Extracting Information on Folding from the Amino Acid Sequence: Accurate Predictions for Protein Regions with Preferred Conformation in the Absence of Tertiary Interactions

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

Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, Avenue P. Hêger, 1050 Brussels, Belgium

Received January 17, 1992; Revised Manuscript Received June 12, 1992

ABSTRACT: A recently developed procedure to predict backbone structure from the amino acid sequence [Rooman, M., Kocher, J. P., & Wodak, S. (1991) *J. Mol. Biol, 221,* 961–979] is fine tuned to identify protein segments, of length 5–15 residues, that adopt well-defined conformations in the absence of tertiary interactions. These segments are obtained by requiring that their predicted lowest energy structures have a sizable energy gap relative to other computed conformations. Applying this procedure to 69 proteins of known structure, we find that regions with largest energy gaps—those having highly preferred conformations—are also the most accurately predicted ones. On the basis of previous findings that such regions correlate well with sites that become structured early during folding, our approach provides the means of identifying such sites in proteins without prior knowledge of the tertiary structure. Furthermore, when predictions are performed so as to ignore the influence of residues flanking each segment along the sequence, a situation akin to excising the considered peptide from the rest of the chain, they offer the possibility of identifying protein segments liable to adopt well-defined conformations on their own. The described approach should have useful applications in experimental and theoretical investigations of protein folding and stability, and aid in designing peptide drugs and vaccines.

Clear evidence is emerging from experimental analyses of kinetic folding intermediates in complete proteins (Roder et al., 1988; Udgaonkar & Baldwin, 1988; Baum et al., 1989; Matouschek et al., 1989; Bycroft et al., 1990; Hughson et al., 1990), and from peptide models (Oas & Kim, 1988; Staley & Kim, 1990), that, in the early stages of folding, regions of the polypeptide can adopt their native fold in the absence of interactions with other structural elements that are formed only later (Wright et al., 1988). This is further confirmed by the finding that relatively small peptides, some of which correspond to early folding regions, can have well-defined conformations in solution (Brown & Klee, 1971; Bierzynski et al., 1982; Shoemaker et al., 1985, 1987; Eisenberg et al., 1986; Marqusee & Baldwin, 1987; Dyson et al., 1988a,b; Marqusee et al., 1989), leading to the suggestion that local effects may dominate in some regions of the protein and thereby play an important role in determining the folding pathway (Wright et al., 1988). Several theoretical analyses also seem to support this view. They indicate that notwithstanding the overall low success score of secondary structure prediction methods (Kabsch & Sander, 1983a), these perform much better than average in such early folding regions and for peptides with well-defined structure (Rooman & Wodak, 1991). Those methods would thus yield physically meaningful information, which could be very useful when incorporated into folding simulations. Classical secondary structure prediction algorithms are however ill suited for this purpose, mainly because they do not provide a detailed description of the backbone conformation and often lack convenient means of evaluating the reliability of the predictions.

This has prompted us to develop an alternative approach to predict backbone structure from local sequence information (Rooman et al., 1991). In this approach, the backbone conformation is described as a combination of a small number of discrete states, each characterized by a single value of the dihedral angles ϕ, ψ, and ω, representing allowed conformation of the isolated dipeptide. Using this description, a knowledge-based force field is derived from a database of well-resolved and refined protein structures, by combining principles of the GOR method (Garnier et al., 1978; Gibrat et al., 1987) with those of Rooman and Wodak (1988, 1991) and Sippl (1990). This force field is restricted to contributions from local interactions along the sequence—involving residues removed by at most 8 positions from each other—and discards side chain degrees of freedom. Assuming that conformations of individual residues are independent from one another, lowest energy structures for a given sequence are then computed by determining for each residue the structural states with lowest energy. This does not require searching conformation space and is therefore extremely fast. It is moreover possible to compute not only the minimum energy conformation but any number of low-energy structures whose preference can be evaluated from the relative values of the computed energies.

We have shown that this procedure, when applied to peptides that adopt well-defined conformations in solution or to segments belonging to early folding regions, yields lowest energy structures that agree rather well with the observed conformations. These predicted lowest energy structures display moreover a significant energy gap relative to other predicted conformations, an indication that they may be dominant. A very similar approach has been taken by Kang et al. (1992). They derived ϕ–ψ probability tables from known protein structures and suggested that these tables can be used to evaluate the conformational entropy along the polypeptide chain from sequence information. Evidence was

‡ Chargée de Recherches at the Fonds National Belge de la Recherche Scientifique.

§ Chercheur Associé at the Laboratoire d'Oncologie et de Chirurgie Expérimentale. J.-P.A.K. acknowledges support from the Association Belge contre le Cancer and the Fondation Lefèbvre.

0006-2960/92/0431-10226$03.00/0

then presented that regions of low entropy correlate well with early folding sites.

In the present study, we probe further into the ability of procedures for predicting backbone structure from local sequence information, to identify regions of the polypeptide chain that adopt well-defined conformations in the absence of tertiary interactions. For that purpose, we systematically investigate the correspondence between the preferred conformations of protein segments predicted by our method and those observed in the native state, for 69 well-resolved and refined protein structures. Our results show that protein regions with highly preferred conformations—those that display the largest energy gaps relative to other computed structures—are also the most accurately predicted ones, suggesting that our procedure is able to recognize the relevant sequence features that impart greater than average local conformational robustness. Considering the proposals that such regions correlate well with early folding sites (Rooman et al., 1991; Kang et al., 1992), this offers the possibility of identifying those sites in proteins without prior knowledge of the tertiary structure and should have useful applications both in experimental analyses of protein folding and stability and in folding simulations.

MATERIALS AND METHODS

(*a*) *Protein Structural Data.* Our study uses a dataset of 69 protein crystal structures from the Brookhaven Databank (Bernstein et al., 1977), consisting of refined structures, determined to at least 2.5-Å resolution, and displaying at most 20% sequence identity with each other (see list in the legend of Figure 1). Various other data needed in this work, such as polypeptide chain ends, secondary structure defined by DSSP (Kabsch & Sander, 1983b), backbone dihedral angles, and assignments of residue backbone conformations to allowed regions in the Ramachandran map (Ramachandran & Sasisekharan, 1968), are extracted from the relational database SESAM (Huysmans et al., 1991).

(*b*) *Representation of the Backbone Structure.* Protein backbones can be adequately represented using fixed bond lengths and valence angles, whose values correspond to averages computed from known structures, and the three dihedral angles ϕ, ψ, and ω of each residue. This conveniently reduces the number of independent degrees of freedom of the backbone to only three per residue. Accessible conformational space can be further reduced if one considers that backbone dihedral angles tend to cluster into a small number of allowed regions in the Ramachandran (ϕ–ψ–ω) map (Ramachandran & Sasisekharan, 1968). Following our previous work (Rooman et al., 1991), we consider 7 allowed regions: 6 regions with the trans conformation of the peptide bond ($\omega = 180°$), denoted by the assignments A, C, B, P, G, and E, and one region with the cis-peptide conformation ($\omega = 0°$), which is denoted by O. The precise limits of the 7 regions are depicted in Figure 1.

By assigning each residue along the chain to the region to which its (ϕ, ψ, ω) angles belong, a succession of structural states is obtained describing the backbone conformation. This representation can in turn be used to reconstruct the complete backbone by associating to each structural state a single average value of the (ϕ, ψ, ω) angles (Figure 1). Given the rather drastic simplifications, i.e., the use of a single point per allowed region and average bond distances and angles, the reconstruction technique leads to local structural deformations that usually do not compensate each other. Complete protein backbones reconstructed in this way can hence substantially

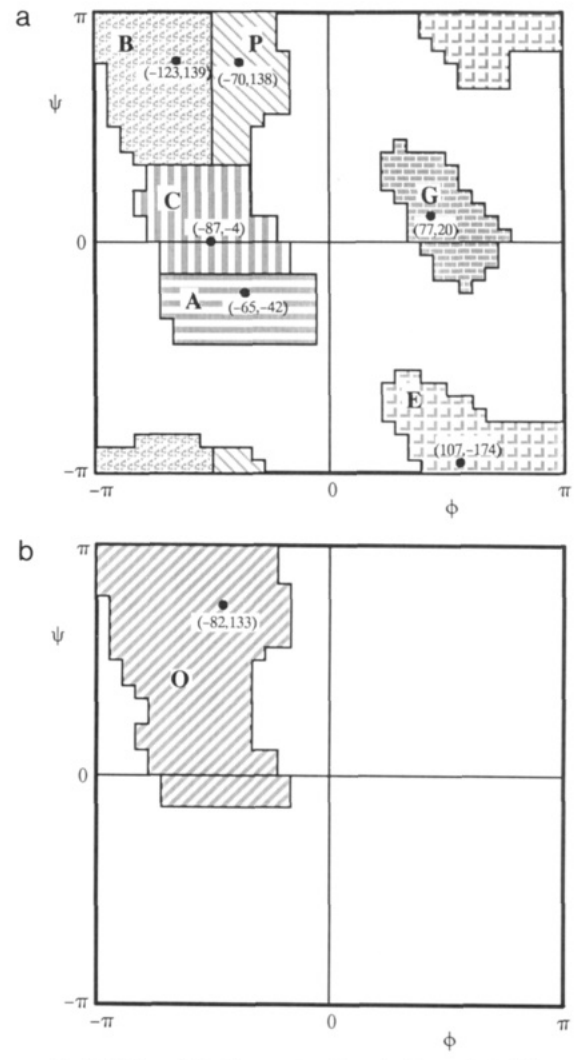

FIGURE 1: Division of the Ramachandran ϕ–ψ map into 7 domains, 6 for $\omega \approx 180°$ (panel a) and 1 for $\omega \approx 0°$ (panel b). For convenience, ω is defined as the angle about the peptide bond preceding the residue. The representative value of each domain is indicated. It is computed as the average of the ϕ–ψ values occurring in the domain in the proteins of our dataset. All ω values lower than −150° and higher than 150° are set equal to 180°, and all ω values between −30° and 30° are set to 0°. Our dataset includes the following 69 proteins from the Brookhaven Databank (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, 2YHX, 3ADK, 3DFR, 3FXC, 3GRS, 3TLN, 3WGA, 451C, 4FXN, 4HHB, 4MDH, 4RXN, 5CPA, 5PTI, 6LDH, 7RSA, 8ADH, 8CAT.

deviate from the original crystal structure (Liquori, 1969). Yet, local deformations along the polypeptide chain remain limited, and thus, the reconstruction technique yields reasonable backbone conformations for short protein segments. For example, the average rms deviation between reconstructed and observed backbone atoms (N, Cα, C, O) in 8-residue segments is 1 Å (Rooman et al., 1991), which is regarded as acceptable for the purpose of this study.

(*c*) *Structure Prediction Algorithm.* The algorithm used here to predict the lowest energy backbone structures of protein segments from their amino acid sequence has been described in detail previously (Rooman et al., 1991). In the following, therefore, only the essential features are given.

Probabilities for a given residue i in the sequence to be found in each of the 7 (ϕ, ψ, ω) regions are computed from

the database of 69 protein structures described above. They take into account influences of single residues and residue pairs (including gaps) removed from i by at most 8 positions along the chain. The structural preferences of individual residues being considered as independent, the overall probability for a sequence to adopt a given conformation is expressed as the product of individual residue preferences.

After normalization, which consists in dividing the probability for a particular sequence to adopt a given structure by the average probability of any sequence to adopt the same structure, probabilities are translated into potentials of mean force, assuming a Boltzmann distribution. The normalization is required to obtain potentials that, like the commonly used empirical energy functions, are not biased toward the most abundant conformational states in the database (Sippl, 1990; Rooman et al., 1991). These potentials correspond to *free energies* when considered as describing ensemble averages of all side-chain conformations and subsets of backbone structures. However, when considered as characterizing discrete backbone states, they correspond to *energies*. These energies are equivalent to the *net energies* previously defined (Sippl, 1990; Rooman et al., 1991).

To translate the probabilities into energies, it is in principle required to evaluate the partition function. But since the latter appears as an additive term that does not depend on the conformation, it may be omitted provided one compares energy values for different conformations of the *same* sequence (Sippl, 1990; Rooman et al., 1991), which is the case throughout this paper.

Using these potentials of mean force, the lowest energy conformations of a given amino acid sequence can be readily computed. The residue conformations being considered as noncorrelated, no conformational search is needed, and the prediction algorithm is extremely fast. It yields not only the lowest energy conformation, corresponding to the most favorable combination of structural states, but any number of lowest energy conformations ranked in order of increasing energies. These are obtained by successive combinations of increasingly less favorable structural states.

To avoid overestimating the performance of our predictions, the *jack knife* procedure (Efron, 1982) is used. Whenever the protein to be predicted displays more than 20% sequence identity with a protein from our dataset, it is omitted in the computation of the mean force potentials.

(*d*) *Detecting Protein Regions with Preferred Structure.* We have shown previously (Rooman et al., 1991) that this prediction algorithm, which considers only contributions from local interactions along the chain, can be used to identify short protein segments that have well-defined conformational preferences. These segments are characterized by an energy gap between their predicted most probable structure and the next best ranking predicted structures in the ordered list.

Identifying such segments requires specifying the minimum energy gap that will in general be sufficient to ensure that their lowest energy conformation is preferred over all others. Since energy values computed by our prediction procedure cannot be readily compared to those obtained with conventional force fields, the task of objectively defining the minimum gap is not straightforward and requires empirical calibrations. Such calibrations could among other things consider experimentally observed structural propensities. Here, this gap is set to 0.5 kcal/mol, and the effect of using larger values is tested.

One finds furthermore that successive entries in the list of predicted low-energy structures often display very similar conformations, differing for instance by a single structural state. The relevant energy gap, required to exceed the threshold defined above, is therefore taken as that between the lowest energy conformation and the next one in the rank that has a significantly different structure. To determine structural similarity, backbones are constructed from the computed structural assignments using standard bond lengths and angles. Structures are considered as significantly different if the rms deviation of their backbone atoms (N, Cα, C), after coordinate superimposition (Kabsch, 1978), is larger than $\rho = (0.5 + L \times 0.05)$ Å, where L is the length of the considered segments, taken here to span the range of 5–15 residues. This rms threshold is significantly below the average rms value between all possible conformations of segments with length L generated by our 7 structural states. For instance, for $L = 5$, $\rho = 0.75$ Å, while the average rms value is 1.9 Å.

Our prediction procedure thus requires defining three parameters: the minimum energy gap, a criterion for structural similarity, and the length L of the segments considered. With those defined, it proceeds by sliding a window of fixed length L along the sequence one residue at a time and computes the lowest energy conformations for the protein segment delimited by the window. Only segments whose lowest energy structure displays an energy gap equal to or larger than the gap threshold are recorded.

Two versions of this protocol are used in this study. In one version, the predictions for the segment delimited by each window include contributions from flanking residues outside the window. Those residues are removed by at most 8 positions along the sequence from either end of the window. This amounts to viewing the segment as covalently linked to the rest of the chain, while being subjected to local influences, a situation akin to the one that may occur during the early stages of protein folding. This version is hence used to identify regions of the polypeptide likely to adopt a preferred conformation early during folding. In the second version, the predictions are performed ignoring altogether the contributions from residues outside the window. This would mimic the case in which the considered segments are excised from the protein and is used to predict whether the isolated peptide would adopt a well-defined conformation in solution.

(*e*) *Predicting the Protein Backbone Structure.* Applied to a given protein, the procedure described above identifies a set of segments that have a well-defined preferred conformation. Some of these segments are expected to overlap along the sequence, thereby providing redundant information. To obtain a consensus prediction for the protein backbone, the predictions from overlapping segments are combined using the rule of simple majority vote. The selected structural state at each position along the sequence is thus the one most frequently predicted in the collection of overlapping segments. In fact, this applies only to segments identified with the second version of our procedure, where contributions from flanking residues are ignored. Those identified by including contributions from flanking residues always provide identical predictions at the sequence position onto which they map, because this position is always in the same sequence environment. Note that, in general, portions of the sequence remain unpredicted, as segments with preferred conformations are not identified everywhere.

(*f*) *Scoring the Predictions.* To assess the performance of our method, the predicted structural states are compared to the observed ones. The latter are derived from the protein crystal structures using the procedure described in section b above. The percent of correctly predicted structural states constitutes the score. Residues that are not part of segments

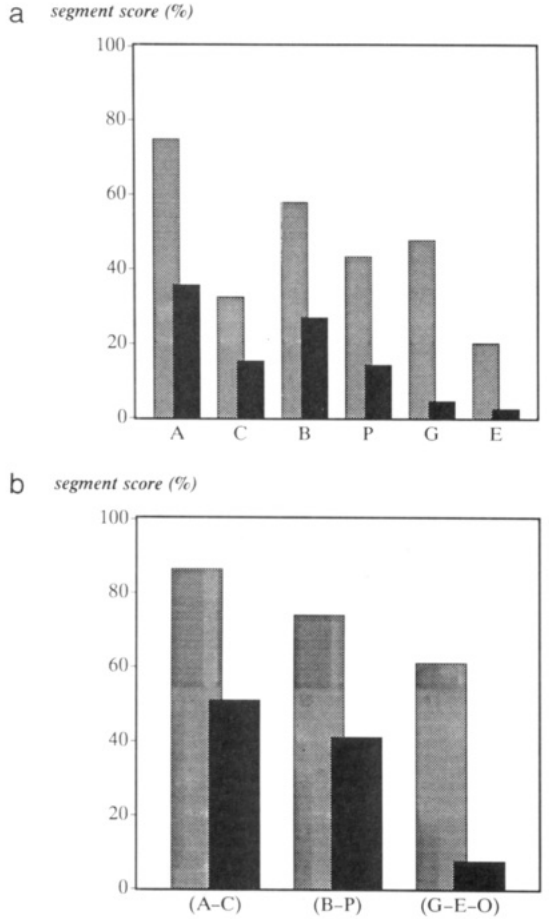

FIGURE 2: Segment scores as a function of the structural state. The gray bars represent segment scores, and the black bars represent the frequency of the corresponding structural state in the dataset. The structural state O does not appear, because it is never predicted in the segments. (a) 7 structural states; (b) 3 structural states.

predicted as having preferred conformations, the unpredicted ones, are not included in the score evaluation. When the score is computed for predictions provided by individual segments, we ignore the fact that these segments may overlap, and refer to it as the *segment score*. When the score is computed for backbone predictions, where predictions from individual segments are combined by the majority vote rule, we refer to it as the *prediction score*. This latter score is more directly comparable to scores computed for the more commonly used secondary structure prediction methods.

RESULTS

(*a*) *Detecting Protein Regions with Preferred Conformation.* Our procedure has been applied to search the entire dataset of 69 proteins for polypeptide segments that have a well-defined conformational preference. When residues in the chain flanking the segments are included in the prediction, this leads to identifying 9610 segments that sometimes overlap and that cover 5994 out of 13 562 residues. The segment score for these predictions is 68% when 7 structural states are considered. When these are grouped into only 3 states, (A and C), (B and P), and (G, E, and O), which are reminiscent of, but not equivalent to, the 3 secondary structures α-helix, β-strand, and coil, this score rises to 83%. The average rms deviation after superimposition of N, Cα, and C main-chain atoms, between the predicted structures of the 5–15-residue segments and those observed in the original protein, is 1.0 Å. Predictions performed by ignoring, for any given segment, contributions from flanking residues yield a slightly different set of segments,

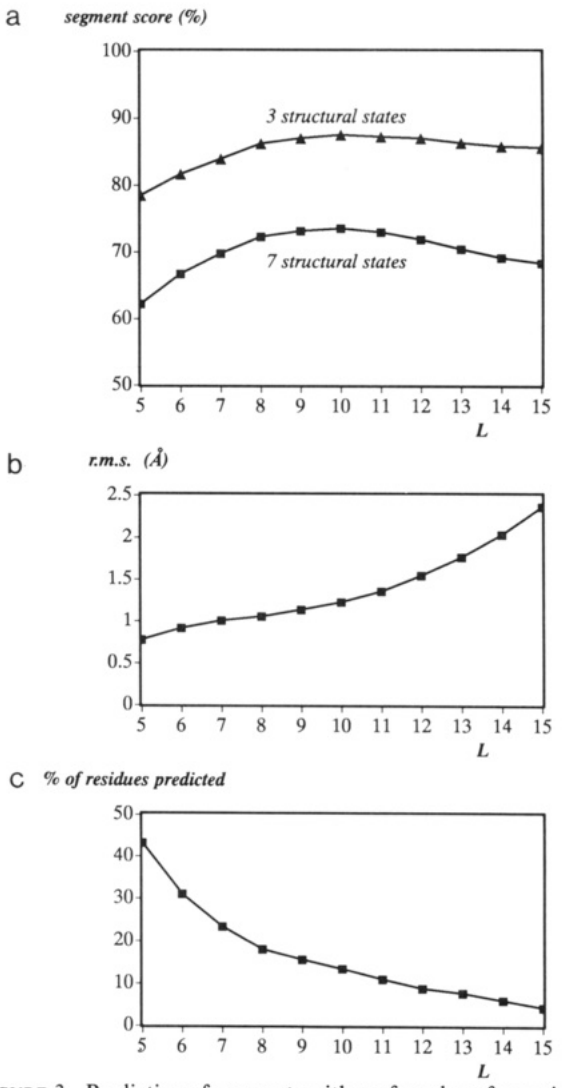

FIGURE 3: Prediction of segments with preferred conformation, as a function of their length L. (a) Average segment scores as a function of L. Squares correspond to the 7 structural states, triangles to the 3 states. (b) Average rms deviation between predicted lowest energy conformation and observed one, as a function of L. (c) Percent of predicted residues, as a function of L.

but their number and scores are quite similar and are therefore not given explicitly.

Yet higher scores are obtained if one takes into account that our procedure yields for each segment not only the lowest energy structure, but a family of closely resembling structures whose energies rank better than the first significantly different conformation encountered down the rank. This requires defining a somewhat different segment score, namely, the percent of structural states correctly predicted in any member of the family of closely resembling low-energy structures. This family segment score is 78% for 7 structural states, and 87% when these are grouped into 3 states. The average rms deviation, computed as that between the observed structure and the best fitting segment of each family, is 0.9 Å.

We see that predictions provided by segments with preferred conformation map on the average onto 44% of the protein sequence. However, the proportion of predicted residues varies substantially in different proteins. Some proteins are predicted to have nearly no segments with preferred conformation. This is the case, for instance, of neurotoxin B and of wheat germ agglutinin. In the former, not a single segment is identified when the contribution from residues flanking the segments is included, and only 9% of its residues are predicted when

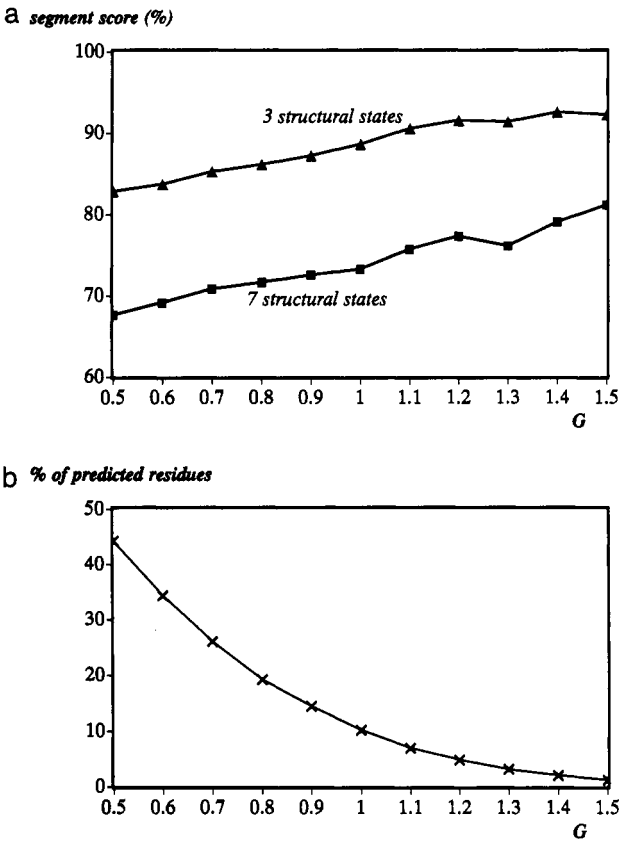

FIGURE 4: Prediction of protein segments with preferred conformation, as a function of the minimum energy gap G, between the best conformation and the first structure down the rank that is significantly different. (a) Segment scores as a function of G. Squares correspond to 7 structural states, triangles to 3 states. (b) Percent of predicted residues, as a function of G.

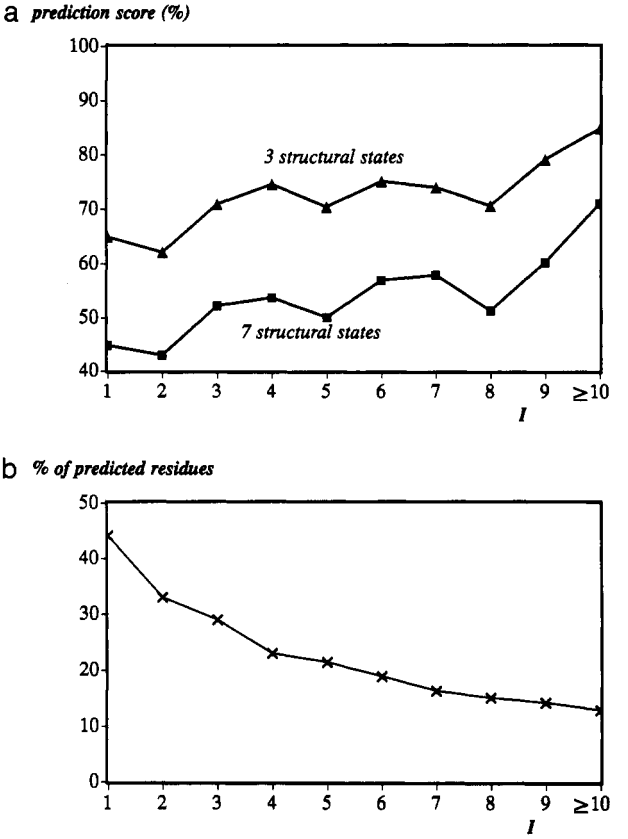

FIGURE 5: Backbone predictions, using segments with preferred conformations, as a function of I, the number of segments mapping onto the same residue. (a) Prediction scores as a function of I. Squares correspond to 7 structural states, triangles to 3 states. (b) Percent of predicted residues, as a function of I.

flanking residues are ignored. The opposite is found for wheat germ agglutinin. Only 9% of its residues are predicted including the contributions from flanking residues, and none when this contribution is ignored. If our interpretation that segments with preferred structure correspond to regions formed early during folding is correct, every protein that folds spontaneously should have some. This may not be true for small proteins containing disulfide bridges, a category to which the above two cases belong. But otherwise, the reason for the absence of segments with preferred conformation in certain sequences remains unclear.

Among the proteins whose sequence is most extensively covered by our predictions is cytochrome c' (2CCY): 79% of its residues are predicted when flanking residues are considered, with a segment score of 92% on 7 structural states. Another well-predicted protein is 434 CRO (2CRO), with 68% of the residues predicted and a score of 85%. These results could be taken to mean that, in proteins such as these, local interactions along the sequence play a significant role in defining the global fold.

To gain insight into the factors that might influence the predictions, we proceed to analyze the dependence of the number of predicted segments and their segment scores on several parameters: the predicted structural state, the segment length, and the minimum energy gap between the best and next-best ranking structures with different conformations.

(*i*) *Segment Scores for Different Structural States.* The segment scores computed separately for each of the structural states considered here are given in Figure 2. For the two cases studied, the one with 7 structural states and that with only 3, the scores are clearly much higher than the frequency of the corresponding states in our protein dataset and are hence much higher than what would be expected for random predictions. As expected on statistical grounds, the highest scores are obtained for the most frequent helical states (A) or (A–C). Interestingly, the score for state G is higher than that for states C and P, which are however significantly more frequent, a possible indication that the subdivisions between A and C and between B and P states are somewhat artificial.

(*ii*) *Influence of Segment Length on Segment Scores.* The segment score also varies with the length L of the segments considered, as depicted in Figure 3a. Highest scores are obtained for 10-residue segments, especially when 7 structural states are considered. This suggests that in the shorter segments intrasegment interactions are more readily overridden by those with the rest of the protein. The somewhat lower scores for the longer segments could be due to the lesser validity of the approximation of noncorrelation between structural states. The average rms deviation between observed and predicted structures (including correct and incorrect predictions) is given in Figure 3b as a function of L. This deviation clearly increases with L, from 0.8 Å for $L = 5$ to 2.4 Å for $L = 15$, reflecting the fact that a single incorrectly predicted structural state in the middle of a 15-residue segment leads to much larger deformation than in a 5-residue segment.

The influence of segment length on the fraction of predicted residues is illustrated in Figure 3c. This fraction is seen to decrease drastically with L. Gaps in energy of 0.5 kcal/mol or more are thus much more frequent in the shorter segments of the considered dataset, but the corresponding average segment score is lower, as discussed above. The former trend is a direct consequence of including a measure of structural similarity in defining the energy gap (see section d under

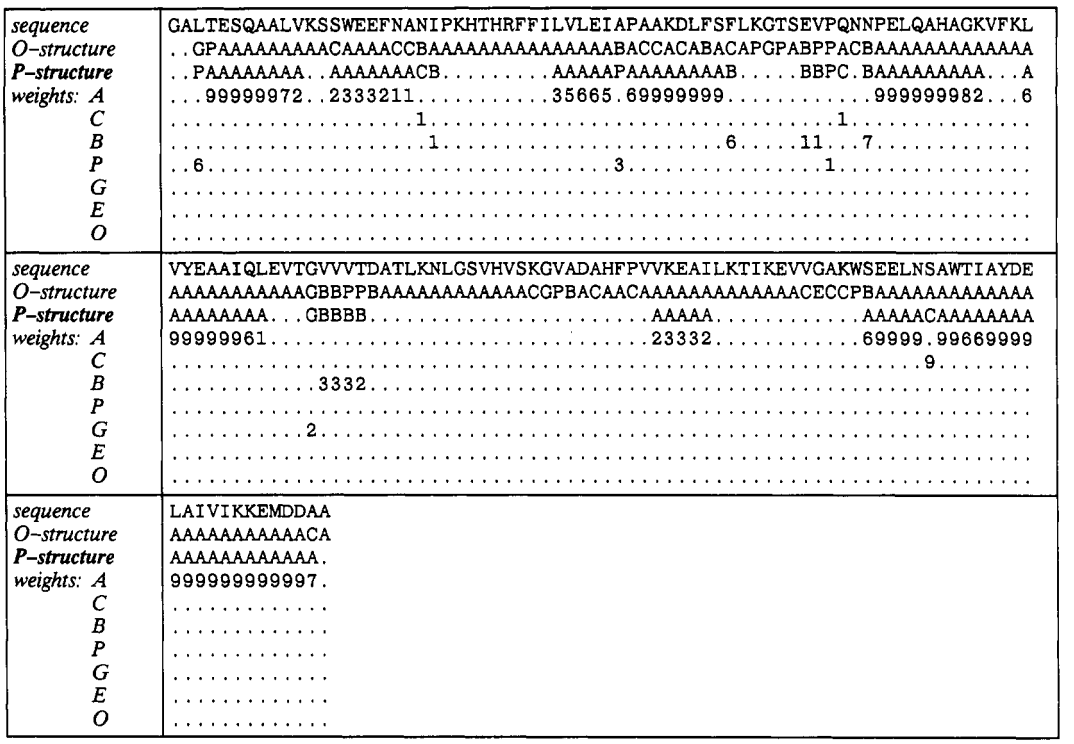

FIGURE 6: Backbone structure predictions of leghemoglobin (2LH4), provided by segments with preferred conformations. The predictions include contributions from flanking residues. The segments were identified using a minimum energy gap $G \geq 0.5$ kcal/mol and their length L spans 5–15 residues. O- and P-structures refer respectively to observed and predicted structures. Observed structures marked as dots indicate that the ϕ, ψ, or ω angles of the corresponding residues are undetermined. Dots in the predicted structure indicate that the corresponding residue is unpredicted. The weights of each structural state are given at each position. It is the number of different segments that map onto that position. Weights higher than 9 are represented by the digit 9, with dots indicating the absence of segments. 61% of the residues are predicted, with a prediction score of 78% on 7 structural states and 92% on 3 states.

Materials and Methods). Indeed, the chance that the best ranking conformations, occurring before the one with a sizable energy gap, display similar structures decreases with increasing segment length. The prevailing situation in longer segments is the one in which small differences in energy are accompanied by large differences in structure.

(*iii*) *Influence of the Energy Gap on Segment Scores.* A crucial parameter influencing segment scores is the gap in energy between the structure with lowest energy and the next-best ranking one that displays a significantly different conformation. Figure 4a shows how the segment scores vary with the values of the energy gap G. When considering 7 structural states, the segment score is only 68% for gaps $\geq$0.5 kcal/mol. With gaps of 1.5 kcal/mol or more, the score increases to 81%, reaching values as high as 92% with 3 structural states. Unfortunately, such large gaps are rather infrequent, and the predictions made by using them cover no more than 1.2% of the residues (Figure 4b).

(*b*) *Predicting the Backbone Structure.* To obtain genuine predictions for the protein backbone, predicted structures of individual segments are combined. As noted in section e under Materials and Methods, segments identified by including the influence from flanking residues always provide consistent predictions. Interestingly, we see that overlapping segments identified by ignoring these contributions seldom provide inconsistent predictions. This is encountered in our entire protein dataset at only 12 sequence positions when 7 structural states are considered, and nowhere when 3 states are considered. Everywhere else, the predicted structural states, which are sometimes obtained from more than 10 segments, are in perfect agreement. This is not surprising since we deal here with segments whose conformations display energy gaps: even when the predicted lowest energy structures of overlapping segments are different, the chances that both display gaps are low.

The average prediction scores obtained by combining predictions from segments considered with their flanking residues are 56% with 7 states and 73% with 3 states. Here too, predictions cover 44% of the protein sequences. The influence on the prediction score of the number I of segments mapping onto a given residue is depicted in Figure 5a. When 10 or more overlapping segments predict the same structural state at a given position, the prediction score is as high as 71% for 7 states and 85% for 3 states. Unfortunately, the average sequence fraction covered by such reliable predictions is only 13% (Figure 5b).

Neglecting the contributions from flanking residues yields slightly lower scores: 55% for the 7 structural states and 71% for 3 states, and the predictions cover only a slightly larger fraction (45%) of the sequence. When predictions performed with and without flanking residues are simply added up using the majority vote rule, the predicted sequence fraction increases to 60%, while the prediction score drops only slightly to 53% and 70%, for 7 and 3 states, respectively. Leghemoglobin is an example of a protein well predicted by the combined approach (Figure 6). Its prediction score (leaving out unpredicted residues) is 78% considering 7 states and 92% considering only 3 states.

Average prediction scores are thus over 10% lower than the segment scores discussed in the previous section. This drop results from the fact that, on the average, segments providing correct predictions tend to overlap more than those providing incorrect predictions (see Figure 5a). The former will hence contribute less to the prediction score, where predictions from individual segments are combined, than to the segment score, where they are considered individually.

(*c*) *Segments with Preferred Conformations and Reliable Sequence–Structure Associations.* In previous studies (Rooman & Wodak, 1988; Rooman et al., 1990), we identified in proteins of known structure a number of amino acid sequence patterns of the type AAXXK, where X denotes any residue, that reliably characterize a given secondary structure motif, such as XHHHX, where H denotes helix and X any structure. Since these reliable sequence–structure associations were identified considering only local interactions along the polypeptide chain, it was suggested (Rooman et al., 1990) that they could correspond to protein regions that adopt a preferred conformation on their own, and/or early during folding. In formulating this hypothesis it was implicitly assumed that no other reliable associations, predicting a different secondary structure, occur in the same region. As a corollary, incorrect predictions in certain regions were interpreted by invoking the important influence of tertiary interactions or the compensating (contradictory) effects of local interactions along the chain.

The identification of reliable sequence structure association, though different from the approach taken here, relies nevertheless on the same premises. It seemed worthwhile therefore to compare the conformations predicted by segments in this study to the secondary structures predicted by reliable sequence–structure associations.

To perform a fair comparison, associations were rederived from the database of the present study. They are defined as follows. The sequence patterns have 3 specified amino acids over a sequence stretch of 7 residues. The associated secondary structure motifs each correspond to a given secondary structure at a specific single position along the chain relative to the sequence pattern. The secondary structures are α-helix, β-strand, or coil, as defined by DSSP (Kabsch & Sander, 1983b). The distance between any of the specified amino acids in the sequence pattern and the secondary structure motif may not exceed 7 consecutive sequence positions. To be qualified as reliable, patterns characterizing α-helix and β-strand must occur at least 10 times, and those characterizing the more frequent coil, at least 15 times. The proportion of tolerated counterexamples, incidences in which the observed secondary structure is different from that defined by the structure motif, must not exceed 20% for α-helix and β-strand and 10% for coil. Note that the reliable associations identified here are slightly different from those obtained in Rooman and Wodak (1988) and Rooman et al. (1990), essentially due to differences in the protein dataset from which the associations are derived. The present set excludes protein sequences with more than 20% identity, while the previous set included proteins with a somewhat higher sequence identity. Consequently, some sequence patterns occurring in related proteins at homologous positions are now absent.

Searching our present database for such reliable sequence–structure association yields a total of 172 associations, occurring 11 times on the average. The secondary structure of individual residues predicted by means of these associations is then compared to the structural states predicted by the segments. Consistent predictions are defined as follows: for α-helix, β-strand, and coil associations, respectively, at least one structural state (A–C), (B–P), or (G–E–O) must be predicted by segments with preferred conformations. This state must occur no more than 2 positions away from the residue predicted by the associations.

A subset of the 172 reliable sequence–structure associations is given in Table I. This subset corresponds to patterns that characterize α-helix and β-strand and that occur at least 10 times in the database without a single counterexample, or 11 times with 1 counterexample. Their observed secondary structure and (ϕ, ψ, ω) conformation, as well the conformation obtained from segment prediction, are listed for all individual hits of the sequence patterns in the database. We see that, in most of the listed examples, the predictions obtained by means of reliable sequence–structure associations and segments with preferred conformations are consistent.

When the full set of 172 reliable associations is considered, the two prediction methods are consistent in 78% of the cases. Hence, the two different methods agree rather well in identifying sequence features that may impart local conformational preferences. In most of these cases (73%), predicted structures agree with those observed in the native protein. In the remaining consistent cases (5%), predicted structures are different from the native ones. These discrepancies with the observed structures could be ascribed to the local interactions being overridden by tertiary influences upon folding.

In 22% of the cases predictions provided by the two methods are not consistent. In about half of those (10%) the structure predicted by segments agrees with the native one, while the structure predicted by the associations does not. This might be due to the fact that associations do not incorporate contradictory influences from local interactions along the chain, while segment predictions do. Finally, in the remaining inconsistent predictions (12%), only those provided by the associations are correct. But in the majority of these (11%) the corresponding sequence regions are not covered at all by segments, which may indicate that parameter thresholds of both methods are not completely compatible. For example, lowering the minimum energy gap for segments could reconcile them. Finally, the remaining inconsistent predictions (1%) and probably also some of the other cases rationalized above could well reflect imperfections of both prediction methods.

DISCUSSION

In the method presented, a potential of mean force that considers local influences of residues along the chain is used to predict the structure of the protein backbone. Its main interest lies with the fact that it can be fine tuned to predict, with high accuracy and at very modest costs in computer time, the three-dimensional structure of short segments of the polypeptide with well-defined conformational preferences. Under certain stringent conditions, such as specifying a minimum energy gap of 1.5 kcal/mol between the lowest energy structure of each segment and the next-best ranking distinctly different structure, no more than 1.2% of the sequence is predicted on the average, with a segment score of 81% for 7 structural states (92% for 3 states). Computing these accurate predictions for a large number of proteins offers the opportunity to further investigate their physical significance.

Much larger fractions of the sequence (60%) can on the average be predicted, when the minimum energy gap is lowered to 0.5 kcal/mol, and when combining predictions that include, with those that ignore, the contributions from the flanking residues of each segment. But then, the segment score drops to 68% for 7 structural states (83% for 3 states). The prediction score computed by combining overlapping segments is even lower, with 53% on 7 states (70% on 3 states). Hence our method, when used to predict the protein backbone, encounters the well-known limitations of prediction methods that consider contributions from local sequence information alone. Due to the neglect of tertiary interactions, prediction scores for the

Table I: Comparison of Structure Predictions Provided by Reliable Sequence–Structure Associations and by Segments with Preferred Conformation[a]

protein	position	amino-acid sequence	observed secondary structure	predicted φ-ψ-ω conformation	observed φ-ψ-ω conformation
		ALK	H HHH		
1GD1	019	GRNVFRAALKN	HHHHHHHHHLLL	xxxxAAAAACCP	AAAAAAACCCB
1GD1	0258	TVEEVNAALKA	LHHHHHHHHHHH	BAAAAAAAAAAA	BAAAAAAAAAA
1HIP	16	ADNATAIALKY	LLLHHHHHHLL	AAAAAAAAAAA	ACPAAAAACGP
2CCY	A106	AAKAGPDALKA	HHHHLHHHHHH	AAACEAAAAAA	AAAAEAAAAAA
2CRO	12	RLKKRRIALKM	HHHHHHHHLLL	AAAAAAAxAAA	AAAAAAAACGB
2CTS	337	YTCQREFALKH	HHHHHHHHHHH	xxAAAAAAxxx	AAAAAAAAAAA
2CYP	27	QKVYNAIALKL	HHHHHHHHHHH	AAAAAAAAAAx	AAAAAAAAAAA
2LBP	264	ANQGIVDALKA	LLHHHHHHHHH	CCCxxAAAAAC	ACAAAAAAAAA
4MDH	A167	NRAKAQIALKL	HHHHHHHHHHH	AAAAAAAAAAC	CAAAAAAAAAC
5CPA	229	VAKSAVAALKS	HHHHHHHHHHL	AAAAAAAAAAB	AAAAAAAAACC
		AA K	HHHHHH		
1CC5	79	ELKAAIGKMSGL	HHHHHHHHHHLL	AAAAAAExxxxx	AAAACAACAAxx
1RNT	21	TAQAAGYKLHED	HHHHHHHHHHHH	AAAAAGBxxxxC	AAAAAAAAAAAC
2CCY	A97	TKLAAAAKAGPD	HHHHHHHHHLHH	AAAAAAAACEAA	AAAAAAAAAEAA
2CCY	A112	AQAAATGKVCKA	HHHHHHHHHHHH	AAAAACxxxxxx	AAAAAACAAAAA
2CDV	91	GADAAKKKELTG	LLLHHHHHHHHL	AAAAAAAAAACx	ECPAAAAAACCP
2LH4	8	ESQAALVKSSWE	HHHHHHHHHHHH	AAAAAAAxxAAA	AAAAAAAAACAA
3GRS	344	VAIAAGRKLAHR	HHHHHHHHHHHH	AAAACAAAAAAA	AAAAAAAAAAAA
4HHB	A12	NVKAAWGKVGAH	HHHHHHHHHLLL	xAAAAAxxxxxx	AACCACAAAExC
4MDH	A113	CQGAALDKYAKK	HHHHHHHHHLLL	AAAAAAAAAAAx	AAAAAAAAAPPC
4MDH	A231	QRGAAVIKARKL	HHHHHHHHHLLL	AAxxAAAAAAAA	CAACAAAAACGB
		A A K	H		
1GD1	0111	AKHLEAGAKK	LHHHHLLLLE	AAAAAAGAAx	CAAAACGPAB
1HIP	14	ADNATAIALK	LLLHHHHHHL	AAAAAAAAAA	ACPAAAAACG
2CCY	A97	ESTKLAAAAK	HHHHHHHHHH	AAAAAAAAAA	AAAAAAAAAA
2CDV	89	LETAGADAAK	HHHHLLLHHH	AAAPAAAAAA	AAAAECPAAA
2CYP	25	QKVYNAIALK	HHHHHHHHHH	AAAAAAAAAA	AAAAAAAAAA
2LBP	60	PKQAVAVANK	HHHHHHHHHH	AAAAxxxxxx	AAAAAAACAC
2LH4	37	LVLEIAPAAK	HHHHHLLLLL	AAAAAPAAAA	AAAAABACCA
2LH4	62	NPELQAHAGK	LHHHHHHHHH	BAAAAAAAAA	BAAAAAAAAA
2SNS	130	LRKSEAQAKK	HHHHHHHHHH	AAAAAAAAAA	CAAAAAACAC
3GRS	448	MLQGFAVAVK	HHHHHHHHHH	AAAxxAAAAA	AAAAAAAAAA
		AKK	HH		
1BP2	55	HDNCYKQAKK	HHHHHHHHLL	xxxxAAAAxx	AAAAAAAACC
1GD1	0113	AKHLEAGAKK	LHHHHLLLLL	AAAAAAGAAx	CAAAACGPAB
1PFK	A315	KGDWLDCAKK	LHHHHHHHHH	xxxxxCAAAx	CAAAAAAAAA
1PRC	C30	LHPATVKAKK	ELHHHHHHHH	xBAAAAAAAA	BPAAAAAAAA
1RHD	234	ELRAMFEAKK	HHHHHHHHLL	AAAAAAAAAA	AACAAAAACG
2CDV	92	ETAGADAAKK	HHHLLLHHHH *	AAPAAAAAAA	AAAECPAAAA
2SNS	69	TKKMVENAKK	HHHHHHLLLL	AxAAAAACxx	AxAAAACPCB
2SNS	132	LRKSEAQAKK	HHHHHHHHHH	AAAAAAAAAA	CAAAAAACAC
4MDH	A119	GAALDKYAKK	HHHHHHHLLL	AAAAAAAAAx	AAAAAAAPPC
5CPA	83	QATGVWFAKK	HHHHHHHHHH	AACGxxxxxx	AAAAAAAAAA
8ADH	337	KLVADFMAKK	HHHHHHHLLL	xAAAAAAAAA	AAAAAAACGA
		K DA	H		
1CC5	40	SAAWKTRADA	HHHHHHHHHH	CAAAAAAAAA	AAAAACAAAA
1HIP	18	AIALKYNQDA	HHHHLLLLLL	AAAAAAxxCC	AAACGPBABA
1PFK	A91	ENLKKRGIDA	HHHHHLLLLL	AAAAACGAxB	AAAAACGPAB
1PRC	C31	TVKAKKERDA	HHHHHHHHLL	AAAAAAAAAA	AAAAAACACC
2ABX	A26	LCYRKMWCDA	LLLEELLLLL *	xxxxxxxxxx *	xBPPBExCxx
2CCY	A101	AAAAKAGPDA	HHHHHHLHHH	AAAAACEAAA	AAAAAAEAAA
2LH4	100	VHVSKGVADA	HHHHLLLLLL	xxxxxxxxxx *	AAAACGPBAC
2LH4	147	IVIKKEMDDA	HHHHHHHHHH	AAAAAAAAAA	AAAAAAAAAC
3GRS	100	WRVIKEKRDA	HHHHHHHHHH	xxxAAAAxAA	AAAAAAAAAA
3TLN	307	VASVKQAFDA	HHHHHHHHHH	xxxxxxxxxx *	AAAAAAAAAA
4HHB	A60	KGHGKKVADA	HHHHHHHHHH	xxxxxxxxAA *	AAAAACACAA

Table I (Continued)

protein	position	amino-acid sequence	observed secondary structure	predicted ϕ-ψ-ω conformation	observed ϕ-ψ-ω conformation
		A K K	HH		
1CCR	13	KAGEKIFK	HHHHHHHH	Axxxxxxx *	AAAAAAAA
1GOX	13	IAKQKLPK	HHHHHLLH	AAAAABPx	AAACAPPA
1PRC	C26	PATVKAKK	HHHHHHHH	AAAAAAAA	AAAAAAAA
2CAB	153	EANPKLQK	LLLLLLHH *	xxxAAAAA	PPBCCCAA
2CCY	A113	AATGKVCK	HHHHHHHH	AACxxxxx	AAACAAAA
2LBP	301	LALVKDLK	HHHHHHHH	xAAAAAAA	AAAAAAAA
2LH4	113	EAILKTIK	HHHHHHHH	AAAxxxxx	AAAAAAAA
2SNS	130	EAQAKKEK	HHHHHHLL	AAAAAAAA	AAACACAG
4MDH	A103	KANVKIFK	LLLHHHHH	ACAAAAAA	CAAAAAAA
4MDH	A114	AALDKYAK	HHHHHHLL	AAAAAAAA	AAAAAAPP
4MDH	A232	AAVIKARK	HHHHHHLL	xAAAAAAA	CAAAAACG
		V V G	EE		
1GD1	O57	DAEVSVNG	LLLEEEEL	xxxxxxCG *	BBBBBBBE
1PFK	A166	ISVVEVMG	EEEEEELL	BBBBxxxx	BBBBBBPx
1PHH	97	GKTVTVYG	LLLEEELL	xBBBBBBE	GBBBBBAE
1PRC	C59	YKNVKVLG	LLLLLLLL *	xxxxxxxx *	BABBAACC
2ALP	120J	SSFVTVRG	LEEEELLL	xBBBBBBG	CBBBPPCE
2FB4	L105	GTKVTVLG	LEEEEELL	xxBBBBBG	EBBBBPPE
2LBP	5	DIKVAVVG	EEEEEEEE	xxxxxxxx *	BPPBBBBE
2SOD	O27	GDTVVVTG	LLEEEEEE	xBBBBBBG	xGBBBBBE
3TLN	253	HYGVSVVG	ELLEEELL	xxGBBBBE	BGGBBBPP
4MDH	A6	PIRVLVTG	LEEEEEEL	xxxxxxxx *	PBBBBBPG
4MDH	A125	SVKVIVVG	LLEEEELL	PBBBBBBE	CPBPBBAP

[a] The listed reliable associations are a subset of those considered in the text. They are defined as the associations characterizing α-helices and β-strands, that occur at least 10 times in the database without a single counterexample, or at least 11 times with 1 counterexample. Column 1 lists the proteins in which the association occurs, using their Brookhaven Databank code (Bernstein et al., 1977). Column 2 gives the position of the first amino acid in the sequence pattern, denoted by its Brookhaven Databank number. Column 3 gives the amino acid sequence in the neighborhood of the pattern. The sequence pattern is indicated above and is underlined in the amino acid sequences. Column 4 lists the observed secondary structure, with the letters H, E, and L representing, respectively, α-helix, β-strand, and coil (or loop). The secondary structures characterized by the sequence patterns are indicated above. Note that secondary structure motifs associated with the same sequence pattern are combined here. Their hits in the observed structures are underlined, and the counterexamples are marked by an asterisk. Column 5 gives the conformation predicted by means of segments, based on 7 (ϕ-ψ-ω) structure assignments. "x" indicates that no conformation is predicted at the corresponding position. Underlined positions indicate that these are in agreement with the secondary structure characterized by the sequence pattern, and asterisks mark predictions that are inconsistent with the sequence–structure associations. Column 6 lists the observed structural states. "x" indicates that the assignment of the observed (ϕ-ψ-ω) conformation to a structural state could not be made due to missing atomic coordinates.

three secondary structure classes α-helix, β-strand, and coil are indeed not expected to exceed the limit of about 70% for the entire sequence (Kabsch & Sander, 1984; Rooman & Wodak, 1991). Moreover, only scores of 60–63% are presently attained owing to the limited database size, which certainly hampers accurate evaluation of the force fields (Rooman & Wodak, 1988, 1991). Clearly, when 7 rather than 3 structural states are considered, as in the present study, still lower scores are expected. Note also that the 70% prediction score on 3 states, obtained here not counting unpredicted residues, indicates that the accuracy of our method is comparable to that of secondary structure predictions, where lower scores of 60–63% are obtained considering the entire sequence.

Identifying protein segments with well-defined conformational preferences in the absence of tertiary interactions may have interesting applications. As suggested earlier (Rooman & Wodak, 1991; Rooman et al., 1991), these segments could correspond to regions of the protein that adopt well-defined conformations early during folding, or to peptides with relatively well-defined structures in aqueous solution. This may appear to be a tenuous hypothesis, knowing that the conformation of any protein segment is the result of a delicate balance between interactions within the segment and those it makes with its environment. In the folded protein, the latter consist primarily of other protein atoms, while in the unfolded state, or in isolation, they consist mostly of solvent molecules. In addition, the interactions within a segment can themselves be altered by the environment. A case in point are those between polar and charged atoms, which should be significantly influenced by the dielectric properties of the surrounding medium.

These considerations do not preclude the possibility that the polypeptide chain contains a limited number of regions with sequence features that impart greater than average local conformational robustness. This may, among other things, serve the very important purpose of limiting the number of degrees of freedom during the folding process (Wright et al., 1988). One way of achieving such robustness is by increasing the contribution from local, intrasegment terms, so as to make the overall free energy balance in the region less sensitive to contributions from interactions with the solvent or the protein environment. Such intrasegment contributions could act to lower conformational entropy or to provide particularly favorable interaction enthalpies, or do both. On the basis of these considerations, protein segments most likely to qualify as being robust are those for which the preferred conformation

Table II: Examples of Protein Segments Likely To Be Part of Early Folding Intermediates[a]

protein	residues	sequence observed structure predicted structure	*G* (kcal/mol)	7-score (%)	3-score (%)	rms (Å)
1ACX	76–80	GTPVG GPPAE EBPBx	1.9	25	75	0.7
1BP2	35–39	GTPVD EPPCB GBPBx	2.1	25	75	0.7
1CCR	6–10	EAPPG CPPPB CPPCx	2.1	75	75	0.2
1CPV	10–21	ADIAAALEACKA AAAAAAAACCAB AAAAAAAAAAAx	1.4	82	100	0.6
1CPV	64–73	LKLFLQNFKA AACCACCCBA AAAAAAAAAx	1.2	33	89	1.4
1CRN	33–37	IIIPG BBBCP BBBPx	1.7	75	75	0.3
1GD1	104–113	DAAKHLEAGA AACAAAACGP AAAAAAAAGx	1.2	78	100	0.9
1GD1	192–201	KDLRRARAAA APCACAPPAC AAAAAAAAAXx	1.2	44	67	2.3
1GD1	257–266	NAALKAAAEG AAAAAAAACx AAAAAAAAAx	1.2	89	100	0.2
1GOX	8–16	NEYEAIAKQ AAAAAAAAC AAAAAAAAx	1.2	100	100	0.5
1GOX	143–150	RRAERAGF AAAAACGP AAAAAAGx	1.3	86	100	0.2
1GOX	331–339	MRDEFELTM AAAAAAAAA AAAAAAAAx	1.2	100	100	0.2
1PFK	227–234	DVDELAHF BAAAAAAA AAAAAAAx	1.3	86	86	0.5
1PHH	236–245	DERFWTELKA AACAAAAAAC AAAAAAAAAx	1.2	89	100	0.2
1RHD	93–100	THVVVYNG PPBBBBBC BBBBBBCx	1.2	57	86	0.8
1RHD	227–236	ELRAMFEAKK AACAAAAACG AAAAAAAAAx	1.2	78	100	0.3
2CAB	163–170	DALQAIKT AACACCAB AAAAAAAx	1.3	57	100	1.0
2CCY	39–50	DAAQRAENMAMV CAAAAAAAAAAA AAAAAAAAAAAx	1.2	91	100	0.4
2CCY	91–98	TESTKLAA AAAAAAAA AAAAAAAx	1.3	100	100	0.2
2CRO	3–13	TLSERLKKRRI BAAAAAAAAAA AAAAAAAAAAx	1.2	90	90	0.5
2FB4	16–23	QRVTISCS PBBBBBBB BBBBBBBx	1.2	86	100	0.6
2FB4	121–125	LFPPS BBPPP BBPPx	1.8	100	100	0.5
2TS1	4–8	LAELQ AAAAA AAAAx	1.7	100	100	0.1
2TS1	279–290	EALEQELREAPE AAAAAAAAABCC AAAAAAAAAAAx	1.2	82	91	1.1
3GRS	37–44	RRAAELGA AAAAACGP AAAAAAAx	1.3	71	86	0.5
3TLN	252–256	GVSVV GBBBP GBBBx	1.9	100	100	0.5

Table II (Continued)

protein	residues	sequence observed structure predicted structure	G (kcal/mol)	7-score (%)	3-score (%)	rms (Å)
451C	40–47	AEAELAQR AAAAAAAA AAAAAAAx	1.2	100	100	0.2
451C	60–64	PMPPN PBPPB PPPPx	1.9	75	100	0.3
4MDH	161–168	NRAKAQIA CAAAAAAA AAAAAAAx	1.2	86	100	0.2
8ADH	333–342	ADFMAKKFAL AAAACGABBC AAAAAAAAAx	1.2	56	67	3.3
average			**1.4**	**78**	**93**	**0.6**

[a] Subset of segments predicted as adopting a preferred conformation while including contributions from flanking residues. These segments are computed specifying either an energy gap $G \geq 1.7$ kcal/mol and a length $5 \leq L \leq 7$ residues or an energy gap $G \geq 1.2$ kcal/mol and a length $8 \leq L \leq 12$ residues. Columns 1 and 2 list the proteins and the positions in the sequence, using respectively Protein Databank codes and numbers (Bernstein et al., 1977), in which the segments occur. Column 3 indicates the amino acid sequence, as well as the observed and predicted conformations, denoted by the 7 (ϕ–ψ–ω) structural states. The last structural state of the predicted conformation is always undetermined ("x"), since the last Cα-position is determined by the state of the preceding residue (assuming the usual trans conformation). When several segments overlap in a given region, only the longest one is indicated, and when segments are of the same length, the one occurring first from the N-terminus is listed. Column 4 contains the value of the energy gap G for the corresponding segment. Columns 5 and 6 list the segment scores for 7 and 3 structural states, respectively, denoted as 7-score and 3-score. Column 7 contains the rms deviation between predicted and observed backbones (N, Cα, C atoms). The last row contains the average values of the rms deviation and of the segment scores, or equivalently of the prediction scores since overlapping segments have been removed from the table.

predicted by our method (or any method that adequately takes into account local interactions) is the same as in the native protein. One can argue indeed that the conformation of such segments is locally determined and remains essentially unaffected by interactions made in the folded protein.

Examples of segments that could correspond to robust structures likely to form early during folding in the 69 proteins of our dataset are given in Table II. They are a small subset of the 9610 segments identified by our method, representing the most reliable predictions, and consist of two categories: (1) segments with an energy gap of 1.7 kcal/mol or more whose lengths span 5–7 residues and (2) longer segments (8–12 residues) with a somewhat lower energy gap (≥1.2 kcal/mol). The average segment score of this subset is quite high, with 78% on 7 structural states and 93% on 3 states. Though many of these segments are helical, some are extended, and others contain turns.

To verify our hypothesis, it would be of interest to compare these segments with experimental data on the conformation of early folding intermediates. Unfortunately, such data are available only on a handful of proteins (Roder et al., 1988; Udgaonkar & Baldwin, 1988; Baum et al., 1989; Matouschek et al., 1989; Bycroft et al., 1990; Hughson et al., 1990) and the information they provide is often hard to interpret in terms of the stability of individual segments of the size analyzed here. In the few cases previously analyzed, cytochrome c, myoglobin, and α-lactalbumin (Rooman & Wodak, 1991; Rooman et al., 1991), at least one of the helices shown to form early in the experiments is predicted to be preferred in our procedure. The only reason that these helices are not among the examples listed in Table II is that they happen to be predicted with a somewhat lower energy gap than that used to compute those examples. They are nevertheless among the 9610 segments predicted with an energy gap of at least 0.5 kcal/mol.

Establishing a comprehensive list of segments with preferred conformation requires calibrating further the parameters of our procedure, especially the minimum value of the energy gap, against experimental results on early folding intermediates. This must however await the availability of a larger body of experimental data. Information on the presence and nature of nonnative conformations during the early folding stages would be particularly valuable for comparisons with our predictions for such conformations.

In the meantime, to probe further into the probable link between predicted segments with preferred conformation and early folding intermediates it would be of interest to determine the extent to which predictions are conserved in families of homologous proteins. Available evidence seems to suggest (Kim & Baldwin, 1982; Jaenicke, 1987) that closely related proteins follow similar folding pathways. Hence, consensus predictions of segments with preferred conformations in these proteins could correspond to regions that play an important role in folding. This question is dealt with in detail in the following paper (Rooman & Wodak, 1992).

Examples of segments predicted to have preferred conformations when excised from the protein, and thus likely to correspond to peptides with relatively well-defined structures in aqueous solution, are given in Table III. They were computed requiring an energy gap of 1.3 kcal/mol or larger and a length span of 5–15 residues. Once more, the segment score is fairly high, 77% and 87% on 7 and 3 structural states, respectively. Clearly, to have a well-defined structure in aqueous solution, the corresponding peptides should satisfy additional requirements, not directly monitored by our procedure. An important one is that they be soluble in water, which is not always the case since the peptides originate from larger "precursors". For the same reason, the possible influence of the newly created chain ends on the peptide conformational preferences should also be investigated. It is furthermore important to realize that we have at present no way of quantitatively relating predicted conformational preferences to experimentally determined structural propensities. Following such additional considerations, peptides in Table III, and others with somewhat lower energy gaps, could be

Table III: Examples of Protein Fragments Likely To Have Preferred Conformations on Their Own[a]

protein	residues	sequence observed structure predicted structure	G (kcal/mol)	7-score (%)	3-score (%)	rms (Å)
1CCR	83–87	YIPGT ABPGP BBPGx	1.3	75	75	0.8
1CPV	10–20	ADIAAALEACK AAAAAAAACCA AAAAAAAAAAx	1.3	80	100	0.5
1PFK	99–103	VVIGG BBBBB BBBEx	1.4	75	75	0.3
1PFK	189–193	FVVVP BBBBC BBBBx	1.4	100	100	0.2
1PRC	973–977	GYPLV GBOPB EBOPx	1.3	75	100	1.0
1RHD	226–236	EELRAMFEAKK AAACAAAAACG AAAAAAAAAAx	1.4	80	100	0.3
1RHD	281–285	RAPPE APPAC CPPPx	1.4	50	75	0.4
1SN3	57–61	TYPLP BPOPC BBOPx	1.4	75	100	0.6
2ACT	164–168	IVIVG BBBCE BBBBx	1.4	75	75	0.4
2CAB	194–198	TYPGS BBPBB BBOEx	1.4	50	50	2.2
2CTS	113–120	EWAKRAAL AAAACPBB AAAAAAAx	1.3	57	71	1.2
2LZM	118–125	LRMLQQKR AAAAACGB AAAAAAAx	1.4	71	86	0.6
2ADK	12–17	IFVVGG BBBBBP BBBBEx	1.3	80	80	0.6
3GRS	33–44	LASARRAAELGA AAAAAAAAACGP AAAAAAAAAAGx	1.3	91	100	0.3
451C	40–47	AEAELAQR AAAAAAAA AAAAAAAx	1.4	100	100	0.2
4FXN	13–21	EKMAELIAK AAAAAAAAA AAAAAAAAx	1.3	100	100	0.1
average			**1.4**	**77**	**87**	**0.6**

[a] Subset of segments predicted as adopting a preferred conformation when excised from the protein. They have an energy gap $G \geq 1.3$ kcal/mol and are of length $5 \leq L \leq 15$ residues. See Table II footnote for details.

considered as candidates for synthesis and experimental structure determination, with possible applications in the area of vaccine and drug design.

These expectations are supported by results obtained previously (Rooman et al., 1991) from comparing predictions obtained by our procedure with the solution structures of peptides with substantial helical propensity (Brown & Klee, 1971; Bierzynski et al., 1982; Shoemaker et al., 1985, 1987; Eisenberg et al., 1986; Marqusee & Baldwin, 1987; Dyson et al., 1988a,b; Marqusee et al., 1989). We verified here that, in all these sequences, a large number of segments with preferred conformation is identified. Moreover, the energy gaps for these segments are always larger than or equal to 1 kcal/mol, sometimes reaching 2 kcal/mol, and thus of the same order as those observed in the examples listed in Table III.

Another interest of our procedure resides in the possibility offered to assign an objective measure of confidence to each predicted structural state, using both the specific minimum energy gap, and the number of consistent predictions provided for this state by individual segments. This information could be used in folding simulations that include both local and nonlocal interactions along the chain, to limit or direct the search for stable conformations, thereby alleviating some of the computational bottlenecks associated with such simulations.

ACKNOWLEDGMENT

The prediction programs described in this paper (Prelude and Fugue) are available from the authors on request. We are grateful to J. Richelle for useful discussions and for use of the SESAM database.

REFERENCES

Baum, J., Dobson, C., Evans, P., & Hanley, C. (1989) *Biochemistry 28*, 7–13.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanoushi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.

Bierzynski, A., Kim, P., & Baldwin, R. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 2470–2474.

Brown, J., & Klee, W. (1971) *Biochemistry 10*, 470–476.

Bycroft, M., Matouschek, A., Kellis, J. T., Jr., Serrano, L., & Fersht, A. R. (1990) *Nature 346*, 488–490.

Dyson, J. H., Rance, M., Houghton, R. A., Lerner, R. A., & Wright, P. E. (1988a) *J. Mol. Biol. 201*, 161–200.

Dyson, J. H., Rance, M., Houghton, R. A. Wright, P. E., & Lerner, R. A. (1988b) *J. Mol. Biol. 201*, 201–217.

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) *Proteins 1*, 16–22.

Garnier, J., Osguthorpe, D., & Robson, B. (1978) *J. Mol. Biol. 120*, 97–120.

Gibrat, J.-F., Garnier, J., & Robson, B. (1987) *J. Mol. Biol. 198*, 425–443.

Hughson, F., Wright, P., & Baldwin, R. (1990) *Science 249*, 1544–1548.

Huysmans, M., Richelle, J., & Wodak, S. (1991) *Proteins 11*, 59–76.

Jaenicke, R. (1987) *Prog. Biophys. Mol. Biol. 49*, 117–237.

Kabsch, W. (1978) *Acta Crystallogr. A34*, 827–828.

Kabsch, W., & Sander, C. (1983a) *FEBS Lett. 155*, 179–182.

Kabsch, W., & Sander, C. (1983b) *Biopolymers 22*, 2577–2637.

Kabsch, W., & Sander, C. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1075–1078.

Kang, H. S., Kurochkina, N. A., & Lee, B. (1992) *J. Mol. Biol.* (in press).

Kim, P., & Baldwin, R. (1982) *Annu. Rev. Biochem. 51*, 459–489.

Liquori, A. (1969) *Q. Rev. Biophys. 2*, 65–92.

Marqusee, S., & Baldwin, R. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 8898–8902.

Marqusee, S., Robbins, V., & Baldwin, R. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 5286–5290.

Matouschek, A., Kellis, J., Serano, L., & Fersht, A. (1989) *Nature 340*, 122–126.

Oas, T., & Kim, P. (1988) *Nature 336*, 42–48.

Ramachandran, G., & Sasisekharan, V. (1968) *Adv. Protein Chem. 23*, 283–437.

Roder, H., Elöve, G., & Englander, W. (1988) *Nature 335*, 700–704.

Rooman, M., & Wodak, S. (1988) *Nature 335*, 45–49.

Rooman, M., & Wodak, S. (1991) *Proteins 9*, 69–78.

Rooman, M. J., & Wodak, S. J. (1992) *Biochemistry* (following paper in this issue).

Rooman, M., Rodriguez, J., & Wodak, S. (1990) *J. Mol. Biol. 213*, 337–350.

Rooman, M. J., Kocher, J.-P., & Wodak, S. J. (1991) *J. Mol. Biol. 221*, 961–979.

Shoemaker, K., Kim, P., Brems, D., Marqusee, S., York, E., Chaiken, I., Stewart, J., & Baldwin, R. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 2349–2353.

Shoemaker, K., Kim, P., York, E., Stewart, J., & Baldwin, R. (1987) *Nature 326*, 563–567.

Sippl, M. (1990) *J. Mol. Biol. 213*, 859–883.

Staley, J. P., & Kim, P. S. (1990) *Nature 344*, 685–688.

Udgaonkar, J., & Baldwin, R. (1988) *335*, 694–699.

Wright, P. E., Dyson, J., & Lerner, R. (1988) *Biochemistry 27*, 7167–7175.