energies on the potential surface $E(S,C)$ as a function of the conformational variables of C. In the case of oligopeptides the problem can be solved efficiently. Our strategy uses a pool of conformations C_p, $p = 1,\ldots,N$ prepared from the proteins in the data base. A protein of sequence length L contains $L - l + 1$ fragments of length l, so that for small l we obtain a considerable number of different conformations. For $l < 10$ the current data base yields on the order of $N \approx 20{,}000$ fragments.

Once the pool is constructed from the data base, probable conformations are identified as follows. The sequence of interest, S, is mounted on all conformations $C_p, p = 1,\ldots,N$ in the pool and the associated total net energies $\Delta E(S,C_p)$ are calculated (Equation 5). Now the pool is equivalent to a statistical ensemble of identical peptides folded up in a variety of conformations. We obtain the most probable conformations of S by skimming off the conformations of lowest energy from the pool. In a last step the ensemble of most probable conformations is clustered in terms of conformational similarity. For ease of reference the summary of these procedures required to calculate ensembles of probable conformations is called the Boltzmann Device (Sippl, 1990).

The data base of protein conformations used to compile the potentials and to prepare the pool of conformations consists of 98 individual protein chains extracted from the Brookhaven Protein Data Bank. The potentials of mean force are sampled as discrete functions. In view of the sparse data sets, we used a grid of 20 intervals, which is sufficient to sample the main features of the potentials. As in our previous study, σ, the weight of one measurement, was set to 1/50, and $RT = 0.512$ kcal/mol (Hendlich et al., 1990; Sippl, 1990).

In the present study the molecular force field consists of the 25 interactions among the atoms N, C^α, C′, O, and C^β. Because there are 400 pairs of amino acids, we have a total of $25 \times 400 = 10{,}000$ potentials on a particular level k. Polypeptides of sequence length l consist of $l - 1$ structural levels k, so that energy calculations on hexapeptides involve on the order of 50,000 individual potentials.

With the exception of the C^β atoms, the force field used in this study does not contain any side chain atoms and is therefore incomplete. However, the mean force potentials of the backbone atoms to a considerable extent capture the behavior of the associated side chain interactions. Nevertheless, the neglect of explicit side chain interactions must be taken into account when interpreting the results.

Strategies to connect fragments contained in ensembles of overlapping peptides

The basic assumption in our approach is that the structures contained in the ensembles calculated from the Boltzmann Device reflect the conformational preferences of short fragments. Combinations of such fragments along the entire amino acid sequence should therefore yield a model structure that reflects the local preferences,

i.e., α-helix, β-strand, turns, and irregular structures. However, it is clear at the outset that in general such models will not describe the overall fold because nonlocal interactions are neglected.

Nevertheless, locally optimized models constitute the first step toward the calculation of complete tertiary structures, provided that such models to a considerable extent resemble the local backbone structures found in native conformations. Obviously, subsequent refinements of locally optimized models will be most successful if large parts of the local backbone conformations are correct and if, in the absence of the known tertiary structure, unreliable parts of a model can be identified.

Therefore, the goals of the present study are (1) the design of strategies that can be used to assemble complete backbone conformations from ensembles of overlapping fragments, (2) the development of criteria that are helpful in identifying incorrect or unreliable parts of the model without reference to known native folds, and (3) the evaluation of the success of the methods in the calculation of locally optimized models for myoglobin, lysozyme, and thymosin β_4.

Before we start, we have to recall the complex structural features of individual ensembles (see Sippl [1990] for the terminology used). Stable fragments contain a single type of conformation, whereas flip-flop, metastable, and unstable fragments contain a range of conformations. In addition, the behavior of fragments can change dramatically when sliding along the amino acid sequence by a single residue, i.e., two adjacent fragments can have completely different structural preferences.

Many of these features are illustrated by the ensembles calculated for 10 consecutive overlapping hexapeptide fragments of thymosin β_4 corresponding to the sequence Glu 21-Thr-Gln-Glu-Lys-Asn-Pro-Leu-Pro-Ser-Lys-Glu-Thr-Ile-Glu-Gln 36 (Fig. 1). Fragments 21 and 30 produce stable α-helical conformations, whereas the ensembles obtained for fragments 22–29 are flip-flop (e.g., fragment 22 preferring turnlike and extended conformations), metastable, and unstable. The ensembles contain a variety of conformations including β-strands (fragment 25) and turn-structures (e.g., fragments 27 and 28). The conformational preferences of fragments 29 and 30 are quite incompatible, although they overlap by five residues.

In view of this complexity many strategies are conceivable, which could be used to construct complete backbone conformations from these ensembles. The design of such strategies basically involves two operations that have to be defined: (1) extraction of structural information contained in an individual ensemble and (2) combination of this information for consecutive fragments.

Both operations can be defined in a large number of ways and they can be implemented as interactive or automatic procedures. Here we present two fully automatic versions based on different definitions of the basic operations. The first strategy uses rather crude and simple rules and as a by-product yields a measure of the local variability of the resulting models. The second is a refinement of the former.

The procedures will be presented along the following lines: We define the basic operations for the first strategy, illustrate the calculations involved using thymosin β_4 as an example, and compare the results to model conformations derived from NMR measurements. In addition we calculate models for myoglobin and lysozyme and compare the results to the known X-ray conformations. Finally, we use an alternative set of rules, which yields an improvement in the quality of the local backbone conformations.

Basic operations

The first strategy uses the largest cluster of an ensemble to represent the structural preferences of a particular fragment. In the case of thymosin β_4 residues 21–36, these are the clusters in the left column of Figure 1. Individual clusters contain several closely related conformations. We define the most typical or average conformation as the central conformation in each cluster. The central conformation is identified as that structure that has minimal root mean square (rms) error of spatial superimposition (Sippl & Stegbuchner, 1991) to all other conformations in the cluster. Hence, the central conformation in the largest cluster is used to represent the conformational preference of individual fragments.

This definition yields a unique structure for each fragment, and it is a reasonable choice in the case of stable or metastable fragments. However, using this definition, a large amount of structural information on flip-flop and unstable fragments is discarded.

The second operation involves the combination of successive fragments along the chain. Two adjacent hexapeptide fragments overlap by five amino acid residues. Therefore, any two adjacent fragments can be joined to yield a consistent structure for the resulting heptapeptide, provided the conformations contained in the respective ensembles are superimposable in the overlapping regions. Once the heptapeptide is constructed we can proceed in either the N- or C-terminal direction by adding the next hexapeptide along the sequence.

Figure 2 shows the spatial orientation of the central conformations obtained from the largest clusters of fragments 21–30 (Fig. 1) after superimposition of the overlapping parts. The structures do not fit exactly on each other, but it is obvious that the superimposed aggregate starts with an α-helix turn, continues into a relatively irregular extended structure, and terminates in an α-helical turn. Figure 4 shows the complete aggregate of thymosin β_4 assembled from the largest clusters of all ensembles. Although the aggregate consists of individual hexapeptide fragments, an overall model for the thymosin β_4 conformation is discernible. The structure starts with an N-ter-

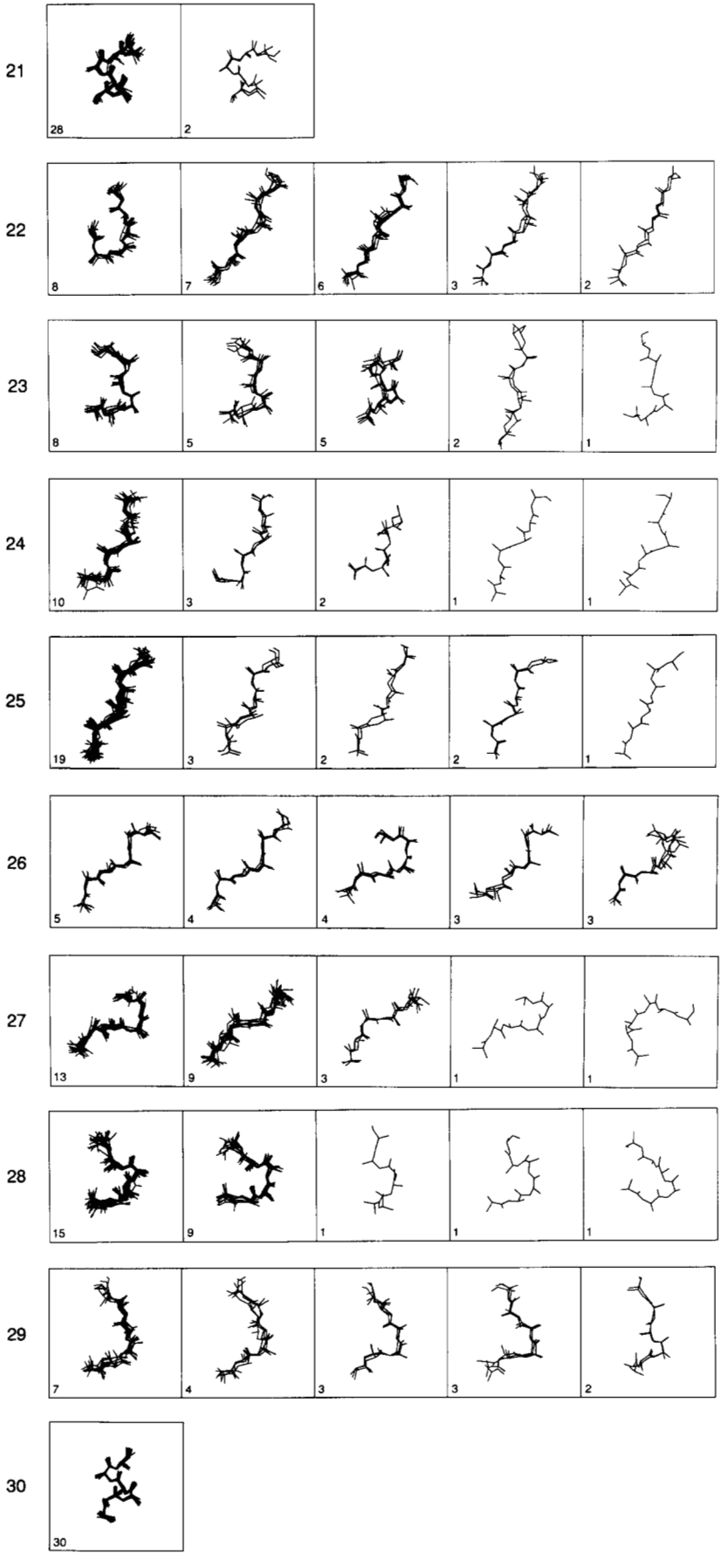

Fig. 1. Ensembles of hexapeptide fragments 21–30 of thymosin β_4 calculated from the Boltzmann Device (Sippl, 1990). The fragments cover the sequence Glu 21-Thr-Gln-Glu-Lys-Asn-Pro-Leu-Pro-Ser-Lys-Glu-Thr-Ile-Glu-Gln 36. Fragment 21, for example, corresponds to the hexapeptide Glu 21-Thr-Gln-Glu-Lys-Asn 26. Any two adjacent hexapeptides overlap by five residues. The ensembles contain the 30 conformations of lowest net energy and they are represented as clusters of similar conformations. The clusters are ordered from left to right by decreasing size. Only the five largest clusters are shown. Note that some ensembles contain a single type of conformation (e.g., ensemble 30), others contain several clusters of rather similar conformations (e.g., ensembles 21 and 25), and some contain conformations of quite distinct type (e.g., ensembles 22 and 27).

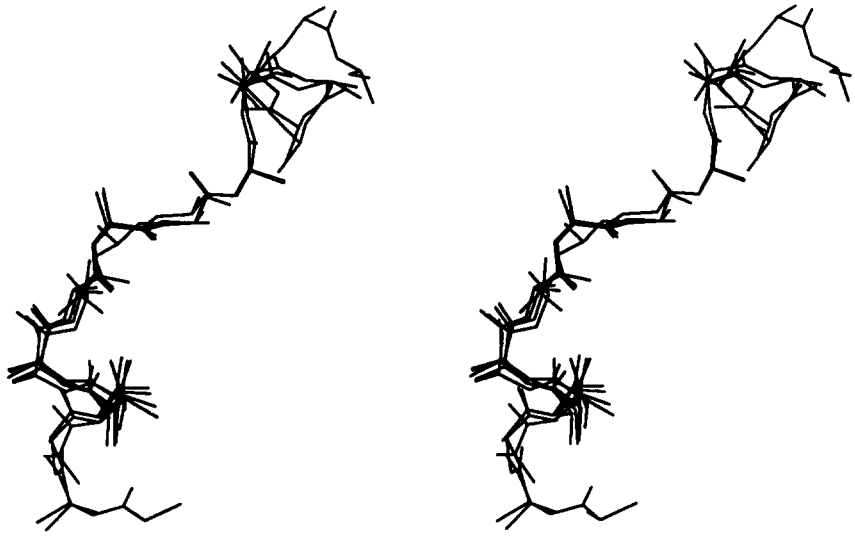

Fig. 2. Superimposed aggregate of all fragments shown in Figure 1. The aggregate is formed from the overlapping fragments by superimposing the central conformations of the largest clusters in each ensemble (right column in Fig. 1). The N-terminus (residue 21) is at the bottom.

minal arm, followed by a long kinked α-helix. Then the conformation continues into an extended conformation, and residues 30–43 form a C-terminal α-helix. As can be inferred from Figure 4 some parts of the superimposed fragments are quite incompatible, especially in the region of the helix kink of the N-terminal helix.

In general the aggregate obtained by superposition will contain regions of incompatible adjacent fragments. It is therefore necessary to apply a suitable averaging procedure to obtain a consistent model from the superimposed aggregate. One possibility is to average over equivalent atoms of the individual fragments. However, in regions of incompatible fragments this results in gross distortions of covalent bonds and valence angles. The local geometry can be preserved by averaging over dihedral angles along the polypeptide backbone and by rebuilding the chain from the average angles. With the exception of the N- and C-terminal residues a particular backbone dihedral, ϕ, ψ, or ω, is found in six overlapping hexapeptides.

Mean values are obtained by averaging equivalent angles over the individual fragments. Because angles are periodic quantities, it is necessary to take the average over the sine and cosine functions, which can be calculated in the following way. We seek an average over n angles $\alpha_1, \alpha_2, \ldots, \alpha_n$. Each angle can be represented as a two-dimensional vector whose components are $x_i = \sin(\alpha_i)$ and $y_i = \cos(\alpha_i)$. Because $\sin^2(\alpha_i) + \cos^2(\alpha_i) = 1$, the (x_i, y_i) are unit vectors originating from the center of the unit circle. Then

$$X = \frac{1}{n}\sum x_i \qquad \text{and} \qquad Y = \frac{1}{n}\sum y_i \tag{6}$$

represent the coordinates of the average vector. In general $X^2 + Y^2 < 1$, so that we have to normalize (X, Y)

$$\bar{X} = \frac{X}{L} \qquad \text{and} \qquad \bar{Y} = \frac{Y}{L}, \tag{7}$$

where

$$L = \sqrt{X^2 + Y^2}. \tag{8}$$

From this we get the average angle as

$$\bar{\alpha} = \text{sign}(X/L)\arccos(Y/L). \tag{9}$$

The procedure yields a unique average angle $\bar{\alpha}$ as long as $X^2 + Y^2 \neq 0$. If $X^2 + Y^2 = 0$ the individual unit vectors add up to the zero vector, and the average angle is undefined. This happens for example if we average over the three angles −120, 0, 120. The result is satisfying because in this case there is no value that can be considered as a meaningful average over these angles.

The length L of the average vector provides a natural estimate of the quality of the average. If $L = 1$, then all angles for which the average is calculated must be equal. On the other hand, if $L = 0$, the angles are completely unrelated, and the average angle can be chosen randomly. We define the variability v of $\bar{\alpha}$ as

$$v = 1 - L. \tag{10}$$

The average structures obtained from the superimposed aggregates of Figures 2 and 4 are shown in Figures 3 and 5. The consensus structures are quite similar to the superimposed aggregates, the main difference being the more pronounced kink between the two N-terminal helices (Figs. 4, 5).

As a by-product, averaging yields the variability v of individual dihedrals. High variabilities point to unreliable or undefined regions in the model conformations. As

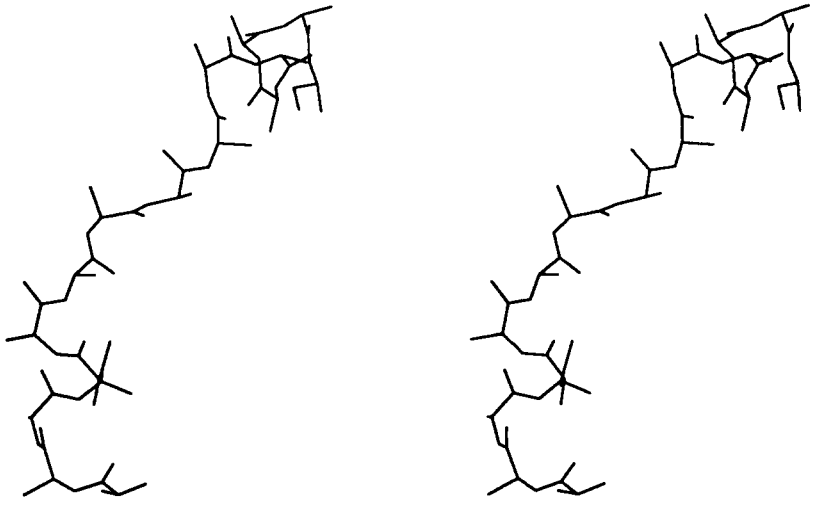

Fig. 3. Average conformation of the superimposed aggregate shown in Figure 2. The average conformation was calculated by averaging over equivalent backbone dihedrals derived from the central conformation of the largest cluster of the individual ensembles (Fig. 1) and by rebuilding the chain from the average angles.

Fig. 4. Superimposed aggregate obtained by superimposing all hexapeptide fragments of thymosin β_4. The N-terminus is at the bottom.

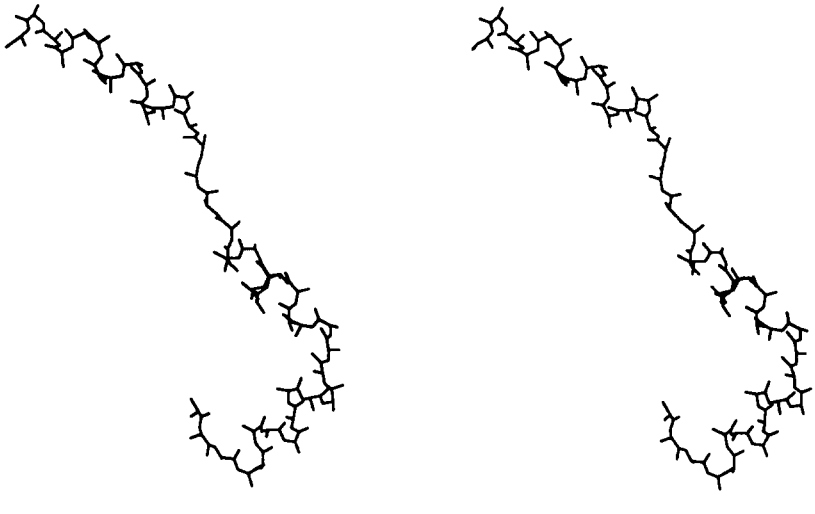

Fig. 5. Average conformation obtained from the fragments shown in Figure 4.

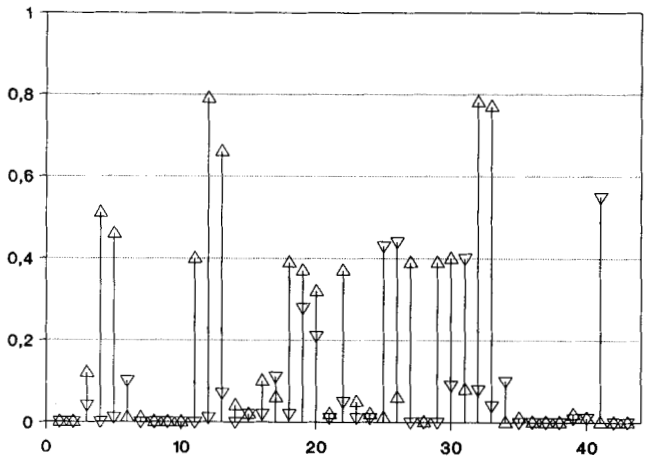

Fig. 6. Variability of the backbone dihedrals ϕ and ψ of the calculated thymosin β_4 model shown in Figures 4 and 5 ($\phi = \nabla$, $\psi = \Delta$).

shown in Figures 2 and 3, some consecutive fragments have quite different conformations in their overlapping parts resulting in a large rms error of superposition. Hence, it is immediately clear that in this part of the chain the resulting consensus structure is quite unreliable. This is also evident from the variability v of the backbone angles obtained by averaging over the fragments. Regions of high confidence have variabilities $v < 0.25$, whereas for $v > 0.5$ the dihedrals are largely undetermined.

Most of the backbone angles of the thymosin β_4 model shown in Figure 5 have low variability (Fig. 6). Angles of high variability are concentrated at the helix kink (residues 11–13) in the extended part of the conformation (residues 19–30) and at the start of the C-terminal helix (residues 31–33).

Variabilities of backbone angles indicate the local reliability of constructed models. They are evaluated independently of any experimental evidence so that this is a most important tool in assessing the quality of a model for a yet undetermined structure. In addition, regions of high uncertainty may actually correspond to regions of high flexibility in the molecule and may indicate the dynamical behavior of the polypeptide chain. Regions of low variability are not necessarily correct. In the following section we compare the calculated thymosin β_4 model with conformations obtained from experimental information.

Comparison of the calculated thymosin β_4 model with structures derived from NMR measurements

Zarbock et al. (1990) have investigated the conformation of thymosin β_4 in alcoholic solution. Figure 7 shows the variabilities of backbone angles calculated from five models of thymosin β_4 derived from NMR constraints. The average angles and variabilities are calculated from the equivalent angles in the five models. The constraints used to calculate the NMR-derived structures consisted of a set of 180 approximate interproton distance constraints, 33 ϕ constraints obtained for $J_{NH\alpha}$ coupling data and 23 ψ dihedral angles identified on the basis of short-range nuclear Overhauser effects (NOEs). These constraints were used in a combined distance geometry energy minimization calculation (Holak et al., 1989a,b). All calculated structures satisfy the distance constraints. The structures contain two helical regions extending from residues 4 to 16 and 30 to 40. The relative orientation of the two helices could not be determined due to the lack of long-range NOEs.

The variability is very low in the α-helical regions (Fig. 7) similar to the low variability in the corresponding regions of our model (Fig. 6). The variability of the NMR models is high in the regions between the helices (residues 20–30) and at the N-terminus of the C-terminal helix. This agrees to a considerable extent with the variabilities obtained from our model, the main differences being the low variability of our model at the N- and C-termini, which is high in the NMR-averaged structures and the high variability of our model at residues 11–13, which is low in the NMR structures.

Figure 8 shows the deviation of the average angles of our model from the NMR-derived structures. The differences are small in the α-helical regions. Large differences occur in the variable regions between the helices and at the N- and C-termini of the molecule. In summary, our model and the NMR models agree in regions of high confidence, the large differences being concentrated in regions of low reliability.

The NMR studies on thymosin β_4 were carried out in solutions of 60% trifluorethanol-d_3 and 50% hexafluorisopropyl-d_2. In aqueous solutions thymosin β_4 seems to behave like a random coil. The CD and NMR spectra in water show no obvious regular structure. Addition of al-

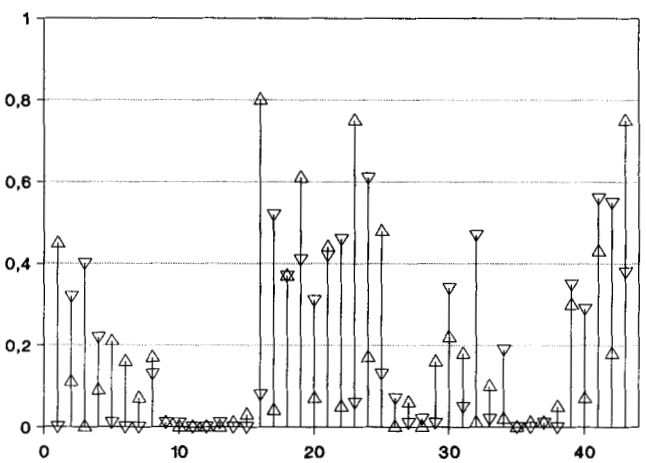

Fig. 7. Variability of the backbone dihedrals ϕ and ψ obtained by averaging over five conformations obtained from NMR constraints ($\phi = \nabla$, $\psi = \Delta$).

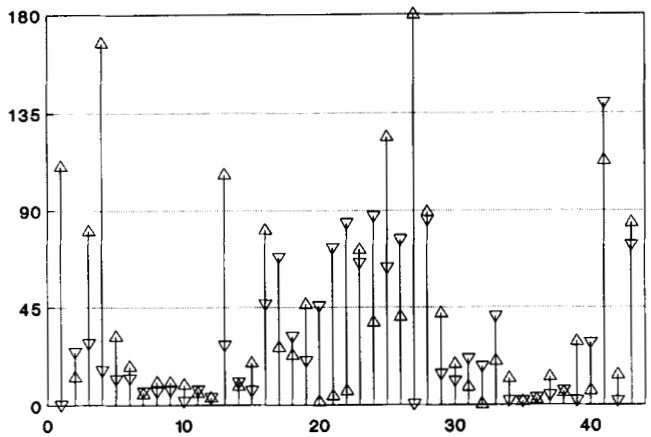

Fig. 8. Deviation of average ϕ and ψ angles of the model of thymosin β_4 from the respective average angles calculated from five NMR-derived structures ($\phi = \nabla$, $\psi = \Delta$).

cohol stabilizes the secondary structure of the molecule. On the other hand, the potentials used in this study are derived from data obtained from protein crystals, and it is an interesting question why the calculated models resemble conformations stabilized in alcohols. We will return to this question in a later section.

Calculated models for myoglobin and lysozyme backbone conformations

Using this same strategy, we calculated backbone conformations for sperm whale myoglobin (1MBA) and hen egg white lysozyme (6LYZ). We emphasize that in these calculations we removed all globins and lysozymes from the data base. Hence the data base did not contain closely or distantly related members of the respective protein families.

Figures 9 and 10 show the differences in backbone angles of the calculated myoglobin and lysozyme conformations and the variabilities obtained from averaging. In most cases, large errors in the models correspond to high variabilities, but it should be noted that overall the variabilities are rather conservative in the sense that the variability is also high in some regions of rather small errors (e.g., between residues 50–60 of myoglobin).

The errors in ϕ and ψ in the model myoglobin conformation with respect to the X-ray structure, when averaged over the whole sequence, are 21 and 41 degrees, respectively. For lysozyme these numbers are 33 and 45. The average error in ϕ is generally much smaller as compared to ψ, which is due to the restricted range of ϕ as can be inferred from a Ramachandran map.

As discussed above, the strategy used is rather crude because by using only the most prominent cluster a large amount of structural information is discarded from the ensembles and is not available in assembling the final model. We therefore investigated a different strategy,

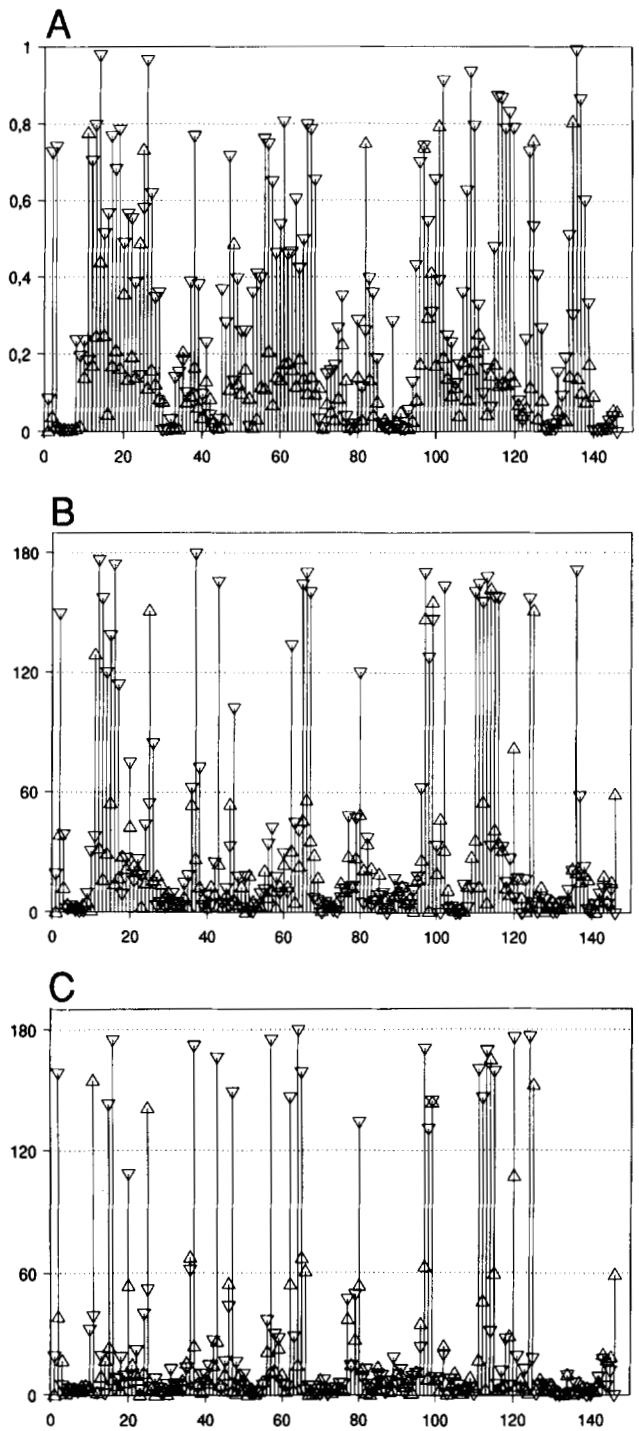

Fig. 9. **A**: Variability of backbone ϕ and ψ angles of the calculated model of myoglobin 1MBA. **B**: Deviation from the respective angles calculated from the X-ray coordinates. **C**: Deviation of ϕ and ψ of the model generated using the improved strategy from the respective X-ray values ($\phi = \nabla$, $\psi = \Delta$).

which compares the variabilities of ϕ and ψ angles in the ensembles. A particular ϕ (or ψ) along the amino acid sequence is contained in several overlapping fragments (i.e., six in the case of hexapeptides). We calculate the

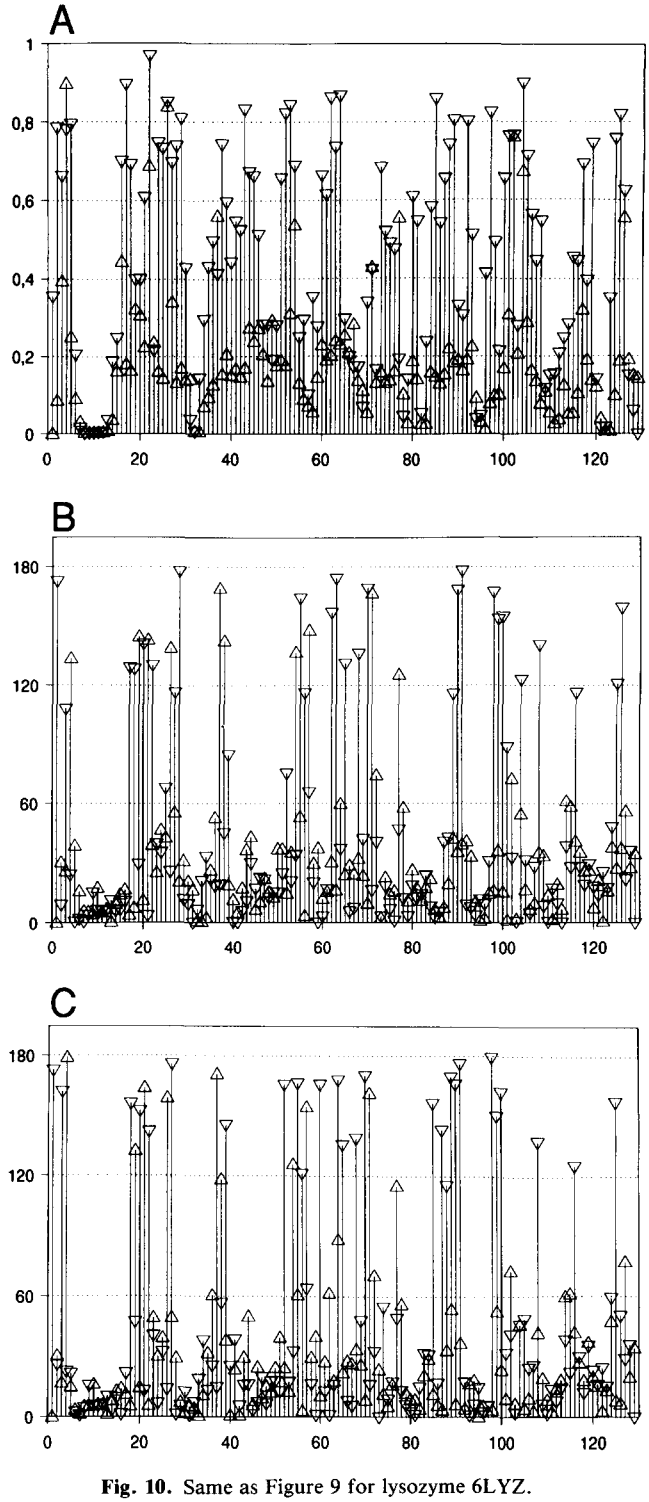

Fig. 10. Same as Figure 9 for lysozyme 6LYZ.

variability of this angle by averaging over all structures contained in each of the ensembles. In the case of hexapeptides this results in six average values for a particular ϕ or ψ, and we choose the average angle of least variability when assembling the model. By averaging over all conformations in the ensembles we retain all structural information, and by choosing the angle of least variability we use the strongest structural determinant for this angle.

Using this method, the average errors in ϕ and ψ being 17 and 32 (1MBA) and 32 and 48 (6LYZ), respectively, are smaller as compared to the former method. Figures 9C and 10C show the errors in these model structures as compared to the respective X-ray structures. The model calculated for thymosin β_4 shows a few changes in regions of high variability, but these differences are indistinguishable from the former model when compared to the averaged structures derived from NMR studies.

The errors plotted in Figures 9 and 10 express the quality of the calculated models in quantitative terms. To present the quality of the local structures in a more comprehensive way we show a few detailed comparisons of calculated structures and X-ray conformations in Figures 11–14. Figure 11 shows the calculated structure for residues 31–50 of myoglobin superimposed on the X-ray structure. This region contains the C-helix, which is correctly predicted. The major difference between predicted and X-ray structure is found at both helix-termini where the calculated conformation protrudes into different directions due to a few large errors in dihedral angles (see corresponding region in Fig. 9C). It is noteworthy that the conformations on both ends of the fragment shown in Figure 11 agree very well with the X-ray conformation.

Figure 12 shows the calculated structure for residues 79–108 of 1MBA. Again the calculated local conformation agrees fairly well with the X-ray conformation. In Figure 13 we show the calculated region 1–40 of lysozyme superimposed on the known conformation. The major error in this region is found at the kink in front of the second helix. Finally, Figure 14 shows the results obtained for the β-strand region 38–57 of lysozyme, which has a distorted angle in each β-strand.

Ranking the quality of a number of different models

If we use different strategies a number of models can be generated for a single amino acid sequence. The ϕ–ψ variabilities indicate the degree of uncertainty of these models but they are of little help in estimating the relative quality of several models.

Hendlich et al. (1990) have recently demonstrated that the total net energy calculated from the mean force potentials can be used to identify native protein folds among a large number of incorrect conformations. The same approach can be used to estimate the quality of individual models of unknown protein folds. Models that resemble the native fold should have low energy with respect to all structures in a large pool of conformations.

We demonstrate the application of this approach in the case of thymosin β_4. A pool of conformations is con-

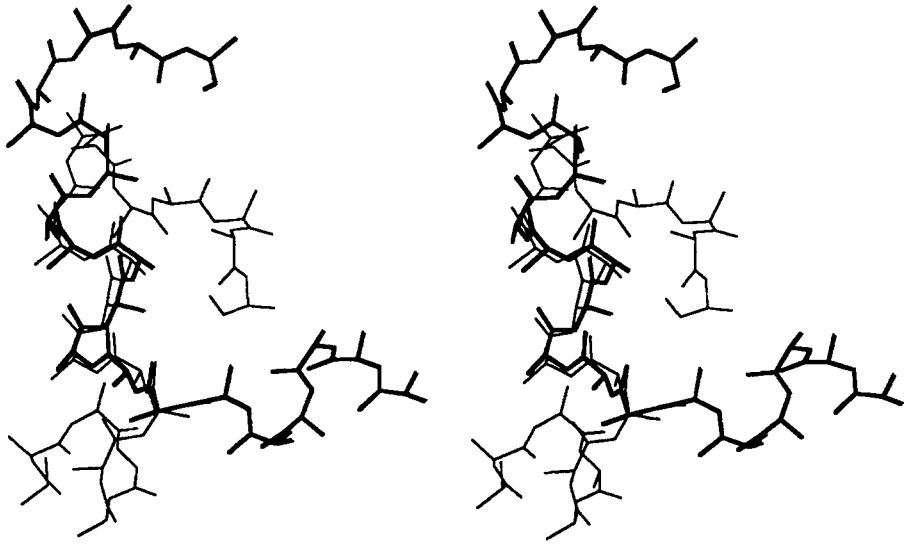

Fig. 11. Calculated backbone structure for residues 31–50 of myoglobin 1MBA (bold lines) superimposed on the X-ray conformation.

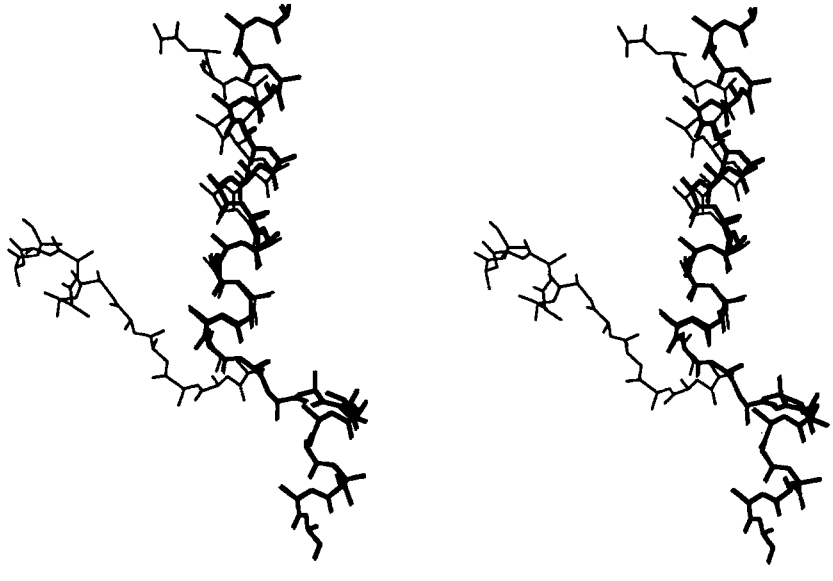

Fig. 12. Calculated backbone structure for residues 79–108 of myoglobin 1MBA (bold lines) superimposed on the X-ray conformation.

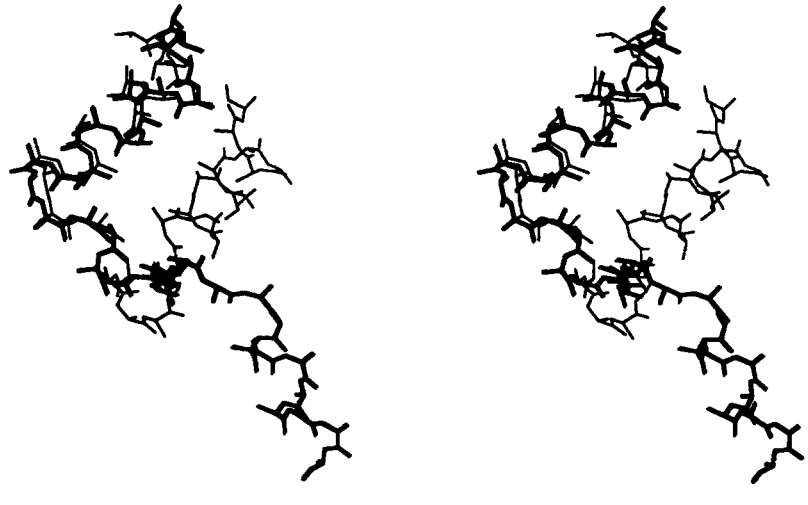

Fig. 13. Calculated backbone structure for residues 1–40 of lysozyme 6LYZ (bold lines) superimposed on the X-ray conformation.

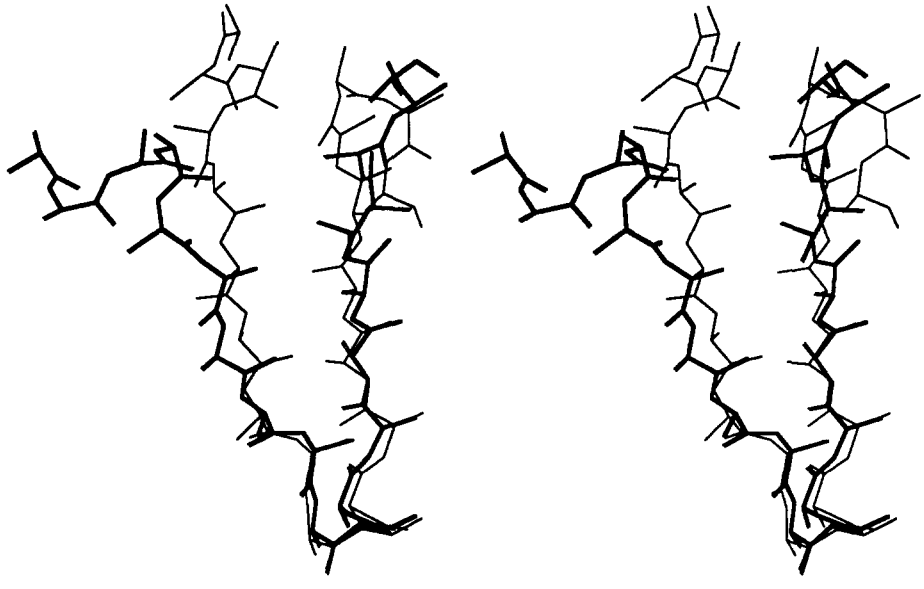

Fig. 14. Calculated backbone structure for residues 38–57 of lysozyme 6LYZ (bold lines) superimposed on the X-ray conformation (bold lines).

structed from the data base of known protein folds as reported previously (Hendlich et al., 1990). In the case of thymosin β_4 consisting of 43 amino acids, we obtain a pool of 18,552 fragments of length $l = 43$ from the current data base. We add the calculated model and five models derived from NMR constraints (Zarbock et al., 1990). Then the total net energy (Equation 5) of the thymosin β_4 sequence is evaluated with respect to all conformations in the pool, and the conformations are ranked with respect to their total net energy.

The data in Table 1 show the ranking of the individual conformations in the pool. The structure of lowest total net energy is fragment 80–122 from 1PP2-R (rattlesnake phospholipase A_2). The calculated model ranks at position 8. In addition, Table 1 shows the positions of the structures obtained from NMR constraints, ranging from 77 (B4TC15) to 377 (B4TC02). Thus, in terms of the total net energy, the calculated structures as well as the models derived from NMR measurements belong to the preferred conformations of thymosin β_4.

The conformation of lowest total net energy 1PP2-R-80 has many features in common with the model structures but the C-terminal part is quite irregular, containing only a single helix turn (Fig. 15). In Table 2 the conformations in the pool are ranked with respect to the short-range energy contributions calculated for $k = 1, \ldots, 10$. 1PP2-R-80 has considerably higher energy in this range as compared to the calculated model. The low total net energy of 1PP2-R-80 is mainly due to favorable medium- and long-range contributions. When ranked with respect to the short-range energy, 1PP2-R-80 appears on position 495 (Table 2).

Top positions with respect to the short-range contributions are occupied by fragments from hemerythrin (1HMQ-A-40) and myohemerythrin (2MHR-40). As shown in Figure 16 the conformation of 1HMQ-A-40 consists of two α-helices separated by a kink. The conformation is similar to the models derived from NMR measurements and mean force calculations, but the α-helices are longer, they are more regular, and they are closely packed in an antiparallel configuration. With respect to the short-range energies, the calculated model oc-

Table 1. *Total net energies of several models of thymosin β_4*

Model[a]		Total net energy		Short-range energy[b]	
		Rank	$\Delta E(S,C_p)$	Rank	$\Delta E(S,C_p)$
1PP2-R	80	1	−102.5	494	−35.7
1PP2-R	76	2	−99.3	893	−29.2
1LDX	221	3	−97.2	351	−38.5
4ADH	323	4	−95.9	301	−40.1
1PP2-R	77	5	−95.4	75	−43.3
1LDX	220	6	−94.8	993	−28.1
1PP2-R	79	7	−93.2	521	−35.2
Construct	1	8	−93.1	24	−53.9
1HMG-B	74	9	−93.1	160	−44.3
B4TC15	1	76	−78.8	1,495	−23.2
B4TC28	1	129	−74.0	702	−32.0
B4TC04	1	177	−70.1	414	−37.3
B4TC22	1	249	−65.9	2,011	−19.0
B4TC02	1	376	−60.5	748	−31.4

[a] Conformation "construct" was calculated from the mean field as discussed in the main text. B4TC02, B4TC04, B4TC15, B4TC22, and B4TC28 were derived from distance constraints (Zarbock et al., 1990); the remaining codes are identical to the codes used by the Brookhaven Protein Data Bank (Bernstein et al., 1977). The numbers refer to the first residue in the respective fragment (i.e., 1PP2-R-80 corresponds to residues 80–122 of 1PP2-R).

[b] Net energy $\Delta E(S,C_p)$ calculated for sequential separations $k = 1, \ldots, 10$.

Table 2. *Short-range net energies of several models of thymosin β_4*[a]

Model		Short-range energy		Total net energy	
		Rank	$\Delta E(S,C_p)$	Rank	$\Delta E(S,C_p)$
1HMQ-A	40	1	−65.0	2,184	−26.9
2MHR	40	2	−62.9	1,464	−35.5
2MHR	38	3	−62.9	1,156	−40.3
2HHB-B	94	4	−60.8	1,090	−41.5
1PRC-M	114	5	−60.2	1,690	−32.5
1PHH	328	6	−59.8	513	−55.5
2CCY-A	7	7	−58.6	3,040	−18.5
1PHH	326	8	−57.9	429	−58.9
2HHB-B	96	9	−57.8	2,287	−25.9
Construct	1	24	−53.9	8	−93.1
B4TC04	1	414	−37.3	177	−70.1
B4TC28	1	703	−32.0	129	−74.0
B4TC02	1	748	−31.4	376	−60.5
B4TC15	1	1,494	−23.2	76	−78.8
B4TC22	1	2,010	−19.0	249	−65.9

[a] See footnotes to Table 1.

cupies position 25. The short-range positions of the models derived from NMR constraints range from 414 (B4TC04) to 2,010 (B4TC22).

In Table 3 the total net energy is split into the contributions of the various structural levels. This allows the identification of those structural levels that are most favorable for a particular model. Among the conformations of lowest total net energy, the calculated model has the most favorable short-range energy (i.e., levels 1–5) of −39.9 kcal/mol. The short-range energy of 1PP2-R-80 being −10.4 kcal/mol is quite unfavorable. The stabilizing contributions in 1PP2-R-80 occur in the medium and long range. In contrast, the medium- and long-range energies of the calculated model contribute little to its low total net energy. This clearly reflects the fact that the model was built from local interactions alone.

Similarly, 1HMQ-A-40, a conformation of favorable short-range energy (Table 2), has a rather high total net energy of −26.9 kcal/mol due to unfavorable medium- and long-range interactions. Obviously, the close association of the helices in this conformation is favored by interactions with respect to the original amino acid sequence of hemerythrin. In contrast, the sequence of thymosin β_4 folded up in the 1HMQ-A-40 conformation produces quite unfavorable medium- and long-range interactions, although both sequences favor the same local fold.

The analysis of the energy contributions clearly shows that the most favorable conformations, although having comparable total net energy, are stabilized by interactions from different structural levels. The calculated model has a low short-range energy but rather high medium- and long-range contributions. On the other hand, 1PP2-R-80 and other conformations of low total net energy are stabilized by medium- and long-range interactions (Tables 1, 3).

Similar calculations show that the models obtained for myoglobin and lysozyme are top models in terms of the short-range energy, whereas the medium- and long-range energies are unfavorable. These results indicate that it should be possible to considerably refine the models by optimizing the nonlocal interactions.

We are now in a position to discuss possible causes for the similarity of the calculated model with conformations derived from NMR measurements in alcoholic solutions. There are two features of our approach that we have to emphasize when discussing this issue. First, our approach calculates ensembles of conformations for oligopeptide fragments that are assembled to a complete model. If a large number of the ensembles is unstable, i.e., contains a range of different conformations, and if many overlap-

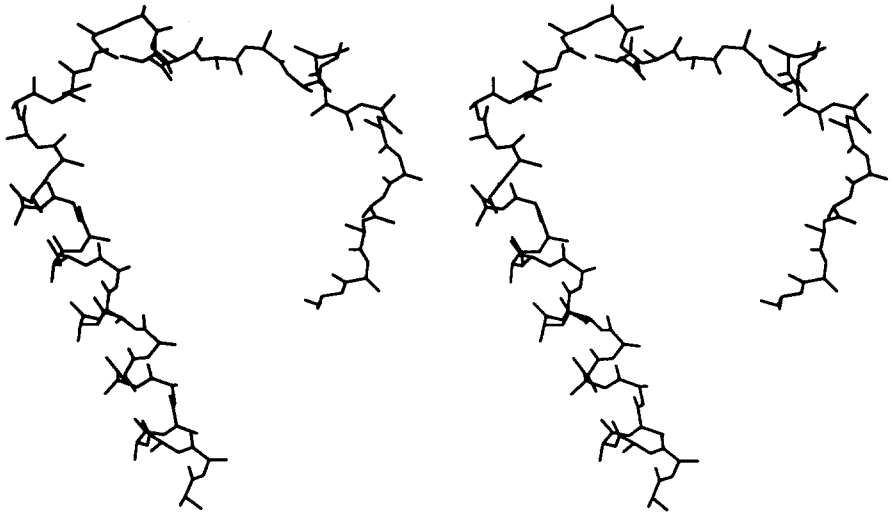

Fig. 15. Conformation of fragment 1PP2-R-80 of phospholipase A2. This conformation has the lowest total net energy with respect to the thymosin β_4 amino acid sequence.

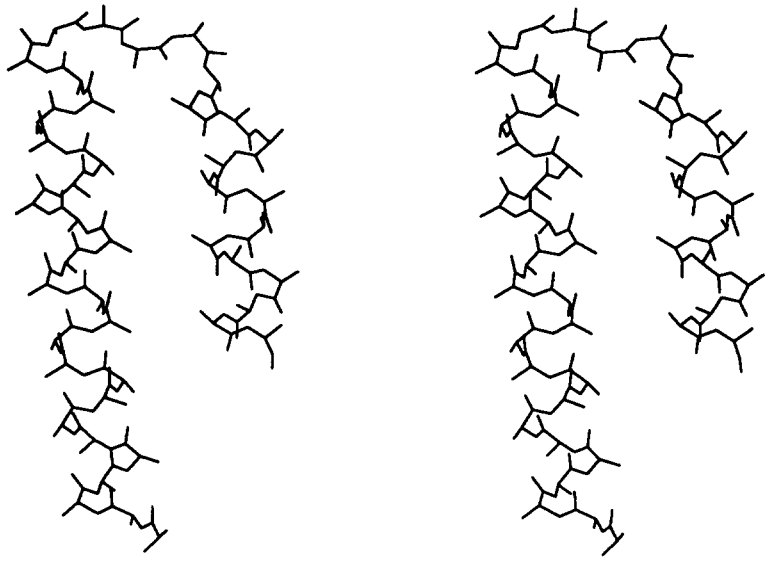

Fig. 16. Conformation of fragment 1HMQ-A 40-82 from hemerythrin. This conformation has the lowest net energy in the short range $k = 1, \ldots, 10$ with respect to the thymosin β_4 amino acid sequence.

ping fragments are incompatible, then the variability of most of the dihedral angles will be very high. In this case the results would indicate that the molecule does not have a preferred conformation behaving like a random coil. Most of the ensembles corresponding to the α-helical regions in the model of thymosin β_4 are stable and compatible (see for example fragments 21 and 30 in Fig. 1). The high variabilities are concentrated in the irregular region, residues 20-30. Hence, in principle the method is able to identify polypeptides that behave like random coils, but the calculations clearly favor conformations consisting of two α-helices.

Second, the potentials of mean force are compiled from a data base of globular protein structures determined by X-ray analysis. The potentials contain all kinds of forces that contribute to the stability of native protein folds including solvent effects. Hence, the potentials are averages over different environments of the individual interactions ranging from the hydrophobic interior of protein molecules to the fully exposed surface.

We may argue that the environment in globular proteins on average is hydrophobic, and therefore, the force field should reflect the interactions in organic solvents more closely as compared to aqueous solutions. This would explain the agreement between our model and the conformations obtained from NMR studies in alcoholic solution. At first sight this explanation seems reasonable. Note however that by calculating the net potentials we

Table 3. *Contribution of individual topological levels to the total net energy in several models of thymosin* β_4[a]

Model		Rank	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	Total
1PP2-R	80	1	−10.4	−23.1	−13.5	−17.2	−14.9	−13.6	−5.6	−4.0	−102.5
1PP2-R	76	2	−26.0	−17.4	−3.7	−14.8	−17.2	−8.5	−6.8	−4.1	−99.3
1LDX	221	3	−32.8	−11.2	−3.8	−10.6	−14.2	−13.0	−8.8	−2.7	−97.2
4ADH	323	4	−30.2	−10.2	−0.6	−8.9	−18.8	−12.3	−9.3	−5.8	−95.9
1PP2-R	77	5	−22.2	−15.5	−6.1	−14.8	−15.8	−10.6	−5.0	−4.2	−95.4
1LDX	220	6	−25.1	−13.4	−7.5	−11.0	−12.6	−11.5	−8.8	−4.4	−94.8
1PP2-R	79	7	−16.5	−19.2	−8.3	−14.0	−14.8	−10.1	−5.0	−4.7	−93.2
Construct	1	8	−39.9	−13.9	−10.0	−13.8	−9.6	−2.0	−3.2	−0.9	−93.1
1HMG-B	74	9	−32.3	−19.2	−5.6	−20.2	−9.5	−2.6	−2.9	−0.6	−93.1
B4TC15	1	76	−8.8	−14.4	−4.8	−12.4	−19.6	−8.9	−6.2	−3.7	−78.8
B4TC28	1	129	−19.5	−12.5	−4.6	−5.4	−11.5	−10.5	−6.7	−3.4	−74.0
B4TC04	1	177	−25.0	−12.3	+3.3	−7.7	−15.0	−7.5	−4.8	−1.0	−70.1
B4TC22	1	249	−20.0	+1.0	−1.9	−3.3	−17.3	−10.8	−7.0	−5.7	−65.9
B4TC02	1	376	−22.1	−9.3	+3.8	−6.3	−10.7	−8.5	−5.6	−1.7	−60.5

[a] The total net energy is split into contributions from several ranges. The header of each column (e.g., 1-5) defines the respective k-range. See footnotes to Table 1 for the remaining symbols.

subtract the reference state from the mean force potentials. The reference state corresponds to the mean force potential obtained by averaging over all amino acid pairs so that much of the average background is removed from the potentials.

The energy contributions from the different structural levels indicate an alternative explanation. All conformations that have very low short-range energies (Table 2) do have rather high energies in the medium and long range (Tables 1, 3). This can be explained if we assume that the amino acid sequence of thymosin β_4 produces a conflicting force field. The short-range contributions favor two α-helices (see Figs. 5, 16) but the medium- and long-range forces favor more extended structures (Fig. 15), i.e., the two helices interact unfavorably via the medium- and long-range forces.

In fact, in the conformation of lowest net energy, 1PP2-R-80, the C-terminal helix is reduced to a single helix turn (Fig. 15). The comparatively high short-range energy of this conformation is balanced by the gain in the medium and long ranges. In aqueous solution the medium- and long-range forces may be too strong to allow the formation of α-helices, whereas the addition of alcohol may screen these forces inducing the formation of α-helices. Because our model was built solely from short-range interactions, the unfavorable medium- and long-range forces did not enter the calculation. If this explanation is correct it should be impossible to considerably improve the medium- and long-range energies and at the same time retain the favorable short-range energies of the model. We are investigating this issue in our current studies on energy minimization.

Discussion

We presented a method that can be used to calculate models for unknown protein folds from amino acid sequences having favorable local energies. The procedure uses the most probable conformations for short fragments, which are calculated from a knowledge-based mean field. The overlapping fragments are assembled to a complete model by averaging over dihedral angles along the polypeptide. Substantial parts of the calculated models agree fairly well with the corresponding local structures known from X-ray or NMR studies. It is noteworthy that these results are obtained using a data base devoid of even distantly related proteins so that the knowledge-based molecular force fields for myoglobin, lysozyme, and thymosin β_4 do not contain any specific information linking their amino acid sequences with the respective structures.

When designing a model for a protein fold it is very important that the qualities of the proposed conformations can be judged without reference to the native fold, because in many cases no such information will be available. The procedures presented in this work yield at least three criteria that are independent of any experimental information on the native conformation of the molecule. (1) The variabilities of individual angles provide an estimate of the local reliability and flexibility of the models, (2) the comparison of the total net energy of a model with the energies obtained from a large pool of conformations shows whether the proposed model is among the most favorable structures, and (3) the decomposition of the total net energy into contributions from various structural levels identifies those parts of a model that are energetically unfavorable.

Only a (yet unknown) fraction of all possible amino acid sequences adopt stable three-dimensional folds. Hence, in the prediction of native protein conformations from amino acid sequences one faces two problems. The first problem is to locate the accessible global minimum on an energy surface, which reasonably models the protein solvent system. The second concerns the stability of the global minimum. If the minimum is wide and shallow or if there are several or many minima of comparable energy scattered in conformation space, then the molecule will be unstable. Driven by thermal collisions, the molecule will fluctuate among many different conformational states.

No long-range NOEs have been observed in the NMR experiments on thymosin β_4. This is compatible with the view that the molecule has ordered local structure but an ill-defined tertiary fold. The mean force calculations yield essentially the same conclusions as those inferred from experimentation: To a considerable extent the molecule has ordered backbone structure but no preferred spatial arrangement of local structural elements.

Figures 11–14 show that a single large error in ϕ or ψ sends the chain into the opposite direction from the X-ray structure. Expressed in energetic terms there are two possible causes for large errors. (1) The mean field favors a particular fold that differs from the native conformation (large error and low variability), or (2) the mean field does not favor a particular fold (large error and high variability). Most of the large errors in the calculated models are associated with average angles of high variability (Figs. 9, 10), i.e., there are several conformational states of comparable energy.

The models generated in this study are optimized with respect to local interactions. Large (local) errors in ϕ or ψ have a strong effect on the overall fold as well as on the nonlocal, i.e., medium- and long-range, energies. It should be possible to refine and considerably improve the models by including nonlocal forces and by minimizing the total net energy as a function of dihedral angles. The first step in such refinements will concentrate on angles of high variability, keeping angles of high confidence fixed. Compared to the variation of all backbone dihedrals in protein molecules, this approach reduces the complexity of the search problem in conformational space by several orders of magnitude.

We note that the procedures used in this work are un-

refined prototypes. There is a vast number of possibilities for the design of more powerful procedures. We mention a few points that are likely to improve the calculations. (1) In the present study we reported calculations using hexapeptides as the basic building blocks. Obviously, the calculations can be carried out using a number of different fragment sizes. Our preliminary results show that the models can be improved using larger fragments. (2) Models calculated for different fragment sizes can be combined, which should be helpful in removing some of the uncertainties of the local folds. (3) In the case of flip-flop, metastable, and unstable fragments the ensembles contain a large amount of information on the structural preferences. It is conceivable that more sophisticated rules will yield more accurate models.

The strategies presented in this work are fully automatic and unambiguous. This is a desired feature in terms of repeatability of the calculations. The drawback is that automatic procedures often fail in situations where a skilled user easily finds a useful solution. If we are interested in the conformation of a particular protein it is clear that we will try to incorporate all available information on that molecule. For this reason we are currently implementing an interactive version of the program.

Finally we want to emphasize that the force field used in this study is not complete. We used only interactions among the backbone atoms (including C^{β}) and the models were built from short-range interactions $k < 6$ only. The interactions among the remaining side chain and backbone atoms as well as the interactions corresponding to large sequential separations (i.e., nonlocal forces) have been neglected. These interactions are very important for the stability of protein conformations, and it is indeed surprising that many of the structural features of protein conformations can be captured and reproduced by the reduced set of interactions used in this study.

Acknowledgments

The coordinates for the five structures derived from NMR constraints (labeled B4TC02, B4TC04, B4TC15, B4TC22, B4TC28) were kindly supplied by Tad Holak, Max Planck Institut für Biochemie, Martinsried bei München, Germany. We are indebted to all X-ray crystallographers who submitted coordinates to the Brookhaven Protein Data Bank. This work was supported by the Fonds zur Förderung der wissenschaftlichen Forschung under project number P8361-CHE.

References

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science 181*, 223–230.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer based archival file of macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem. 4*, 187.

Burkert, U. & Allinger, N.L. (1982). *Molecular Mechanics.* American Chemical Society, Washington, D.C.

Carson, M. & Hermans, J. (1985). Molecular dynamics workshop laboratory. In *Molecular Dynamics and Protein Structure* (Hermans, J., Ed.), pp. 165–166. University of North Carolina, Chapel Hill.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., & Sippl, M.J. (1990). Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol. 216*, 167–180.

Holak, T.A., Gondol, D., Otlewski, J., & Wilusz, T. (1989a). Determination of the complete three-dimensional structure of trypsin inhibitor from squash seeds in aqueous solution by nuclear magnetic resonance and a combination of distance geometry and dynamical simulated annealing. *J. Mol. Biol. 210*, 635–648.

Holak, T.A., Nilges, M., & Oschkinat, H. (1989b). Improved strategies for the determination of protein structures from NMR data: The solution structure of acryl carrier protein. *FEBS Lett. 242*, 218–242.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol. 213*, 859–883.

Sippl, M.J. & Stegbuchner, H. (1991). Superposition of three-dimensional objects: A fast and numerically stable algorithm for the calculation of the matrix of optimal rotation. *Computers Chem. 15*, 73–78.

Sippl, M.J. & Weitckus, S. (1991). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* (in press).

van Gunsteren, W.F., Berendsen, H.J.C., Hermans, J., Hol, W.G.J., & Postma, J.P.M. (1983). Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. *Proc. Natl. Acad. Sci. USA 80*, 4315–4319.

Weiner, P.K. & Kollman, P.A. (1981). AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Comp. Chem. 2*, 287–299.

Zarbock, J., Oschkinat, H., Hannappel, E., Kalbacher, H., Voelter, W., & Holak, T.A. (1990). Solution conformation of thymosin β_4: A nuclear magnetic resonance and simulated annealing study. *Biochemistry 29*, 7914–7821.