PROTEINS: Structure, Function, and Genetics 5:355–373 (1989)

A 3D Building Blocks Approach to Analyzing and Predicting Structure of Proteins

Ron Unger,[1] David Harel,[1] Scot Wherland,[2] and Joel L. Sussman[3]

[1]Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel; [2]Department of Chemistry, Washington State University, Pullman, Washington 99164–4630; and [3]Department of Structural Chemistry, The Weizmann Institute of Science, Rehovot 76100, Israel

ABSTRACT **A new approach is introduced for analyzing and ultimately predicting protein structures, defined at the level of C_α coordinates. We analyze hexamers (oligopeptides of six amino acid residues) and show that their structure tends to concentrate in specific clusters rather than vary continuously. Thus, we can use a limited set of standard structural building blocks taken from these clusters as representatives of the repertoire of observed hexamers. We demonstrate that protein structures can be approximated by concatenating such building blocks. We have identified about 100 building blocks by applying clustering algorithms, and have shown that they can "replace" about 76% of all hexamers in well-refined known proteins with an error of less than 1 Å, and can be joined together to cover 99% of the residues. After replacing each hexamer by a standard building block with similar conformation, we can approximately reconstruct the actual structure by smoothly joining the overlapping building blocks into a full protein. The reconstructed structures show, in most cases, high resemblance to the original structure, although using a limited number of building blocks and local criteria of concatenating them is not likely to produce a very precise global match. Since these building blocks reflect, in many cases, some sequence dependency, it may be possible to use the results of this study as a basis for a protein structure prediction procedure.**

Key words: protein, structure, prediction, primary, secondary

INTRODUCTION

Understanding the relationship between the three-dimensional structure of proteins and their one-dimensional amino acid sequence is still one of the most fundamental unsolved problems in physical biochemistry. Much attention has been given lately[1–4] to the relationship between structure and sequence of short oligopeptides. The idea was to see to what extent identical or similar sequences imply similar structures. We would like to suggest a different approach to this problem. We want to identify, with high accuracy, all the possible conformations of short oligopeptides and then analyze the sequences they can cope with. It is well known that proteins are composed of secondary structural elements such as helices, sheets, and various types of turns. These classifications are crude, they describe only local regions of the structure, and do not lead in any simple way to a prediction of tertiary structure. Recent work, by Jones and Thirup[5,6] and Blundell et al.[7], have shown the usefulness of a library of structural motifs in fitting structural models to electron density maps.

We suggest that an extended library of well-chosen short structural motifs can be extracted from known structures and used in analyzing and predicting protein structure. In deriving these representative motifs we consider only the C_α coordinates defining the structure and do not use any current definitions of secondary structure elements. Throughout this study we used a set of 82 well-refined protein structures (according to criteria that will be discussed later). These contain about 13,000 different overlapping hexamers (that is, oligopeptides of six amino acid residues). Specifically, we address the three following questions:

1. Is it possible to divide these 13,000 hexamers into a reasonable number of really different structural shapes? We thus want to know whether the conformations of hexamers vary continuously or can be separated into disjoint clusters.
2. Can we use the "library" of these different shapes, which we call *building blocks*, to reconstruct the structure of proteins?
3. Do the building blocks "carry" some sequence specificity that will enable us to assign building blocks from sequences, and thus use them in a three-dimensional prediction scheme?

Below we discuss these three questions. We describe in detail the methods we used and the results

Received November 29, 1988; revision accepted March 27, 1989.

Address reprint requests to Joel L. Sussman, Department of Structural Chemistry, The Weizmann Institute of Science, Rehovot 76100, Israel.

TABLE I. Refined Brookhaven Peptides*

1APR	1BP2	1CC5	1CCR	1CPP	1CPV	1CRN	1CTF
1ECA	1FB4h	1FBJl	1FC2d	1FDX	1GAPa	1GCR	1HIP
1HMQa	1INSa	2INSb	1LH1	1LZ1	1LZT	1MBD	1NXB
1PCY	1PP2r	1PPD	1PPT	1SBT	1SN3	1TGSi	2ABXa
2ACT	2ALP	2APP	2AZAa	2CAB	2CCYa	2CDV	2CTS
2CYP	2ESTe	2FD1	2GN5	2INSa	2LHB	2LZM	2OVO
2PABa	2PKAa	2PKAb	2RHE	2SGA	2SNS	2SODo	351C
3C2C	3DFR	3ICB	3PGM	3PTP	3RP2a	3RXN	3SGBe
3TLN	4ADH	4APE	4ATCa	4ATCb	4CYTr	4DFRa	4FXN
4HHBb	4HHBc	4HHBd	4SBVa	5CPA	5LDH	5PTI	5RSA
5RXN	7CAT						

*The first four characters are the Brookhaven Data Bank file name (release of January 1987), and the chain indicator, when required, is the fifth lower case character.

we obtained, and first show that hexamers can indeed be divided into distinct structural clusters. We then describe an algorithm for reconstructing protein structure and discuss its performance and limitations. As to the third question, about the sequence specificity of the building blocks, we present some preliminary results indicating the feasibility of the proposed scheme.

DATA BASE CREATION

Our structural data base was taken from the Brookhaven Protein Data Bank, as released in January 15, 1987 (354 polypeptide chains, 61,064 residues). We extracted the sequence records, separated each record into different polypeptide chains, and found the polypeptides that share identical dodecamer sequences. Only one representative of each such set was retained, in order to eliminate trivially homologous peptides. This procedure still retains a very few peptides with high sequence homology and 3D structural similarity. We retained only those structures for which X-ray data had been collected to 3.0 Å or higher resolution, and which had been refined against the observed X-ray data to an R factor of less than 30%. This left us with a library of 82 peptides (12,973 residues), which we will refer to as the "refined Brookhaven" data base (see Table I.) Actually, all 82 structures had been solved to a resolution of 2.8 Å or higher and 68 structures had a resolution of 2 Å or better.

DEFINING THE DISTANCE BETWEEN PROTEIN STRUCTURES

The distance (or similarity) between two structures is not easy to define. We basically used the following well-accepted definition. The RMS deviation distance between two structures s and t is measured by first aligning them to the greatest possible extent using the BMF (best molecular fit) algorithm of Nyburg[8] or Kabsch, [9,10] and then calculating the difference in the positions of the corresponding C_α atoms as a normalized root mean square deviation

$$RMS = \left[\frac{\sum_{i=1}^{n}(r_i^s - r_i^t)^2}{n-2}\right]^{1/2}$$

One must keep in mind that this definition does not always capture the intuitive notion of similarity. First, it is highly scale dependent, i.e., two structures with a similar shape but different sizes will show no similarity. Second, it is not sensitive to the geometrical and topological properties of the structures, see for example Figure 1. Third, and possibly most important, it is sensitive to insertions and deletions since it is based on the distance between corresponding atoms in the two structures. However, for short structures such as hexamers, the RMS distance seems to be a good measure of similarity.

We are exploring several different approaches for measuring the distance between longer structures, both local ones (having to do with properties of substructures) and global ones (having to do with properties of the structure as a whole, for example its curvature).

MEASURING THE RANDOM DISTANCE BETWEEN PROTEIN STRUCTURES

In order to evaluate the statistical significance of our results, we need a framework of measurements of random distances between protein fragments of various lengths. We define random distance as the expectation of the RMS distance between a pair of fragments, of given length, drawn at random from our refined library. We calculate this distance by choosing a few sample proteins of different types, "extracting" all of their overlapping fragments, and calculating the average distance between them. In this study, we initially used four proteins: 4HHBb (human deoxyhemoglobin, β-chain), 5PTI (bovine pancreatic trypsin inhibitor), 1BP2 (bovine pancreatic phospholipase A2), and 1PCY (oxidized poplar plastocyanin), with total length of 426 residues. These structures had been very accurately determined, and they represent different structural classes of proteins. (Repeating the experiment using different sets of initial proteins yielded similar results). We calculated the RMS distance between

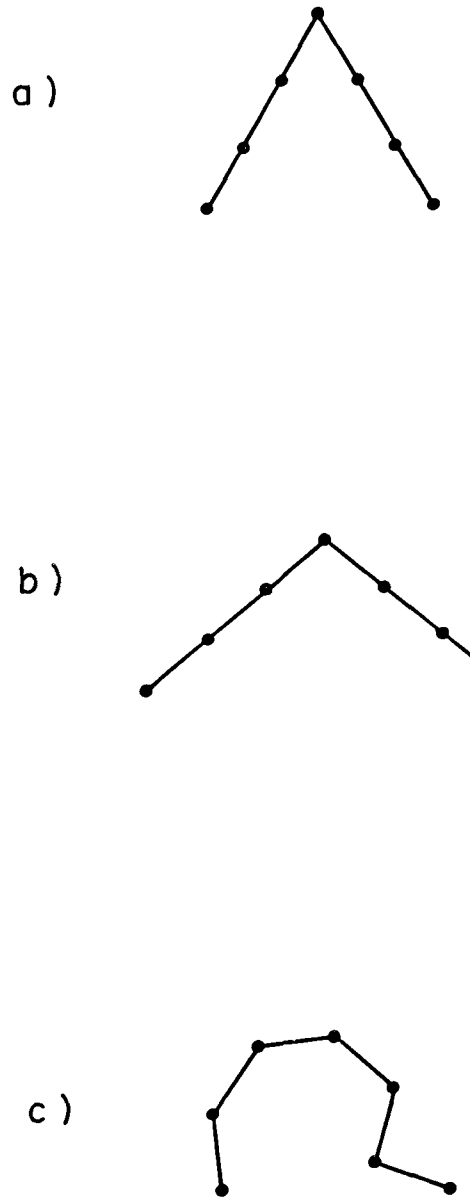


Fig. 1. The RMS distance between (**a**) and (**b**) is larger than the distance between (**a**) and (**c**). But, the shape of (**a**) is more similar to (**b**) than to (**c**).

any two fragments, whose lengths L range from 6 to 38. The results for the average distance and the standard deviation are shown in Table II. The L dependence of the distance in this range is linear and can be expressed by the formula:

$$RMS = 0.21L + 2$$

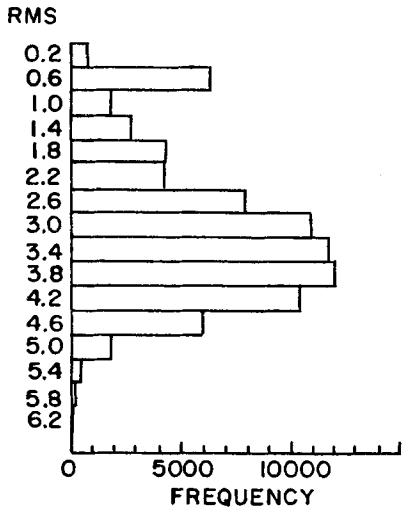
This linear formula fits the data with a standard deviation of 0.2. It differs from the one given by Remington and Matthews,[11] which is $RMS = 1.55\sqrt{L}$. Remington and Matthews' estimation was interpolated for the wide range of 10–130 residues, which may possibly explain the difference in the formulae. For the range of the lengths we considered, this difference is not significant relative to the standard deviations.

Figure 2 presents a histogram of distances between all pairs of hexamers in the four sample proteins. The bimodal distribution shows a fairly sharp division between similar and dissimilar pairs of hexamers. The separation between the two peaks occurs at about 1 Å RMS distance. Again, repeating the experiment for different proteins yields similar results. The existence of so many pairs of similar hexamers is the first indication that we may expect a pronounced clustering of their conformations.

SELECTING THE BUILDING BLOCKS BY A CLUSTERING ALGORITHM

Our first question was whether conformations of oligomers of a given size vary continuously or can be separated into a resonable number of distinct clusters. Throughout this analysis we used hexamers, fragments of length 6, which appear to be long enough to carry structural meaning. The detailed reasons for the selection of this length will be elaborated in the discussion. In our approach, a cluster is defined as a set of structures with the property that the RMS deviation between members, or alternatively, from some typical member, is less than some fixed value. Since Figure 2 indicates that 1 Å seems to be the separation point between similar and nonsimilar hexamers, 1 Å was selected as the threshold value for the clustering process.

Initially, we used the same four sample proteins as above. Each protein was divided into overlapping hexamers; thus, for a protein of length N there were N–5 hexamers, and for these four proteins a total of 406 hexamers. The RMS distance between each of the 82,215 pairs of hexamers was calculated (this number is simply $[K(K-1)]/2$ for $K=406$). The clustering procedure consisted of two stages. In the first a variant of the K-nearest neighbor clustering algorithm[12] was used (a schematic illustration of the process, in the two dimensional case, is shown in Fig. 3): A hexamer is selected to be the first member in the first cluster, and all other hexamers closer to this first member than the 1 Å threshold value are assigned to the same cluster. Each member of the cluster then serves as a new source to add all of its sufficiently close neighbors to the cluster. This "annexation" process is repeated until no further hexamers can be added to the cluster. A new hexamer is then selected as the first member of the next cluster, which is constructed in the same way. The procedure is terminated when all the hexamers have been assigned to clusters. This part of the algorithm is deterministic, i.e., the assignment to clusters is independent of the order in which the hexamers are used, and it has a run-time complexity that is quadratic in the number of elements to be clustered.



RMS	FREQ.	CUM. FREQ.	PERCENT	CUM. PERCENT
0.2	805	805	0.98	1
0.6	6498	7303	7.90	9
1.0	1932	9235	2.35	11
1.4	2824	12059	3.43	15
1.8	4405	16464	5.36	20
2.2	4377	20841	5.32	25
2.6	8029	28870	9.76	35
3.0	10912	39782	13.27	48
3.4	11722	51504	14.26	63
3.8	12065	63569	14.67	77
4.2	10387	73956	12.63	90
4.6	5926	79882	7.21	97
5.0	1872	81754	2.28	99
5.4	373	82127	0.45	100
5.8	96	82223	0.12	100
6.2	2	82225	0.00	100

Fig. 2. The histogram of the distribution of distances between each pair of hexamers extracted from the proteins: 4HHBb, 5PTI, 1BP2, and 1PCY. The bimodal distribution can be interpreted to imply that hexamer pairs can be divided into two categories: related pairs with RMS distance below 1 Å and unrelated pairs with distance normally distributed about an average around 3.6 Å.

TABLE II. Random RMS Distance Between Oligomers*

Oligomer length	No. of pairs	Average RMS	Standard deviation
6	82215	2.8266	1.2056
8	79003	3.4654	1.2294
10	75855	4.0251	1.2761
12	72771	4.5490	1.3428
14	69751	5.0726	1.4327
16	66795	5.5642	1.5195
18	63903	6.0315	1.6078
20	61075	6.4722	1.6860
22	58311	6.8841	1.7482
24	55611	7.2645	1.7894
26	52975	7.6312	1.8241
28	50403	7.9955	1.8642
30	47895	8.3634	1.9142
32	45451	8.7279	1.9682
34	43071	9.0802	2.0128
36	40755	9.4120	2.0392
38	38503	9.7220	2.0504

*The random RMS distance between oligomers, of length ranging from 6 to 38, is calculated by averaging the distances between all such pairs of oligomers extracted from four proteins: 4HHBb, 5PTI, 1BP2, and 1PCY.

In the process described above, the *diameter* of a cluster (i.e., the greatest distance between any pair of members) can grow significantly. Consequently, in the second stage of the procedure we applied a second algorithm to obtain a finer subclustering assignment. We wanted to divide each cluster into subclusters, each of which contains a member whose distance from any other member is not more than a threshold value, and again we used 1 Å. This member is called the *center* of the cluster. Optimal subclustering, in the sense that the number of subclusters is minimal, is a hard problem that belongs to a class of problems in computer science, called the NP-complete problems, for which no efficient (polynomial-time) algorithm is known. Thus we apply the following heuristic procedure: For each cluster, the member having the maximum number of neighbors is chosen as the center of a new subcluster containing those neighbors as members. The process is repeated for the other unassigned hexamers until all the hexamers of the cluster are assigned to subclusters. In this way the first stage of the clustering process resolved the major classes of hexamer structures, while the second stage located representative hexamers within these classes. Other clustering algorithms, when provided with parameters such as the density, number of members, etc., might yield a similar set of clusters in a single step, but this would not provide the distinction between major and subclusters.

Applying this algorithm to our hexamers yielded the following result: the 406 hexamers were first clustered into 55 distinct clusters. The fact that our annexation algorithm did not produce a small number of large clusters showed that indeed the structure of hexamers does not vary gradually and they are readily separable into distinct clusters. Those clusters were then further subdivided as described above to give a total of 103 subclusters. The central hexamer of each subcluster was selected to be a standard *building block*.

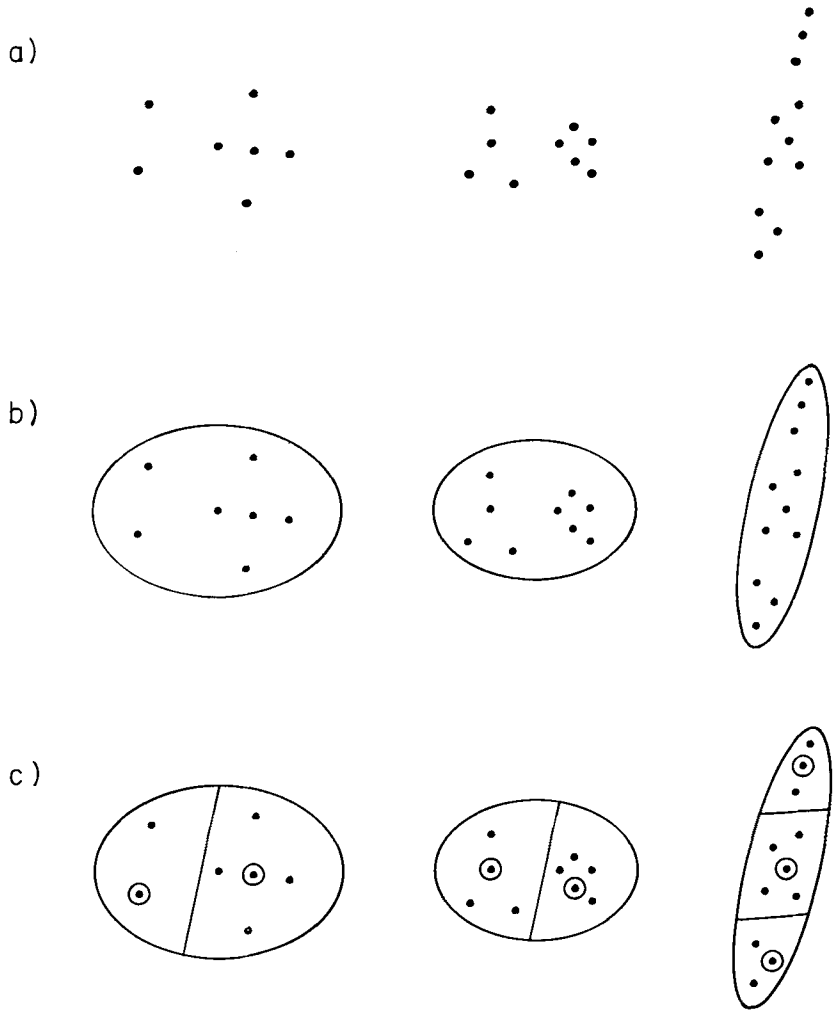


Fig. 3. The two stages of the clustering process (in a two dimensional example). **a**: The data elements. **b**: The elements are separated into clusters such that each element has at least one close neighbor (in distance less than some threshold) in the cluster. **c**: The clusters are further divided into subclusters, such that for each subcluster there exist a central member whose distance to any other member is less than some threshold. In the real analysis this central member was chosen as a building block.

In order to estimate how well these building blocks can represent any hexamer found in proteins, we tested each of the 12,973 hexamers in our refined Brookhaven data base. We found that 76% of them had a distance of less than 1 Å from at least one of the standard building blocks, and 92% had a distance of less than 1.25 Å from one of them. The average distance between a hexamer and its closest building block (including the 24% of the hexamers whose closest building block was at a distance greater than 1 Å) was 0.74 Å.

The building blocks seem to be distributed well in the conformational domain, as most of the hexamers can be represented only by one or two building blocks. Of the hexamers that can be replaced with an error of less than 1 Å, 65% could fit only to a single building block, and 95% could fit to no more than two.

Since the hexamers overlap, each residue is part of six consecutive hexamers (fewer for the terminal residues). We counted the residues that belong to at least one hexamer that is within 1 Å of one of our building blocks, thus trying to "tile" the proteins with overlapping building blocks, all with distance less than 1 Å from the corresponding hexamers. This check showed that we were able to cover about 99% of residues.

The fact that the standard building blocks can fulfill this job although they were extracted from only four proteins shows that they are really common and standard. Since the clustering algorithm is sensitive to each individual hexamer, it is clear that using too many proteins as a basis, besides being computationally time consuming, might blur the results. A single hexamer, one that is either very rare or not determined exactly, may link between two or more

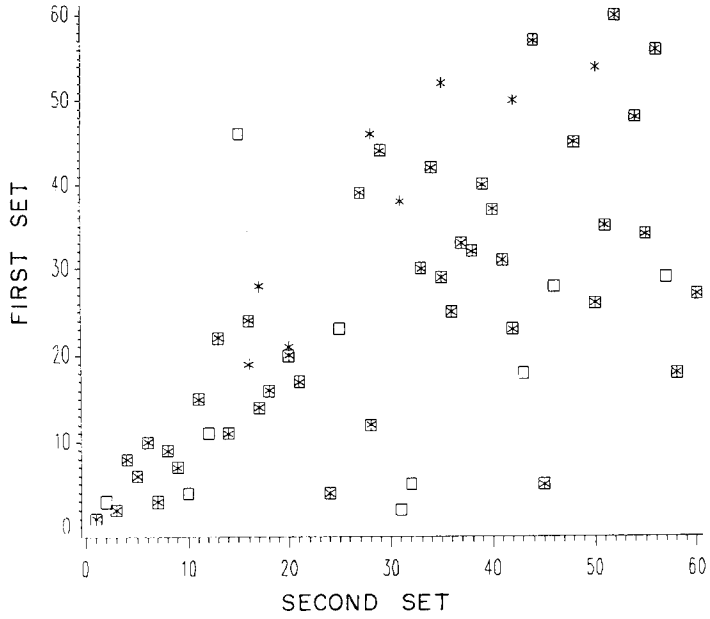


Fig. 4. The 60 most popular building blocks from the first set of proteins (4HHBb, 5PTI, 1BP2, 1PCY) were matched with the 60 most popular ones from the second set (4CYT, 2SGA, 1LZ1, 4FXN). For each building block from the original set (on the vertical axis), a star indicates its closest match of distance less than 1 Å from the second set (on the horizontal axis). Similarly, for each building block in the second set, a square indicates its closest match in the original set. Thus, a star inside a square indicates a one-to-one mapping between the two building blocks; 100 matches of distance less the 1 Å were found, and 82 in one-to-one matches. The matches tend to concentrate along the main diagonal of the plot.

really distinct clusters in the first stage of the clustering process.

The next stage in our work was to show that our analysis is not overly dependent on the initial set of proteins that we used. We repeated the experiment using a different set of four proteins, extracted a new set of building blocks, and then examined the correspondence between the two sets. We used the following proteins: 4CYT (cytochrome *c*, albacore tuna heart), 2SGA (proteinase A, *Streptomyces griseus*, strain k1), 1LZ1 (lysozyme, human), and 4FXN (flavodoxin, *Clostridium* MP), with a total length of 552 amino acids (532 hexamers). These proteins yielded 67 distinct clusters, which were further resolved into 144 subclusters. These numbers are slightly higher than the 55 and 103 found previously. We then determined the similarity between the two sets of building blocks. We took the 60 most frequently occurring building blocks from each set and ordered them by the number of occurrences. (An occurrence is a hexamer in the refined data base that selected the building block in question as its closest neighbor with a distance of less than 1 Å.) We then mapped each of the two sets of 60 building blocks onto the other, finding for each building block its closest neighbor in the other set. Again, we only considered matches with a distance of less than 1 Å. Of the 60 building blocks in each set, 41 had one-to-one mappings with building blocks from the other set, i.e., we had 41 pairs of original and new building blocks that selected each other as their closest match. The mapping is shown in Figure 4. There is a clear tendency of the matching pairs to concentrate near the main diagonal of the plot. Since the two sets were sorted by the number of occurrences of the building blocks, the tendency toward the diagonal indicates that the these two different sets constitute good representatives of virtually the same clusters. As a control, we compared our original set to a random set of 60 hexamer conformations chosen from the second sample of four proteins. We mapped the original set of building blocks onto this random set and found that only 75 pairs out of 120 had a distance of less than 1 Å, compared with 100 pairs in the previous case. In addition, only 14 elements from the first set had a one-to-one mapping with a random element, and no diagonal tendency was seen. This random matching is shown in Figure 5.

We consider the number of one-to-one matches as an important indication of the level of similarity between two sets of building blocks. A large number of simple matches from either set to the other may just reflect the fact that many building blocks were chosen from highly occupied subdomains of the conformational space, but the sets of building blocks may not be sufficiently representative and some less occupied, but important, domains are left unsampled. On the other hand, a high number of one-to-one matches indicates that the two sets of building blocks sample the whole conformational space in a similar

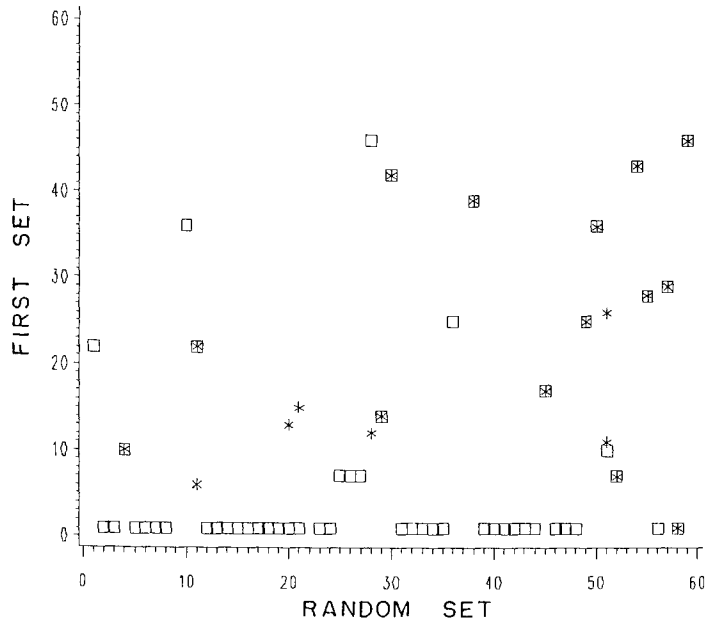


Fig. 5. The original set of building blocks (vertical axis) was matched to a random set of 60 hexamers (horizontal axis). Only 75 matches of a distance less than 1 Å were found. Fourteen one-to-one mappings can be seen, and the tendency to concentrate in specific rows and columns shows that the random set of hexamers does not have the same expressive power as the real building blocks.

way, and that highly occupied subdomains are not unnecessarily overrepresented. Thus, the fact that the two real sets have 41 one-to-one matches, while 14 one-to-one matches were found with the random set, shows the significant improvement achieved by our automatic clustering procedure.

Another experiment was to derive the building blocks from an extended set of proteins. We combined the two initial sets of four proteins each to get a set of eight proteins with a total of 958 hexamers. Again, we used the two-stage clustering process to divide this large set of hexamers into subclusters. As suggested above, with a large number of proteins the first stage of the clustering process is likely to produce huge clusters. Indeed, after the first stage we got one very big cluster of 846 members plus 53 very small clusters. In the second stage of the clustering process, dividing the clusters into subclusters with radii of 1 Å, 170 subclusters emerged that very much corresponded to our original set. This correspondence was measured again by matching the 60 most popular building blocks from this new set against the original set. This time, 44 elements of the original set had one-to-one mapping with elements of the new set, indicating their high similarity.

As mentioned above, 76% of the 12,973 hexamers in the refined Brookhaven data base fit one of the building blocks with a distance of less than 1 Å. As this figure applies to the total data base, it was interesting to find if there was a big difference in the way different proteins could be matched with the standard building blocks (see Fig. 6). Some proteins have almost all of the hexamers matched with the library of building blocks. For example, 92% of the hexamers of 1ECA (erythrocruorin) and 93% of the hexamers of 1CTF (ribosomal protein L12/L7) have RMS distance less than 1 Å to one of the building blocks. On the other hand, 2 out of the 82 proteins performed very poorly in this test. For 2FD1 (azotobacter ferredoxin), *only 14%* of its hexamers could fit our building blocks. For 2ABXa (α-bungarotoxin, chain a), only 26% of the hexamers were close enough to one of the building blocks. We suggest that such a surprisingly low rating can be explained either by an incorrectly determined structure or by a very unusual structure. As for 2FD1, a recent redetermination by Stout et al.[13] showed the original structure to be incorrect. The 2ABXa case still has to be checked. Both structures have many ϕ and ψ angles different from the allowed values as determined by Ramakrishnan and Ramachandran.[14]

ANALYZING THE BUILDING BLOCKS

Analyzing the structure of these building blocks shows their natural relationship to the well known secondary structure elements. By using Kabsch and Sander's DSSP program,[15] we obtained the secondary structure assignment for each building block in the context of the protein in which it was found. (Two peptides, 1GAPa and 1CPP, have only C coordinates available, and thus no secondary structure assignment.) Many of them fall into the categories of helices, sheets, and turns, but naturally with a fine resolution within these categories. In addition, some

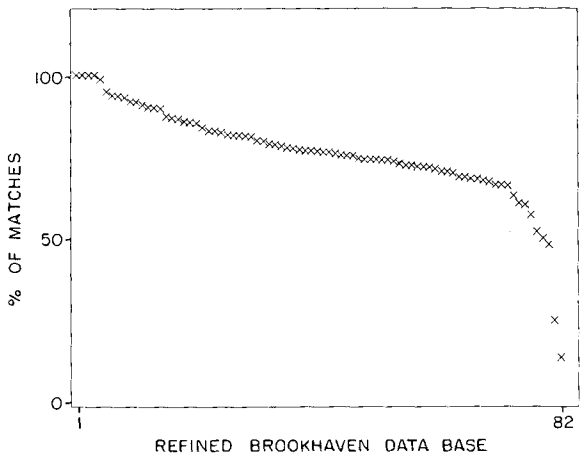


Fig. 6. For each protein in the refined Brookhaven data base (horizontal axis) the percentage of its hexamers that can be represented with distance less than 1 Å by the set of 103 standard building blocks is shown. The proteins are ordered (left to right) according to their match with the set of building blocks. The low matching of 2ABXa (26%) and 2FD1 (14%) is clearly outstanding.

building blocks show the structural ways in which these secondary-structure elements are connected. A list of the 30 most frequently occurring building blocks is shown in Table III. (The clusters are designated as decimal numbers, the cluster number to the left of the decimal point and the refined subcluster number to its right. The numbers themselves were determined arbitrarily, according to the order in which they were found by the clustering algorithm program.) A few examples are:

1. Building block No. 1 (Cluster 4.01), sequence ICFSKV starting at residue 104 of 1BP2, which is by far the most frequent building block. This building block was found to represent many helices found in protein structure. Surprisingly, its own secondary structure assignment is not to a perfect helix but rather to a region where a helix ends in a turn (Fig. 7a).

2. Building block No. 2. (Cluster 1.01), sequence NEITCS starting at residue 80 or 1BP2, which is a typical β-sheet (Fig. 7b).

3. Building block No. 19. (Cluster 17.01), sequence CSSENN starting at residue 84 of 1BP2 is a turn that connects two adjacent extended β-strands (Fig. 7c).

Each building block is associated with a sequence distribution matrix. This matrix was obtained simply by counting, for all of the hexamers in the cluster represented by this building block, the distribution of amino acids at each one of the six positions along the hexamers. These matrices can be normalized against the distribution of the different amino acids in the data base. In a similar way, we obtained the matrices of the distribution of the secondary structure elements (as assigned by the DSSP program). These secondary structure assignment matrices enable one to follow, position by position, the specific relationship between the building blocks and secondary structure elements. Many of the building blocks show sequence distribution matrices that deviate considerably from a normal distribution, indicating a possible linkage between sequence and structure. Of special interest are the building blocks in which a large variance was found in the frequency of occurrence of the same amino acid along the different positions. For example, the building block from cluster 17.01 has a very strong preference for glycines in the fourth position while glycines are avoided in other positions. Indeed, the fourth position is assigned to turns that connect secondary structure elements. Tables IV, V, and VI show, for the three building blocks described above, their distribution matrices (with and without normalization). A study on the properties of these building blocks, from structural and energetic points of view, is in progress.

As expected, many of the building blocks closely represent well-known secondary-structure elements. These building blocks emerged from the process *without* any assumptions of secondary structure. In addition, our clustering algorithm recognizes some structural motifs that standard techniques, as represented by the DSSP algorithm, do not find. Two examples are shown in Figures 8 and 9. In each figure, we show a similar structural motif that repeats in four different proteins. The four hexamers have no obvious sequence similarity and their DSSP secondary structure assignments are quite different. Nevertheless, by our process, these hexamers were assigned to the same cluster and, indeed, the figures show how similar they are. In Figure 8 we show four out of 83 hexamers from cluster 34.01:

1. Residues 7–12 of 1PCY (oxidized poplar plastocyanin) with sequence ADDGSL and DSSP secondary-structure assignment -TT--S (The DSSP symbols are given in the legend of Table III). This hexamer is used as the building block representing this cluster.

2. Residues 25–30 of 2OVO (silver pheasant ovomucoid, third domain) with sequence GSDNKT and DSSP assignment ETTS-E. This hexamer has a distance of 0.47 Å from its building block.

3. Residues 98–103 of 1LH1 (leghemoglobin) with sequence VSKGVA and DSSP assignment HHTT--. Its distance from its building block is 0.45 Å.

4. Residues 17–22 of 1CRN (crambin) with sequence RLPGTP and DSSP assignment HTTT--. Its distance from its building block is 0.38 Å.

Figure 9 shows four of the 34 hexamers in cluster 4.11:

1. Residues 116–121 of 4HHBb (human deoxyhe-

TABLE III. The 30 Most Popular Building Blocks*

No.	Cluster	Protein	Residue	Sequence	Secondary structure	No. of occurrences
1	4.01	1BP2	104	ICFSKV	HHHHTS	2908
2	1.01	1BP2	80	NEITCS	TEEEE-	901
3	1.02	5PTI	31	QTFVYG	EEEEE-	403
4	3.01	1BP2	57	KLDSCK	T-HHHH	319
5	1.05	1BP2	108	KVPYNK	TS---G	286
6	1.08	1BP2	74	SYSCSN	-EEEET	283
7	5.01	4HHBb	54	VMGNPK	HHH-HH	226
8	2.02	4HHBb	2	HLTPEE	---HHH	194
9	2.01	1PCY	40	VFDEDS	EE-TTS	151
10	1.03	4HHBb	1	VHLTPE	----HH	148
11	4.02	4HHBb	82	KGTFAT	HHHHHH	145
12	1.04	4HHBb	16	GKVNVD	TT--HH	135
13	1.11	5PTI	30	CQTFVY	EEEEEE	135
14	4.03	1PCY	85	SPHQGA	GGGTTT	132
15	4.05	1BP2	18	PLLDFN	HHHHTT	132
16	4.04	1PCY	86	PHQGAG	GGTTTT	119
17	37.01	1PCY	10	GSLAFV	--S-EE	116
18	1.12	5PTI	27	AGLCQT	TTEEEE	100
19	17.01	1BP2	84	CSSENN	E-TT--	99
20	26.01	1BP2	61	CKVLVD	HHHTT-	95
21	1.06	5PTI	5	CLEPPY	GGS---	91
22	6.01	4HHBb	32	LVVYPW	HHHSGG	88
23	22.01	1PCY	72	VALSNK	EE--S-	88
24	2.03	1PCY	2	DVLLGA	EEEES-	88
25	1.07	1PCY	66	KGETFE	TT-EEE	85
26	21.01	1BP2	30	GLGGSG	SS---S	85
27	34.01	1PCY	7	ADDGSL	-TT--S	83
28	12.02	1BP2	56	KKLDSC	TT-HHH	73
29	1.10	4HHBb	95	KLHVDP	TT---T	72
30	5.03	1PCY	53	SKISMS	HHHS--	68

*The cluster column gives the identification of each building block in terms of cluster (left of the decimal point) and refined subcluster (right of the decimal point). The protein and residue columns give the location of the first residue of the building block in the data base. The sequence column shows the sequence associated with the building block. The secondary structure column shows the assignment of that sequence to secondary structure elements by the program DSSP.[15] The secondary structure elements are H, α-helix; B, residue in isolated β-ladder; E, extended strand; G, 3_{10}-helix; T, H-bonded turn; S, bend; "-", unassigned residue. The last column shows the number of members of each cluster, i.e., the number of occurrences of hexamers that have a distance of less than 1 Å from the building block.

moglobin, β-chain), with sequence HHFGKE and DSSP secondary structure assignment HHHTTT. This hexamer is used as the building block representing this cluster.

2. Residues 170–175 of 3PTP (bovine trypsin) with sequence SAYPGQ and DSSP assignment HHSTTT. This hexamer has a distance of 0.50 Å from its building block.

3. Residues 14–19 of 1PCY (oxidized poplar plastocyanin) with sequence FVPSEF and DSSP assignment EESSEE. Its distance from its building block is 0.79 Å.

4. Residues 93–98 of 1CTF (ribosomal protein L12/L7) with sequence ALKEGV and DSSP assignment EEEEEE. Its distance from its building block is 0.97 Å.

Many of the building blocks reflect the structural ways in which standard secondary-structure elements tend to be connected to form the full three-dimensional structure of a protein. We have identified structural motifs by which, for example, a helix ends a turn, two helices are connected, etc. These suggested structural motifs have some sequence preference and should be analyzed carefully, from geometric and energetic aspects, to confirm their stability and importance.

RECONSTRUCTING PROTEINS BY STANDARD BUILDING BLOCKS

Having established a library of standard structural building blocks (i.e., the set of 103 building blocks described above) we then attempted to use it to reconstruct protein structures. We used the following procedure: First, we replaced each original hexamer of the protein by its closest (in terms of RMS distance) standard building block. Then, since the building blocks overlap, we could align every two consecutive building blocks by using the BMF algorithm. Onto the suffix (the last five residues) of the first building block we fitted the prefix (the first five residues) of the next building block. Thus, the 3D position of the last residue of the latter hexamer was determined and added to the growing chain. This process was repeated until the whole protein

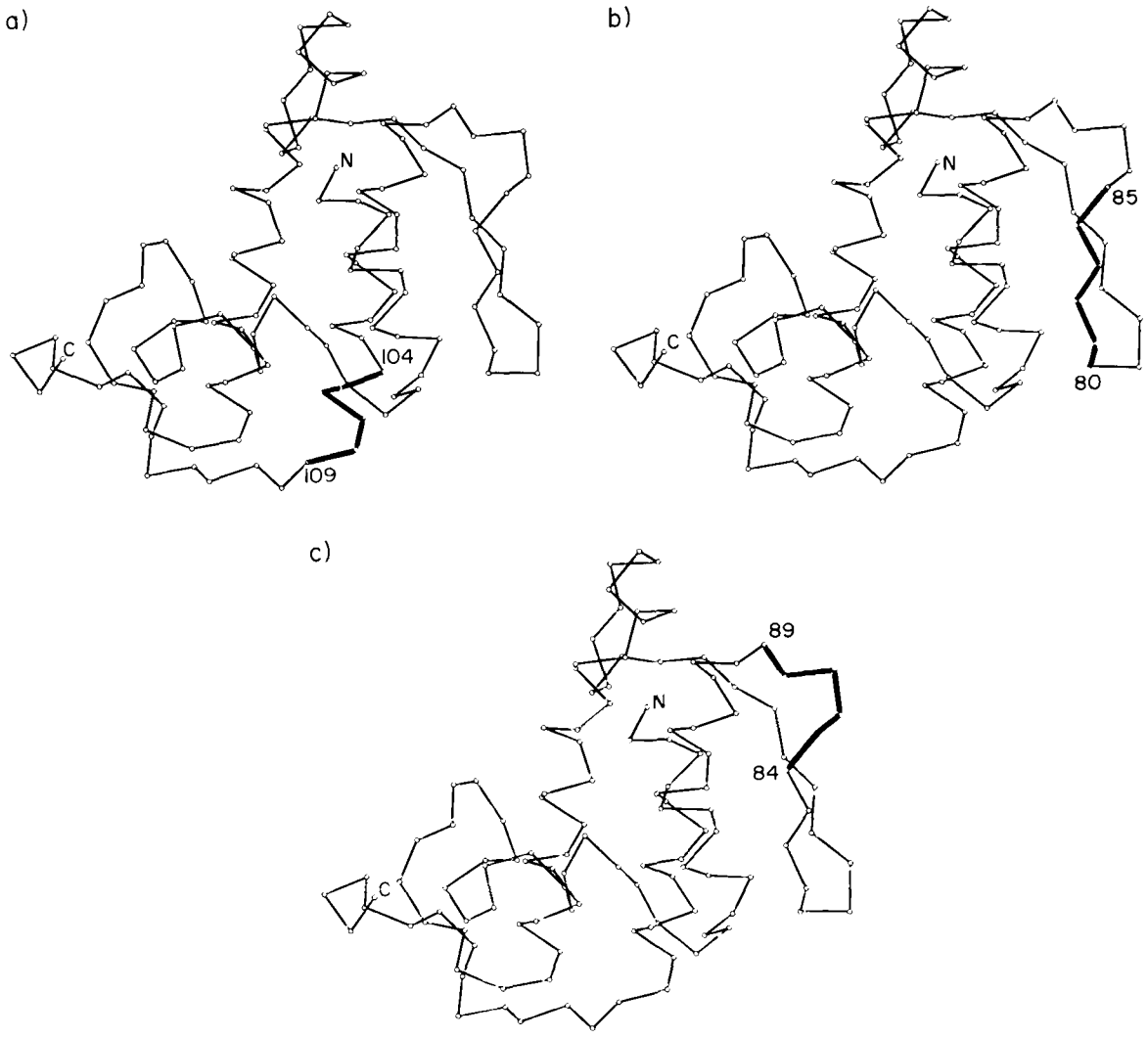


Fig. 7. Three building blocks (taken from 1BP2) are shown. **a**: The most typical helical motif (starting at residue 104 of 1BP2). The amino acid sequence of this hexamer is "ICFSKV." **b**: The most typical extended strand (starting at residue 80 of 1BP2). The amino acid sequence of this hexamer is "NEITCS." **c**: A building block that is used to connect secondary structure elements (starting at residue 84 of 1BP2). The amino acid sequence of this hexamer is "CSSENN." In 1BP2 it is used to connect two extended strands but it is used more often to connect helices.

was reconstructed. The process can begin at either end of the protein or at any point within the structure and proceed to both ends.

Evaluating the performance of this reconstruction procedure is not a trivial task. Only the first 60 residues of each protein were used in order to have a standard (not too long) protein length on which to test our approach. From the refined Brookhaven data base, we took all of the proteins (71) of length greater than 60 and used only their first 60 residues. One of the simplest ways to measure the similarity between the original and reconstructed structure is to calculate the RMS distance between them. These distances were compared to the average "random" distance between structures, which we took to be the average distance measured between any pair of truncated proteins in our library. Thus, we had 2485 pairs (from the 71 proteins of length 60) with average distance of 12.85 Å and SD of 2.12 Å. This measurement nicely fits Remington and Matthews' estimation, according to the formula given above, of 12 Å. (Our linear formula is valid only in the range of 6–38 residues, but extrapolating it to length 60 still gives a reasonable estimation of 14.6 Å.) The average distance between the original proteins and the reconstructed ones is 7.3 Å, which is 2.6 SD lower than the random average. Of the proteins 28% had been reconstructed with an RMS of less than 5 Å, but 25% had RMS distances greater than 9 Å. Re construction of the protein backward, from C to N terminals, gave roughly the same overall performance. It is interesting to note that the four original

TABLE IV. The Sequence and Secondary-Structure Distribution in Cluster 4.01*

TABLE IVa

Amino acid	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6	Standard deviation
A	11.62	12.90	12.59	13.17	12.76	12.45	0.4853
C	1.93	2.24	2.13	2.20	2.17	2.03	0.1065
D	7.29	5.81	5.33	4.61	3.61	4.13	1.2082
E	6.88	7.32	7.05	6.46	4.95	5.02	0.9508
F	4.23	4.71	5.09	4.99	4.85	4.37	0.3129
G	5.40	4.26	4.30	4.06	3.78	6.40	0.9102
H	2.06	2.13	2.48	2.41	2.92	2.85	0.3256
I	4.95	5.57	5.61	5.57	5.81	5.09	0.3056
K	6.22	6.57	7.60	8.05	8.60	8.84	0.9722
L	8.80	10.45	11.14	12.35	12.59	10.63	1.2655
M	1.65	2.06	2.20	2.37	2.48	2.27	0.2667
N	4.44	3.16	3.30	3.37	3.61	4.40	0.5158
P	3.61	3.16	1.31	0.79	0.79	1.58	1.1130
Q	4.06	4.26	4.61	4.68	4.57	4.54	0.2191
R	3.65	3.68	3.58	3.78	4.06	3.54	0.1719
S	6.40	4.57	4.81	4.47	5.33	6.02	0.7261
T	5.43	4.57	4.68	4.40	4.71	4.71	0.3235
V	6.98	7.63	7.53	7.50	7.60	6.77	0.3326
W	1.65	2.10	2.03	1.86	1.72	1.27	0.2728
Y	2.75	2.82	2.65	2.92	3.09	3.09	0.1673

TABLE IVb

Amino acid	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6
A	11.357	15.978	14.854	16.977	15.479	14.355
C	−0.615	−0.290	−0.398	−0.326	−0.362	−0.507
D	4.120	0.425	−0.778	−2.583	−5.075	−3.786
E	5.253	6.460	5.717	4.139	0.055	0.240
F	1.474	3.004	4.206	3.878	3.441	1.911
G	−7.486	−9.934	−9.860	−10.379	−10.973	−5.335
H	−1.911	−1.510	0.495	0.094	3.101	2.700
I	0.359	1.795	1.875	1.795	2.354	0.678
K	0.278	0.836	2.508	3.233	4.124	4.515
L	4.569	10.795	13.388	17.928	18.836	11.443
M	0.902	6.499	8.365	10.697	12.096	9.297
N	−0.415	−2.169	−1.980	−1.885	−1.553	−0.463
P	−3.224	−4.547	−10.044	−11.571	−11.571	−9.230
Q	0.834	1.510	2.637	2.862	2.524	2.411
R	0.917	1.031	0.688	1.375	2.291	0.573
S	−6.538	−15.650	−14.447	−16.166	−11.868	−8.429
T	−1.835	−3.512	−3.311	−3.848	−3.244	−3.244
V	−0.360	0.343	0.232	0.195	0.306	−0.582
W	0.932	4.942	4.325	2.783	1.549	−2.462
Y	−7.936	−7.425	−8.703	−6.658	−5.379	−5.379

TABLE IVc

Secondary structure	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6
B	0.21	0.03	0.00	0.00	0.03	0.17
E	0.62	0.00	0.00	0.03	0.14	0.34
G	1.44	1.75	1.99	2.06	1.82	1.93
H	78.54	88.79	88.93	87.35	81.98	72.70
S	2.37	0.76	0.17	0.28	1.24	2.96
T	1.96	1.72	1.96	3.27	6.36	10.63
—	7.91	0.00	0.00	0.07	1.48	4.33

*Table IVa: The amino acid distribution (in percentage) along the 2908 hexamers that are members of cluster 4.01 (see Table III and Fig. 7a). The standard deviation column shows the variance of this value along the different positions of each amino acid. Table IVb: The normalized amino acid distribution. The normalization was carried out relative to the frequency of the different amino acids in the refined Brookhaven data base. The values are the number of standard deviation from the averaged frequency. In this case, we can see, for example, the preference of the helical building block to use alanine and leucine. Table IVc: The distribution of secondary structure assignment along the positions of all of the hexamers that are represented by this building block. The assignment was carried out by the DSSP program[15] and the legend to the different element symbols appears in Table III. In this case, the helical nature of the building block is clearly shown here (see Fig. 7a).

TABLE V. The Sequence and Secondary-TABLE Va Structure Distribution in Cluster 1.01*

TABLE Va

Amino acid	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6	Standard deviation
A	6.44	6.99	6.55	5.22	6.66	5.99	0.5710
C	1.11	2.55	3.11	3.11	3.11	2.89	0.7147
D	3.66	2.89	2.00	3.00	4.33	5.77	1.2035
E	4.11	3.88	2.89	2.55	3.66	4.33	0.6421
F	3.55	4.11	5.99	6.77	4.66	3.66	1.2012
G	14.32	4.66	2.11	3.77	4.66	9.88	4.2026
H	1.66	2.66	2.00	2.33	2.66	2.89	0.4234
I	5.66	7.55	8.77	9.32	4.11	4.77	1.9747
K	4.55	5.77	4.88	4.00	5.11	5.22	0.5537
L	5.66	6.99	9.10	8.88	7.21	8.21	1.1909
M	2.33	1.78	1.55	1.22	0.67	1.22	0.5183
N	3.77	2.55	2.22	3.22	4.33	4.33	0.8173
P	4.77	3.44	3.33	1.33	7.10	4.66	1.7549
Q	3.55	4.33	3.33	2.22	3.00	2.77	0.6579
R	3.33	3.66	3.22	4.33	4.11	3.33	0.4202
S	7.88	6.77	7.44	5.11	8.55	8.55	1.1915
T	8.10	10.88	8.66	6.10	9.43	6.55	1.6320
V	8.77	11.21	12.99	16.54	10.43	9.66	2.5694
W	1.22	1.66	3.11	1.89	2.11	1.11	0.6618
Y	5.55	5.66	6.77	9.10	4.11	4.11	1.7134

TABLE Vb

Amino acid	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6
A	−7.476	−5.461	−7.073	−11.910	−6.670	−9.088
C	−1.473	0.044	0.628	0.628	0.628	0.394
D	−4.945	−6.887	−9.106	−6.610	−3.281	0.325
E	−2.227	−2.827	−5.523	−6.422	−3.426	−1.628
F	−0.681	1.082	7.079	9.548	2.846	−0.329
G	11.752	−9.077	−14.584	−10.992	−9.077	2.175
H	−4.234	1.589	−2.293	−0.352	1.589	2.884
I	2.003	6.381	9.214	10.502	−1.602	−0.057
K	−2.434	−0.456	−1.895	−3.334	−1.535	−1.355
L	−7.284	−2.261	5.692	4.855	−1.424	2.343
M	10.127	2.600	−0.411	−4.927	−12.453	−4.927
N	−1.329	−3.012	−3.471	−2.094	−0.564	−0.564
P	0.215	−3.728	−4.056	−9.970	7.114	−0.114
Q	−0.825	1.721	−1.552	−5.189	−2.643	−3.371
R	−0.134	0.975	−0.504	3.193	2.454	−0.134
S	0.882	−4.667	−1.338	−12.991	4.211	4.211
T	3.372	8.786	4.455	−0.526	5.971	0.340
V	1.563	4.190	6.100	9.921	3.354	2.518
W	−2.923	1.059	14.003	3.051	5.042	−3.919
Y	12.873	13.698	21.951	39.283	2.144	2.144

TABLE Vc

Secondary structure	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6
B	1.11	1.66	2.22	1.78	1.44	1.66
E	46.06	64.37	76.69	77.14	64.59	47.28
G	0.44	0.11	0.00	0.00	0.44	1.78
H	0.44	0.11	0.00	0.00	1.11	3.00
S	12.10	3.22	0.00	0.11	5.66	10.77
T	10.65	0.33	0.00	0.00	3.44	9.99
—	26.30	27.30	18.20	18.09	20.42	22.64

*Table Va: The amino acid distribution (in percentage) along the 901 hexamers that are members of cluster 1.01 (see Table III and Fig. 7b). Table Vb: The normalized amino acid distribution. Note the avoidance of glycine in the middle positions and the preference of tyrosine in this extended building block. Table Vc: The distribution of secondary structure assignment shows that this is an extended strand building block.

proteins, from which the building blocks had been extracted, were among the group of proteins that were successfully reconstructed but they did not have the best scores in this group, indicating that even a good level of hexamers matching does not guarantee a perfect reconstruction. This is due to the fact that only 103 out of 406 hexamers were chosen as building blocks, and the other hexamers were re-

TABLE VI. The Sequence and Secondary-Structure Distrubtion in Cluster 17.01*

TABLE VIa

Amino acid	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6	Standard deviation
A	7.07	6.06	15.15	5.05	12.12	7.07	3.6264
C	3.03	6.06	3.03	1.01	7.07	1.01	2.3144
D	2.02	22.22	5.05	10.10	5.05	6.06	6.6151
E	3.03	5.05	6.06	4.04	5.05	2.02	1.3573
F	11.11	0.00	1.01	5.05	4.04	1.01	3.7644
G	4.04	3.03	8.08	16.16	2.02	1.01	5.1725
H	5.05	10.10	0.00	2.02	2.02	5.05	3.2470
I	11.11	2.02	2.02	0.00	0.00	3.03	3.7795
K	5.05	6.06	4.04	6.06	5.05	9.09	1.5882
L	6.06	1.01	4.04	1.01	6.06	3.03	2.0824
M	2.02	0.00	1.01	0.00	2.02	2.02	0.9066
N	7.07	9.09	4.04	11.11	6.06	6.06	2.2898
P	5.05	12.12	15.15	0.00	0.00	8.08	5.7041
Q	2.02	2.02	3.03	3.03	5.05	11.11	3.1764
R	2.02	0.00	2.02	7.07	4.04	4.04	2.2143
S	5.05	11.11	10.10	17.17	9.09	7.07	3.7982
T	5.05	3.03	11.11	9.09	11.11	12.12	3.3880
V	11.11	0.00	5.05	1.01	2.02	6.06	3.7531
W	1.01	0.00	0.00	0.00	5.05	3.03	1.9121
Y	2.02	1.01	0.00	1.01	7.07	2.02	2.2898

TABLE VIb

Amino acid	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6
A	−5.176	−8.844	24.172	−12.513	13.166	−5.176
C	0.546	3.733	0.546	−1.578	4.795	−1.578
D	−9.050	41.436	−1.477	11.145	−1.477	1.047
E	−5.133	0.321	3.048	−2.406	0.321	−7.860
F	23.343	−11.968	−8.758	4.082	0.872	−8.758
G	−10.417	−12.596	−1.701	15.730	−14.775	−16.954
H	15.505	44.951	−13.941	−2.162	−2.162	15.505
I	14.651	−6.444	−6.444	−11.132	−11.132	−4.100
K	−1.624	0.013	−3.261	0.013	−1.624	4.925
L	−5.775	−24.822	−13.394	−24.822	−5.775	−17.203
M	5.915	−21.485	−7.785	−21.485	5.915	5.915
N	3.217	6.002	−0.961	8.787	1.824	1.824
P	1.038	21.967	30.936	−13.912	−13.912	10.007
Q	−5.843	−5.843	−2.533	−2.533	4.087	23.947
R	−4.497	−11.227	−4.497	12.329	2.234	2.234
S	−13.266	17.036	11.985	47.337	6.935	−3.165
T	−2.582	−6.523	9.243	5.301	9.243	11.214
V	4.084	−7.870	−2.436	−6.783	−5.697	−1.350
W	−4.814	−13.876	−13.876	−13.876	31.432	13.309
Y	−13.371	−20.882	−28.393	−20.882	24.185	−13.371

TABLE VIc

Secondary structure	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6
B	7.07	7.07	0.00	0.00	5.05	4.04
E	17.17	3.03	0.00	0.00	5.05	19.19
G	2.02	0.00	2.02	2.02	2.02	1.01
H	1.01	0.00	0.00	0.00	4.04	8.08
S	17.17	10.10	22.22	15.15	16.16	8.08
T	11.11	1.01	72.73	72.73	1.01	10.10
—	41.41	75.76	0.00	7.07	62.64	46.46

*Table VIa: The amino acid distribution (in percentage) along the 99 hexamers that are members of cluster 17.01 (see Table III and Fig. 7c). Table VIb: The normalized amino acid distribution. Note, for example, the importance of glycine in the fourth position relative to avoiding glycine in any other position. Proline is avoided in the fourth and fifth positions but is preferred elsewhere. Table VIc: The distribution of secondary structure assignment shows that the third and fourth positions are turns, the second and fifth positions are unassigned, and the first and last positions begin to show the β-sheets. Thus, we can conclude that this building block is typically used to connect two extended strands.

placed by the building blocks. Hence, the overlapping concatenation of the building blocks was not perfect, resulting in some deviation of the reconstructed protein from the original. When viewed subjectively, many of the structures that have a large RMS from the original protein show some resemblance to the original protein (see Fig. 10).

Following the approach that was taken by Mat-

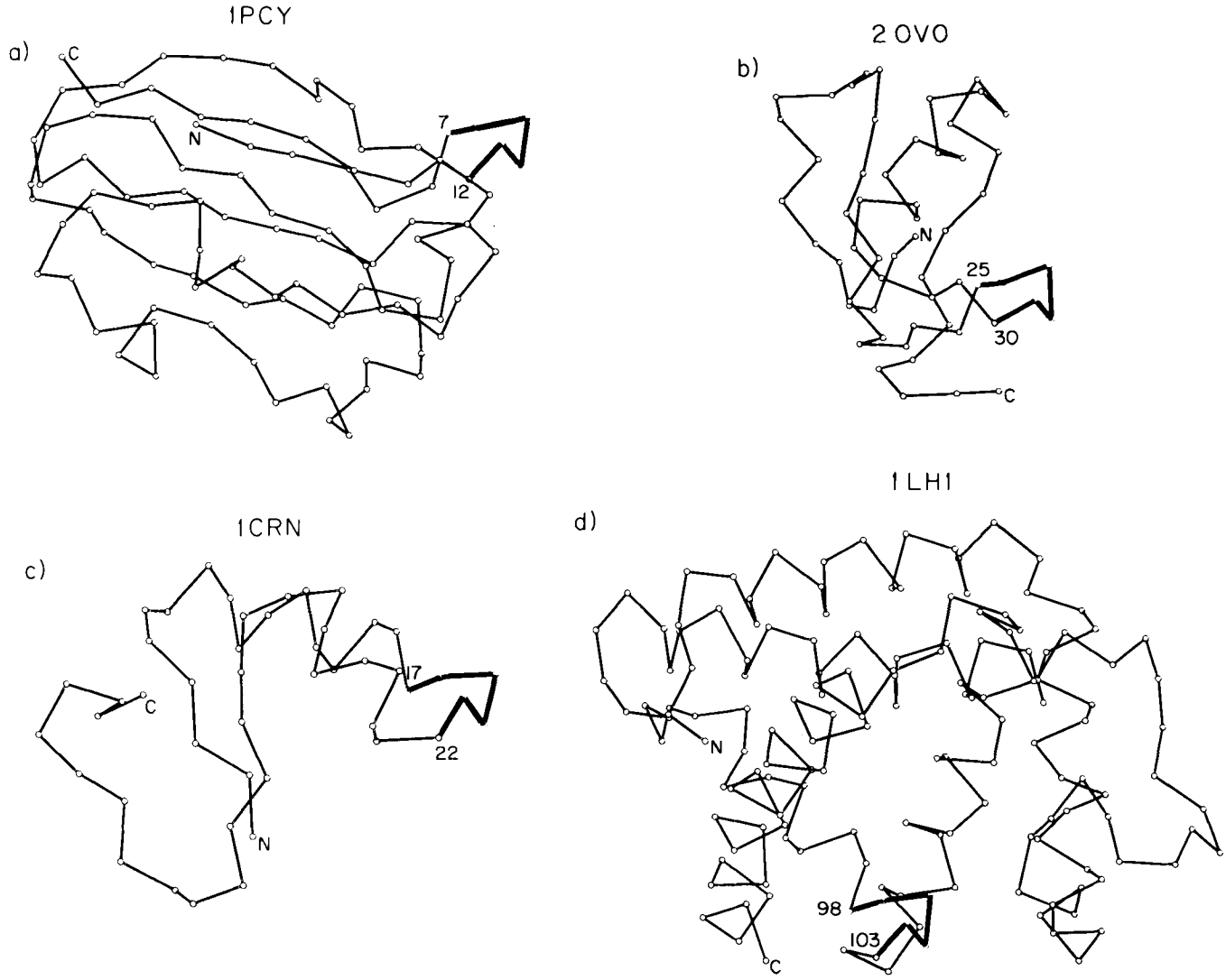


Fig. 8. Conformationally similar hexamers that occur in four different proteins. These hexamers were assigned to the same cluster although their DSSP secondary-structure assignments and their sequences are quite different. **a**: 1PCY, residues 7–12, sequence ADDGSL, DSSP assignment -TT--S. **b**: 2OVO, residues 25–30, sequence GSDNKT, DSSP assignment ETTS-E. **c**: 1CRN, residues 17–22, sequence RLPGTP, DSSP assignment HTTT--. **d**: 1LH1, residues 98–103, sequence VSKGVA, DSSP assignment HHTT--. The proteins have been oriented so that the similarity between the hexamers can be clearly seen.

thews and Rossmann,[16] we tried to capture this "qualitative" similarity of the structures by comparing them as fragments. We measured the distance between the two structures when breaks are allowed. The intuitive concept behind this approach is that the reconstruction process may have made a few "mistakes," especially at hinge points, which may result in a displacement of large parts of the structure with respect to each other. These may be corrected by introducing breaks in the structures.

For example, we tried each C_α position as a single breakpoint of the structure (of length L) into two substructures (of lengths i and L–i). Since we did not face problems of insertions and deletions here we used the same breakpoints for the two structures. The distance was calculated by choosing the breakpoint for which the larger normalized distance of the two substructures obtained by this break is minimal. Their normalized distance was accordingly calculated as:

$$\mathrm{RMS}_{1\#} = \min_i\{\max[NRMS(s_1 \ldots s_i, t_1 \ldots t_i), NRMS(s_{i+1} \ldots s_L, t_{i+1} \ldots t_L)]\}$$

where $NRMS$ (s,t) is defined as:

$$NRMS(s, t) = \frac{RMS(s, t) - E[RMS(s, t)]}{SD[RMS(s, t)]}$$

This is the RMS distance between s and t, normal-

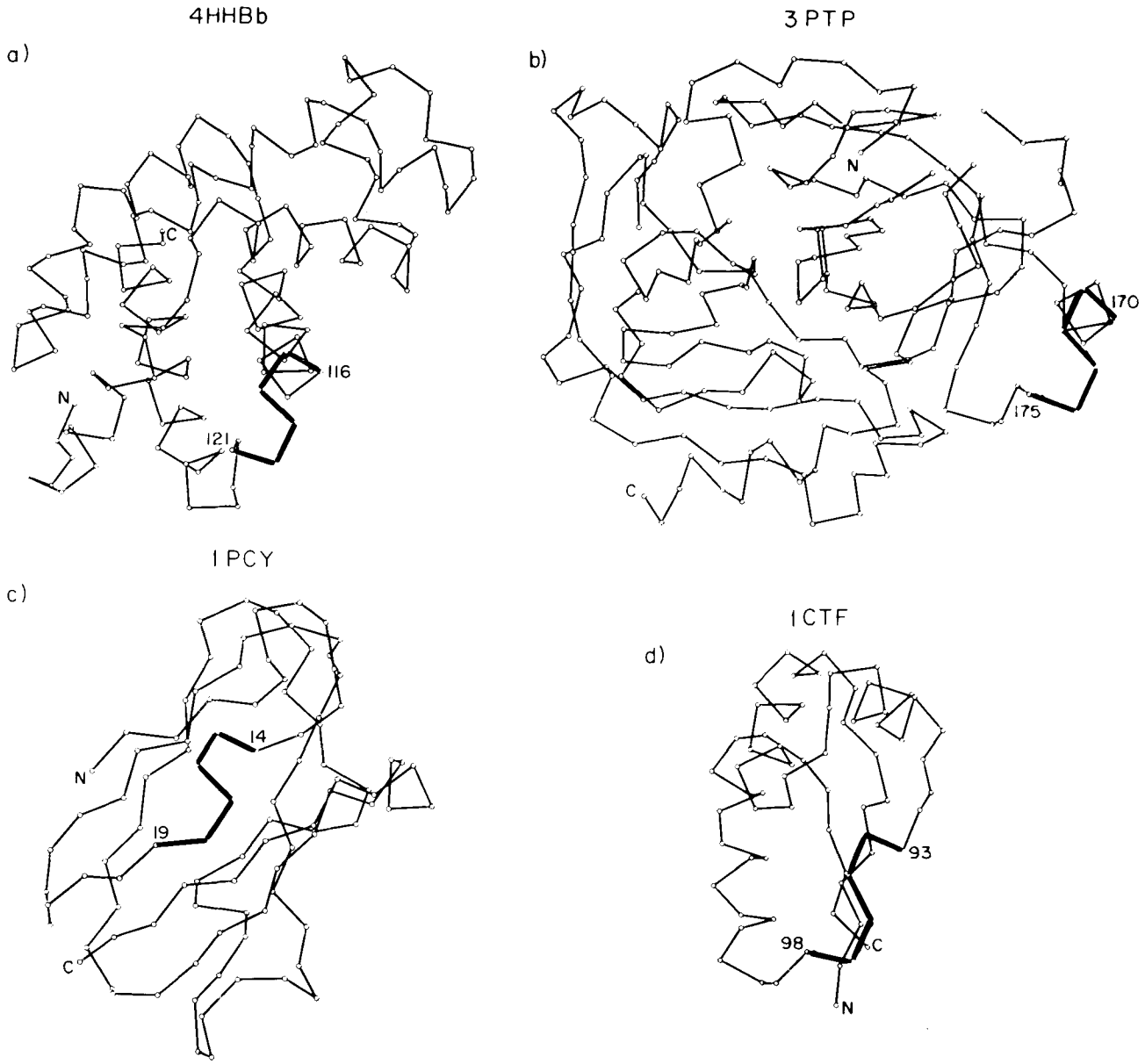


Fig. 9. Similar hexamers that repeat in different proteins. These hexamers were assigned to the same cluster, although their DSSP secondary-structure assignments and their sequences are quite different. **a**: 4HHBb residues 116–121, sequence HHFGKE, DSSP assignment HHHGGG. **b**: 3PTP, residues 170–175, sequence SAYPGQ, DSSP assignment HHSTTT. **c**: 1PCY, residues 14–19, sequence FVPSEF, DSSP assignment EESSEE. **d**: 1CTF, residues 93–98 sequence ALKEGV, DSSP assignment EEEEEE. It is interesting to see, for example, the similarity in structure between (**a**) and (**d**), although their secondary-structure DSSP assignments are so different.

ized by the expected distance and standard deviation for this length, as shown in Table II.

This min–max function was used, rather than averaging the results for the two segments, in order to avoid situations where one pair of substructures fits very well while the second pair fits poorly. A similar definition was used for $RMS_{2\#}$, where two breaks were allowed.

The results for our reconstructed set of 71 proteins of length 60 were compared to the results of applying the same process to measure the distance for 1000 pairs of structures (of length 60) drawn at random from this set. The results are summarized in Table VII. With no breaks, a clear majority of the proteins were well reconstructed, but for 27 of the 71 proteins the improvement, compared to a random pair of structures, was less than 2 SD. When one break was allowed, in almost all cases (68 out of 71 proteins of length 60) the distance, as defined above, is at least 2 SD less than expected. Allowing two breaks did

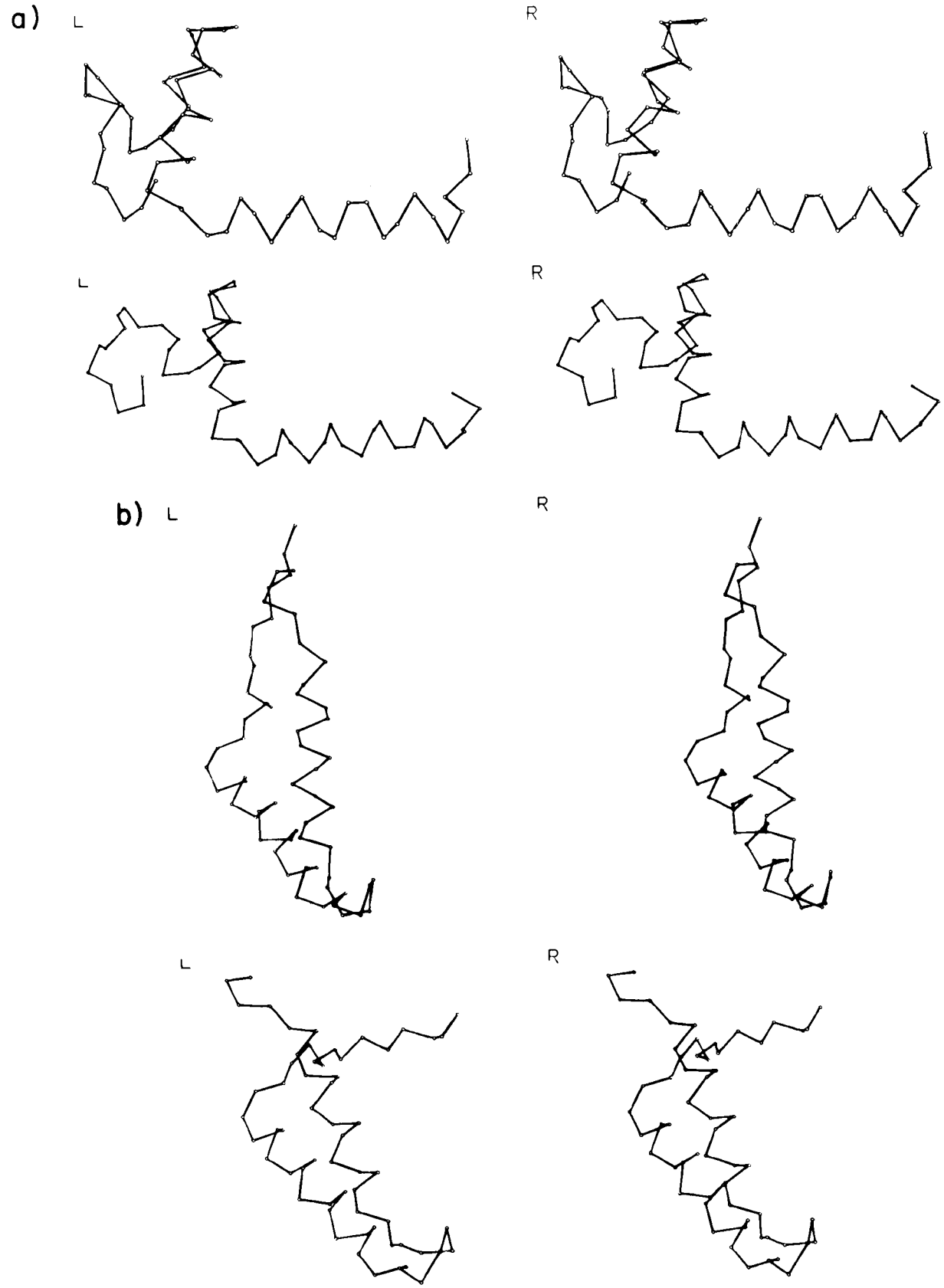


Fig. 10. The original and reconstructed proteins. **a**: The first 60 residues of 1LH1. Top: A stereo plot of the original protein; bottom: the reconstructed protein. The similarity is evident and the RMS between the structures is small (4.4 Å). **b**: The first 60 residues of 2CYP. Top: A stereo plot of the original protein; bottom: the reconstructed protein. Although the RMS between the structure is rather large (8.4 Å), the resemblance between them can be seen. The difference is mainly in incorrect orientation of the upper arms.

TABLE VII. The Success of the Reconstruction Algorithm*

	No breaks (%)	One break (%)	Two breaks (%)
Reconstructed vs. original proteins	61.9	95.7	94.3
Two randomly chosen proteins	1.7	1.8	0.7

*The percentage of proteins that have been reconstructed successfully (the distance from the original protein being 2 SD or more, lower than expected for a random pair of structures). When one break is allowed, thus enabling some correction in the reconstruction process, the performance improved. Allowing a second break had no significant effect. As a control the distance between 1000 pairs of proteins was calculated in the same way. It was very rare to find among them a pair whose distance was 2 SD less than the average. In the random case, introducing breaks did not make any difference.

not change the results significantly. In the random control, allowing breaks was not significant, since only in 1.7% of the cases was the normalized distance 2 SD lower than the expected value.

ASSIGNMENT OF HEXAMERS TO BUILDING BLOCKS BY THEIR SEQUENCE

As mentioned above, each building block is associated with a sequence distribution matrix. These matrices reflect the sequences of the hexamers that are represented by each building block. They can be used to assign a hexamer sequence to its corresponding building block. The sequence of each hexamer is matched against these matrices and assigned to the building block whose matrix it fits best. Preliminary results indicate that, after certain normalizations, this assignment by sequences often correctly predicts the assignment according to structural proximity. In one experiment, using the 40 most popular building blocks, about 38% of 8000 hexamers were assigned by their sequence to the same building block that was assigned by matching the known three-dimensional coordinates. About 67% of the hexamers were assigned to one of the five nearest building blocks according to the structural match. We plan to perform a more detailed analysis of these sequence distribution matrices, extracting the important features and applying careful normalizations, in order to improve this approach.

DISCUSSION

Why hexamers?

In this research we concentrated on an analysis of hexamers as building blocks. It is clear that there is no magic in the number six. We considered hexamers as a starting point for our research for the following pragmatic reasons:

- Kabsch and Sander[1] observed that the same pentamer sequence can be found in totally different conformation in different proteins. Thus, pentamers by themselves seem to be too short to carry structural stability.
- The length of secondary structure elements is usually in the range of 4–16, e.g., the length of a turn is usually 4–6 amino acids; helices and sheets have larger ranges but they are still usually less than 16 residues. For example, in the refined Brookhaven data base (according to the DSSP assignments) the average length of helices is 10 residues. Thus, using longer oligomers (say more than 8) would preclude the identification of short structural elements.
- When using longer oligomers, their corresponding sequences tend to the average distribution and their sequence dependency seems to fade out. Hence, longer building blocks may be less sequence specific, and we should use as short fragments as possible.

We believe that even better results may be obtainable with variable length building blocks in the range of 6–16 residues. As the analysis of building blocks with variable length is more complicated, this work is still in progress.

CONCLUSION

The aim of this research was to show the existence of a manageable-sized set of standard building blocks that has sufficient expressive power to replace most of the oligomers in known structures. We can thus answer our first question in the affirmative; the fact that 55 disjoint clusters were formed indicates that the known hexamers can be divided into distinct structural motifs. These structural motifs include the well-known types of secondary-structure elements with finer resolution, and many standard units that connect them. The classification into a set of a few dozen building blocks seems to be more meaningful than the crude classification to very few secondary-structure elements. We want to stress that our building blocks are not secondary structure elements; they are short *3D* motifs. Even if the secondary structure assignment of a protein is known, it is not clear how to assign three-dimensional structure to the protein. However, because our building blocks reflect three-dimensional information (for example, the direction of a turn after a helix) they easily lend themselves to the reconstruction process.

Thus, we can give a positive answer to the second question: The simple reconstruction algorithm that we applied yields good results in simulating the

original proteins. Recent work by Jones and Thirup[5,6] and Wodak[17] shows that the structure of proteins can be reconstructed with very high fidelity by fragments of various lengths taken from the entire library. This fact demonstrates that structural motifs tend to repeat in different proteins, but the approach cannot be used as the basis of a prediction method. In contrast, using a limited number of building blocks, we were able to associate many of them with a statistically significant sequence-distribution matrix. Thus, with respect to the third important question, about the ability to use these observations as a basis for structure prediction, we have a good indication of the feasibility of the method. Our analysis suggests that there are only about 60 truly different dominant hexamer classes, further divided into about 100 structures. Thus, the sequence analysis need only lead to the assignment of each hexamer to one of these 100 structures. Our preliminary results indicate that, after certain normalizations, the sequence distribution matrices of the building blocks can be used to assign many hexamers to their correct structural clusters. Naturally, in many cases the sequence is not sufficient to determine the structural classification uniquely, and, thus, in cases of ambiguity several alternative building blocks are suggested.

The prediction scheme suggested here is based on assigning each hexamer sequence to its corresponding building block and then smoothly concatenating these building blocks into a full predicted protein structure. In this report we have concentrated on showing that each step in this scheme is feasible. We are currently trying to incorporate these steps into a fully integrated system.

Work in a parallel direction was reported recently by Vásques and Scheraga.[18] In their research, the structures of the short oligomers are determined by energy calculations, and the concatenation into a full structure is backed up by NMR measurements. In the case of BPTI, this procedure is reported to be very successful.

Our procedure uses the fast (perhaps not fast enough) growing body of information about the 3D structure of proteins to predict the conformations of short fragments. The next step in our research will deal with multiple ways to correctly combine these short fragments into a full protein. It is well known that the structure of a protein is determined by a combination of local and global interactions. While our method is capable of dealing with the short-range behavior, it does not yet consider the more global effects. This is why our reconstruction process ended in large RMS distances for some of the proteins. Introducing one break, which corresponds to correcting incorrect modeling of one hexamer, significantly improved the quality of many of the matches. Detecting and correcting these few wrong decisions should be possible by use of additional information. Thus, it is necessary to incorporate any available additional long-range constraints, like matching regions of β structure to form sheets, forming the correct disulfide cross links, enabling linkages to known prosthetic groups and metal-binding ligands, keeping N-terminal to C-terminal distances, and being consistent with NMR measurements. A structure prediction can then be built by combining the predicted oligomer structures into a full structure under these additional long-range constraints. At this stage, the ambiguity in sequence-based assignment of local structures should be resolved by the choice of building blocks that can be joined "smoothly" while obeying the long-range constraints.

We believe that the work described here is a first step toward a system that would be able to predict reasonable 3D models of the backbone of a protein structure.

ACKNOWLEDGMENTS

We would like to thank John Moult, Fred Hirshfeld, Alex Wlodawer, and Peter Stern for helpful and stimulating discussions. We would also like to acknowledge the extremely careful and constructive comments of the referees. This work was supported by Grant DAJA 45–86–C–0016 from the U.S. Army Research Office (through its European Research Office) to JLS, and by the Kimmelman center for biomolecular structure and assembly.

REFERENCES

1. Kabsch, W., Sander, C. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. U.S.A. 81:1075–1078, 1984.
2. Wilson, I.A., Haft, D.H., Getzoff, E.D., Tainer, J.A., Lerner, R.A., Brenner, S. Identical short peptides in unrelated proteins can have different conformations: A testing ground for theories of immune recognition. Proc. Natl. Acad. Sci. U.S.A. 82:5255–5259, 1985.
3. Argos, P. Analysis of sequence-similar pentapeptides in unrelated protein tertiary structure: Strategies for protein folding and a guide for site-directed mutagenesis. J. Mol. Biol. 197:331–348, 1987.
4. Rooman, M. Wodak, S.J. Identification of predictive sequence motifs limited by protein structure data base size. Nature (London) 335:45–49, 1988.
5. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. Embo J. 5:819–822, 1986.
6. Jones, T.A. Computer graphics in structure analysis. Acta Crystallogr. A43:Supple. ML 18–1, 1987.
7. Blundel, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. Nature (London) 326:347–352, 1987.
8. Nyburg, S.C. Some uses of a best molecular fit routine. Acta Crystallogr. B30:251–253, 1974.
9. Kabsch, W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. B32:922–923, 1976.
10. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A34: 828–829, 1978.
11. Remington S.J., Matthews, B.W. A Systematic approach to the comparison of protein structures. J. Mol. Biol. 140:77–99, 1980.
12. Fu, K.S., Lu, S.Y. A sentence to sentence clustering procedure for pattern analysis. IEEE Trans. Syst. Man Cybern. SMC-8:381–389, 1978.

13. Stout, G.H., Turley, S., Sieker, L.C., Jensen, L.H. Structure of ferrodoxin I from Azotobacter vinelandii. Proc. Natl. Acad. Sci. U.S.A. 85:1020–1022, 1988.
14. Ramakrishnan, C., Ramachandran, G.N. Streochemical criteria for polypeptide and protein conformation: 2. Allowed conformation for a pair of peptide units. Biophys. J. 5:909–933, 1965.
15. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
16. Matthews, B.W., Rossmann, M.G. Comparison of protein structures. Methods Enzymol. 115:397–421, 1985.
17. Wodak, S.J. Personal communication.
18. Vásquez, M., Scheraga, H.A. Calculation of protein conformation by the build-up procedure. Application to BPTI using limited simulated NMR data. J. Biomolec. Struc. Dyn. 5:705–755, 1988.