. Mol. Biol. (1991) 221, 961–979

Prediction of Protein Backbone Conformation Based on Seven Structure Assignments

Influence of Local Interactions

Marianne J. Rooman, Jean-Pierre A. Kocher and Shoshana J. Wodak

Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles CP 160, Av. P. Héger, 1050 Brussels, Belgium

(Received 10 January 1991; accepted 29 May 1991)

A method is developed to compute backbone tertiary folds from the amino acid sequence. In this method, the number of degrees of freedom is drastically reduced by neglecting side-chain flexibility, and by describing backbone conformations as combinations of only seven structural states. These are characterized by single values of the dihedral angles ϕ, ψ and ω, representing allowed conformations of the isolated dipeptide. We show that this restrictive model is none the less capable of describing native backbones to within acceptable deviations. Using our backbone description, potentials of mean force are derived from a database of known protein structures, based on statistical influences of single residues and residue pairs on the conformational states in their vicinity along the chain. This yields the force-field component due to local interactions, which is then used to predict lowest-energy conformations from any given amino acid sequence. The prediction algorithm does not require searching conformational space and is therefore extremely fast. Another important asset of our method is that it is able to compute not only the minimum energy conformation, but any number of lowest energy structures, whose relative preferences can be determined from the corresponding computed energy values. The performance of our procedure is tested on short peptides that are likely to be stabilized by local interactions. These include several helical structures and a hexapeptide with a β-bend conformation, corresponding to peptides shown to have relatively well-defined conformations in aqueous solution, and to protein segments believed to adopt their native conformation early during folding. In addition, several flexible peptides are analysed. Except for the problems encountered in predicting observed disulphide bridges in two of the flexible peptides, and in a somewhat larger fragment comprising residues 30 to 51 of bovine trypsin inhibitor, prediction results compare very favourably with experimental data. Potential applications of our procedure to protein modelling and its extension to protein folding are discussed.

Keywords: peptide conformation; structure-prediction; protein folding

1. Introduction

In the absence of structural information on homologous proteins, prediction of protein tertiary structure from the amino acid sequence involves such a large number of variables that drastic approximations are unavoidable. A commonly used approach consists in predicting secondary structure first (for a review, see Schulz, 1988; Fasman, 1989; Garnier *et al.*, 1990), and then assembling the secondary structure elements into a three-dimensional fold, following rules derived from known protein structures (Cohen *et al.*, 1979; Ptitsyn, 1981; Murzin & Finkelstein, 1988). The main problem with this strategy resides in the relatively poor performance of secondary-structure prediction methods (Kabsch & Sander, 1983*a*). This has been attributed in part to the limited amount of protein structural data (Rooman & Wodak, 1988), but more importantly, to the neglect of non-local interactions, between residues far removed along the sequence, but close in space (Kabsch & Sander, 1984; Rooman & Wodak, 1991). It has been suggested, however, that whereas the native secondary structure is clearly influenced by non-local effects, local interactions may dominate in regions of the protein that adopt stable conformations early during folding (Wright *et al.*, 1988). Thus, secondary-structure predictions would perform much better than average in these regions (Rooman & Wodak, 1991), and would yield a physically meaningful starting structure, whose

0022-2836/91/190961-19 $03.00/0

evolution towards the native fold, essentially due to non-local interactions, could conceivably be simulated.

A quite different approach entirely bypasses secondary structures, by proceeding directly to predict the three-dimensional conformation of lowest energy from the amino acid sequence (for reviews, see Nemethy & Scheraga, 1977; Skolnick & Kolinski, 1989*a*). This is usually achieved by more or less randomly sampling the conformations of the entire polypeptide chain (Paine & Scheraga, 1987; Li & Scheraga, 1987) and evaluating generated structures by energetic criteria. Here, the major difficulty lies in determining the level of detail with which the system must be described to yield meaningful results, without prohibitive computations. Detailed atomic models being too complex to be useful, except for very short peptides, approximations are made to reduce the number of independent parameters, while maintaining access to relevant regions of conformational space. These include simplified side-chain and backbone representations (Levitt, 1976; Wilson & Doniach, 1989), coarse sampling of backbone dihedral angles (Miyazawa & Jernigan, 1982), the use of regular lattice models (Ueda *et al.*, 1978; Skolnick & Kolinski, 1989*b*), or limiting interatomic distances to a discrete set (Sippl, 1990). Energetic criteria, used to evaluate generated conformations, are based on more-or-less simplified force-fields, either empirical (Levitt, 1976; Paine & Scheraga, 1987) or derived from known protein structures (Miyazawa & Jernigan, 1985; Wilson & Doniach, 1989; Sippl, 1990). To limit the possibilities further, simulations are often biased in favour of the native fold, by assigning native secondary structure propensities (Skolnick & Kolinski, 1989*b*), or by requiring that generated conformations lie preferably within the spatial envelope of the native structure (Miyazawa & Jernigan, 1982). Thus, although these procedures provide valuable insight, it is often hard to evaluate how relevant and successful they may be for actual prediction of a protein structure.

Here, we lay the ground for a method that attempts to combine the advantages of the two approaches described above. Aspects of the methodology are borrowed from secondary structure predictions, but they are incorporated into a procedure that determines tertiary structure from sequence in a single step. First, we derive a backbone representation that effectively reduces the allowed conformational space of the protein backbone, while uniquely defining atomic co-ordinates for the entire main chain. Every residue along the amino acid sequence is assigned to one of seven structural states, corresponding to allowed conformations of the isolated dipeptide (Liquori, 1969; Scheraga, 1971; Ralston & De Coen, 1974). Each state is represented by a single value of the backbone dihedral angles.

After verifying that this parsimonious representation is none the less capable of describing the native backbone sufficiently well, we turn to the problem of predicting stable backbone folds based on sequence information. This requires, among other things, a force-field that describes interactions between protein atoms and residues in a physically meaningful way. In particular, it should adequately describe contributions from local, as well as non-local interactions. On the basis of the assumption that these contributions are additive, we focus here on driving and testing the local force-field component. This component is derived by a knowledge-based approach, in which a database of well-resolved and refined protein structures is surveyed. Probabilities for a given residue to adopt the seven discrete conformational states are computed, considering influences of neighbouring residues along the sequence. This is done by combining principles of the GOR† method (Garnier *et al.*, 1978; Gibrat *et al.*, 1987) with those of Rooman & Wodak (1988, 1990). Side-chain degrees of freedom are not explicitly considered. Probabilities are then translated into potentials of mean force, following the approach of Sippl (1990). Prediction of lowest energy conformations for a given sequence is achieved by determining for each residue the structure assignments with lowest energy contributions. This procedure does not require searching conformational space, and is therefore extremely fast.

To test our method, predictions are performed on short peptides or protein regions that are stabilized mainly by local interactions. The chosen examples include protein segments believed to adopt their native conformation early during folding (Baum *et al.*, 1989; Matouschek *et al.*, 1989; Roder *et al.*, 1988), as well as several short peptides shown experimentally to adopt either a well-defined conformation in aqueous solution (Brown & Klee, 1971; Bierzynski *et al.*, 1982; Shoemaker *et al.*, 1985, 1987; Eisenberg *et al.*, 1986; Marqusee & Baldwin, 1987; Marqusee *et al.*, 1989; Reed *et al.*, 1988) or to be flexible (Meraldi *et al.*, 1977; Hallenga *et al.*, 1980; Shoemaker *et al.*, 1985; Oas & Kim, 1988). Results of the calculations are presented, and compared to the structural information available from experiment. Application of the procedure to protein modelling and the promises it may hold for tackling the folding problem are discussed.

2. Materials and Methods

(a) *Protein database*

Information on known protein structures used in this work is extracted from the database SESAM (Huysmans *et al.*, 1991). This includes amino acid sequence, secondary structure obtained by DSSP (Kabsch & Sander, 1983*b*), atomic co-ordinates and dihedral angle values of all proteins of the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). The proteins considered here belong to a subset of the full database. It comprises 69 highly-resolved

† Abbreviations used: GOR, Garnier–Osguthorpe–Robson; DSSP, dictionary of protein secondary structure; r.m.s., root mean square; c.p.u., central processing unit; BPTI, bovine pancreatic trypsin inhibitor.

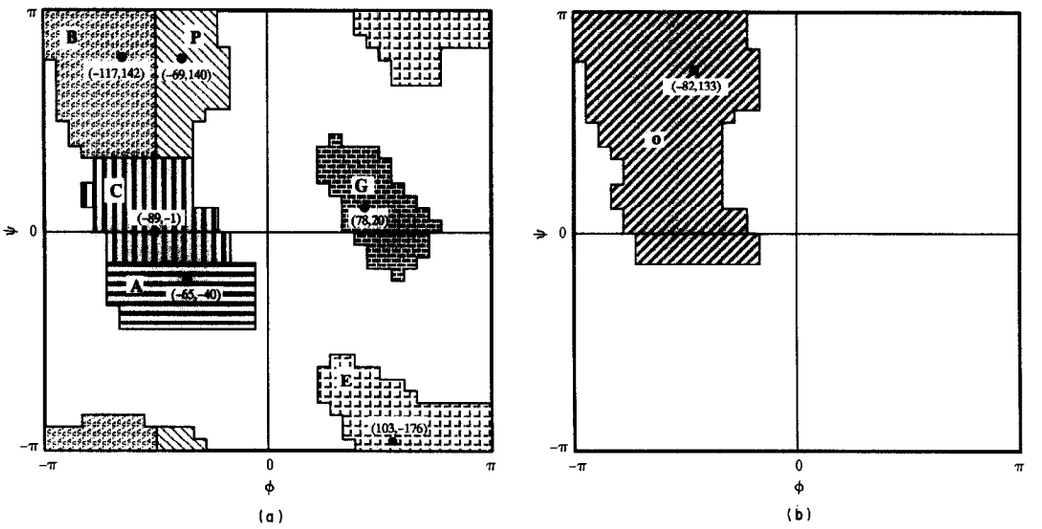

Figure 1. Division of the Ramachandran ϕ, ψ map into 7 domains, (a) 6 for $\omega \approx 180°$ and (b) 1 for $\omega \approx 0°$. These domains are represented by the structure assignments A, C, B, P, E, G and o. The representative value of each domain is indicated (see Materials and Methods, section (b)). All ω values lower than $-150°$ or higher than $150°$ are considered as equal to $180°$, and all ω values between $-30°$ and $30°$ are assigned the value of $0°$. Residues with ω values outside these ranges are probably due to experimental inaccuracy and were not taken into account. The protein set considered consists of the following 69 proteins, given by their Brookhaven Protein Data Bank codes (Bernstein *et al.*, 1977): 1ACX, 1BP2, 1CC5, 1CCR, 1CPV, 1CRN, 1CSE, 1FD2, 1GCR, 1GD1, 1GOX, 1GP1, 1HIP, 1HMZ, 1HOE, 1LZ1, 1MLT, 1NXB, 1PCY, 1PFK, 1PHH, 1PPT, 1PRC, 1RHD, 1RNT, 1SN3, 1TIM, 1TPP, 1UTG, 2ABX, 2ACT, 2ALP, 2APR, 2B5C, 2CAB, 2CCY, 2CDV, 2CNA, 2CRO, 2CTS, 2CYP, 2FB4, 2GN5, 2LBP, 2LH4, 2LZM, 2PAB, 2SNS, 2SOD, 2STV, 2TS1, 2VHX, 3ADK, 3DFR, 3FXC, 3GRS, 3TLN, 3WGA, 451C, 4FXN, 4HHB, 4MDH, 4RXN, 5CPA, 5PTI, 6LDH, 7RSA, 8ADH, 8CAT.

($\leq 2{\cdot}5$ Å resolution; 1 Å = 0·1 nm) and refined proteins, with less than 20% sequence identity, listed in the legend to Fig. 1.

(b) *Representation of backbone conformation*

A well-known way of reducing the number of parameters needed to describe the conformation of a protein backbone is to consider that bond lengths and angles are fixed. Then only 3 internal degrees of freedom, the usual backbone dihedral angles ϕ and ψ, and ω, defined here as the angle about the peptide bond preceding the residue, are sufficient to describe the conformation fully. Searching conformational space as a function of these internal degrees of freedom remains nevertheless computationally prohibitive, even for a small peptide. Further approximations are therefore required.

Due to the influence of covalent and non-covalent interactions, the backbone dihedral angles cannot adopt any value, but tend to be close to a certain number of ideal angles. These correspond to energetically favourable conformations for the dipeptide, which are separated from each other by energy barriers (Ramachandran & Sasisekharan, 1968). This feature is apparent when values of the ϕ and ψ angles observed in known protein structures are plotted on the Ramachandran map. They tend to cluster in domains, the number and limits of which differ somewhat from author to author. An obvious approximation, compatible with the physical properties of real backbones, consists of restricting the values of backbone dihedral angles to a discrete set that corresponds to ideal conformations characterizing the allowed domains (Liquori, 1969; Scheraga, 1971; Ralston & De Coen, 1974).

Inspired by Wilmot & Thornton (1990), we consider a total of 7 domains in (ϕ–ψ–ω) space. These include 6 domains for the *trans* peptide conformation ($\omega \approx 180°$; Fig. 1(a)). Domains A and C represent α and 3_{10}-helix conformations, respectively, and domains B and P correspond to extended structures, with B representing β-strand conformations. The domains of positive ϕ, G and E, are weakly populated, and occur essentially in glycine residues. The 7th domain, termed o, merges all ϕ, ψ values corresponding to *cis* peptide conformations ($\omega \approx 0°$) (Fig. 1(b)). These occur seldom, and almost exclusively in proline residues. Each of the 7 domains is then represented by a single (ϕ, ψ, ω) value. This value is computed as the average of the observed (ϕ–ψ) angles in the database of known structures that fall within the domain. Separate averages are computed for $\omega \approx 0°$ and $180°$. Moreover, to allow for adequate reconstruction of secondary structure elements, only the α-helix and β-strand portions of known protein structures are used to compute the average values that represent regions A and B, respectively. A succession of structure assignments A, C, B, P, E, G and o, one for each residue along the polypeptide chain, then defines a specific backbone structure. Each of the assignments corresponds indeed to a well-defined value of the backbone dihedral angles, i.e. the representative value of the corresponding domain. From them, a complete 3-dimensional structure can be constructed, given standard bond lengths and valence angles.

(c) *Testing the feasibility of the backbone representation*

An important point to investigate at the onset is whether the 7 chosen structure assignments contain sufficient information to reconstruct any protein backbone to within an acceptable root mean square (r.m.s.) deviation from the crystal structure. It is well known that back-

bones reconstructed using the dihedral angles from the crystal structure, and standard values for bond lengths and angles, may depart significantly from the native fold. They tend to be similar to the native backbone locally, but small, local deformations are introduced, which can propagate to significantly alter the direction of the chain. With a reduced set of allowed backbone dihedral angle values, as in our model, the problem only becomes worse, in agreement with previous observations (Liquori, 1969). Thus, a straightforward procedure, whereby the backbone is reconstructed by assigning to each residue the representative (ϕ, ψ, ω) value of the domain to which its observed dihedral angles belong, cannot yield reasonable structure descriptions for average size proteins. It is, on the other hand, quite adequate for short protein fragments.

An alternative approach consists of searching for combinations of structure assignments that yield an accurate overall description of the observed backbone structure, irrespective of the native (observed) (ϕ, ψ, ω) values. This is done here by directing the choice of the conformational state assigned to each residue, so as to yield backbones that are most similar to the native 3-dimensional structures, and that conserve its physically important features, such as hydrogen bonds and disulphide bridges.

Similarity between backbone conformations is estimated from the r.m.s. deviation between superimposed N, C^{α} and C backbone atoms, computed using the algorithm U3BEST (Kabsch, 1978). Hydrogen bonds and disulphide bridges in the native backbone are determined according to criteria defined in the molecular modelling package BRUGEL (Delhaise *et al.*, 1985). In the reconstructed backbone, hydrogen bonds are considered as identifiable when O and H backbone atoms are closer than 4 Å. For disulphide bridges to be identifiable, C^{β} atoms are required to be closer than 6·5 Å. Values for the standard bond lengths and angles are taken as the averages computed from the proteins of our database.

The search algorithm systematically generates structure assignments for each residue, chosen among the 7 discrete (ϕ, ψ, ω) states described above. On the basis of these assignments, full backbones are constructed. The 100 structures with lowest r.m.s. deviation relative to the native structure, in which all native disulphide bridges are identifiable, are then retained. From these, conformations that maximize the number of identifiable hydrogen bonds in secondary structure elements, α-helices and β-sheets, but not in loops, are selected. This is done to obtain an improved description of the backbone in regions of the protein core, mainly composed of regular secondary structures. Larger departures from the native structure are tolerated in loop regions, in agreement with experimental evidence that these regions are usually more flexible (Frauenfelder *et al.*, 1979; Petsko & Ringe, 1984).

To speed up the search procedure, only a subset of all possible backbone conformations is actually considered. This is achieved by imposing an upper threshold on the allowed r.m.s. deviation. Structural states are systematically assigned to each residue, starting from the N terminus, with the initial r.m.s. threshold set to 3 Å. Each time a structural state is assigned to a residue, we compute the r.m.s. deviation between the segment, composed of all residues up to the current one, and the equivalent native fragment. If this r.m.s. deviation is lower than the threshold value, the segment is lengthened. If it is not, the program backtracks. It generates alternative conformational states for each residue, and tests the r.m.s. deviation of the corresponding backbones, until another backbone segment with r.m.s. deviation lower than the threshold value is found. Whenever a complete backbone structure within the allowed r.m.s. deviation is obtained, it is inserted into the list of 100 retained solutions, and the r.m.s. threshold is updated to the worst r.m.s. deviation in this list. This ensures that the portion of conformation space to be searched in subsequent trials remains large enough. The entire procedure is repeated, until all solutions complying with the imposed threshold are found.

Clearly, this algorithm does not guarantee to find the backbone conformation with the smallest r.m.s. deviation from the native one. It is indeed conceivable that such an optimally fitting backbone may contain segments with r.m.s. deviations higher than the threshold value, which are rejected by our procedure. Our algorithm has, however, a fair chance of finding the optimal solution, since not 1 but 100 solutions are retained.

The sole purpose of this purely geometrical procedure has been to establish whether there exist combinations of structural states that provide an adequate overall description of observed backbones. Therefore, little attempt has been made at this stage to increase its computational efficiency, and it has been applied to only 11 low molecular weight proteins (<80 residues). For a 26 residue protein (melittin), the procedure requires of the order of 100 c.p.u. min on a single Silicon Graphics 340 processor. During this time, it generates 2123 complete backbones consistent with the specified r.m.s. threshold. For a protein 3 times that size, such as the α-amylase inhibitor, the time required is multiplied by a factor of about 200, and the number of backbones scanned is 6886.

(d) *Method for predicting the 3-dimensional backbone structure*

Our approach consists of deriving potentials of mean force from probabilistic relations between the amino acid sequence and the backbone structure, described by combinations of the 7 discrete states. These potentials are then used to predict the lowest energy conformations of a given amino acid sequence. This study is devoted to implementing and testing the force-field component due to local interactions. Non-local interactions, between residues far removed along the chain but close in space, are essentially neglected, and added only occasionally as simple filters on predicted structures.

Our method is inspired by 3 different structure prediction techniques. The 1st is the secondary structure prediction method GOR (Garnier *et al.*, 1978; Gibrat *et al.*, 1987), which computes the probabilistic influences of single residues or residue pairs on their own conformation. The 2nd procedure (Rooman & Wodak, 1988, 1991; Rooman *et al.*, 1990) generalizes the concept of sequence-structure association, taking into account the fact that residues can influence the conformation of other neighbouring residues along the chain. For translating probabilities into potentials of mean force, we rely on basic concepts of statistical mechanics, and on specific concepts developed by Sippl (1990) for tertiary structure prediction purposes, which we adapt to our model.

(i) *Computing probabilistic relations between sequence and structure*

To compute the probabilistic relations between the amino acid sequence and the backbone conformations, by the combined GOR (Garnier *et al.*, 1978; Gibrat *et al.*, 1987) and Rooman & Wodak (1988, 1991) approaches, we use the backbone structure assignment technique that

yields reasonable local, but not global, structure representations. It consists in assigning to each residue the (ϕ, ψ, ω) value of the domain to which its observed dihedral angles belong. Since we focus on the force-field components due to local interactions, this assignment should be more appropriate than that obtained by the directed-search method. The latter have been derived by applying global geometric constraints, and could therefore be viewed as representing backbone conformations generated by prediction procedures that take into account both local and non-local effects.

Owing to limited database size, approximations must be made in computing the probability $P^U(\Xi)$ that a given amino acid sequence U adopts the conformation Ξ, with Ξ being described by a succession of structure assignments. Following Garnier *et al.* (1978) and Gibrat *et al.* (1987), residue influences on different structure assignments, s_k, along the polypeptide chain are considered to be independent. As a result, the probability $P^U(\Xi)$ of a given backbone conformation is expressed as a product of the probabilities $P^U(s_k)$ of structure assignments of individual residues.

Furthermore, only the influences of single residues or residue pairs along the sequence are considered. Indeed residue triplets occur, on average, 1·8 times in our 69 protein database of 14,204 residues, which is obviously too little to yield reliable statistics. We know from previous analyses (Rooman & Wodak, 1988) that single residue influences lead, on average, to worse predictions than influences from residue pairs, since they contain less specific structural information. It was therefore tempting to consider here only residue pairs. However, our tests show that they yield poor statistics. Pairs are indeed more frequent that triplets, but apparently still not frequent enough. They occur, on average, 36 times in the database, but some, like M–W, appear only once. We choose, therefore, to combine statistical influences from both singlets and pairs. We find, indeed, that multiplying the corresponding probabilities leads to slightly better prediction scores than when only residue pairs are considered. Note, however, that the specific choice of the approximation depends critically on database size. In larger databases, our approximation will probably be no longer optimal. It may then become preferable to completely neglect single residue influences, and to consider instead influences from triplets.

Thus, single residue probabilities $P^{u_i,u_i}_{i-k,i-k}(s_k)$ and pair probabilities $P^{u_i,v_j}_{i-k,j-k}(s_k)$ are computed from our protein subset. $P^{u_i,u_i}_{i-k,i-k}(s_k)$ is the probability that a single amino acid u at position i along the polypeptide chain is associated with a structure assignment s at position k, with $i-8 \le k \le i+8$. $P^{u_i,v_j}_{i-k,j-k}(s_k)$ is the probability that an amino acid u at position i and an amino acid v at position j, with $i \neq j$, is associated with a structure assignment s at position k, with $i-8 \le k \le i+8$ and $j-8 \le k \le j+8$. It is noteworthy that, as the distance between specified positions increases, the probability, P, becomes less dependent on the residue type, and tends towards the frequency of the corresponding structure assignment in the database. It is hence worthless to consider residues and assignments separated by more than 8 positions.

(ii) *Deriving potentials of mean force from probabilities*

To determine the conformation Ξ with highest probability $P^U(\Xi)$, for a given amino acid sequence U, is equivalent to finding the lowest energy conformation $E^U(\Xi)$. Assuming a Boltzmann distribution, energies and probabilities are indeed related by Boltzmann's law, here in its discrete state space version:

$$E^U(\Xi) = -kT(\ln P^U(\Xi) + \ln Z^U)$$
$$\text{where } Z^U = \sum_{\Xi} e^{-E^U(\Xi)/kT}. \quad (1)$$

Z^U denotes the partition function, k Boltzmann's constant and T the temperature. We take $T = 293$ K, assumed to be roughly representative of protein crystallization temperatures.

According to the approximations described above, the total energy $E^U(\Xi)$ of a sequence U comprising N residues, may be decomposed into energy contributions $E^U(s_k)$ from individual structure assignments s_k, that are assumed to be non-correlated. These are in turn computed as sums-of-energy contributions resulting from influences of all neighbouring residues and residue pairs on the conformation s_k, as follows:

$$E^U(\Xi) = \sum_{k=1}^{N} E^U(s_k) = -kT \sum_{i,j,k=1}^{N} \frac{1}{\zeta_k} \times (\ln P^{u_i,v_j}_{i-k,j-k}(s_k) + \ln Z^{u_i,v_j}_{i-k,j-k}), \quad (2)$$

with $k-8 \le i,j \le k+8$. The normalization factor ζ_k represents the number of times that the influence of each residue contributes to the probability of the conformation s_k. Its value here is 17, corresponding to the width of the window considered around each residue, everywhere except near chain ends.

In general, the proportion of each structure assignment predicted from the minimum value of the energy $E^U(\Xi)$ does not coincide with the observed one, the most frequent conformations being favoured. This is best illustrated by the simple example of a residue pair that is most often, say, in 70% of its occurrences, associated with a given conformation s. Every time this residue pair occurs in any protein sequence, it will be predicted to adopt the conformation s. As a result, its frequency in the predicted assignments will be 100%, and not the observed frequency of 70%, as it should be. The procedure used in Garnier *et al.* (1978) and Gibrat *et al.* (1987) to obtain roughly the correct assignment frequency, consists in determining preference parameters ("decision constants") favouring certain assignments. We choose instead to use the approach of Sippl (1990), which does not require evaluating these *ad hoc* parameters. Following this approach, we compare the probability $P^{u_i,v_j}_{i-k,j-k}(s_k)$ that a residue pair (u_i, v_j) adopts a given conformation s at position k, to the average probability $P_{i-k,j-k}(s_k)$ of finding the conformation s_k, irrespective of the residues at positions i and j. This amounts to computing the difference between $E^U(\Xi)$ and its average, $E(\Xi)$, over all amino acid sequences, yielding the net energy $\Delta E^U(\Xi)$:

$$\Delta E^U(\Xi) = E^U(\Xi) - E(\Xi) = -kT \sum_{i,j,k=1}^{N} \frac{1}{\zeta_k} \times \left(\ln \frac{P^{u_i,v_j}_{i-k,j-k}(s_k)}{P_{i-k,j-k}(s_k)} + \ln \frac{Z^{i_i,v_j}_{i-k,j-k}}{Z_{i-k,j-k}} \right). \quad (3)$$

The only quantities on the right-hand side of eqn (3) that are not straightforward to compute are the partition functions Z. Their evaluation is necessary for comparing conformational energies of different amino acid sequences. For a given sequence, however, they may be omitted, because they are conformation independent and only contribute to the net energy as additive constants (Sippl, 1990). In what follows, the expression "net energy" will refer to the 1st term in eqn (3).

(iii) *Prediction method*

Predictions are performed by determining for a given amino acid sequence, the conformations of lowest net energy, defined by eqn (3) with the partition function term omitted. These conformations can be derived directly, from the probabilities $P^{u_i,v_j}_{i-k,j-k}(s_k)$ and $P_{i-k,j-k}(s_k)$, without performing any kind of conformational search. Deriving the structure with lowest total energy is particularly straightforward. Indeed, since the conformations of successive residues along the chain are considered as non-correlated, it is obtained by simply combining the lowest energy structural assignments of individual residues.

The algorithm for determining the next-best ranking conformations is somewhat more tricky. It starts by computing, for each residue along the chain, the net energy differences $\Delta\Delta E^U(s_k)$, between its most probable conformation and the 6 others. These differences are merged in a single list irrespective of their position in the sequence, and then sorted in order of increasing $\Delta\Delta E^U(s_k)$ values. From this sorted list, any number of lowest energy conformations can be quickly derived, using pointers to the position in the sequence and the type of structural state. For example, the conformations with 2nd and 3rd-best ranking overall energies, each differ from the minimum energy structure by only a single assignment, referring respectively to the 1st and 2nd $\Delta\Delta E^U(s_k)$ values in the sorted list. The next-best ranking conformation differs from the minimum energy structure by 2 assignments that correspond to the 2 lowest $\Delta\Delta E^U(s_k)$ values, provided that they refer to different residues, and that their sum is lower than the 3rd $\Delta\Delta E^U(s_k)$ in the list. Otherwise, it differs by only 1 assignment, corresponding to the 3rd-lowest $\Delta\Delta E^U(s_k)$ value.

To avoid considering conformations with overlapping atoms, a simple steric hindrance test is added to the generated structures. Each combination of the 7 structural states obtained by the algorithm described above, is translated into a full 3-dimensional backbone structure, and the distances between C^α atoms are computed. Whenever any of these distances is closer than 2·5 Å, the structure is rejected. This constraint is not too permissive, considering that our method generates what could be qualified as low-resolution backbones.

(iv) *Correcting for contributions from rare events*

Predictions based on probabilities derived from known structures are clearly affected by limited database size, although the situation has been somewhat improved by combining approximations from pairs and from single residues. Probabilities $P^{u_i,v_j}_{i-k,j-k}(s_k)$ should indeed be quite reliable if they are computed from single residues or from frequent residue pairs, but not if they are based on sequence patterns that occur only a few times in the database. To correct selectively for influences from rare events, different weights are attached to computed preferences, according to their number of occurrences, following Sippl (1990). More precisely, $P^{u_i,v_j}_{i-k,j-k}(s_k)$ in eqn (3) is transformed into:

$$P^{u_i,v_j}_{i-k,j-k}(s_k) \rightarrow \frac{1}{\sigma + m^{u_i,v_j}_{i-k,j-k}} \times (\sigma P_{i-k,j-k}(s_k) + m^{u_i,v_j}_{i-k,j-k} P^{u_i,v_j}_{i-k,j-k}(s_k)). \quad (4)$$

where $m^{u_i,v_j}_{i-k,j-k}$ is the number of observations from which $P^{u_i,v_j}_{i-k,j-k}(s_k)$ has been computed, and where σ is a constant, σ^{-1} representing the weight of an observation relative to the average probability $P_{i-k,j-k}(s_k)$. Eqn (4) reduces to an identity when the number of observations, $m^{u_i,v_j}_{i-k,j-k}$, tends to infinity, the rate of convergence depending on the chosen value of σ. With a zero value of σ, all sequence patterns have equal predictive power, irrespective of their frequency. As σ increases, the weight of patterns that

Table 1

Prediction results as a function of σ

σ	Prediction score (%)	Number of predicted structures						
		A	C	B	P	G	E	o
0	43·10	6910	1210	4276	781	6	0	0
1	45·38	5128	1995	4116	1412	475	57	0
2	45·39	4893	2057	3984	1499	639	111	0
3	45·39	4771	2084	3910	1521	737	160	0
4	45·41	4653	2104	3864	1567	790	205	0
5	45·48	4573	2125	3819	1599	828	239	0
6	45·54	4506	2135	3793	1629	855	265	0
7	45·50	4467	2150	3754	1650	863	299	0
8	45·32	4424	2161	3740	1666	875	317	0
9	45·20	4380	2171	3739	1688	870	335	0
10	45·07	4345	2183	3725	1700	874	356	0
20	44·52	4181	2224	3641	1793	849	480	15
50	43·15	4002	2215	3484	1829	871	673	109
10^6	31·26	3186	1368	2315	1048	534	1610	3122
Number of observed structures		4724	2033	3527	1893	623	349	34

The average prediction score is obtained by the jack knife test, performed on the 69 proteins of our sample, as described in the text. Column 1 lists values of σ (eqn (4)) used in individual prediction experiments. For each of the 7 classes of structural states, labelled according to the definitions in Fig. 1, the number of predicted and observed structure assignments is given. The maximum score of 45·5% is obtained for $\sigma = 6$. This score may seem low, compared, say, to the score of secondary structure predictions which is about 60%. But the probability of correctly predicting 3 secondary structure classes, assuming that their relative average frequency is conserved, is expected to be higher than that of predicting 7 structural classes from simple statistical considerations. Random predictions yield scores of 38% for the 3 secondary structure classes, and of only 25% for our 7 states.

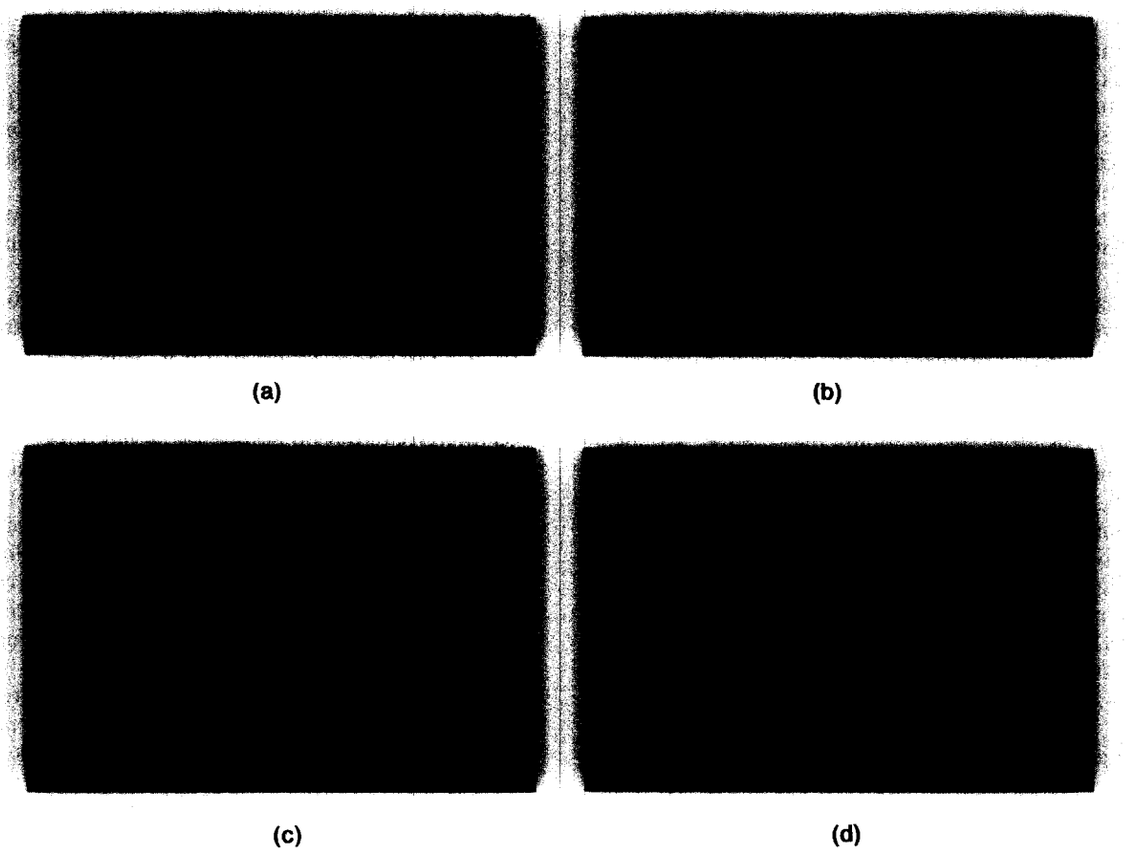

Figure 2. Ribbon drawings of superimposed native (blue) and reconstructed (red) backbone structures based on 7 structure assignments and using the directed search method described in Materials and Methods, section (c), for (a) crambin (1CRN), (b) trypsin inhibitor (5PTI), (c) uteroglobin (1UTG) and (d) α-amylase inhibitor (1HOE).

occur only once decreases. The value of σ is determined here so as to obtain the highest overall score in predictions of backbone assignments for proteins of our database.

To measure the overall accuracy of a prediction method, it must be applied to proteins that exhibit low sequence identity with those used for deriving prediction rules. In the absence of a large independent set of test proteins, the jack knife test (Efron, 1982) constitutes a reliable procedure. It consists of dividing the whole database into a learning set, comprising all but one protein, and a test set, that contains the removed protein. The statistical parameters $P^{u_i,v_j}_{i-k,j-k}(s_k)$ are compiled from the learning set, and applied to the test protein so as to predict its lowest energy conformation. A prediction score is then computed. It is defined as the percentage of predicted structure assignments that coincide with the observed ones. To have a reliable evaluation of prediction accuracy, each protein of the database is, in turn, chosen as test protein. The final prediction score of our method is computed as the average of the scores obtained for individual proteins, weighted by their number of residues.

The prediction score computed from the jack knife test is given as a function of σ in Table 1. The highest average prediction score is obtained for $\sigma = 6$, which is the value used throughout this study. With this value, the number of predicted and observed structure assignments is, moreover, roughly identical. Note that determining the value of σ from predictions performed on structures belonging to the learning set would not make sense. Indeed, optimal predictions inside the learning set are obviously obtained by giving highest weights to sequence patterns that occur seldom, i.e. by setting $\sigma = 0$, while the opposite is expected for predictions outside the learning set.

3. Results

(a) *Assessment of the backbone representation*

Structure assignments, based on the seven (ϕ, ψ, ω) states, have been derived using two procedures: a systematic search that directs the choice of local conformations towards the native global fold, and a local technique that assigns each residue to the domain in which its (ϕ, ψ, ω) angles occur. Results obtained for 11 proteins are given in Table 2.

The r.m.s. deviations between native structures and backbones generated by the directed-search procedure lie between 1·18 Å and 1·67 Å. The fraction of the native hydrogen bonds that can be identified in the rebuilt backbones varies between 50% and 100%. The reconstructed backbone of neurotoxin B (1NXB) departs most from the native structure, with a r.m.s. deviation of 1·67 Å. Inspection of the structure assignments obtained by the local technique, shows that 15 of the 62 residues have angles outside allowed domains. Such unusual (ϕ, ψ, ω) values occur only three times in all other

proteins examined here, suggesting that the coordinates of this specific Brookhaven Protein Data Bank entry may not be reliable, despite the high resolution (1·3 Å) of the diffraction data. While occasional unfavourable dihedral angle values can in principle be attributed to the influence of tertiary interactions, this cannot apply here, since nearly a quarter of the residues are involved.

Another poorly reconstructed protein backbone is that of wheat germ agglutinin (3WGA), which displays an r.m.s. deviation of 1·54 Å from the native structure. Here, this could be attributed to its irregular structure, which contains virtually no α-helix or β-sheet structure, but mainly loops and turns. Those are less precisely described by the seven structure assignments used here. All other reconstructed backbones display deviations lower than 1·5 Å relative to the corresponding crystal structures, with at least 63% of the native hydrogen bonds identified.

Figure 2 illustrates the superimposition of native and reconstructed backbone structures of crambin (1CRN), trypsin inhibitor (5PTI), uteroglobin (1UTG) and α-amylase inhibitor (1HOE). We see that the fit is particularly good in the core regions that include most of the secondary structure elements. Proteins with high β-strand content, such as α-amylase inhibitor (1HOE), are somewhat less-well reconstructed, since β-strand conformations are usually more diverse, and thus more difficult to characterize with a single (ϕ, ψ, ω) value. Helices, on the other hand, are usually less distorted and closer to ideal structures.

Considering that our representation has only seven degrees of freedom per residue, the fit between reconstructed and native backbones is rather satisfactory. By emphasizing resemblance to the native conformation in regions of regular secondary structure, it ensures furthermore that, at least in these regions, the deviations from the native (ϕ, ψ, ω) values are comparable to the departures from local energy minima encountered in real proteins. We find, in addition, that each of the assignments of backbone structure listed in Table 2 is not the only acceptable solution to the description of the native backbone. Among the 100 best-fitting structures, many other combinations of structure assignments that fit nearly equally well can be found. These solutions differ from one another only in small segments, that may be substituted without affecting the overall fit to the native structure, and could therefore be taken to characterize conformational flexibility.

Furthermore, it is interesting to compare the obtained backbone descriptions with those derived by the local assignment method. We observe that the assignments derived by the two techniques are usually quite similar in secondary structure regions. In helices, the differences are essentially limited to switches between A (α-helix) and C (3_{10}-helix) states. In β-strands, changes occur mainly between B and P, which represent similar extended conformations, except in cases where β-strands depart significantly from ideal conformations. In loop regions, on the other hand, assignments obtained by both procedures may be quite different, which is not unexpected, since the backbone construction method preferentially concentrates local deformations in these regions, in favour of a better global fit to the native fold.

Further comparison reveals that segments of backbones generated by both procedures exhibit, on the average, the same degree of similarity with respect to the native structure. For instance, the average r.m.s. deviations of eight-residue fragments are, in both cases, equal to 1·0 Å. But as expected, the average deviation of the native (ϕ, ψ, ω) values from the assigned discrete values is much lower for assignments based on closest (ϕ, ψ, ω) domain (13°, 14° and 4° for ϕ, ψ and ω, respectively) than for those based on best global fit (40°, 38° and 5° for ϕ, ψ and ω, respectively). This confirms the choice of using the former assignments for the prediction method developed in this study, which is restricted to local influences. It would in fact be interesting to compare predictions obtained with parameters derived with both assignment methods. This is unfortunately impossible at present, since the directed search procedure is too time consuming to be applied to all the proteins in our database (see Materials and Methods, section (c)). It could, however, be argued that differences between statistical preferences computed from the two types of assignment methods would vanish in sufficiently large databases, in particular since the frequency of each conformation is roughly equal with both assignment techniques.

(b) *Prediction of three-dimensional backbone structures*

Having shown that the seven structure assignments defined in Figure 1 contain enough information to represent adequately the three-dimensional structure of a protein backbone, we proceed to predict, from the amino acid sequence, lowest energy conformations expressed as combinations of these assignments. Our procedure being restricted to local interactions, we test it on short peptides and protein segments that are mainly stabilized by these interactions. Segments of the polypeptide chain that form early during the folding process, as well as small peptides that adopt well-defined conformations in aqueous solution, have a high probability of belonging to this category (Wright *et al.*, 1988; Rooman & Wodak, 1991).

Results obtained for a number of such peptides are summarized in Table 3. Peptides 1 to 4 in Table 3 were shown, by experiment (Brown & Klee, 1971; Bierzynski *et al.*, 1982; Shoemaker *et al.*, 1985, 1987; Eisenberg *et al.*, 1986; Marqusee *et al.*, 1989; Marqusee & Baldwin, 1987), to adopt stable, fully helical conformation in water. All of them are correctly predicted as α-helices by our procedure. Moreover, relevant information about the extent to

Table 2

Assessment of the backbone representation

Protein	Amino acid sequence, secondary structure, structure assignments based on closest ϕ–ψ–ω domain, structure assignments based on best global fit	r.m.s. deviation	% H-bonds
1MLT	GIGAVLKVLTTGLPALISWIKRKRQQ CHHHHHHHHHTTHHHHHHHHHHHHC XAAAAAAAAABYAAAAAAAAAAAAP PAAAAAAAAAACAAAAAACACAAAAX	1·18	100
1PPT	GPSQPTYPGDDAPVEDLIRFYDNLQQYLNVVTRHRY CCCCCCCCCTTSCHHHHHHHHHHHHHHHHHHTTCCC XPPPPPPCEACPPAAACAAAAAAAAAAAAAACGBCA APBPBPPAECCEoAACAAAAAAAAAAACACACGBCX	1·18	100
3WGA	XRCGEQGSNMECPNNLCCSQYGYCGMGGDYCGKGCQDGACWTSK CCCBGGGTTBCCGGGCEECTTSCEECSHHHHSTTCCSSSCSSCC XBBEACCGGPBPPGGPPBPCCGBPBBECCACPPGPABECPABPP PBBEACCGGPBPPGGPPAEACGAGAEEAACBEPGCGEEAPEPPX	1·54	50
1CRN	TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN CEECSSHHHHHHHHHHHTTTCCHHHHHHHHHSCEECSSSCCCGGGCC XBBBCBAAAAAAAAAAACACGPPAAAAAAACGPBBBCPCBPPCCBCC BBBBCBACAAAACACACACGAEAAACAACBCBBPBAPAPBPPGCEX	1·22	100
4RXN	MKKYTCTVCGYIYDPEDGDPDDGVNPGTDFKDIPDDWVCPLCGVGKDEFEEVEE CCCEEETTTCCEECTTTCBGGGTBCTTCCGGGSCTTCBCTTTCCBGGGEEECCC XBPBBPCACGPBPBACAEBACCGBPPGPPACCPPCCPBPAACGPPCCCPBPPBA CPBBBPACBCBBPBACAEAGGCGBPPGAEACCBPAAPBBEAPCAECACBBPPCX	1·33	100
5PTI	RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA CCGGGGSCCCCCSCCCCEEEEEEETTTTEEEEEEECSSSCCSSCBSSHHHHHHHHCCC XPACCCPPPBAECPCPBBPBBBPBCACGBPBPBBBCGBGPBAPBBCBAAAAAAACPYA CBCACCBPPBAEAPCPPBCPBBPPABEBEBBPBPPBPEPPCEGBBABAACACAABPBX	1·42	85
1NXB	RICFNQHSSQPQTTKTCSPGESSCYHKQWSDFRGTIIERGCGCPTVKPGIKLSCCESEVCNN CEEECCSTTCSCCEEECCTTCCCEEEEEEEETTEEEEEEEESCCCCCSSCCEEEECSTTTTC XBBYCBPYYYPPBBBYPPPYPCPBBBBBBBEAYEBBBBBBBEYYPYYYGYGBBBBYBPGACY ABBPAEGGGAAGBBBPCECAGAEBPBEEEEEPGEBBBPCPAPAEAEEEABPBBBBCPAAECX	1·67	76
1SN3	KEGYLVKKSDGCKYGCLKLGENEGCDTECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSC CCEECBCTTTCCBCBCSSCBSCHHHHHHHHHSTTTCCSEEEEETTEEEEESCCTTSCCSSCTTCCC XBEPPBPAACGBPBPBABPEPBCAAAAAACPCAAGPABBBBBGGBPBBBGPPACPPBPoPCABPA AEEPPBBAABCBPPBBAPBEPEGGACAAACBEABBCGEEEEEACEBBPBGAECABPCEGPAAPBX	1·45	67
2CRO	MQTLSERLKKRRIALKMTQTELATKAGVKQQSIQLIEAGVTKRPRFLFEIAMALNCDPVWLQYGT CCSHHHHHHHHHHHHTTCCHHHHHHHHHTSCHHHHHHHHHTTCCSSCTTHHHHHHHHTTSCHHHHHHCC XCBAAAAAAAAAAACGBBAAAAAAACGBPCCAAAAAACGAPABPACAAAAAAACGBPAAAAACGX ACBAACAAAAACACBCPPAACACAACGPPAAAAAAAACGABAGABGCAAAAACCGEGAACAABAX	1·38	100
1UTG	GICPRFAHVIENLLLGTPSSYETSLKEFEPDDTMKDAGMQMKKVLDSLPQTTRENIMKLTEKIVKSPLCM CCCHHHHHHHHHHHHHSCHHHHHHHHHHTTCCCHHHHHHHHHHHHHHHHTTSCHHHHHHHHHHHHHHHHTSGGGC XBPAAAAAAAAAAAACBPAAAAAAAAACACGPPAAAAAAAAAAAAAAAACCPPAAAAAAAAAAAAAAAACCPACCP CEPAAAAAAAAAAAABECGACAAAACACAGAEAAACACAAAAAAACCBEPACACAACACAAAAACBEAACX	1·32	100
1HOE	DTTVSEPAPSCVTLYQSWRYSQADNGCAETVTVKVVYEDDTEGLCYAVAPGQITTVGDGYIGSHGHARYLARCL CCCCSCBCCTTEEEEECSSEEEEEECSSSCEEEEEEETTSCBCCCEEECTTEEEEEEECTTSTTCSEEEEEECC XBCABPPPPACBBBBBBACBBBBBPCBCBPBBBBPBPACYPBPPBBPBPPGBBBPCBPPCCPCCEPPABBBPBB CAGAPPPAEPGAEBAEBEAPBBBBAGGBABBAGPBBBACEEBBPBBPBCEGEPCEEEBPAABGEAGCPPPPAEX	1·49	63

Results are given for 11 proteins of fewer than 80 residues. The proteins are denoted by their Brookhaven Protein Data Bank codes.

Amino acid sequences are given in the first row, using the 1-letter code. The second row contains the secondary structure assignments, computed by the DSSP procedure (Kabsch & Sander, 1983*b*). H and G denote respectively α- and 3_{10} helices, E and B extended and isolated β structures, T turn, S bend and C coil. The third row contains the backbone structure assignments, obtained by assigning to each residue the ϕ–ψ–ω domain, defined in Fig. 1, in which its observed backbone dihedral angles occur. Y indicates that the dihedral angles of the corresponding residue occur outside the permitted domains. X indicates that the values of the residue dihedral angles ϕ, ψ or ω are undetermined, which is always the case for the assignment of the 1st residue in each sequence, as its ϕ and ω angles are undefined. The fourth row gives the backbone assignments obtained by the directed search procedure described in Materials and Methods (section (c)). The conformation of the last residue is left unassigned (X), since its C^{α} position is determined by the dihedral angles of the preceding residue. The r.m.s. deviation between the native conformations and the backbones reconstructed by this procedure, and the percentage of the native hydrogen bonds that can be identified in these backbones are given in columns 3 and 4, respectively.

Table 3

Predicted lowest net energy conformations of short peptides with relatively well-defined conformations in aqueous solution

No.	Peptide	Amino acid sequence	Predicted 3-dimensional structure	Experimentally observed structure
1a	C-peptide 1RN3[a,b]	KETAAAKFERQHM	AAAAAAAAAAAAX	fully helical
1b	peptide II 1RN3[b]	AETAAAKFERAHA	AAAAAAAAAAAAX	
1c	peptide III 1RN3[c]	AETAAAKFLRAHA	AAAAAAAAAAAAX	
2	α-1 peptide[d]	ELLKKLLEELKG	AAAAAAAAAAAX	fully helical
3a	3K(I)[e]	AAAAKAAAAKAAAAKA	AAAAAAAAAAAAAAAX	fully helical
3b	3K(II)[e]	AKAAAAKAAAAKAAAA	AAAAAAAAAAAAAAAX	
3c	4K[e]	AKAAKAAAAKAAAAKA	AAAAAAAAAAAAAAAX	
3d	6K(I)[e]	AKAAKAKAAKAKAAKA	AAAAAAAAAAAAAAAX	
3e	3E[e]	AEAAAAEAAAAEAAAA	AAAAAAAAAAAAAAAX	
3f	(*i*+3)E, K[e,f]	AEAAKAEAAKAEAAKA	AAAAAAAAAAAAAAAX	
3g	(*i*+3)K, E[f]	AKAAEAKAAEAKAAEA	AAAAAAAAAAAAAAAX	
4a	6K(II)[e]	AKAAAKKAAAKKAAAKA	AAAAAAAAAAAAAAAAX	fully helical
4b	(*i*+4)E, K[e,f]	AEAAAKEAAAKEAAAKA	AAAAAAAAAAAAAAAAX	
4c	(*i*+4)K, E[f]	AKAAAEKAAAEKAAAEA	AAAAAAAAAAAAAAAAX	
5	β-bend hexapeptide[g]	GRGDSP	GPECPX	β-bend

The peptides are listed by the Brookhaven Protein Data Bank code (Bernstein *et al.*, 1977) of the proteins from which they originate, or by their name in the literature. As in Table 2, the conformation of the last residue is considered as undetermined, and denoted as X. Note that limits of observed structures are not precisely determined. Peptides 1a, 1b and 1c are fragments of ribonuclease A (1RN3): point mutations have been generated in 1b and 1c. Since our database contains a ribonuclease A (7RSA), the parameters used for predicting the structures of these fragments are computed with 7RSA removed from the learning set. Peptides 2 to 5 are synthetic peptides. References describing observed structures are as follows:

[a]Brown & Klee (1971), Bierzynski *et al.* (1982).
[b]Shoemaker *et al.* (1985).
[c]Shoemaker *et al.* (1987).
[d]Eisenberg *et al.* (1986).
[e]Marqusee *et al.* (1989).
[f]Marqusee & Baldwin (1987).
[g]Reed *et al.* (1988).

Table 4

The 10 conformations of lowest net energy predicted for the synthetic helical peptide (i+4)E, K

Peptide	Amino acid sequence	Predicted conformations	Net energy
(*i*+4)E, K	AEAAAKEAAAKEAAAKA	AAAAAAAAAAAAAAAAX	−16·87
		AAAAAAAAAAAAAAACX	−16·04
		CAAAAAAAAAAAAAAAX	−15·87
		AAAAAAAAAAAAAAPAX	−15·72
		AAAAAAAAAAAAAACAX	−15·66
		PAAAAAAAAAAAAAAAX	−15·54
		AAAAAAAAAAAACAAAX	−15·53
		ACAAAAAAAAAAAAAAX	−15·49
		AAAAACAAAAAAAAAAX	−15·46
		AAAAAAAAAAAAAAABX	−15·44

The conformations of (*i*+4)E, K (peptide 4b of Table 3) are defined by the corresponding structural assignments and by their computed net energy values. The latter are obtained by eqn (3), with the partition function term omitted, and refer to the 1st term in this equation.

Figure 3. Ribbon drawing of superimposed backbone conformations corresponding to the 10 structures of lowest net energy predicted for the synthetic helical peptide $(i+4)$E, K.

which the helical state is preferred over other conformations, can be obtained by analysing, for each of the peptide sequences, a set of best-ranking, lowest net energy conformations. Indeed, if a given structure is preferred, its net energy should be significantly lower than the net energy of less probable conformations. We find that this is the case for all α-helical peptides that we considered. As an illustration, Table 4 lists the ten lowest-energy conformations obtained for the synthetic peptide $(i+4)$E, K (Marqusee & Baldwin, 1987). Although the absolute net energy values listed are not readily comparable to energy values from other sources, either experimental or theoretical, a clear gap between the lowest net energy and the net energy of other generated conformations can be detected. In addition, the somewhat less probable conformations are not very different from the preferred one. In particular, the first or last residues are changed into 3_{10}-helix (designated by C), which constitutes only a minor change, as illustrated in Figure 3.

A quite interesting result is obtained for the only available non-helical peptide, the hexapeptide GRGDSP (the last peptide in Table 3). Nuclear magnetic resonance data (Reed *et al.*, 1988) suggest that this peptide adopts a type III, III′ or I turn, on the basis of the scalar J-couplings between nuclear spins of bonded atoms. While the presence of one or more β-bends is likely, its specific type is, however, not clearly established, since J-couplings do not yield precise structural information, especially not when several conformations co-exist (Hallenga *et al.*, 1979). Our method predicts the structure depicted in Figure 4(a). In this structure, a backbone hydrogen bond between the residues 2 and 5 can be readily identified: N–H and C–O bonds are nearly co-planar, and the O–H distance is equal to 4 Å. Replacing all residues except glycine and proline

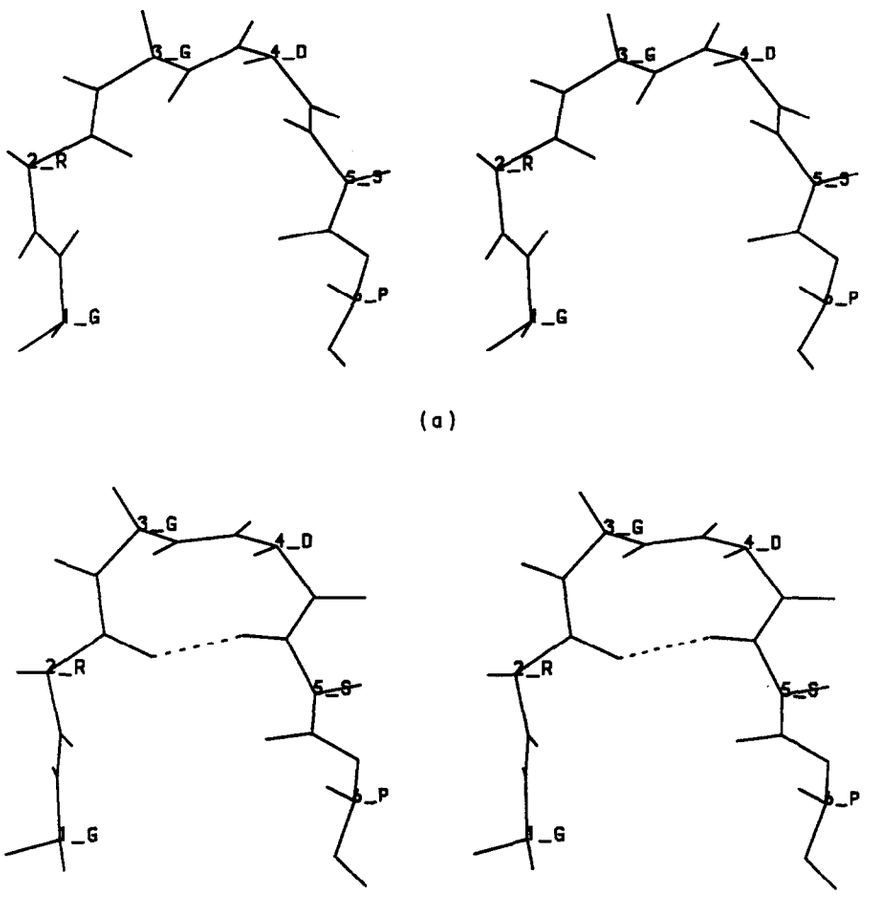

Figure 4. Stereo view of the predicted 3-dimensional backbone structure of the β-bend hexapeptide GRGDSP. (a) Predicted conformation, (b) conformation after 100 steps of restrained molecular dynamic simulation (with a time step of 2 fs).

Table 5

Predicted lowest net energy conformations for four flexible peptides

Peptide	Amino acid sequence	Predicted 3-dimensional structure	Net energy	Observed structure
S-peptide 1RN3	KETAAAKFERQHMDSSTSAA	AAAAAAAAAAAAACAAAPAX	−5·47	Helical up to residue 13[a]
		AAAAAAAAAAAAACCAAPAX	−5·46	
		AAAAAAAAAAAAACAAAAAX	−5·45	
		AAAAAAAAAAAAACCAAAAX	−5·44	
		AAAAAAAAAAAABCAAAPAX	−5·41	
		AAAAAAAAAAAABCCAAPAX	−5·39	
		AAAAAAAAAAAABCAAAAAX	−5·39	
		AAAAAAAAAAAABCCAAAAX	−5·37	
		AAAAAAAAAACAACAAAPAX	−5·35	
		AAAAAAAAAACAACCAAPAX	−5·33	
5PTI 30-51	CQTFVYGGCRAKRNNFKSAEDC	GBBBABGEBAPPAAGCPPAACX	−0·42	Extended from residues 1 to 6, helical from 19 to 22 and disulphide bridge between Cys1 and Cys22[b]
		GBBBABGEBAPPACGCPPAACX	−0·40	
		GABBBBGEBABAAACCPPAACX	−0·31	
		PBBBABGEBAPPACGCPPAACX	−0·28	
		GBBBABGEBAPPAAGCPPAAAX	−0·28	
		GBBBABGEBAPPACGCPPAAAX	−0·27	
		GABBBBGABAPAAACCPPAACX	−0·26	
		GABBBBGEBABAAPGCPPAACX	−0·25	
		GBBBABGAAAAAAAAGCPPAACX	−0·24	
		GBBBABGEPBPPAAGCPPAACX	−0·24	
Oxytocin	CYIQNCPLG	BCBPGPPPX	−1·82	Flexible, with disulphide bridge between Cys1 and Cys6[c]
		BCBPGBPPX	−1·71	
		BCBPGPPCX	−1·70	
		BBABGPPPX	−1·68	
		BABPGPPPX	−1·65	
		CGBPGPPPX	−1·62	
		BCBPGBPCX	−1·59	
		BBABGBPPX	−1·57	
		BBABGPPCX	−1·56	
		BABPGBPPX	−1·53	
Somatostatin	AGCKNFFWKTFTSC	CGCACBBBCBACCX	−0·96	Flexible, with disulphide bridge between Cys3 and Cys14[d]
		CGPACBBBCBACBX	−0·86	
		CGPACBAABABCCX	−0·79	
		CGCACBBBCBACAX	−0·75	
		CGPACBBACPACPX	−0·74	
		CGPCCABACBCCCX	−0·74	
		CGPACBAABPACCX	−0·74	
		CGPACBBAABCCCX	−0·73	
		PGCACBBBCBACCX	−0·71	
		CGPBCBAABPBCCX	−0·71	

Peptides are listed with Brookhaven Protein Data Bank codes used to designate protein fragments. The structure assignments of the 10 predicted lowest energy conformations are listed in column 3. The net energies of predicted conformations, defined in eqn (3) and in Materials and Methods, section (d), are also listed. The first 2 peptides are protein fragments, comprising respectively residues 1 to 20 of ribonuclease A (1RN3) and residues 30 to 51 of the bovine trypsin inhibitor (5PTI). Since our data base contains these 2 proteins, the parameters used for predicting the structures of these fragments are computed with either 5PTI or 7RSA (which refers to the same protein as 1RN3) removed from the learning set. Structural information available from experiment is given in the last column.

[a]Shoemaker *et al.* (1985).

[b]Oas & Kim (1988).

[c]Meraldi *et al.* (1977).

[d]Hallenga *et al.* (1980).

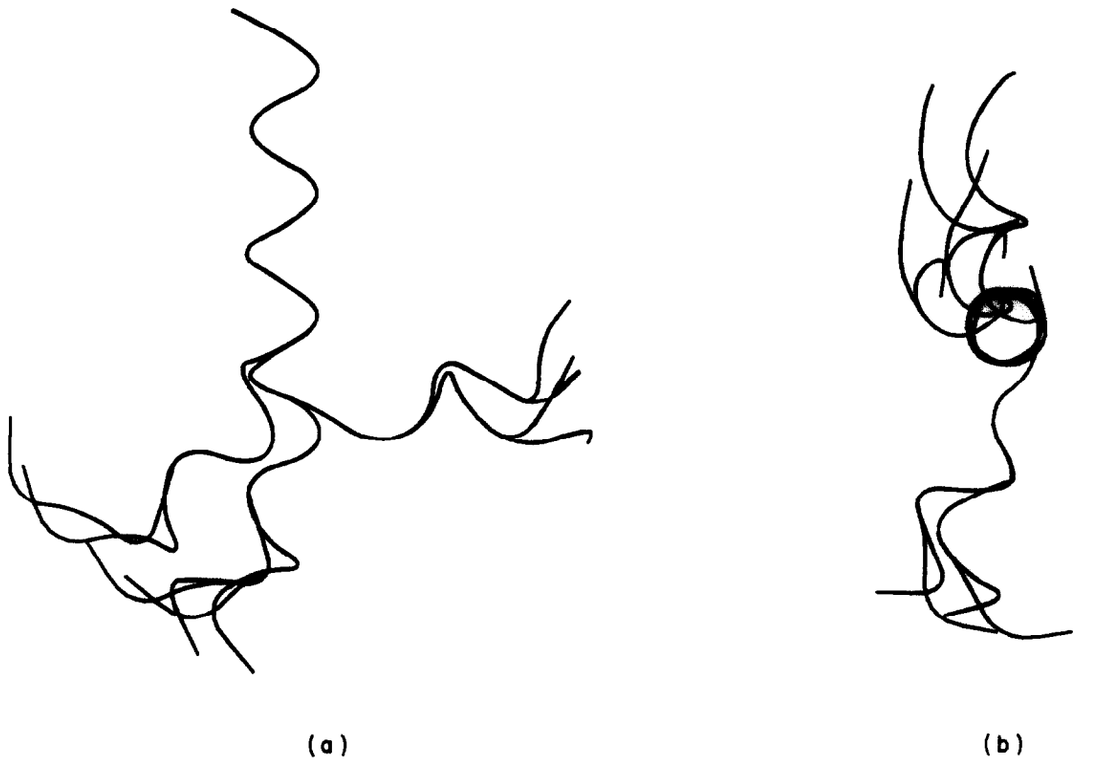

Figure 5. Ribbon drawing of the superimposed backbone structures of the S-peptide of ribonuclease A, corresponding to the 10 predicted conformations of lowest net energy. (a) Front view; (b) side-view.

with alanine, this structure was subjected to a short vacuum molecular dynamics simulation, with restraints on the hydrogen bond, followed by energy minimization (using the BRUGEL package; Delhaise *et al.*, 1985). This leads to the structure depicted in Figure 4(b), the conformation of which corresponds approximately to a type II′ β-turn, in close enough agreement with the experimentally characterized conformation. It should be added that the difference between the net energy values of most probable and less probable conformations is somewhat less marked in this case than in the other investigated peptides. This indicates that the conformation of the lowest net energy is somewhat less preferred, and that other conformations may co-exist, in agreement with conclusions drawn from experiments (Reed *et al.*, 1988).

Since our prediction method provides information about the relative preferences of peptide structures, it is interesting to consider not only peptides that adopt well-defined conformations in water, but also peptides that adopt flexible conformations. Prediction results obtained for four such peptides are given in Table 5. The first peptide corresponds to the 20 residues at the N terminus of ribonuclease A. It should be recalled that a somewhat shorter peptide containing the first 13 residues is predicted to be helical by our method, in agreement with experiment (Table 3). Experimental evidence suggests, moreover, that even in the 20-residue peptide, these 13 residues remain α-helical (Shoemaker *et al.*, 1985). Our predictions agree with these conclusions. Indeed, the ten structures of lowest net energy, listed in Table 5, are all helical to approximately residue 13. The conformations of the C-terminal portion of the peptide are, on the contrary, quite different. This is most clearly seen in Figure 5, in which all ten predicted structures of this peptide are superimposed. Moreover, the net energies of all these structures are very similar (Table 5), which indicates that there is no clear preference for any one of them, and suggests that the C terminus of this peptide is flexible, in good agreement with experimental evidence.

The second peptide in Table 5 is a fragment of the bovine pancreatic trypsin inhibitor (BPTI), composed of residues 30 to 51. Evidence based on nuclear magnetic resonance data suggests that, under oxidizing conditions, when the disulphide bridge between residues 30 and 51 is formed, this fragment conserves its native secondary structure (Oas & Kim, 1988). The lowest net energy conformation calculated by our procedure has the latter feature: a β-strand extending over the first six residues of the fragment, and a helix extending over the last four residues. It lacks, however, the disulphide bridge, which is not unexpected, since at this stage only local interactions are taken into account in the predictions. To obtain structures consistent with this bridge, we searched for the ten conformations of lowest net energy that satisfy, in addition, the constraint that the C^{β} atoms of both cysteine

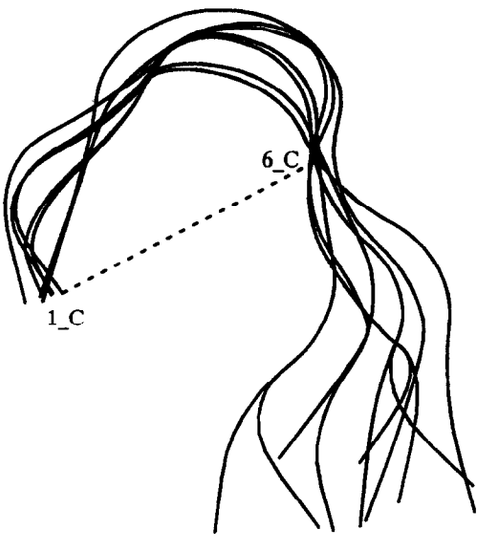

Figure 6. Ribbon drawing of superimposed backbone conformations of oxytocin, corresponding to the 10 predicted structures of lowest net energy, that satisfy the constraint of forming the observed disulphide bridge (broken lines) between Cys1 and Cys6.

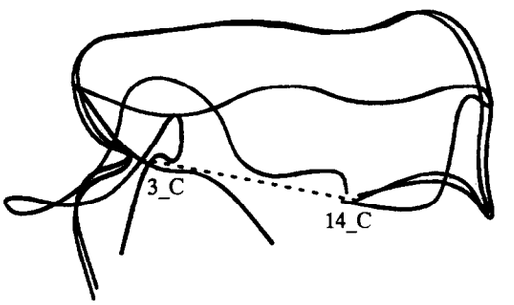

Figure 7. Ribbon drawing of superimposed backbones of somatostatin, representing the 4 predicted structures with lowest net energy that are consistent with the observed disulphide bridge (broken lines) between Cys3 and Cys14.

residues must be within a distance of 3 to 6·5 Å. These conformations are listed in Table 5. We see that both the β-strand and the helix are well conserved in all these structures, with only a few switches occurring from A (α-helix) to C assignments (3_{10}-helix). None the less, these structures differ in overall shape from each other as well as from the conformation adopted by the fragment in the native structure. Once more, the energies of these predicted structures are quite similar (Table 5), which may be taken to mean that all these conformations, and many others that conserve the disulphide bridge and secondary structure, are roughly equally probable. However, inspection of the crystal structure suggests that non-local interactions, in particular packing between helix and β-strand, should have an important stabilizing role. It would thus be necessary to include these interactions in order to yield correct predictions.

The next peptide in Table 5 is oxytocin, a peptide of nine residues shown experimentally to have a disulphide bridge between residues one and six, but otherwise to be rather flexible (Meraldi *et al.*, 1977). The lowest net energy conformation predicted by our procedure corresponds to the following assignments: BBBCGPPPX, which is inconsistent with the formation of the above-mentioned disulphide bridge. Imposing a constraint to form this bridge on the generated structures, leads to the ten lowest net energy conformations listed in Table 5, and superimposed in Figure 6. All these structures represent quite different conformations that have comparable net energies. They differ both in the loops connecting the two cysteines and in the C-terminal tail, in agreement with the observed flexibility of this peptide.

Similar conclusions can be drawn for somatostatin (Hallenga *et al.*, 1980), the flexible structure of which is well predicted here. The ten conformations of lowest net energy that satisfy the constraint to form the observed disulphide bridge between Cys3 and Cys14, are given in Table 5. They all have comparable net energy values, but their structures are quite different. This flexibility is illustrated in Figure 7, which shows the superimposition of the four lowest net energy conformations.

Finally, we proceed to predict lowest net energy conformations of early folding intermediates, the conformations of which have been trapped in nuclear magnetic resonance experiments. This is based on the assumption that protein segments that adopt their native structure early during folding are essentially stabilized by local interactions. Only a few examples of very early folding intermediates are presently known. A near exhaustive list is given in Table 6. In predicting the structures of the fragments, the local context in the native protein is taken into account, i.e. the influence of residues outside the fragment, which are at most eight residues from the ends of the fragment, is also considered.

We find that the helix comprising residues 108 to 115 of α-lactalbumin (LCA$CAVPO), shown by Baum *et al.* (1989) to form early during folding, is well predicted. As to the fragment consisting of residues 6 to 18 that adopts an α-helical conformation in barnase (RNBR$BACAM), Matouschek *et al.* (1989) suggest that its C-terminal portion, but not its N-terminal portion, forms early. This feature is well represented in our predictions, which assign a helical structure from residue 12 up to 17. To be consistent with experiment, His18 should in fact also be part of the helix. But considering that the definition of secondary structure limits is not unique (Presta & Rose, 1988; Richardson & Richardson, 1988), this may not be a problem. We find, furthermore, that the helical regions of these two protein fragments are conserved in the ten predicted conformations of lowest net energy, while adjacent regions along the sequence display different conformations,

indicating that the helical conformation is clearly preferred.

In horse cytochrome *c* (CYC$HORSE), experimental evidence (Roder *et al.*, 1988) suggests that both N- and C-terminal helices form early. These helices are packed against each other in the native structure. Results in Table 6 show that the C-terminal helix is well predicted, while the N-terminal one is not. Moreover, inspection of the ten predicted lowest net-energy conformations for each of the segments shows that the C-terminal helix is conserved in all these solutions, while the N-terminal segment is seen to adopt different conformations, suggesting that it has no preferred structure. This feature is conserved in other cytochromes *c* in our database where the two terminal helices display a similar spatial arrangement, and which have been suggested to follow the same folding pathway (Kim & Baldwin, 1982; Roder *et al.*, 1988). These include 1CCR, 3C2C and 155C. Note that the N terminus of 155C is predicted to be helical, but the helix is only partly conserved in the ten solutions of lowest net energy.

Exactly the same results have been obtained from secondary structure prediction (Rooman & Wodak, 1991). This strengthens the interpretation that the incorrect prediction of the N-terminal helix is not due to an error in our prediction methods, but suggests rather that the formation of the N-terminal helix may occur after docking of the N- and C-terminal segments, which could possibly be mediated by interactions with the prosthetic group. This is consistent with the observation that the simultaneous formation of the two helices, on the same rapid time scale, is unlikely (Roder *et al.*, 1988).

The fact that experimental data on early folding intermediates do not contain information on packing between structure elements makes them particularly difficult to interpret. Other early folding data, where packing seems likely, were therefore not analysed. These include observations of the early formation of the β-sheets in barnase (Matouschek *et al.*, 1989) and ribonuclease A (Udgaonkar & Baldwin, 1988). Similarly, the three myoglobin helices shown to form early (Hughson *et al.*, 1990), probably already interact with one another. Scenarios analogous to the folding of the terminal helices of cytochrome *c* can then be envisaged. Indeed, we find that two of the early formed helices in myoglobin, A and H, are well predicted, while the third one, G, is predicted to adopt an extended conformation near its C terminus. More detailed experimental data are clearly needed to verify predictions such as those made here.

Finally, noteworthy results concern the computa-

Table 6

Prediction results for protein fragments corresponding to experimentally-detected early folding intermediates

Protein fragments	Amino acid sequence	Predicted 3-dimensional structure	Observed early structure
LCA$CAVPO 108–115	IMCVKKIL	AAAAAAAA	Fully helical[a]
RNBR$BACAM 6–18	TFDGVADYLQTYH	BACGBPAAAAAAB	Helical near C terminus of fragment[b]
CYC$HORSE 88–101	KTEREDLIAYLKKA	AAAAAAAAAAAAAC	Fully helical[c]
1CCR 96–109	PQERADLISYLKEA	PAAAAAAAAAAAAC	—
3C2C 98–108	DDEIENVIAYL	CCAAAAAAAAA	—
155C 107–117	QADVVAFLAQD	AAAAAAAACCC	—
CYC$HORSE 3–13	VEKGKKIFVQK	ACCGAAAABAB	Fully helical[c]
1CCR 11–21	PKAGEKIFKTK	PAPAPBAPAAB	—
3C2C 4–10	AAAGEKV	AAAEAAA	—
155C 6–12	AKGEKEF	AAAAAAA	—

Proteins are denoted either by their Brookhaven Protein Data Bank code (Bernstein *et al.*, 1977) or by their SwissProt code, followed by the residue numbers defining the fragment limits. Their amino acid sequences, the 3-dimensional structures predicted by our method and the information about the structures observed during the 1st stages of folding are given. For cytochrome *c*, structurally homologous proteins are added, although no experimental data are available. The protein of our database that exhibits the highest sequence identity to horse cytochrome *c* (CYC$HORSE) is rice cytochrome *c* (1CCR): 56% using FASTA (Pearson & Lipman, 1988). Two other cytochromes *c*, 3C2C and 155C, display reasonable structural homology with 1CCR, as obtained by the automatic C^{α} superimposition algorithm AutoFit (J. Richelle, M.-E. Ochagavia & S. Wodak, unpublished results), and are also mentioned. The sequences and predicted structures of all these cytochromes *c* are aligned according to FASTA for 1CCR and to AutoFit for 3C2C and 155C. A dash in the last column indicates that no experimental data for the corresponding proteins exist. Note that the structure prediction for the protein fragments has been performed in the context of the full protein, taking into account the influence of residues outside the fragment and situated, at most, 8 residues from the fragment ends in the protein sequence. Moreover, whenever a fragment is part of a protein that is contained in the database or that exhibits more than 20% sequence identity with a database protein, the parameters used for predicting its structure are computed with that protein removed from the learning set.

[a]Baum *et al.* (1989).

[b]Matouschek *et al.* (1989).

[c]Roder *et al.* (1988).

tional speed of the prediction procedure. On a single Silicon Graphics 340 processor, only 9 to 15 seconds of c.p.u. time are required to determine the 20 lowest energy structures for the sequences considered. When filters are added, c.p.u. times span 9 to 42 seconds. Furthermore, computer time only modestly increases for longer sequences and for larger numbers of ranked lowest energy conformations.

4. Discussion

This study describes the first steps in developing a method for predicting the tertiary structure of a protein backbone from its sequence. An important point has been obtaining a representation of the protein backbone that drastically reduces the conformational space, while remaining capable of adequately describing native protein structures. By conserving backbone atomic detail and allowing dihedral angles to adopt a discrete set of conformations, described by seven structural states, our representation carries much more information than secondary structure assignments or the four (ϕ, ψ) states defined by Gibrat *et al.* (1991). Unlike these other two representations, it yields a global and complete description of three-dimensional structures. It also carries more information than the interatomic-distance-based model of Sippl (1990), and is, moreover, much easier to handle. It has the advantage of uniquely specifying co-ordinates of all main-chain atoms. This cannot be readily achieved with sets of distances, since those can often not be represented in three dimensions, and are, moreover, invariant to mirror reflections.

Our backbone model can be thought of as reaching the level of simplification of regular lattice models, while conserving important geometric and physical features of real structures. For instance, α-helices can be built in all spatial directions, which is impossible with regular lattices that have of the order of seven degrees of freedom per residue. These attributes are confirmed by our feasibility tests, in which the native backbones of 11 proteins are reproduced with our discrete backbone description to within 1·2 to 1·5 Å r.m.s. deviation. The proteins tested include all-α, all-β, $\alpha+\beta$, as well as irregular topologies. It is therefore reasonable to assume that these conclusions also apply to other proteins, in agreement with preliminary results obtained for triose phosphate isomerase.

This level of accuracy is quite acceptable, if we take as a premise that our backbone model is not destined to represent the native structure too faithfully, but rather one which is reminiscent of the hypothetical molten globule state. This state has been described as a loosely packed and collapsed state of the polypeptide chain, which has native-like backbone conformation, but flexible side-chains (Ptitsyn, 1987; Shakhnovich & Finkelstein, 1989; Finkelstein & Shakhnovich, 1989), and is presumed to precede the native structure along the folding pathway. By analogy, it should be possible to recover the native conformation from predicted backbones, by adding a computational step that would mimic the molten globule → native state transition. Such a step, which we intend to investigate in the near future, would combine modelling techniques with molecular dynamics simulations and energy minimizations, and use detailed co-ordinates for all atoms, including the side-chains.

The other important aspects dealt with in this study concern deriving and testing the force-field component due to influence from local interactions. Our procedure borrows from three structure prediction methods (Gibrat *et al.*, 1987; Rooman & Wodak, 1991; Sippl, 1990). The actual algorithm is, however, quite different. Unlike these methods, it uses energetic criteria to select and rank an arbitrary number of lowest-energy conformations, described by linear combinations of structural states of individual residues. Hence, conformational searches or exhaustive analyses of observed structures are avoided.

These features are an important asset of our method, which thus can provide information not only on the lowest energy conformation, but also on an arbitrary number of low-energy structures, whose relative preferences can be estimated from the computed energy values. For peptides that adopt a well-defined conformation in aqueous solution and for early folding intermediates, a gap is expected between the minimum net energy value and those of other predicted conformations, indicating that the corresponding structure is preferred over the other low-energy solutions. For flexible peptides with no single preferred structure, many quite different conformations, with comparable energy values, should be obtained.

This behaviour is well reproduced in our prediction tests. We see, moreover, that predicted conformations agree well with the observed ones, for seven of eight unrelated sequences that have been shown to be formed early during folding or to adopt well-defined conformations in aqueous solution. These include six helical conformations and a β-bend hexapeptide GRGDSP (Reed *et al.*, 1988). The structure and the partial flexibility of the N-terminal 20-residue peptide of ribonuclease A (1RN3) (Shoemaker *et al.*, 1985) are also quite well predicted.

However, problems are encountered in predicting certain specific features in other flexible peptides that we tested. We find that for oxytocin (Meraldi *et al.*, 1977) and somatostatin (Hallenga *et al.*, 1980) predicted conformations of lowest energy are incompatible with the disulphide bridge known to form in these peptides. In somatostatin, this is not surprising, since the bridging cysteine residues are separated by more than eight residues, thus falling in the realm of non-local interactions, which are not included in our force-field. Incidentally, similar considerations also apply to a different test case, the BPTI-fragment (Oas & Kim, 1988), where the formation of the disulphide bridge is also not predicted. In oxytocin, on the other hand, the

bridging cysteine residues are separated by only four residues. Here, the failure to obtain low-energy structures that contain the bridge may be ascribed to shortcomings in the description of the local force-field and, in particular, to the neglect of correlated influences at consecutive positions along the sequence, which is a serious approximation.

The approximations used, as well as the limited database size, are thus expected to influence the reliability of our prediction method. Detailed assessment of this influence must, however, await further analyses. It is, none the less, reassuring to see that our method is not very sensitive to the values of various adjustable parameters, as predicted structural states result from many different contributions. Indeed, parameters such as σ in equation (4), considering statistical influences of single residues, of pairs or residues, or of both combined, or using different learning sets, do not affect the ranking of lowest net energy structures in peptides or protein fragments with a strongly preferred conformation. They do so only in flexible peptides, where predictions yield several conformations with very similar energy values. The precise ranking of such conformations provided by our method should hence be considered with caution.

This study is thus seen to yield a rather effective tool for predicting the conformation of short peptides, which is, moreover, very fast compared to other procedures, since it does not require that we search conformational space. As such, it should have many useful applications. It could, for example, be used to detect regions of the protein that are locally stable, by requiring that there be a clear gap between the minimum net energy value and those of other generated structures. It could also be applied to model the conformation of protein loops from their amino acid sequence, using geometric constraints provided by the requirement that favourable interactions with the protein framework be formed (Moult & James, 1986; Fine *et al.*, 1986; Lesk & Chothia, 1984).

Moreover, the results obtained here indicate that our procedure provides a good basis for extending the prediction method to larger protein fragments and entire proteins. Since our backbone representation is general enough to handle both local and non-local interactions, this would require mainly that an adequate description for the non-local force-field component be derived, and that the relative influences from local and non-local contributions be correctly accounted for. A convenient way of deriving the non-local energy contributions would be to compute, from known protein structures, statistical preferences of residue-pair interactions, as a function of their spatial distance (Miyazawa & Jernigan, 1985; Wilson & Doniach, 1989; Sippl, 1990). Provided that local and non-local energy contributions are derived in an analogous fashion, the need for applying weighting factors, when adding up the two energy terms, should in principle be eliminated. Furthermore, introducing a distance dependent term in the potential will require searching conformational space, a major bottleneck in folding simulations. However, the use of our backbone model, with only seven states per residue, is expected to increase the efficiency of the conformational search by classical techniques such as simulated annealing (Kirkpatrick *et al.*, 1983). These various aspects are being investigated.

We are grateful to J.-L. De Coen, K. Hallenga, D. Van Belle and M. Prevost for useful discussions, J. Richelle and M. Huysmans for use of the SESAM database, P. Berthet for assistance with computer systems, and P. Delhaise and M. Bardiaux for help with the BRUGEL package (Delhaise *et al.*, 1985) that has been used throughout this study for figures and computer simulations. J.-P.A.K. acknowledges support from the Association Belge contre le Cancer and the Fondation Lefèbvre. M.J.R. is a Chargée de recherches at the Fonds National Belge de la Recherche Scientifique. J.-P.A.K. is a Chercheur associé at the Laboratoire d'Oncologie et de Chirurgie Expérimentale.

References

Baum, J., Dobson, C., Evans, P. & Hanley, C. (1989). Characterization of a partly folded protein by NMR methods: studies on the molten globule state of guinea pig α-lactalbumin. *Biochemistry*, **28**, 7–13.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanoushi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bierzynski, A., Kim, P. & Baldwin, R. (1982). A salt bridge stabilizes the helix formed by isolated C-peptide of RNase A. *Proc. Nat. Acad. Sci., U.S.A.* **79**, 2470–2474.

Brown, J. & Klee, W. (1971). Helix-coil transition of the isolated amino terminus of ribonuclease. *Biochemistry*, **10**, 470–476.

Cohen, F., Richmond, J. & Richards, F. (1979). Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J. Mol. Biol.* **132**, 275–288.

Delhaise, P., Van Belle, D., Bardiaux, M. & Wodak, S. (1985). Analysis of data from computer simulations on macromolecules using the CERAM package. *J. Mol. Graph.* **3**, 116–119.

Efron, B. (1982). *The Jack Knife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia.

Eisenberg, E., Wilcox, W., Eshita, S., Pryciak, P., Peng Ho, S. & De Grado, W. (1986). The design, synthesis, and crystallization of an α-helical peptide. *Proteins*, **1**, 16–22.

Fasman, G. (1989). The development of the prediction of protein structure. In *Prediction of Protein Structure and the Principles of Protein Conformation* (G. Fasman, ed.), pp. 193–316, Plenum Press, New York.

Fine, R., Wang, H., Shenkin, P., Yarmush, D. & Levinthal, C. (1986). Predicting antibody hypervariable loop conformations. II. Minimisation and molecular dynamic studies of MCPC603 from many randomly generated loop conformations. *Proteins*, **1**, 342–362.

Finkelstein, A. & Shakhnovich, E. (1989). Theory of cooperative transitions in protein molecules. II.

Phase diagram for a protein molecule in solution. *Biopolymers,* **28**, 1681–1694.

Frauenfelder, H., Petsko, G. & Tsernoglou, D. (1979). Temperature-dependent X-ray diffraction probe of protein structural dynamics. *Nature (London),* **280**, 558–563.

Garnier, J., Osguthorpe, D. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

Garnier, J., Levin, J., Gibrat, J.-F. & Biou, V. (1990). Secondary structure prediction and protein design. *Biochem. Soc. Symp.* **57**, 11–24.

Gibrat, J.-F., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* **198**, 425–443.

Gibrat, J.-F., Robson, B. & Garnier, J. (1991). Influence of the local amino acid sequence upon the zones of the torsional angles phi and psi adopted by residues in proteins. *Biochemistry,* **30**, 1578–1586.

Hallenga, K., Van Binst, G., Knappenberg, M., Brison, J., Michel, A. & Dirkx, J. (1979). The conformational properties of some fragments of the peptide hormone somatostatin. *Biochim. Biophys. Acta,* **577**, 82–101.

Hallenga, K., Van Binst, G., Scarso, A., Michel, A., Knappenberg, M., Dremier, C., Brison, J. & Birkx, J. (1980). The conformational properties of the peptide hormone somatostatin. (III). Assignment and analysis of the ^{1}H and ^{13}C high resolution NMR spectra of somatostatin in aqueous solution. *FEBS Letters,* **119**, 47–52.

Hughson, F., Wright, P. & Baldwin, R. (1990). Structural characterization of a partly folded apomyoglobin intermediate. *Science,* **249**, 1544–1548.

Huysmans, M., Richelle, J. & Wodak, S. (1991). SESAM, a relational database for structure and sequence of macromolecules. *Proteins,* in the press.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. sect. A,* **34**, 827–828.

Kabsch, W. & Sander, C. (1983*a*). How good are predictions of protein secondary structure? *FEBS Letters,* **155**, 179–182.

Kabsch, W. & Sander, C. (1983*b*). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.

Kabsch, W. & Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Nat. Acad. Sci., U.S.A.* **81**, 1075–1078.

Kim, P. & Baldwin, R. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* **51**, 459–489.

Kirkpatrick, S., Gelatt, C. & Vecchi, M. (1983). Optimization by simulated annealing. *Science,* **220**, 671–680.

Lesk, A. & Chothia, C. (1984). Mechanisms of domains closure in proteins. *J. Mol. Biol.* **174**, 175–191.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

Li, Z. & Scheraga, H. (1987). Monte Carlo-minimisation approach to the multiple minima problem in protein folding. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 6611–6615.

Liquori, A. (1969). The stereochemical code and the logic of a protein molecule. *Quart. Rev. Biophys.* **2**, 65–92.

Marqusee, S. & Baldwin, R. (1987). Helix stabilization by $Glu^{-} \ldots Lys^{+}$ salt bridges in short peptides of *de novo* design. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 8898–8902.

Marqusee, S., Robbins, V. & Baldwin, R. (1989). Unusually stable helix formation in short alanine-based peptides. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 5286–5290.

Matouschek, A., Kellis, J., Serano, L. & Fersht, A. (1989). Mapping the transition state and pathway of protein folding by protein engineering. *Nature (London),* **340**, 122–126.

Meraldi, J.-P., Hruby, V. & Brewster, A. (1977). Relative conformational rigidity in oxytocin and [1-penicillamine]-oxytocin: a proposal for the relationship of conformational flexibility to peptide hormone agonism and antagonism. *Proc. Nat. Acad. Sci., U.S.A.* **74**, 1373–1377.

Miyazawa, S. & Jernigan, R. (1982). Equilibrium folding and unfolding pathways for a model protein. *Biopolymers,* **21**, 1333–1363.

Miyazawa, S. & Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules,* **18**, 534–552.

Moult, J. & James, M. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins,* **1**, 146–163.

Murzin, A. & Finkelstein, A. (1988). General architecture of the α-helical globule. *J. Mol. Biol.* **204**, 749–769.

Nemethy, G. & Scheraga, H. (1977). Protein folding. *Quart. Rev. Biophys.* **10**, 239–352.

Oas, T. & Kim, P. (1988). A peptide model of a protein folding intermediate. *Nature (London),* **336**, 42–48.

Paine, G. & Scheraga, H. (1987). Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. III. Probable and average conformations of enkephalin. *Biopolymers,* **26**, 1125–1162.

Pearson, W. & Lipman, D. (1988). Improved tools for biological sequence analysis. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 2444–2448.

Petsko, G. & Ringe, D. (1984). Fluctuations in protein structure from X-ray diffraction. *Annu. Rev. Biophys. Bioeng.* **13**, 331–371.

Presta, L. & Rose, G. (1988). Helix signals in proteins. *Science,* **240**, 1632–1641.

Ptitsyn, O. (1981). Protein folding: general physical model. *FEBS Letters,* **131**, 197–202.

Ptitsyn, O. (1987). Protein folding: hypotheses and experiments. *J. Protein Chem.* **6**, 273–297.

Ralston, E. & De Coen, J.-L. (1974). Folding of polypeptide chains induced by the amino acid side-chains. *J. Mol. Biol.* **83**, 393–420.

Ramachandran, G. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Advan. Protein Chem.* **23**, 283–437.

Reed, J., Hull, W., von der Lieth, C., Kübler, D., Suhai, S. & Kinzel, V. (1988). Secondary structure of the Arg-Gly-Asp recognition site in proteins involved in cell-surface adhesion. Evidence of the occurrence of nested β-bends in the model hexapeptide GRGDSP. *Eur. J. Biochem.* **178**, 141–154.

Richardson, J. & Richardson, D. (1988). Amino acid preferences for specific locations at the ends of α-helices. *Science,* **240**, 1648–1652.

Roder, H., Elöve, G. & Englander, W. (1988). Structural

characterization of folding intermediates in cytochrome C by H-exchange labelling and proton NMR. *Nature (London)*, **335**, 700–704.

Rooman, M. & Wodak, S. (1988). Identification of predictive sequence motifs limited by protein structure data base size. *Nature (London)*, **335**, 45–49.

Rooman, M. & Wodak, S. (1991). Weak correlation between predictive power of individual sequence patterns and overall prediction accuracy in proteins. *Proteins*, **9**, 69–78.

Rooman, M., Rodriguez, J. & Wodak, S. (1990). Relations between protein sequence and structure and their significance. *J. Mol. Biol.* **213**, 337–350.

Scheraga, H. (1971). Theoretical and experimental studies of conformations of polypeptides. *Chem. Rev.* **71**, 195–217.

Schultz, G. (1988). A critical evaluation of methods for prediction of protein secondary structures. *Annu. Rev. Biophys. Biophys. Chem.* **17**, 1–21.

Shakhnovich, E. & Finkelstein, A. (1989). Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers*, **28**, 1667–1680.

Shoemaker, K., Kim, P., Brems, D., Marqusee, S., York, E., Chaiken, I., Stewart, J. & Baldwin, R. (1985). Nature of the charged-group effect on the stability of the C-peptide helix. *Proc. Nat. Acad. Sci., U.S.A.* **82**, 2349–2353.

Shoemaker, K., Kim, P., York, E., Stewart, J. & Baldwin, R. (1987). Tests of the helix dipole model for stabilization of α-helices. *Nature (London)*, **326**, 563–567.

Sippl, M. (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.

Skolnick, J. & Kolinski, A. (1989*a*). Computer simulations of globular protein folding and tertiary structure. *Annu. Rev. Phys. Chem.* **40**, 207–235.

Skolnick, J. & Kolinski, A. (1989*b*). Dynamic Monte Carlo stimulations of globular protein folding/unfolding pathways. I. Six-member, Greek key β-barrel proteins. *J. Mol. Biol.* **212**, 787–817.

Udgaonkar, J. & Baldwin, R. (1988). NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. *Nature (London)*, **335**, 694–699.

Ueda, Y., Taketomi, H. & Gō, N. (1978). Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A three-dimensional lattice model of lysozyme. *Biopolymers*, **17**, 1531–1548.

Wilmot, C. & Thornton, J. (1990). β-Turns and their distortions: a proposed new nomenclature. *Protein Eng.* **3**, 479–493.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins*, **6**, 193–209.

Wright, P., Dyson, J. & Lerner, R. (1988). Conformation of peptide fragments of proteins in aqueous solution: implications for initiation of protein folding. *Biochemistry*, **27**, 7167–7175.

Edited by R. Huber

Note added in proof. After this paper was completed, experimental data on a number of additional peptides with relatively stable conformation in solution were brought to our attention by Dr P. Wright. These include protein fragments and a series of synthetic peptides adopting essentially an α-helical structure, as well as peptides with a β-turn conformation (Brems *et al.*, 1987; Dyson *et al.*, 1985, 1988*a*,*b*; Gooley & MacKenzie, 1988; Gras-Masse *et al.*, 1988; Jiménez *et al.*, 1987; Merutka & Stellwagen, 1989, 1990; Merutka *et al.*, 1990; Montelione *et al.*, 1984; Padmanabhan *et al.*, 1990). Our procedure was applied to all these peptides yielding results very similar to those reported here. Except for one non-helical peptide (Dyson *et al.*, 1985), the predictions are in good qualitative agreement with the experimental observations. These results will be discussed in detail elsewhere.

Brems, D., Plaisted, S., Kauffman, E., Lund, M. & Lehrman, S. (1987). Helical formation in isolated fragments of bovine growth hormone. *Biochemistry*, **26**, 7774–7778.

Dyson, J., Cross, K., Houghten, R., Wilson, I., Wright, P. & Lerner, R. (1985). The immunodominant site of a synthetic immunogen has a conformational preference in water for a type-II reverse turn. *Nature (London)*, **318**, 480–483.

Dyson, J., Rance, M., Houghten, R., Lerner, R. & Wright, P. (1988*a*). Folding of immunogenic peptide fragments of proteins in water solution. I. Sequence requirements for the formation of a reverse turn. *J. Mol. Biol.* **201**, 161–200.

Dyson, H., Rance, M., Houghten, R., Wright, P. & Lerner, R. (1988*b*). Folding of immunogenic peptide fragments of proteins in water solution. II. The nascent helix. *J. Mol. Biol.* **201**, 201–217.

Gooley, P. & MacKenzie, N. (1988). Location of an α-helix in fragment 96-133 from bovine somatotropin by ^{1}H NMR spectroscopy. *Biochemistry*, **27**, 4032–4040.

Gras-Masse, H., Jolivet, M., Drobecq, H., Aubert, J., Beachey, E., Audibert, F., Chedid, L. & Tartar, A. (1988). Influence of helical organization on immunogenicity and antigenicity of synthetic peptides. *Mol. Immunol.* **25**, 673–678.

Jiménez, M., Nieto, J., Herranz, J., Rico, M. & Santoro, J. (1987). ^{1}H NMR and CD evidence of the folding of the isolated ribonuclease 50-61 fragment. *FEBS Letters*, **221**, 320–324.

Merutka, G. & Stellwagen, E. (1989). Analysis of peptides for helical prediction. *Biochemistry*, **28**, 352–357.

Merutka, G. & Stellwagen, E. (1990). Positional independence and additivity of amino acid replacements on helix stability in monomeric peptides. *Biochemistry*, **29**, 894–898.

Merutka, G., Lipton, W., Shalongo, W., Park, S.-H. & Stellwagen, E. (1990). Effect of central-residue replacements on the helical stability of a monomeric peptide. *Biochemistry*, **29**, 7511–7515.

Montelione, G., Arnold, E., Meinwald, Y., Stimson, E., Denton, J., Huang, S.-G., Clardy, J. & Scheraga, H. (1984). Chain-folding initiation structures in ribonuclease A: conformational analyses of *trans*-Ac-Asn Pro-Tyr-NHMe and *trans*-Ac-Tyr-Asn-NHMe in water and in the solid state. *J. Amer. Chem. Soc.* **106**, 7946–7958.

Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue, T. & Baldwin, R. (1990). Relative helix-forming tendencies of nonpolar amino acids. *Nature (London)*, **344**, 268–270.