Journal of Computer-Aided Molecular Design, 7 (1993) 457–472
ESCOM

J-CAMD 212

The importance of short structural motifs in protein structure analysis

Ron Unger[a,b] and Joel L. Sussman[c]

[a]*Center for Advanced Research in Biotechnology, Maryland Biotechnology Institute, University of Maryland, 9600 Gudelsky Dr., Rockville, MD 20850, U.S.A.*
[b]*Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, U.S.A.*
[c]*Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel*

Received 25 November 1992
Accepted 26 February 1993

Key words. Protein structure; Structural motifs; Secondary structure, Building blocks; Clustering; Structure verification; Structure prediction

SUMMARY

Proteins tend to use recurrent structural motifs on all levels of organization. In this paper we first survey the topics of recurrent motifs on the local secondary structure level and on the global fold level. Then, we focus on the intermediate level which we call the short structural motifs. We were able to identify a set of structural building blocks that are very common in protein structure. We suggest that these building blocks can be used as an important link between the primary sequence and the tertiary structure. In this framework, we present our latest results on the structural variability of the extended strand motifs. We show that extended strands can be divided into three distinct structural classes, each with its own sequence specificity. Other approaches to the study of short structural motifs are reviewed.

INTRODUCTION

Although the revolution in molecular biology has greatly enhanced our understanding of many biological phenomena, the process of folding a linear polypeptide chain into an exquisite three-dimensional (3D) functional protein is still poorly understood. During the past few years, there has been a great deal of excitement and unexpected developments in the area of protein folding [1]. However, some of the most recent studies seem to blur the picture even further. The role of hydrophobicity and its quantitative properties is one example, the elusive role of chaperones is another.

The number of protein structures determined experimentally by X-ray crystallography and NMR spectroscopy is growing rapidly, with over 1000 entries in the current Brookhaven database [2]. However, these are lagging far behind (by a factor of about 50) the number of amino acid sequences that have been determined [3]. This gap is likely to widen even further due to the results

of the Human Genome project and the rapid sequencing technologies that it promotes. In the foreseeable future, this gap can only be closed by improvements in methods of theoretical structural predictions.

There are two main theoretical approaches to protein structural prediction, one based on energetic considerations that tries to directly compute the native 3D structure and another one based on the body of knowledge accumulated in the structural database that could be exploited to facilitate the rational predictions of structures from their amino acid sequence. The most successful application of this latter approach is the emerging field of homologous modeling, in which an unknown structure is predicted from its sequence and from a known 3D structure with a similar sequence. Availability of additional related structures and sequences has been shown to be very useful. (See a recent review by Benner [4]). However, when such additional clues are not available, neither of these approaches have been successful so far in predicting the native 3D structure of a protein. We believe that eventually these approaches will have to be merged in a clever way in order to address the folding problem. However, as these approaches are currently distinct, we will review them in this way here, with special emphasis on aspects of the database approach that we have studied.

The computational approach is based on the assumption that by minimizing the free-energy function associated with a protein, its native structure can be found. This approach must face three main questions: First, the energy function is not well understood. The basic forces and factors are qualitatively known but their absolute and relative quantitative values are not. Second, the proposed functions, even in a simplified form, do not lend themselves to global minimization procedures. Finding a minimum of a function with thousands of variables with high-degree terms is beyond the ability of any optimization procedure. This difficulty leads us to the third and most intriguing problem. There is no proof that the functional (the so-called 'native') conformation is the global free-energy minimum conformation available to the chain. The claim that the native conformation of proteins is the lowest possible free-energy conformation was suggested by Anfinsen (the thermodynamics hypothesis [5]) and has since been accepted as a dogma. Actually this claim is supported by very little direct evidence, and careful consideration of the facts shows that they only indicate that each protein has a unique functional conformation and not that this conformation is the global minimal free-energy conformation.

The computational complexity of searching for the global-free minimum in protein models has been shown [6,7] to be an NP-hard problem (for a good introduction to the subject of computational complexity see Ref. 8). This finding strongly indicates that no efficient algorithm could be designed to find the global free-energy minimum. This theoretical consideration suggests that the natural folding process itself cannot be guaranteed to reach the global free-energy minimum. This is due to the fact that theoretical computational considerations are assumed to bind the behavior of natural systems (Church's hypothesis [9]) especially in cases where the similarity between the model and the natural problem is high. As the most realistic folding model, molecular dynamics (MD) can be simulated on a computer with only a polynomial slowdown; the natural folding process seems to be bounded by the theoretical NP-hard proof and hence cannot be guaranteed to reach an energetically optimal conformation in each instance.

The implication of this finding seems to be that a direct brute-force minimization of the energy function in order to derive the native conformation is neither feasible nor justifiable. One of the advantages of the database approach is that, to a certain degree, it can bypass these unresolved

issues. By inspecting the folds in the database, and by exploring the relationship between the sequence and the conformation of known structures, one can hope to obtain practical achievements even without having highly accurate energy functions, minimizing-prohibiting complicated functions, or committing oneself to any side of the global/local minimum issue.

RESULTS

Structural motifs

When one looks at the database of known protein structures, it becomes clear that proteins utilize recurrent structural motifs on all levels of organization (for a recent survey see Ref. 10). This recurrence is evident at the level of secondary-structure elements, at the level of the local 3D structures, as well as at the level of the whole topologies (folds) of domains of proteins. On the very local level, the recurrence of secondary-structure motifs has been known for a long time. On the global level, the classification of the overall folds and the design of threading algorithms aimed to test the compatibility between a sequence and a potential fold have received a great deal of attention recently. Our focus here will be on the intermediate level, i.e. the short 3D motifs, where we believe a great deal of useful structural information can be found. We start with a brief review of the secondary structure and the global fold level. Following this, we concentrate on describing the research that has been going on in the last few years on defining and analysing short structural motifs. In this framework we will describe our specific approach and some of our latest results.

Secondary structures

The seminal work of Pauling [11,12] predicted that protein structures are composed of standard secondary-structure elements. Interestingly, this work was based on theoretical considerations combined with the X-ray structures of several short peptides and preceded the actual determination of the first known structures of myoglobin and hemoglobin. The secondary-structure elements are mainly defined by the characteristic values of their dihedral angles and by their internal hydrogen-bonding patterns. They consist of helices, strands, turns and random coil. Even at the level of assigning the secondary-structure elements for a known 3D protein structure there are still definition problems. For example, the actual number of meaningful classes varies between different studies. Some methods (for example Ref. 13) consider turns and random coils as different objects, with random coil being the default assignment for non-helix, non-extended and non-turn. Other methods (for example Ref. 14) only deal with three-state systems (helices, strands, and coil which includes turns). Another issue is choosing the 'right' method of assigning a residue in a structure to one of these classes. The first objective automated method is due to Levitt and Greer [15] and is based mainly on values of dihedral angles, but the more widely used method of Kabsch and Sander [16] is based primarily on hydrogen-bonding patterns. Their program DSSP has become the standard method, and its assignments are used now as a benchmark to test the performance of secondary-structure prediction algorithms. DSSP subclassifies the main element types into eight classes (α-helix, 3_{10}-helix, π-helix, isolated β-bridge, extended strand (part of a β-ladder), H-bonded turn, bend, and a random class for residues that do not fit any of the other classes). Still, in the context of secondary-structure prediction, evaluations of these substates are usually merged into just three states, i.e. α, β and coil.

A large number of different algorithms have been proposed to predict the secondary structure for sequences of unknown structure. For a recent review see Sternberg [17]. On average, the methods have shown similar performance and are able to predict, using the three-state classification, the correct assignment for about 65% of the residues. This 65%-barrier seems to be inherent in the system, and to reflect the extent to which the local sequence determines the local structure. The conformation of the other 35% must be strongly influenced by interactions with nonlocal regions of the protein, i.e. the rest of the tertiary structure.

Global folds

Based on the large number of similar folds observed in the structural databse, Chothia [18] has recently suggested that the total number of actual folds is limited to about 1500. This estimation is based on crude approximations and on a vague definition of folds. Nevertheless, the logic behind this calculation implies that there is a limited number of possible folds. The same logic is behind the recent emphasis on 'threading' algorithms. These algorithms try to evaluate the fit of a sequence to a fold. If indeed, as Chothia suggested, the number of possible folds is not enormous, and if those folds will be determined in the next few years, then the protein folding problem can be virtually bypassed. It will be possible to determine the conformation of a new sequence just by checking its compatibility with the library of all possible folds. Thus, prediction could be achieved without a detailed understanding of the folding process.

Bowie et al. [3] published an important work in this direction. The 3D fold is described by a 1D 'environment string' in which each position represents the environment of this residue in the fold. The value of each position is one of 18 possible classes reflecting properties like solvent accessibility, polarity, and secondary structure. Once a fold is converted into a linear sequence of these values, it becomes possible to align this sequence with amino acid sequences. The alignment is done by a standard dynamic-programming method and the scores of matching an amino acid to a certain environment class are calculated from the frequencies of such a match in a database. The gap penalties are determined empirically and are aimed at preserving the boundary of the secondary-structure elements.

This method yielded remarkable results in identifying sequences that are compatible with a given environment. For example, the fold of sperm whale myoglobin was shown to be highly compatible with any myoglobin sequence. Furthermore, 511 of the 544 globins in the sequence database scored better than any non-globins. Yet, the description of the environment is static, while introducing a new sequence actually changes the environment. An arrangement of polar residues in one structure may be replaced by a cluster of hydrophobic residues in another structure while still supporting the same fold. Thus, in cases where the evolutionary distance between the sequences is large, and the fold is maintained by compensating mutations which change the static environment, the method might fail. We suspect that the limitation to a static environment is inherent in the method, and that a different way of representing the fold, rather than the 1D environment string, must be considered to overcome this limitation.

Hendlich et al. [19] suggested a way to extract potentials of mean force from the database and to use them to check a matching between a sequence and a fold. The energy of the matching is calculated as a sum of pairwise interactions. The values of these interactions were calculated using mean force parameters based on counting frequencies in the database, according to (a) amino acid types of the pair; (b) the sequence distance between these residues (i.e. how many residues

apart they are in the sequence); and (c) the Cartesian distance between the positions they occupy. The method was tested by fitting a given sequence to many folds taken from the database. As the method does not allow for gaps, each sequence was fitted to continuous portions of folds with the same length. (A structure of length N can 'supply' N – L overlapping folds of continuous L residues). The method was used successfully, in many cases, to identify the native conformation among a large number of incorrect folds of the same length. Note however that because of the fixed-size limitation no 'threading' can be performed. This is a clear limitation as it is known that many cases of structural homology are only evident when gaps are allowed.

The method of Bowie et al. [3] utilizes the ability of dynamic-programming algorithms to introduce gaps in the appropriate position in the alignment. Hendlich et al. [19] suggested a way to score the fit of a given sequence to a given fold. The next major step is to combine the two approaches. A first step in this direction was recently suggested by Jones et al. [20]. In this method the score of the match is provided by threading the sequence to the fold. This threading is done by sliding the sequence along the fold, while introducing insertions and deletions. In this way a sequence can be fitted to folds with various sizes. For each thread the energy is calculated as a pairwise sum, based on the actual position in which each residue ends up in the fold. This method was able, for example, to point out the structural similarity between C-phycocyanin and the globin folds, although the sequence similarity is very low. The authors mentioned that in some cases the method failed to find the desired similarities. For example, it found a wrong match of carp parvalbumin and T4 lysozyme with the globin fold, and missed the right match between glycolate oxidase and xylose isomerase.

Three problems should be discussed at this point. First of all, it is not clear whether indeed the energy function that is used is sensitive enough to identify the right fold. Secondly, by introducing gaps, the energy calculations are carried out on infeasible structures, in which the distance between two consecutive amino acid fragments separated by a gap will be much greater then the actual distance possible by the chain connectivity. In Ref. 30 this problem is somehow addressed by calculating the energy only for residues in helices and strands (ignoring any loop residues). Gaps are not allowed in helices and strands. This conservative solution, while working fine in some cases, may be the reason for the failure of the method in others. The third problem arises from the algorithm that is used to find the optimal threading. The algorithm of Taylor and Orengo [21] for structural alignment is used as a basis for the current work. It is based on a double pass of the dynamic-programming procedure by finding, at a low level, the optimal value of assigning a specific residue to a specific position, and then using these values as a basis for a global alignment at a higher level. The problem is that there is no mechanism to maintain consistency between the low-level and the high-level alignments. Thus, there is no guarantee that indeed the optimal solution is found. The computational problem of finding the optimal threading seems to be very hard. While the algorithm of Jones et al. [20] can be considered a reasonable heuristic approach, more robust algorithms are still needed.

Short structural motifs

Similar short 3D structural motifs are common to many proteins. Jones and Thirup [22] were the first to suggest a use for them. They noticed that the main chain of a particular protein could be reconstructed by using fragments of the main chain taken from other structures, e.g. the main chain of retinol-binding protein (RBP) could be reconstructed using fragments from only three

other known protein structures, with an rms of 1.0 Å. Their work mainly suggested that the use of known substructures might aid in the initial stages of X-ray structure determination when one has to fit a polypeptide chain to an electron density map. Using the same idea, Claessens et al. [23] suggested the use of recurrent motifs to build a complete main-chain model. Recently, Levitt [24] went a step further and suggested a method to build a full, all-atom model (including side chains), starting from the C_α atoms only (and in some regions of the chain not all of the C_α atoms are required). By pulling the fragments from the database, using sequence considerations in addition to the structural ones, and by averaging the results over many runs, this method shows excellent results. On average, the rms distance between the model and the correct structure is: 0.42 Å for main-chain atoms, 1.72 Å for side-chain atoms, and 1.26 Å for all atoms.

These methods actually do not require a systematic analysis of the recurrence of short structural motifs in the database. As long as each template can pull out a few or even just one good motif from the database, the modeling activities can proceed. We believe that a significant amount of information is contained in short structural motifs. In order to explore this phenomenon we set out to study short 3D motifs in a systematic manner.

We regard the short fragments as the building blocks of proteins in that they have specific 3D conformations and in many cases have some sequence specificity. Actually, we suggest that they are in fact another, intermediate, level of protein structure organization. These building blocks are actually more meaningful than the conventional secondary-structure (2D) level elements. Secondary-structure elements are in general too vague, both in their loose conformation definition and in their weak sequence specificity. As the building blocks are more specific in their structure and sequence, they may serve as an important link in trying to understand the complicated relationships between the primary sequence of proteins and their 3D conformation.

Our first step was to identify the recurrent short structural motifs [25]. We developed an algorithm that enabled us to identify the recurrent structural motifs which we call building blocks. For this study, we used a set of 82 well-refined proteins as our structural database. In order not to include trivially homologous proteins we retained only polypeptides that do not share identical dodecamer sequences. Working at the level of C_α atoms only and using fragments of six residues (hexamers) in length, we identified 81 building blocks that reoccur at least 35 times in the database. The similarity measure defining the recurrence was an rms distance of less than 1 Å. (Note that our rms measure normalizes the score by $(n-2)$ which makes the similarity for hexamers much larger than normalizing just by n, which has been used in several other studies.) Some of the more popular building blocks simply describe the common secondary-structure elements, but many represent structures of more delicate motifs. For example, many building blocks describe specific structural ways to connect helices and strands. Some of those agree with the common definitions of turns, but other hexamers display structural consistencies in fragments that are considered random coils by standard secondary-structure assignments.

The building blocks we have identified are shown in Fig. 1. Some of them may look quite similar to each other, but the rms distance between any pair of building blocks is at least 1 Å. These 81 building blocks were able to represent 76% of all hexamers in our database with a 1-Å rms similarity level, and 92% with an rms of 1.25 Å. In some cases the 3D structural similarity between hexamers appeared to be very high even though their secondary-structure assignments were significantly different. An interesting application of our work can be found in structure verification. Some structures may include a few nonstandard hexamers which are not represented

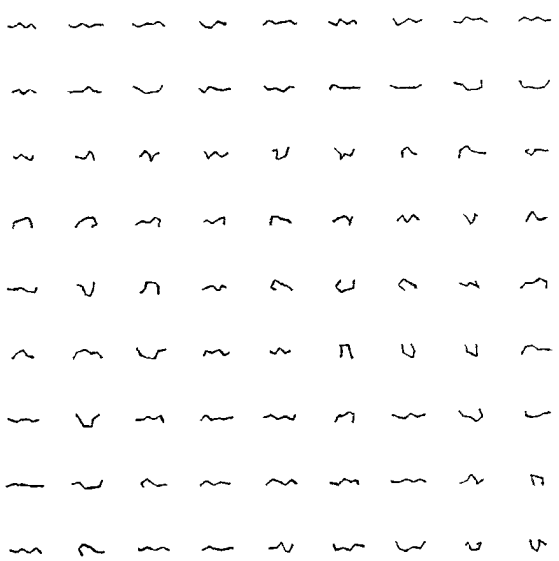

Fig. 1. The 81 most common building blocks in proteins.

by the building blocks. However, their number is usually limited. When we examined the structures in the database we found only two exceptions. For 2FD1 (Azotobacter Ferredoxin) only 14% of its hexamers could fit our building blocks. For 2ABXa (Alpha Bungarotoxin, chain a), only 26% of the hexamers were close enough to one of the building blocks. We suggested that such a low match indicates a severe problem concerning the accuracy of the structure. As for 2FD1, a redetermination of the structure [26a,b] (which can be found as 4FD1 in the PDB database) showed the original structure to be incorrect. The 2ABXa case still has to be checked. Hence, we suggest that our building blocks, derived by automatic clustering of shapes in the database, can be used as a practical first filter in structure verification.

The frequencies for each position of the amino acids of the hexamers that are associated with each building block were counted. This produced a sequence specificity matrix for each building block. Many of these matrices show a nonrandom amino acid distribution, but only few of them had a signal strong enough to be used as the sole source of a prediction scheme. As will be discussed below, Wodak's group decided to concentrate on the fragments that show a strong sequence/structure association.

We suggested [25] that even the weak signals can be useful in predicting the overall conformation. The suggested scheme is based on attaining few alternative building-block selections for each fragment, and then trying to select a consistent subset to cover the whole chain. This can be done by using local smooth connectivity constraints combined with long-range constraints, e.g. forming sheets from strands, forming disulfide cross links, enabling linkage to known ligands, etc.

Our next step was [27] to try to understand the source of the high selectivity of the hexamer population. Many hexamer shapes can be created by a computer program, even when the allowed dihedral angles are restricted to the well-populated regions in the Ramachandran map [28]. We showed that if one builds fragments in this simple way, i.e. keeping the distribution of the dihedral angles similar to what is found in the database, many of the fragments (ranging from over 90%

for hexamers to 65% for fragments with a length of 12 residues) will not contain internal collisions. We showed, however, that many of the fragments produced do not appear in real proteins. Thus, it became clear that the single dihedral-angle distribution is not sufficient to dictate protein-like 3D fragments.

We next constructed fragments of the type that maintains the distribution of consecutive pairs of dihedral angles as found in the database. In other words, if a dihedral-angle pair (ϕ_i, ψ_i) was chosen around one C_α atom, the next dihedral-angle pair was randomly chosen, only from the set of consecutive dihedral angles in which the first values are similar to (ϕ_i, ψ_i). This procedure yielded fragments that were much closer to fragments found in actual proteins. Thus, we concluded that the conformations observed in proteins are not a statistical reflection of sampling the single dihedral-angle distribution, but rather, they reflect the preferred combinations of dihedral-angle pairs. While the local dependency between consecutive pairs of dihedral angles is not large enough to dictate the global conformation of the whole protein, in many cases it may be fundamental in dictating the structure of the short structural motifs.

We focused next on the difference between secondary-structure classification and short structural motifs classification [29]. In Ref. 25 it was noticed that the classification into building blocks crosses the lines of traditional secondary-structure assignment. If, on the other hand, we examine a class of fragments all with the same particular secondary-structure assignment, there is a surprisingly large 3D structural variability amongst them. We present here preliminary results on the structural variability within the extended-strand motif. We have extracted all of the hexamers from our database that have been assigned by the DSSP program [16] to be 'pure' extended strands (i.e., assignment of EEEEEE, where E in Kabsch and Sander's (DSSP) terms denotes 'extended strand, participating in a β-ladder'). Longer fragments were considered in an overlapping manner, e.g., the EEEEEEE heptamer is considered as two consecutive hexamers. Altogether, 539 EEEEEE hexamers were found. The distance between hexamers was calculated as the rms deviation distance. The histogram of all 144 991 rms distances between the EEEEEE hexamers is shown in Fig. 2. The average distance is 1.7 Å. The rms values range, however, from virtually identical 3D conformations to some pairs with an rms distance of more then 4 Å. These numbers are higher than one might expect for hexamers with identical secondary assignments and support our claim that much structural variation is tolerated within the same secondary-structure motif.

We attempted to classify the hexamers into compact disjoint structural classes, such that the sum of distances within each class would be minimal. Formally, we assigned to each class C_i ($i = 1,\ldots n$) a set of elements E_i with centers r_i ($r_i \in E_i$), to minimize the quantity

$$\sum_{i=1}^{n} \sum_{e_j \in E_i} \mathrm{rms}(e_j, r_i) \qquad (1)$$

In order to minimize Term 1 we designed the following heuristic procedure (based on a variation of the K – means algorithm [30]: we chose, randomly, an initial set of elements as centers $r_1,\ldots r_n$ and assigned each of the other hexamers e_j to the set E_i if $\mathrm{rms}(e_j,r_i) < \mathrm{rms}(e_j,r_{k\neq i})$. We then recalculated the set of centers $r_1,\ldots r_n$ as follows: for each set of elements E_i, we selected the center r_i to be the element whose sum of rms distances to all of the elements in E_i is minimal. This process of assigning elements into classes according to their minimal distance to the centers and then recalculating the set of centers was repeated until the set of centers stabilized and did not change

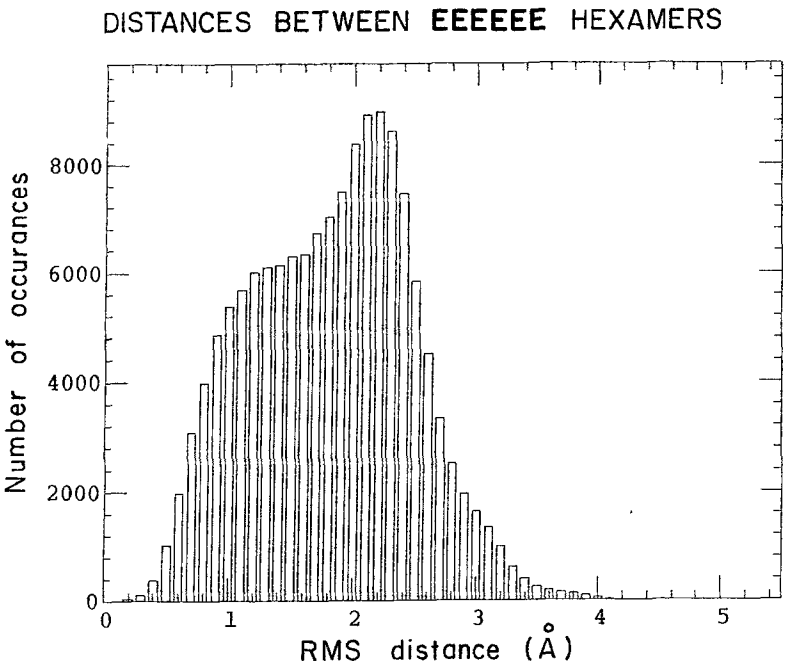

Fig 2. The histogram of the distribution of rms distances between each pair of the 539 EEEEEE hexamers in the database. Note that most of the rms distances are between 1 Å and 3 Å, which is surprising for hexamers with idential secondary assignment.

in consecutive iterations. This algorithm always converges, but is not guaranteed to minimize Term 1. For classification in three subclasses, we ran the algorithm many times with different initial centers and in most cases it converged to the following solution (which was the minimum among all the solutions we obtained): Class I: 192 hexamers centered around hexamer 157-162 of 3RP2a (rat mast cell protease, chain a [31]); Class II: 277 hexamers around hexamer 19-24 of 1FB4h (immunoglobulin fab, chain h [32]); and Class III: 70 hexamers around hexamer 161-166 of 2SGA (Proteinase A [33]). Figure 3 shows a superposition of 30 hexamers from each of the three classes mentioned above, which shows that the classes are structurally homogeneous and quite distinct from each other.

The sequence-specificity matrix of the 539 EEEEEE hexamers is shown in Table 1. The matrix is well correlated with the known amino acid preference of extended sheets [34]. Asp, Glu, Asn and Pro are underrepresented in the extended-sheet population relative to the overall frequencies, while Phe, Ile, Thr, Val and Tyr are overrepresented.

Each class is associated with its own sequence-specificity matrix. The three matrices are shown in Table 2. One can see that the overall matrix (Table 1) is quite flat, namely, for each amino acid there is no significant variation along the positions. On the other hand, the more specific matrices (Table 2) show high variation along the different positions, with some amino acids preferred in specific positions and rejected in others. Another interesting observation is the occurrence of glycines. In the EEEEEE population, Gly is present for 9.3%, quite evenly spread in all positions, which is approximately the observed overall Gly composition (9.2%) in our database. Thus, as a whole, the E residues neither prefer nor reject Gly residues. In the specific classes the situation is

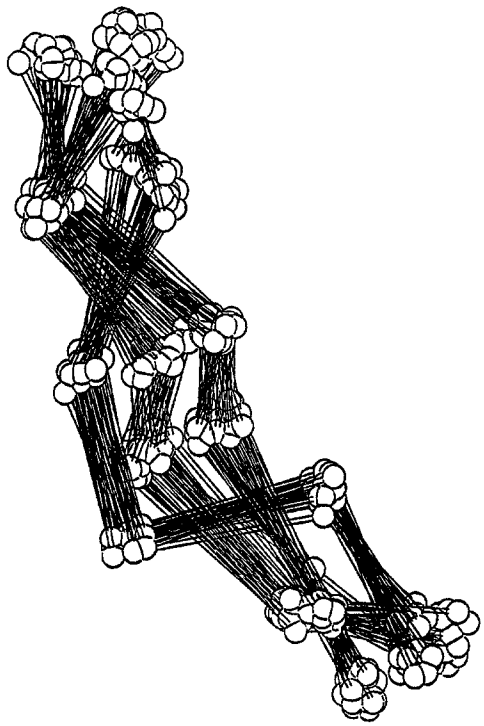

Fig. 3. The three classes of the extended-structure family. Thirty hexamers are shown for each class.

different. Class I has a high content of Gly in all positions, with an average of 12.1%. Class II has a low content of Gly, 6.3% on average. In Class III the average Gly content is the highest, 13.6%, and the variability among the different positions is also very large, with 2.9% in the first position in contrast to 35.7% in the fourth position!

To show the significance of the sequence specificity, we had to show that a random distribution of the hexamers into three classes would not yield any sequence specificity related to the classes. We divided the EEEEEE population, randomly, into three classes of the same size as the real classes and obtained the sequence-specificity matrices. We then calculated the distances between each pair of these matrices. The distance between two sequence-specificity matrices A and B is defined as:

$$\sum_{i=1}^{20} \sum_{j=1}^{6} |A_{i,j} - B_{i,j}| \tag{2}$$

We averaged this value over the three pairs of matrices. Repeating this random distribution a 1000 times we derived a mean value of 3 with standard deviation 0.1. The highest observed value for random distribution was 3.3. Calculating the distance between each pair of the real matrices yielded an averaged value of 4 which is 10 standard deviations more than randomly expected. The strong nonrandom correlation between the sequence and structure subclassifications may enable the use of the sequence signal in a more refined prediction scheme.

In the area of systematic analysis of short structural motifs, much work has been done by Wodak's group. They concentrated on analyzing the relationship between the sequence and the structure of short structural motifs. In the early studies they concluded that the size of the database was too small [35] to derive reliable sequence/structure associations, and that the quality of most derived associations was poor [36]. It was then realized that the poor performance is actually due more to long-range interactions rather than to the size of the database [37]. They suggested [38,39] however, that short structural motifs, which have a high correlation between sequence and structure, can be considered as stable structural units that fold independently of the rest of the structure. In Ref. 39 a method was developed to predict the backbone conformation from the sequence. In this work the emphasis was on identifying the correct values of the dihedral angles (using seven domains in the Ramachandran map [28] as the structural states). For each residue, a prediction was performed independently, calculating the structural state for which the net energy of a local window of length 17 (eight residues on either side of the predicted central residue) will be minimal. The energy terms are based on the potential of mean force, similarly to the one derived by Hendlich et al. [19]. However, since these values are used for structural prediction purposes, the frequencies in the database are calculated based on amino acid types and sequence distance, while the Cartesian distances between positions cannot be considered. The frequencies considered are those of finding a specific amino acid in a specific position inside the

TABLE 1
SEQUENCE-SPECIFICITY MATRIX OF THE EEEEEE HEXAMERS[a]

Amino acid	Pos. 1	Pos. 2	Pos. 3.	Pos. 4	Pos 5	Pos. 6
A	9.1	7.6	7.3	7.8	8.7	6.9
C	2.4	4.3	4.1	3.7	2.6	1.5
D	1.7	2.0	1.9	2 4	2.4	3.5
E	2.6	2.8	2.8	2.2	3.7	3.4
F	4.1	4.1	4.3	6.1	6.5	5.0
G	9.3	9.9	9.7	10.2	9.1	7.4
H	2.8	3.4	4.3	2.2	1.9	2.4
I	7.1	7.1	5.8	5.8	6 1	5.8
K	4.7	3.2	3.4	3.9	4.3	4.5
L	6.7	9.1	8.7	7.8	7.4	7.4
M	1.3	0.6	1.1	0.9	0.9	1.3
N	1.3	1.9	2.2	2.2	2.4	4.5
P	1.1	1.1	0.9	0.6	0.2	1.7
Q	5.2	2.0	3.5	3.2	4.3	3.7
R	3.0	1.9	1.7	2.0	3.9	3.5
S	9.7	8.7	7.6	9.5	8.2	9.9
T	9.7	10.6	11.3	9.5	10.0	9.7
V	10.6	12.3	12.5	13.0	11.3	11.7
W	1.7	1.9	2.0	1.7	1.5	1.7
Y	6.1	5.8	5.2	5.2	4.5	4.7

[a] The normalized distribution of amino acids along the six positions of the 539 hexamers in the database that have a secondary assignment to EEEEEE. The normalization was done by dividing the counts in each entry by the number (539) of hexamers used; the values are given in percentages.

TABLE 2
SEQUENCE DISTRIBUTION MATRICES OF SUBCLASSES OF EXTENDED STRUCTURES[a]

Amino acid	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6
Class I						
A	5.7	6.8	9.4	6.8	11.5	7.3
C	3.6	3.6	3.6	2.6	2.1	0.0
D	0.0	3.1	2.1	2.6	3.6	3.1
E	2.6	3.6	2.1	3.1	4.2	2.1
F	6.3	5.2	5.7	6.8	4.7	6.2
G	13.5	15.6	9.4	12.5	12.0	9.4
H	2.6	3.6	4.2	2.6	1.0	2.1
I	10.4	2.1	3.6	7.3	5.7	6.8
K	3.6	3.1	3.6	3.1	2.6	4.2
L	5.7	7.8	7.8	6.8	5.2	8.3
M	1.0	0.5	0.5	1.6	0.0	0.5
N	1.6	1.0	5.2	0.5	3.6	4.2
P	2.1	0.5	2.1	0.0	0.5	2.6
Q	3.6	3.6	3.1	4.2	4.2	5.2
R	2.1	3.6	1.0	1.0	4.2	3.6
S	10.4	6.8	9.4	7.3	8.3	7.8
T	6.8	14.6	10.9	8.9	14.6	4.2
V	9.9	10.4	10.4	14.1	7.3	15.6
W	3.6	1.0	2.1	1.0	0.5	0.5
Y	4.7	3.1	3.6	7.3	4.2	6.3
Class II						
A	10.5	6.9	6.9	7.6	7.9	5.8
C	2.2	5.4	5.4	4.7	3.6	2.9
D	2.9	1.8	1.1	2.2	1.8	4.3
E	1.4	2.5	3.2	1.8	3.6	4.3
F	3.2	3.6	3.6	6.9	8.7	3.6
G	7.9	6.5	8.3	2.5	6.5	5.8
H	2.9	3.6	4.3	2.2	2.9	2.9
I	5.8	11.2	7.2	5.8	5.4	4.7
K	5.1	3.6	3.2	4.0	5.8	5.4
L	6.9	8.3	9.0	9.7	9.0	7.6
M	1.4	0.7	1.8	0.7	0.7	2.2
N	1.1	2.5	0.7	2.2	1.4	4.3
P	0.8	0.7	0.0	0.7	0.0	1.4
Q	6.5	1.1	3.2	2.9	4.3	3.6
R	3.2	0.7	2.2	2.9	3.6	3.6
S	9.0	9.7	5.8	11.2	7.2	11.2
T	9.4	7.9	11.2	10.5	7.6	10.8
V	10.8	12.3	13.4	14.4	13.7	10.1
W	0.7	1.8	2.5	2.5	2.5	1.8
Y	8.3	9.0	6.9	4.7	3.6	3.6

TABLE 2 (continued)

Amino acid	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos 6
Class III						
A	12.9	12 9	2.9	11.4	4.3	10.0
C	0.0	1.4	0.0	2.9	0.0	0.0
D	1.4	0.0	4.3	2.9	1.4	1.4
E	7.1	1.4	2.9	1.4	2.9	2.9
F	1.4	2.9	2.9	1.4	2.9	7.1
G	2.9	7.1	15.7	35.7	11.4	8.6
H	2.9	1.4	4.3	1.4	0.0	1.4
I	4.3	4.3	4.2	1.4	10.0	7.1
K	5.7	1.4	2.9	5.7	4.3	1.4
L	8.6	15.7	10.0	2.9	7.1	5.7
M	1.4	0.0	0.0	0.0	4.3	0.0
N	1.4	1.4	0.0	7.1	2.9	5.7
P	0.0	5.7	1.4	1.4	0.0	0.0
Q	4.3	1.4	5.7	1.4	4.3	0.0
R	4.3	1.4	1.4	1.4	4.3	2.6
S	10.0	10.0	10.0	8.6	11.4	10.0
T	18.6	10.0	12.9	7.1	7.1	20.0
V	11.4	17.1	15.7	4.3	12.9	7.1
W	0.0	4.3	0.0	0.0	0.0	4.3
Y	1.4	0.0	2.9	1.4	8.6	4.3

[a] Three normalized sequence distribution matrices of subclasses I, II and III of the extended-structure family. Note the large variability of frequencies along the different positions for many of the amino acids.

window, and of finding a specific pair of amino acids in a specific pair of positions inside the window when the central residue is in a specific structural state. This method can be considered to be a structural extension to the GOR secondary-structure prediction [13], as this prediction scheme is able to construct, based on the structural state of each residue, a trace of the protein backbone. The method showed promising results in determining the structure of fragments that were shown experimentally to have a stable structure as peptides. Unfortunately, this study did not report on the relative importance of the different frequencies that were considered for the potential. It would be interesting to know what the importance of the pairs is versus the single amino acids counts, and what the relative importance of different positions, or position pairs, is in the final assignment.

In a very recent set of papers, this group went a step further. In Ref. 40 a method was developed to scan a protein sequence in order to detect fragments for which a specific conformation is highly preferred over alternative conformations. Fragments of preferred conformation are those for which the energy gap between the best conformation and the next distinct conformation is large. As this energy is calculated only from neighboring residues, such fragments are suggested to have a local stable conformation, and thus serve as a nucleation site to initiate the folding process. The concept is similar to that of Moult and Unger [41] in which fragments that locally (without assistance from other regions of the protein) bury a significant portion of hydrophobic surface area from the solvent were suggested to be initiation sites for the folding process. The progress

here is in the attempt to identify these fragments from the sequence alone without additional structural information.

The next question Wodak's group addressed dealt with the conservation of the regions with preferred conformation. If indeed these regions play a significant role in folding and stability then they should be highly conserved within related families of sequences. In this context what should be conserved is not the actual sequence of a region, but rather its predicted preferred conformation. In Ref. 42, fragments that have preferred conformations were traced within their sequence families (i.e. a family of sequences with a high sequence similarity to the sequence containing the fragment). To enable a full analysis, the fragments were taken from known structures. It was found that, in 13 such families, at least one region was predicted to have the same preferred conformation in all the members of the family. The conservation of the conformation is not trivially linked to sequence conservation as the sequence conservation in these regions was not higher than the overall sequence conservation within the family. These regions are suggested to be important to the folding process as they may guide all the members of the family to be folded in a similar pathway. The authors mentioned however that a high level of overall sequence similarity within the family is needed to guarantee the existence of such guiding regions. Thus, it may follows that members of families that share the same fold but have a low overall sequence conservation may be folded using different pathways. This implication still needs more directly convincing support.

In another, more restricted domain, similar results have been obtained. Specifically in the study of antigen-binding loops in immunoglobulins [43,44], it was shown that there are some 'canonical forms' of loops. The specific shape of the loop is associated with specific amino acids that can form those loops.

CONCLUSIONS

During the past few years the role of short structural motifs in our understanding of protein structures has become increasingly important. Computationally, as a basis for a 3D prediction scheme, there is a clear advantage in using the building blocks rather than standard secondary-structure elements. Unlike secondary-structure elements, the building blocks have tertiary meaning, in the sense that concatenating them in an overlapping manner produces a 3D chain. In contrast, secondary-structure elements do not carry a great deal of 3D information; fragments with identical secondary-structure assignment can still be very different from a structural point of view. Even if the secondary-structure elements of a protein are known, it is not obvious how to combine them into a 3D structure. In addition, it appears to be easier to assign building blocks from the sequence rather than secondary-structure elements. The 3D information content of the building blocks leads us to regard them more like units on a '2.5D' level of protein structure organization. The next step should be a procedure that will merge the strong prediction signal from some of the building blocks with the weaker signal of the others. Combined, this information should be further incorporated into a threading algorithm to overcome the limitations of the current algorithms, by providing a way to model the regions of the chain which have to be inserted into a given fold.

ACKNOWLEDGEMENTS

We want to thank Scot Wherland and David Harel for their part in the work on the extended-strands analysis, and John Moult for critical reading of this manuscript.

REFERENCES

1 Levitt, M., Curr. Opin. Struct. Biol., 1 (1991) 224.
2 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
3 Bowie, J.U., Lüthy, R and Eisenberg, D., Science, 253 (1991) 164.
4 Benner, S.A., Curr. Opin. Struct. Biol., 2 (1992) 402.
5 Anfinsen, C.B., Science, 181 (1973) 223.
6 Unger, R. and Moult, J., Bull. Math. Biol., 1993, in press.
7 Ngo, J.T. and Marks, J., Protein Eng., 5 (1992) 313.
8 Garey, M.R. and Johnson, D.S., Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, San Francisco, 1979.
9 Kleen, S.C., Introduction to Metamathematics, Van Nostrand, Princeton, 1952.
10 Branden, C. and Tooze, J., Introduction to Protein Structure, Garland, New York, 1991.
11 Pauling, L., Corey, R.B. and Branson, H.R., Proc. Natl. Acad. Sci. U S.A., 37 (1951) 205.
12 Pauling, L. and Corey, R.B., Proc. Natl. Acad. Sci. U.S.A., 37 (1951) 729.
13 Garnier, J., Osguthorpe, D.J. and Robson, B., J. Mol. Biol., 120 (1978) 97.
14 Chou, P.Y. and Fasman, G.D., Adv. Enzymol., 47 (1978) 45.
15 Levitt, M. and Greer, J., J. Mol. Biol., 114 (1977) 181.
16 Kabsch, W. and Sander, C., Biopolymers, 22 (1983) 2577.
17 Sternberg, M.J.E., Curr. Opin. Struct. Biol., 2 (1992) 237.
18 Chothia, C., Nature, 357 (1992) 543.
19 Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. and Sippl, M., J. Mol. Biol., 216 (1990) 167.
20 Jones, D.T., Taylor, W.R. and Thornton, J.M., Nature, 358 (1992) 86.
21 Taylor, W.R. and Orengo, C.A., J. Mol. Biol., 208 (1989) 1.
22 Jones, T.A. and Thirup, S., EMBO J., 5 (1986) 819.
23 Claessens, M., Van Cutsem, E., Lasters, I. and Wodak, S.J., Protein Eng., 2 (1989) 335.
24 Levitt, M., J. Mol. Biol., 226 (1992) 507.
25 Unger, R., Harel, D., Wherland, S. and Sussman, J.L., Proteins, 5 (1989) 355.
26 a Stout, G.H., Turley, S.T., Sieker, L.C. and Jensen, L.H., Proc. Natl. Acad. Sci. U.S.A., 85 (1988) 1020.
b. Stout, C.D., J. Biol. Chem., 263 (1988) 9256.
27 Unger, R., Harel, D., Wherland, S. and Sussman, J.L., Biopolymers, 30 (1990) 499.
28 Ramakrishnan, C. and Ramachandran, G.N., Biophys. J., 5 (1965) 909.
29 Unger, R., Ph.D. Thesis, Weizmann Institute of Science, Rehovot, Israel, 1991.
30 MacQueen, J., Proceeding of the 5th Berkeley Symposium on Probability and Statistics, University of California Press, Berkeley, 1967.
31 Reynolds, R.A., Remington, S.J., Weaver, L.H., Fisher, R.G., Anderson, W.F., Ammon, H.L. and Matthews, B.W., Acta Crystallogr., B41 (1985) 139.
32 Marquart, M. and Diesenhofer, J., Immunol. Today, 3 (1982) 160.
33 Moult, J., Sussman, F. and James, M.N.G., J. Mol. Biol., 182 (1985) 555.
34 Chou, P Y. and Fasman, G.D., Biochemistry, 13 (1974) 222.
35 Rooman, M.J. and Wodak, S.J., Nature, 335 (1988) 45.

36 Rooman, M.J., Wodak, S.J. and Thornton, J.M., Protein Eng., 3 (1989) 23.
37 Rooman, M.J. and Wodak, S.J., Proteins, 9 (1991) 69.
38 Rooman, M.J., Rodriguez, J. and Wodak, S.J., J. Mol. Biol., 213 (1990) 337.
39 Rooman, M.J., Kocher, J.A. and Wodak, S.J., J. Mol. Biol., 221 (1991) 961.
40 Rooman, M.J., Kocher, J.A. and Wodak, S.J., Biochemistry, 31 (1992) 10226.
41 Moult, J. and Unger, R., Biochemistry, 30 (1991) 3816.
42 Rooman, M.J. and Wodak, S.J., Biochemistry, 31 (1992) 10239.
43 Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davis, D. Tulip, W.R., Colman, P.M., Spinelli, S., Alzari, P.M. and Poljak, R.J., Nature, 342 (1989) 877.
44 Tramontano, A., Chothia, C. and Lesk, A.M., Proteins, 6 (1989) 382.