PROTEINS: Structure, Function, and Genetics 24:35–50 (1996)

RESEARCH ARTICLES

Identification of a Pattern in Protein Structure Based on Energetic and Statistical Considerations

A. Amadei and B. Vallone
Dipartimento di Scienze Biochimiche, Università di Roma "La Sapienza," Rome, Italy

ABSTRACT **We carry out a statistical analysis of the nonbonded interactions in 10 high-resolution nonhomologous protein structures, using original algorithms. We observe a tendency of nonbonded interactions which contribute significantly (i.e., with an energy lower than the average value, referred to as "strong") to protein stability, to be concentrated in clusters of residues that are strongly sequence correlated. We characterize this sequence correlation and subsequently define a "system" as the pattern that describes these clusters. In order to study the distribution of the systems in the proteins we build a matrix for each protein and for each term of the empirical potential function used to compute the nonbonded interactions; each *ij* element is the number of common residues between the systems *i* and *j*. The analysis of the matrices shows the presence of compact blocks that define units in the protein structure which concentrate strong and weak interactions inside the unit itself and display relative independence with respect to the rest of the protein. Comparing the blocks defined by the three nonbonded energy components (electrostatic, hydrogen bonds, and van der Waals interactions) we observe a one-to-one correspondence between the blocks of different energy components with an average overlap of 90% of the residues forming each block.**

Key words: protein structure, protein folding, protein engineering

INTRODUCTION

Since the structures of myoglobin and hemoglobin were first determined in the 1960s, many other protein structures have been solved forming a database of structural information for theoretical studies. In spite of the increase in the number of deposited structures and of many relevant studies based on this database, the principles ruling the phenomenon of protein folding are far from being understood.

Several promising approaches have been developed including simulated annealing,[1] construction of score matrices,[2] search for patterns such as clusters,[3] or specific main chain configuration[4] and patterns in side chain interactions[5] as well as the classical methods to predict secondary structure.[6,7] Simulation of the detailed atomic motions involved in the activity of proteins by means of molecular dynamics is hindered by the vast dimension of the atomic coordinate space which makes it impossible to simulate motions in a time range that overlaps biological functions such as catalysis, structural transitions, or folding (time >>1 ns). The distribution of nonbonded interactions in a protein must be linked to the folding process and to functional properties, therefore the aim of our work was to investigate, with statistical methods, the whole set of nonbonded interactions in 10 protein structures. In this paper we present an extensive statistical analysis resulting in the discovery of a novel pattern of nonrandom relevant interactions organized in a network; this may be useful in understanding the principles which drive protein folding and in providing new information of functional dynamic properties in proteins.[8]

MATERIALS AND METHODS

Protein Structures

We have analyzed the structures of 10 nonhomologous proteins taken from the Brookhaven Protein Data Bank[9] with a resolution varying from 3.0 Å (1PYP) to 1.54 Å (3CPA). These structures belonged to three different secondary structure classes: α, β, and α/β proteins (see Table I).

Energy Calculations

All calculations were performed using the program BRUGEL running in VAX/VMS environ-

Received October 25, 1994; revision accepted April 25, 1995.
Address reprint requests to B. Vallone, Dipartimento di Scienze Biochimiche, Università di Roma "La Sapienza," P. le A. Moro, 5, 00185 Rome, Italy.
Present address of A. Amadei: Department of Biophysical Chemistry, University of Groningen, Groningen, Holland.

TABLE I. Structures Included in the Statistical Analysis*

Protein	PDB file	Secondary structure class	Resolution (Å)	a.a. number
Arabinose binding protein	1ABP	α	2.4	306
Calmodulin	3CLN	α	2.2	148
Concanavalin A	2CNA	β	2.0	237
Carboxypeptidase	5CPA	α/β	1.54	307
Intestine Ca binding protein	3ICB	α	2.3	75
Lysozyme (hen)	3LYZ	α/β	2.0	129
Myoglobin (sperm whale)	1MBD	α	1.4	153
Plastocyanin	1PCY	β	1.6	99
Pyrophosphatase	1PYP	α/β	3.0	281
Rhodanese	1RHD	α/β	2.5	293

*The PDB file name, together with the secondary structure class, structure resolution, and the number of amino acids in the protein are given.

ment,[10] using the potential function reported by Karplus and Petsko.[11]

The potential energy of the structures was analyzed after 100 steps of steepest descent energy minimization in order to optimize small bad contacts present in the initial structures. In no case did the average atom displacement exceed 0.2 Å. This initial minimization was performed in order to eliminate bias due to the heterogeneity of the sample of the chosen proteins.

After this step we computed the nonbonded energy for the 10 protein structures listed in Table I, considering the contribution of each amino acid pairwise interaction.

We have computed the potential energy for each type of nonbonded interactions (van der Waals, electrostatic, and hydrogen bonds) estimating for each protein the average value of an interaction between two residues within a cutoff of 8 Å, and its standard deviation (see Table II). Only the pairwise interaction energies below these average values were considered significant (and from hereon referred to as "strong") and on these data we have performed the statistical analysis that is outlined in the next section, and described in detail in Appendix A.

RESULTS

Statistical Analysis

After computing all the nonbonded residue–residue interactions, we analyze separately for each energy component their distribution along the protein sequence. The scope of our analysis is to evaluate if the distribution of the pairwise interactions between residues in a protein structure can be considered random and if not to try to identify the interactions (residues) responsible for that. We decided to use a general approach based purely on statistics to study the distribution of interactions in protein structures to avoid bias coming from structural and functional principles already described.

The detailed mathematical and statistical procedure that is followed is given in Appendix A to this paper.

TABLE II. Average Values (kcal/mol) for the Residue–Residue Interaction for the Three Components of Nonbonded Potential Energy Calculated Over the Whole Protein Set and the Single Protein Average Standard Deviation Evaluated Over the Same Sample

	x	Sx
Electrostatic	−0.091	0.021
Hydrogen bond	−0.19	0.058
van der Waals	−0.65	0.081

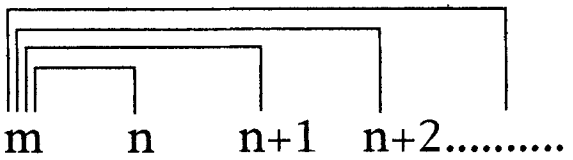
We construct a function that expresses the probability of a number K of pairwise interactions of the generic residue to be distributed into P segments within a sequence window of N residues (see Scheme 1), each segment being separated by at least one weak or null interaction.

Having constructed the probability function we evaluate if the observed distribution of strong interactions computed for our set of proteins corresponded to a random distribution in the sequence, using the statistical variables χ^2 and t.

We observe that the distribution of pairwise strong interactions was not behaving randomly (probability $<<5\%$ that it is really random). Only by excluding from the analysis all strong interactions of an amino acid with a group of three or more residues in sequence ($\alpha \geq 3$), and setting the sequence window $N = 20$ (or close to 20), can we observe that the distribution of the remaining strong interactions has a probability to be random larger that 5%.

We conclude that strong nonbonded interactions of a residue with a stretch of three or more amino acids in sequence ($\alpha \geq 3$) are absolutely not generated by a random distribution while the others seem

Scheme 1. Example of how K = 9 strong interactions can be divided into P = 4 segments in a window of N = 18 residues. A box represents a residue, if it is shaded a strong interaction is present.



Scheme 2. Conditions necessary for 4 (or more) residues to form a system. m: the centre of the system, any residue of the protein; n: the first residue having a strong interaction with m, in any sequence position relative to m; n + 1, n + 2, ...: in order to form a system, at least two residues following n must have strong interactions with m. A line represents a strong interaction between residues.

to be really randomly distributed in N-residues sequence windows.

Having brought into evidence with statistical analysis a pattern of nonrandom strong interactions, we identify and study the sets of residues behaving according to this pattern, i.e., residues having strong nonbonded interactions with at least three other residues in sequence for all the proteins in Table I. We call these sets "systems" and a system consists of four or more amino acids, one of them being the "center" and establishing strong interactions with all the others. The residues different from the center of the system (at least three) must be consecutive over the protein sequence. It has to be stressed that the center of the system is not necessarily in sequence with the others, and in most cases is not (see Scheme 2).

Matrix Construction and Analysis

After identification of the systems, we study the relationship between them within a protein three-dimensional structure, using the same set of 10 proteins (see Table I). First we order the systems according to the sequence number of their centers, defined by the residue which establishes strong interactions with the other members of the system (3 or more, see Scheme 1). After this operation we define a matrix where the ij element is the number of residues in common between system i and system j, the diagonal elements being the number of residues forming each system. The matrices are built for the three components of the nonbonding energy. Figure 1 shows the matrices for ICB; Figure 2 gives the van der Waals (vdW) matrix for LYZ. It appears that most of the matrices are structured in blocks. In Figure 3 we give the vdW matrix for PCY where separation into blocks is not immediately evident.

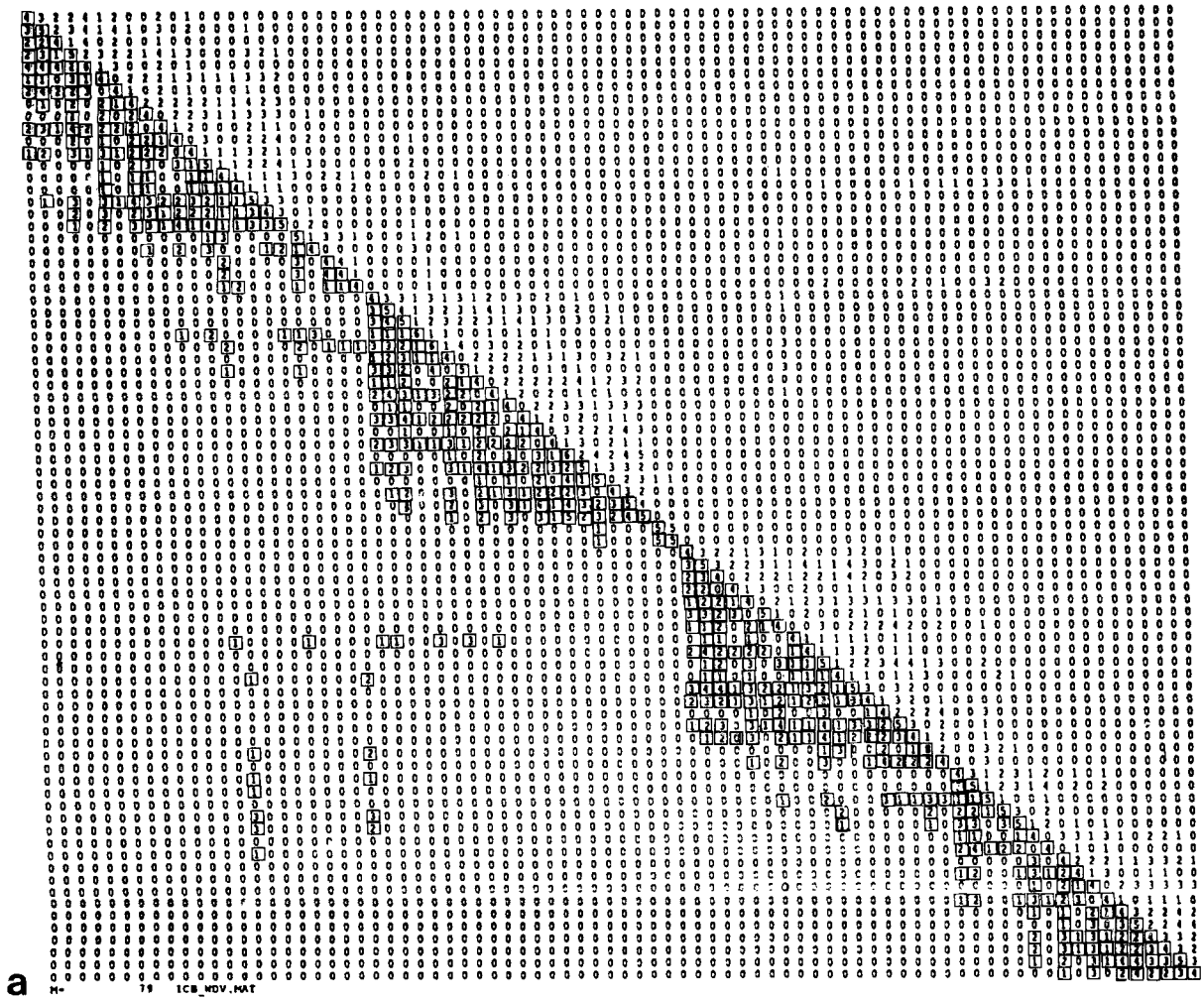
Nevertheless, a small number of permutations allows the same general structure to be identified in this second type of matrices (see Fig. 3). In 3 out of 10 proteins we obtain "scattered" matrices. Without the aid of specific computer programs we could group only the smallest of them, reducing it to a block matrix. The two proteins that generate the large scattered matrices (i.e., PYP and CNA) are not included in the analysis that follows given the fact that we do not identify the blocks.

As far as the other 8 proteins are concerned, we define rigorously the blocks in each matrix (for a total of 24 matrices considering one matrix for each of the three components of nonbonded interaction energy for eight proteins) using the following criteria: (1) we start from the clusters identified by visual inspection and we take as the first member of the visual cluster the system that does not have residues in common with the immediately preceding cluster (block); the last member is chosen with the same criterion (it has no residues in common with the following cluster). (2) We define the internal connectivity of a cluster as the sum of elements of the matrix that have both indices corresponding to members of the cluster, and the connectivity between clusters as the sum of elements in common between the two clusters. When the connectivity between two clusters exceeds 6% of the internal connectivity of at least one of them we join the couple in a single cluster. We iterate this step until it is no longer possible to group two clusters. The clusters built according to this procedure are called "blocks" and we are going to refer to these units using this name throughout this paper.

We could define the blocks for the three components of the nonbonded energy for the 8 proteins on which we carry out the matrix analysis. The following step was to compare for each protein the blocks identified by the van der Waals, hydrogen bond (HB), and electrostatic component. This analysis showed that in a protein the blocks found in the matrix of the van der Waals component of potential energy include those found in the same protein for the matrices arising from the electrostatic and hydrogen bond component with an average overlap of 90% (SD = 16%). Therefore the study of the properties of the blocks is performed only on the blocks evaluated from the van der Waals matrices.

The average number (x) of residues that form a van der Waals block over all 8 proteins is $x = 26$, SD = 13. The average percentage (y) of residues in a protein structure not included in any block is insignificantly small, $y = 4.2\%$ with SD = 4.4%.

Having identified with our criteria these units in protein structure (the blocks) we investigate the relationship occurring between them. We analyze for each block the nonbonded interactions internal to the block and the interactions with the rest of the protein; by this procedure we try to evaluate if the blocks constitute units relatively independent from each other within the protein structure. As a result we find that the internal nonbonded interaction energy of a block represents on average 71% (SD = 8%) of all the total nonbonded interaction energy of the block itself within the protein (including the blocks). With this result we observe the tendency of



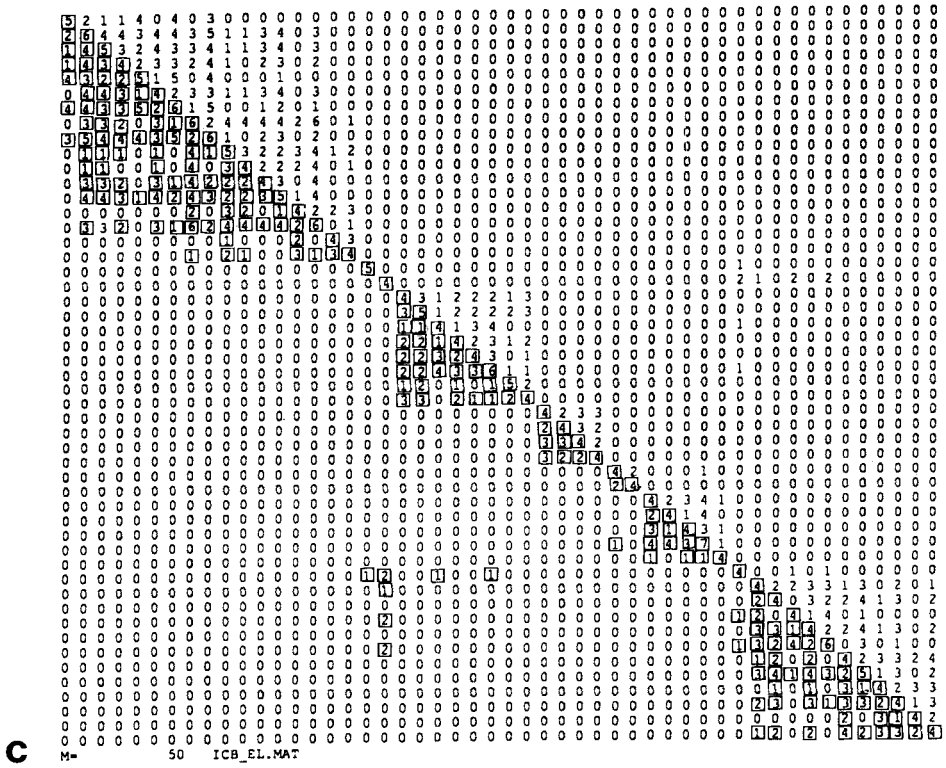


Fig. 1 a and c.

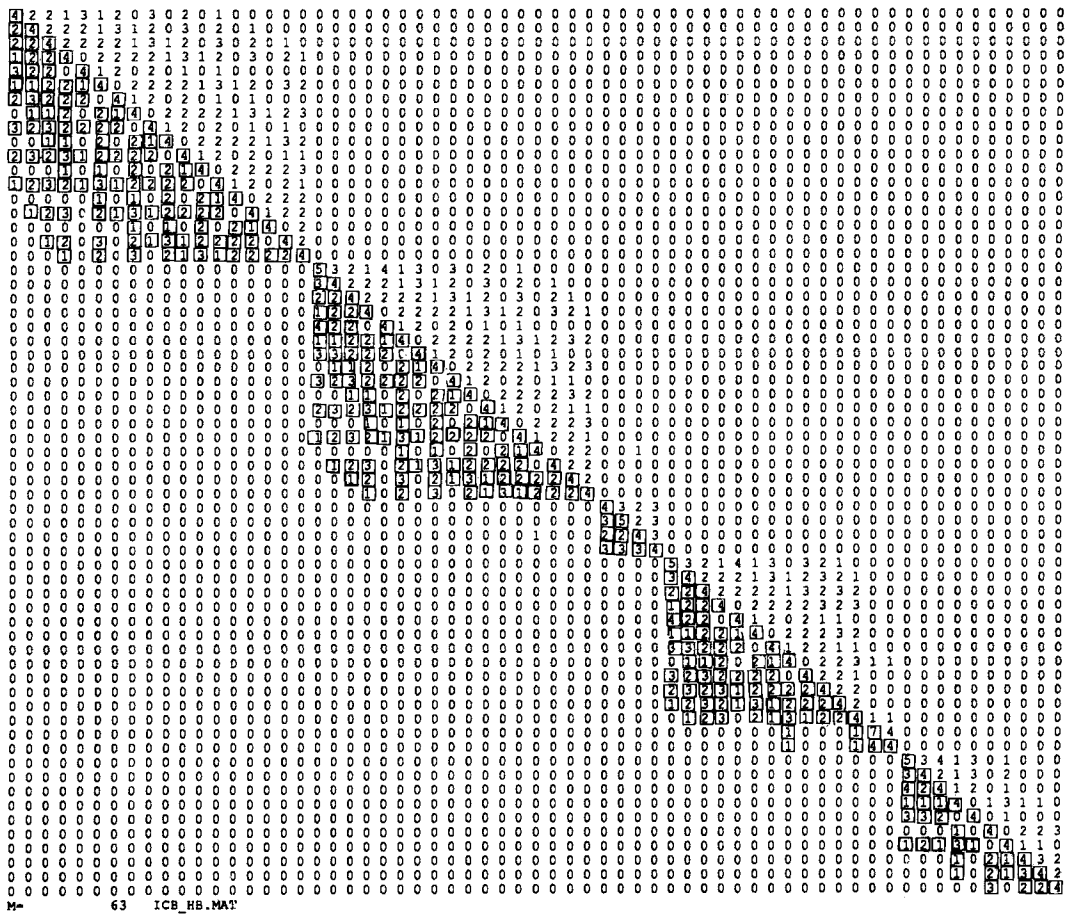


Fig. 1. Matrices displaying the connectivity between systems for intestinal calcium binding protein (ICB); each *ij* element represents the number of common residues between the *i* and *j* systems. (**a**) Connectivity matrix of the systems defined by the strong van der Waals interactions. (**b**) Connectivity matrix of the systems defined by the strong hydrogen bonds interactions. (**c**) Connectivity matrix of systems defined by the strong electrostatic interactions. All the matrices are symmetric; we have highlighted by inclusion in a square the elements different from 0 in one of the two symmetrical halves.

the blocks to represent independent units as far as the nonbonded interaction energy within a protein is concerned.

We compute also the percentage x of the connectivity interactions (strong interactions within the system) over the total number of strong interactions for the three nonbonded energy components: for the vdW interactions $x = 54\%$ (SD = 12%), for the HB interactions $x = 40\%$ (SD = 15%), and for the electrostatic interactions $x = 25\%$ (SD = 4%).

In order to evaluate the relevance of the "strong" nonbonded interactions in stabilizing the three-dimensional structure of a protein, we compute the number of strong interactions as a fraction of the total number of nonbonded interactions and the nonbonded energy involved in strong interactions as a fraction of the total nonbonded energy in the proteins.

We find that for vdW and HB components the strong interactions, although representing only 37% of the total number of interactions, contribute to 74% of the nonbonded energy for both these two components. For the electrostatic component we obtain that strong interactions represent 50% of the total number of interactions of this type and that the distribution function of the electrostatic interactions being a gaussian-like curve with an average value of about −0.1 kcal/mol and an SD of about 0.5 kcal/mol, they contribute to more than 90% of the total attractive electrostatic potential energy.

We visually inspect the blocks in the protein structures in order to evaluate the correlation between secondary structure segments and these units (the blocks). We observed that generally a β-sheet or an α-helix is not truncated, being completely included within a single block; nevertheless a block can be heterogeneous in its secondary structure composition since it often includes a combination of helices, sheets, turns, or coil (see Fig. 4). The blocks may not be continuous (see Fig. 5) as far as sequence is concerned, and there are parts of the protein (an average of 4%) which are not included in the blocks,

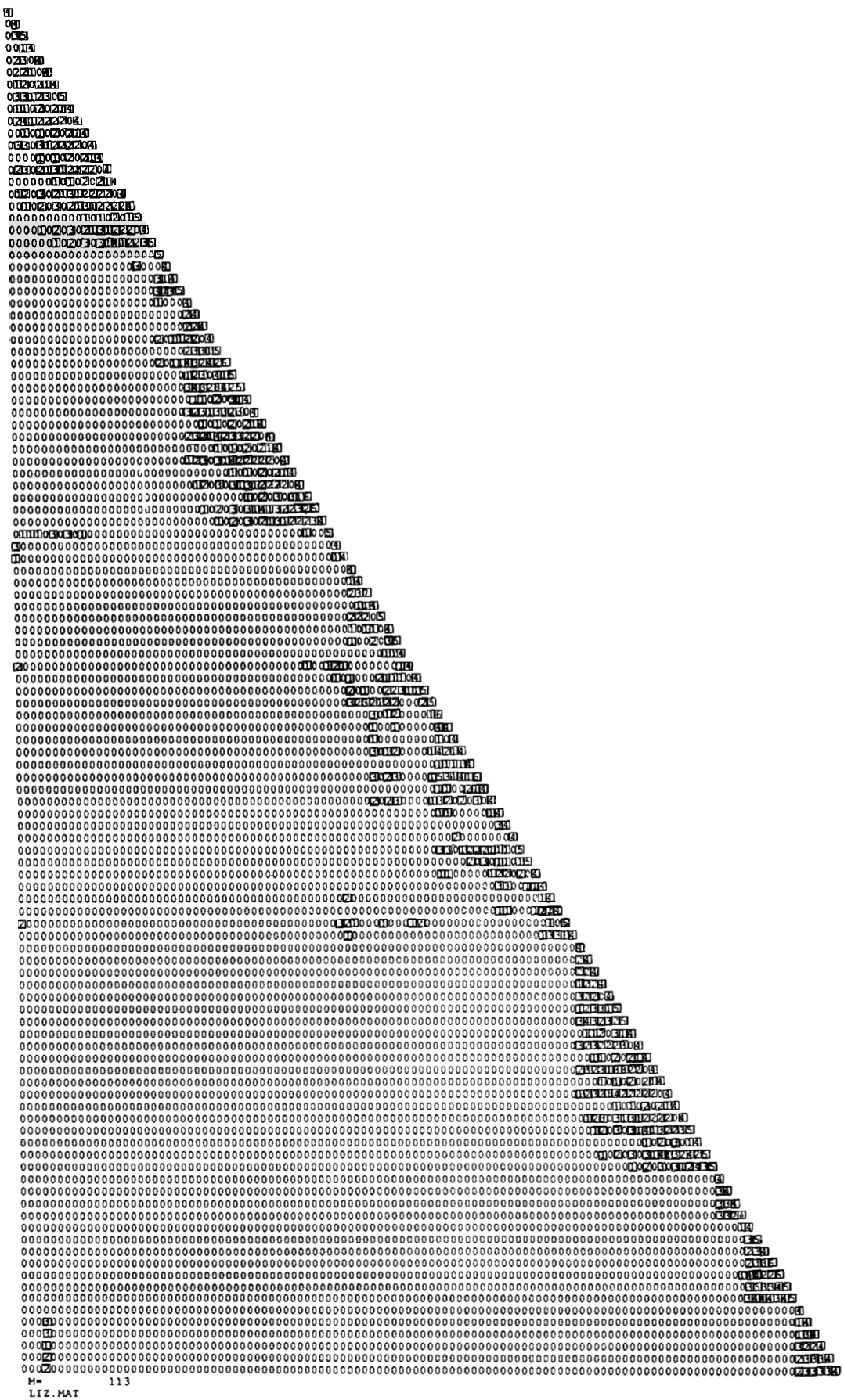


Fig. 2.

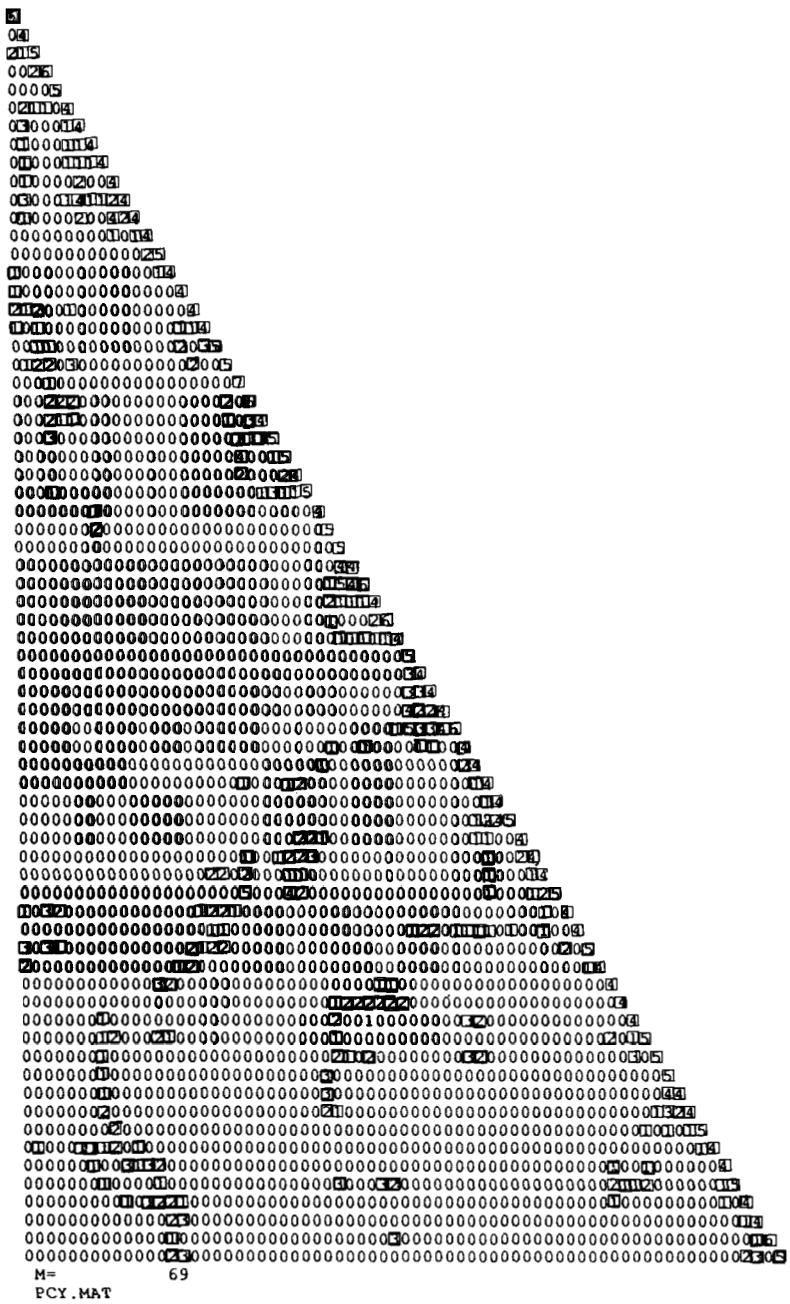


Fig. 3. Matrix displaying the connectivity between systems for plastocyanin (PCY); only the van der Waals component matrix is given (see text). We display only one of the two symmetrical halves or the matrix; the elements different from 0 are included in a square.

and generally also not included in secondary structure segments.

In Figure 5a we display the structure of plastocyanin (PCY). The matrix was originally scattered and we identified the blocks by permutating the systems order in the matrix; we observed that the two matrix blocks define compact regions in the protein. In Figure 5b and c blocks 1 and 2 for PCY are given. In Figure 4a the main chain of lysozyme is shown, displaying its division into five blocks; single blocks are given in Figure 4b–f allowing the variety of the secondary structure composition of the blocks to be shown.

Fig. 2. Matrix displaying the connectivity between systems for lysozyme (LYZ); only the van der Waals component matrix is given. We display only one of the two symmetrical halves of the matrix; the elements different from 0 are included in a square.

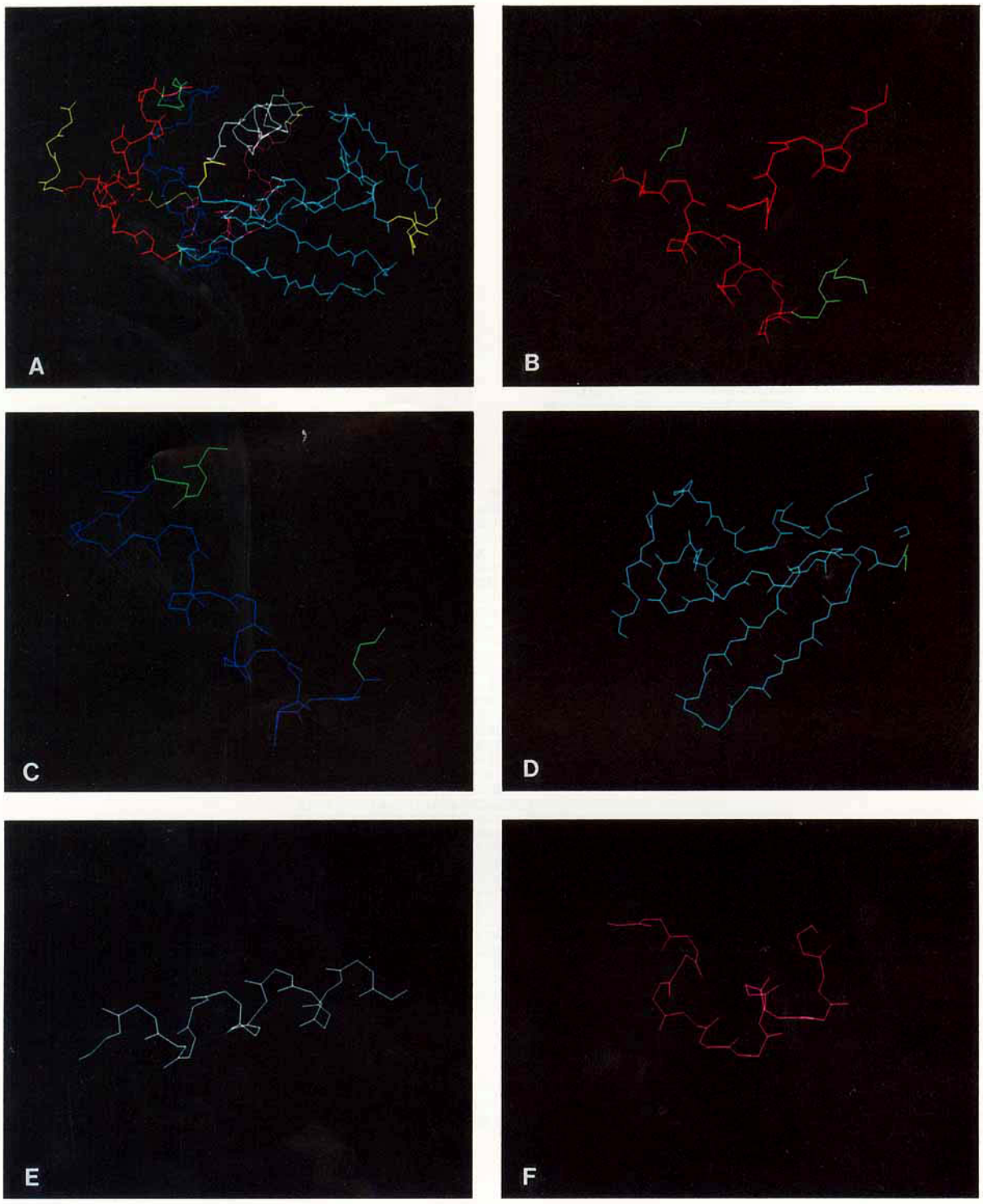


Fig. 4. Three-dimensional structure of lysozyme (LYZ). (**A**) Main chain only, block 1; red, block 2; blue, block 3; light blue, block 4; white, block 5; pink, connectivity bridges; green, residues not belonging to a block. (**B–F**) Structures of blocks 1–5.

Correlation With Structural Features

In Tables III and IV we list the residues included in each block for arabinose binding protein, rhodanese, calmodulin, lysozime, sperm whale, and *Aplysia limacina* myoglobins. In the following section we analyze two points: the correlation between

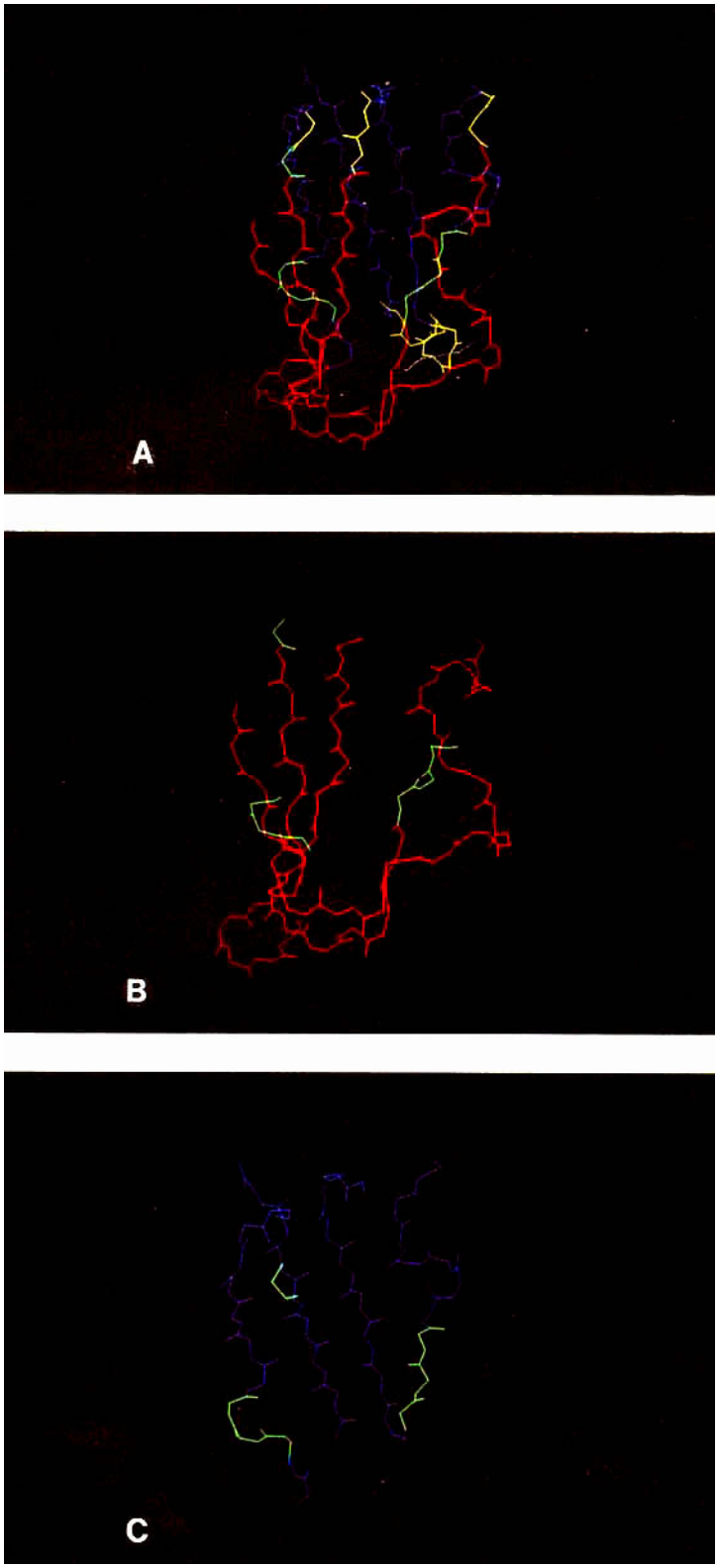


Fig. 5. Three-dimensional structure of plastocyanin (PCY). The main chain of PCY (see Fig. 1) is displayed in **A;** block 1; red; block 2, blue; connectivity bridges, green; residues not belonging to a block, yellow. (**B**) Block 1 main chain. (**C**) Block 2 main chain.

block partitioning and protein domains (for multidomain proteins in the set) and the conservation of blocks in proteins showing the same fold but low sequence homology.

Protein domains

We have analyzed the proteins in our set which contain more than one domain in order to find out whether blocks could be shared between domains or were contained in just one domain.

Three proteins (arabinose binding protein, calmodulin, and rhodanese) show a two domain structure in the stuctural acception, i.e., independent compact globules, whereas in lysozime two "folding" domains have been described.[12] In both cases, as shown in Table III (where domains and blocks are shown and can be compared for these proteins), the division into blocks is not incompatible with domain partitioning.

This appears to be significant, especially for *arabinose binding protein* where the two domains are not structured as "beads on a string" but are discontinuous along the protein chain and one helix is relatively independent from the two domains;[13] in this protein there are 11 blocks, five for each domain and one for the extra helix.

In *calmodulin* where three domains can be identified, the two calcium binding globules and a long α-helix acting as a connection between them,[14] we find seven blocks, three for each globule and one including the long α-helix plus a short loop.

The simplest case appears to be *rhodanese* where two well-defined globular domains are connected by a linker;[15] as expected the linker itself is not included in a block and no block is shared by the two domains.

Extensive studies on folding intermediates have been carried out on *lysozime*[12,16] leading to the definition of two folding domains which become structured on well-separated time scales; in this protein we find that one block includes the central, late folding, β-sheet domain, whereas the "head" and "tail" of the polypeptide which constitutes the early folded part contain the other four blocks. This finding is not inconsistent with the data on folding intermediates since the time required for the structuring of four small units should be shorter than the collapse of the large block constituted by the β-sheet domain.

Fold families

In order to observe whether the partitioning into blocks was preserved in proteins which have the same fold but low sequence homology we have compared sperm whale myoglobin and *A. limacina* myoglobin (sequence homology = 20%). As shown in Table IV, almost the same division was observed, especially as far as the inclusion of helices in the blocks is concerned; on the other hand the inclusion of loops into blocks is less preserved in the two structures. The main difference observed was in block II which in sperm whale myoglobin contains helices B and C, whereas in *A. limacina* myoglobin it is split into two blocks, one for each helix. Nevertheless it appears that even in distantly related proteins (one from a mammal and the other from a mollusc), where sequence homology is low, the distribution of blocks in the fold is conserved pointing toward the conclusion that we are dealing with a feature linked rather to fold architecture than to specific sequences.

The same considerations arise from the comparison of the domains of arabinose binding protein, calmodulin, and rhodanese; these proteins are all composed of two domains that show the same fold, presumably arising from gene duplication and fusion. Each couple of domains can therefore be regarded as an example of two distantly related proteins. In two cases we find that all pairs of domains contain the same number of blocks which include corresponding secondary structure elements in the related folds (see Table III). In arabinose binding protein where the folds of the two domains are not identical, one of the six blocks found for each domain does not observe this correspondence.

DISCUSSION

Having computed the nonbonded interactions within a group of 10 nonhomologous proteins, we first distinguish two classes of nonbonded interactions: the "strong" interactions contributing the largest fraction of the total nonbonded energy in the protein, and the weak ones that we did not further study given their minor relevance. By analyzing the strong interactions we find two groups, one of which includes 46, 60, and 75% of the total number of "strong" interactions, respectively, for the vdW, HB, and electrostatic components, which seem to be randomly distributed. The second group formed by the remaining strong interactions shows a nonrandom distribution, possibly implying a role in the three-dimensional folding of the polypeptide chain. We analyze the distribution of this second class of interactions and we observe that residues establishing these interactions are ordered into small clusters that we call "systems," where a system is defined as a set of residues formed by a "central" residue having strong interactions with at least three other residues in sequence (see Scheme 1). By searching systematically the set of 10 proteins chosen as a sample we find that most of the residues in a protein are members of systems, implying the existence of a network of nonrandom strong interactions within the protein structure. A system in itself has interesting properties, since the formation of one of the interactions between the "central" residue (m) and one of the other three or more ($n,n+1,n+2,\ldots$) residues in sequence favors the formation of the other interactions within the system because of the sequence continuity requirement (in $n,n+1,n+2,\ldots$). In fact, after one of the contacts is established, the others

TABLE III. Distribution of Blocks Within Domains for Arabinose Binding Protein, Rhodanese, Calmodulin, and Lysozyme

Domain	Blocks
Arabinose Binding Protein[13]	
P Domain	
Helices	
I 16–30	(1) 10–31
II 42–57	(2) 42–57
III 70–81	(4) 72–82
IV 257–273	(11) 255–273
Strands	
a 34–39	
b 4–10	
c 59–64	
d 84–89	
e 104–109	(3) 63–70, 83–106
f 281–283	
Q Domain	
Helices	
I 109–129	(5) 111–138
II 146–161	(6) 145–168
III 177–192	(7) 178–196
IV 206–218	(8) 205–231 (includes strand d)
V 233–241	(9) 232–242
Strands	
a 170–172	
b 136–141	
c 199–204	
d 225–232	
e 247–253	
f 287–291	
Helix X (independent from the domains Q and P)	
293–301	(12) 286–301
Rhodanese[15]	
Domain 1 (residues 1–142)	
Helices	
I 11–22	(1) 6–21
II 42–50	(2) 42–50
III 76–87	(4) 65–93
IV 107–119	(5) 99–118
V 129–137	(6) 129–136
	(3) 59–64
Domain 2 (residues 159–293)	
Helices	
I 163–174	(7) 163–174
II 183–189	(8) 183–189
III 224–235	(10) 224–241
IV 251–264	(11) 251–264
V 274–282	(12) 274–287
	(9) 211–223
Hinge	
143–158	
Blocks 3 and 9 contain an homologous segment connecting helices II and III in both domains	
Calmodulin[14]	
Ca-binding domain 1	
Helices	
I 7–19	(1) 5–26
II 29–39	(2) 28–39
III 46–55	(3) 44–62

(continued)

TABLE III. Distribution of Blocks Within Domains for Arabinose Binding Protein, Rhodanese, Calmodulin, and Lysozyme *(continued)*

Domain	Blocks
Ca-binding domain 2	
Helices	
I 102–112	(5) 101–112
II 119–128	(6) 117–135
III 138–148	(7) 137–147
Connection helix	
65–92	(4) 64–99
Lysozyme[16]	
Early folding domain including the four α-helices and one 3_{10}-helix	
Helices	
I 5–15	(1) 4–15
II 25–35	(2) 19–35
III 88–89	(4) 88–101
IV 108–115	
3_{10} 120–124	(5) 103–125
Late folding domain including the β-sheet and the other 3_{10}-helix	
Sheet 40–64	
Loop 65–79	(3) 39–84
3_{10} 80–84	

will form more easily than the first one, since the regions of the protein in which m and $n+1$, $n+2$, $n+3$ are, respectively, situated would be already positioned near to each other. This implies that in the process of folding, a cooperative behavior may be involved in the formation of the interactions defining a system; given the connectivity between systems inside a block, a similar cooperative process may be involved in the building of blocks (a block being defined by clustering of systems).

In the second part of our work we analyze the organization of all the systems within the protein structure, focusing on their reciprocal connections; we build the connectivity matrices aiming to single out the interrelationships between them. These matrices display the number of residues in common between i and j systems and they present a peculiar picture in the distribution of the connectivity within a protein structure (see Figures 1–3), inducing a separation between groups of systems (called blocks) due to the concentration of the system–system connectivity within the block itself. The blocks, once identified, show independence within the protein as far as the noncovalent interactions are concerned, that could lead to the existence of rigid units in the dynamic motions of proteins. The computation of all interactions within a block and between a block and the rest of the protein indicated that most of the nonbonded interactions seem to be concentrated within the blocks (their internal nonbonded energy representing 71% of the total nonbonded energy of a block), indicating that the nonrandom strong interactions control the organization of the whole protein, which seems to be partitioned into the blocks. Other authors have investigated the presence of compact blocks in protein structure[17,18] using a purely geometric criterion. The criteria that led us to the identification of the blocks are of a statistical and energetic nature. We could compare the correspondence of our units (the blocks) with the compact units identified by Go,[17] finding a good correspondence with our results on lysozyme, but not on the globins. Correspondence or discrepancies are difficult to interpret since the criteria used for the identification of units are intrinsically different (i.e., geometry vs energetic and statistical considerations) and in spite of similarities in features of data presentation (i.e., matrices), there is no reason for which units found with our method should overlap those found with geometric criteria (apart from distance dependence of nonbonded interaction).

Our approach, i.e., identification of a pattern of interactions that does not behave according to a random distribution, is comparable with work of Thornton and Singh[5] who scanned all residue–residue interactions in a protein database evaluating if there were geometry of interactions more frequent than expected, extracting structural patterns for residue–residue interactions. Rooman et al.[4] using statistical evaluation tried to extract from a structural data

TABLE IV. Residues Constituting the Blocks, Secondary Structure Elements[19] for *A. limacina* and Sperm Whale Myoglobin

A. limacina Mb			Sperm whale Mb		
Block	Residue	Helix/loop	Block	Residue	Helix/loop
I	1–20	A	I	1–21	A
II	21–28	B	II	21–43	B, C
III	38–50	C, C/D			
IV	51–58	D	III	45–58	C/D, D
V	60–79	E, E/F	IV	59–78	E
VI	81–100	F, F/G	V	79–98	E/F, F, F/G
VII	102–121	G, G/H	VI	100–123	G, G/H
VIII	126–145	H	VII	124–150	H

bank sequence patterns that strongly correlated with certain main chain conformations.

Being that the statistical criterion is the common feature between this paper and the above quoted works, our approach consisting in the use of a general statistical function allowed us to single out nonrandom pairwise interactions in protein structure, with the identification of a general pattern (the "systems"). We subsequently bring into evidence a higher hierarchy organization of the systems, by use of a connectivity matrix (the "blocks").

Our "systems" are somewhat reminiscent of the clusters of Heringa and Argos,[3] but the selection in their work was done on a geometric basis, and with the aim of selecting only a few clusters per protein, whereas in our work the criterion of belonging to a nonrandom network of interactions is priviledged.

In conclusion we put into evidence an intrinsic cooperativity in the construction of the connectivity network and its relative independence on the type of nonbonded interaction leading to almost perfect inclusion of the electrostatic and HB blocks within the vdW ones; moreover the units defined by connectivity between systems (i.e., the blocks) seem to concentrate internally all the nonbonded interactions; these properties suggest that blocks may be folding and functional units.

In fact the finding that the electrostatic blocks are confined within the vdW blocks is not self-evident and may be relevant on the process of folding itself. Since electrostatic interactions are long-range, they are expected to control the intial steps of folding, while the vdW interactions are relevant for the local packing controlling the final folding stages. If initial and final steps of folding would tend to different structural patterns, the folding process could be very slow and inefficient. Proteins should be a selection of polypeptides that exibit fast and efficient folding. We present this conjecture as tentative and worthy of further investigation.

The role of block structure in protein folding and stability could be tested by designing specific mutants (site-directed mutagenesis, truncated proteins) and by investigating the folding of proteins of known structure, with modifications at sites crucial to the structure of a block. We also plan to use molecular dynamics simulations and the new "essential dynamics" approach[8] to evaluate the relationship between concerted and cooperative motions in proteins in structural segments corresponding to the blocks; if this expectation will be fulfilled, we shall proceed to use our data to introduce constraints in molecular dynamics simulations.

At the end of our study we may conclude that by using an unbiased statistical approach it was indeed possible to identify a common pattern in the organization of protein three-dimensional structure (first the "systems" and to a higher hierarchy the "blocks"). The blocks, in spite of having been identified independently from any knowledge of structural features in proteins, seem to show a correlation with fundamental structural properties, such as partioning into domains, secondary strucure elements, and conservation of folds in protein families.

ACKNOWLEDGMENTS

We gratefully thank Prof. M. Brunori and Prof. H.J. Berendsen and the anonymous referees for discussion and stimulating suggestions. We also thank Prof. S. Wodak for allowing us to use the BRUGEL software package. Istituto Pasteur/Fondazione Cenci Bolognetti is acknowledged for support to A. Amadei.

REFERENCES

1. Chou, K.C., Carlacci, L. Simulated annealing approach to the study of protein structures. Protein Eng. 4:661–667, 1991.
2. Luthy, R., MacLachlan, A.D., Eisenberg, D. Secondary structure based profiles: Use of structure-conserving scoring tables in searching protein sequence data bases for structural similarities. Proteins 10:229–2239, 1991.
3. Heringa, J., Argos, P. Side chain clusters in protein structures and their role in protein folding. J. Mol. Biol. 220: 151–171, 1991.
4. Rooman, N.J., Rodriguez, J., Wodak, S.J. Relations between protein sequence and structure and their significance. J. Mol. Biol. 213:337–350, 1990.
5. Singh, J., Thornton, J.M. An automated method for the analysis of the preferred packing arrangements between protein groups. J. Mol. Biol. 211:595–615, 1990.
6. Chou, P.Y., Fasman, G.D. Prediction of protein conformation. Biochemistry 13:222–245, 1974.
7. Garnier, J., Osguthorpe, D., Robson, B. Analysis of the accuracy and implications of simple methods for predicting

the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
8. Amadei, A., Linssen, B.M., Berendsen, H.J.C. Essential dynamics of proteins. Proteins 17:412–425, 1993.
9. Bernstein, T.F., Koetzle, G.J.B., Wiliams, E.F., Meyer, M.D., Jr., Brice, J.R., Rodgers, O., Kennard, T., Shimanouchi, Tasumi, M. The Protein Data Bank: A computer based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
10. Delhaise, P., Bardiaux, M., Wodak, S. Interactive computer animation of macromolecules. J. Mol. Graphics 2:103–106, 1985.
11. Karplus, M., Petsko, G.A. Molecular dynamics simulations in biology. Nature (London) 347:631–639, 1990.
12. Dobson, C.M., Evans, P.A., Radford, S.E. Unnderstanding how proteins fold: The lysozyme story so far. TIBS 19: 31–37 1994.
13. Gilliland, G.L., Quiocho, F.A. Structure of the arabinose binding protein at 2.4 Å resolution. J. Mol. Biol. 146:341–362, 1981.
14. Babu, Y.S., Sack, J.S., Greenhough, T.J., Bugg, C.E., Means, A.R., Cook, W.J. Three-dimensional structure of calmodulin. Nature (London) 315:37–40, 1985.
15. Ploegman, J.H., Drent, G., Kalk, K.H., Hol, W.G.H. Structure of bovine liver rhodanese. I Structure determination at 2.5 Å resolution and a comparison of the conformation and sequence of its two domains. J. Mol. Biol. 123:557–594, 1978.
16. Buck, M., Radford, S.E., Dobson, C.M. A partially folded state of hen egg white lysozyme in trifluoroethanol: Structural characterization and implications for protein folding. Biochemistry 32:669–678, 1993.
17. Go, M. Modular structural units, exons and function in chicken lysozyme. Proc. Natl. Acad. Sci. U.S.A. 80:1964–1968, 1983.
18. Zehfus, M.H., Rose, G.D. Compact units in proteins. Biochemistry 25:5759–5765, 1986.
19. Perutz, M.F., Fermi, G. Hemoglobin and Myoglobin, Atlas of Protein Sequence and Structure, Vol. 2. New York: Oxford University Press, 1981.

APPENDIX A: STATISTICAL EVALUATION

In this section the derivation of the statistical function used to evaluate the distribution of strong interactions within a protein is given.

We have studied separately the three components of the nonbonding energy (i.e., van der Waals, hydrogen bonds, and electrostatic) between all amino acid pairs, defined as the sum over atom pairs. The following analysis has been performed separately on each component.

We define as strong interactions those with an energy lower than the average value, and have studied the distribution of these strong interactions along the sequence. If we consider that the residues which have strong interactions with a given residue are concentrated in a certain number of sequence windows each consisting of N residues, we can evaluate the probability of distributing in one sequence region a number K of strong interactions divided into P segments (a segment is a sequence of adjacent residues in the window having strong interactions and separated from other segments in the window by at least one amino acid having weak interaction, see Scheme 1).

The total number of ways $W(N,K,P)$ of distributing K strong interactions on N ordered residues having P segments can be evaluated as the product of the number of ways of distributing K (strong) interactions over P given segments [that is $(K\text{-}P)$ objects into P boxes, since each segment must include at least one of the strong interactions] and the number of ways of distributing P segments in a window of length N [that is $N\text{-}K\text{-}(P\text{-}1)$ objects into $(P+1)$ boxes because the separation between two segments consists of at least one weak interaction].

In general the number of ways of distributing F equal objects into G boxes is given by the binomial coefficient $(F + G - 1)!/F!(G - 1)!$ so

$$W(N,K,P) = \frac{(K-1)!}{(P-1)!(K-P)!}\frac{(N-K+1)!}{P!(N-K-P+1)!}$$

which is valid if $K \geq 1$; $P \leq K$; $P \leq N + 1 - K$.

If all the interactions of one residue are randomly distributed in sequence regions formed by N residues, then the probability of having K strong interactions divided into P segments in one sequence region formed by N residues is

$$\rho(N; K; P) = p^K q^{(N-K)} W(N; K; P) \qquad (1)$$

where p is the probability that one strong interaction occurs and $q = 1 - p$.

Then if we define $<P>$ as the expectation value for the number of segments in a sequence region of N residues we can write

$$<P> = \sum_{K=1}^{N} \binom{N}{K} p^K q^{(N-K)} \sum_{P=1}^{P'} \frac{W(N; K; P)}{\binom{N}{K}} P \qquad (1')$$

where $P' = \min(K, N + 1 - K)$. Clearly

$$\sum_{P=1}^{P'} W(N; K; P) = \binom{N}{K}$$

and then we can rewrite Eq. (1′) as

$$<P> = \sum_{K=0}^{N} \binom{N}{K} p^K q^{(N-K)} <P>_K \qquad (2)$$

where

$$<P>_K = \sum_{P=1}^{P'} \frac{W(N; K; P)}{\binom{N}{K}} P \qquad \text{with } <P>_{K=0} = 0$$

In order to calculate $<P>$ from Eq. (2) we have to know each $<P>_K$ value.

Since $W(N; K; P)$ is the product of two binomial coefficients depending on the same variables (N, K, and P), we may approximate $<P>_K$ with $P_M(K)$, which is the P value that implies, for a given K, the highest value of $W(N,K,P)$.

Considering also that the *SD of P around $<P>_K$ is very small and that $p << N$, we realize that the probability distribution function of P around $<P>_K$ should be approximately gaussian.

For evaluating $P_M(K)$ we need to calculate the de-

rivative of ln $W(N; K; P)$ with respect to P and to set it equal to zero:

$$\frac{\delta}{\delta P}[\ln W(N;K;P;)] = \ln\left[\frac{(K-P)(N-K-P+1)}{P(P+1)}\right] = 0$$

This implies that

$$\frac{(K-P)(N-K-P+1)}{P(P+1)} = 1 \qquad (3)$$

From Eq. (3) it follows that

$$P_M(K) = \frac{K(N-K+1)}{N+2} \cong <P>_K \qquad (4)$$

From Eqs. (2) and (4) we can consider $<P>$ as the expectation value of a function of K (K = number of strong interactions of a residue in a single sequence region) with a probability distribution of K given by a binomial distribution. Then we can derive $<P>$, which is the expectation value of $<P>_K$, using the standard Taylor method, and obtain

$$<P> \cong \frac{Np(N-Np+1)-Npq}{N+2} \qquad (5)$$

$$<\alpha> = <K/P> \cong \frac{N+2}{N-Np+1} \qquad (6)$$

Here α is the segment length, i.e., a continuous stretch of amino acids having strong interactions with the residue. To evaluate if the total distribution of strong interactions follows the random distribution given by Eq. (1), we can use a χ^2 test on the distribution of K and a Student's t test on the distribution of P, using the statistical variables:

$$\chi^2 = \frac{S_K^2}{\sigma_K^2}(n-1) \quad \text{and} \quad t = \frac{\overline{P}-<P>}{S_p/\sqrt{n}}$$

where S_K is the sample SD of the number of strong interactions in each sequence region formed by N residues, σ_K is the variance of the binomial distribution (approximately gaussian) of K, $\overline{P}$ is the average value of P for the total sample, $<P>$ is the ex-

APPENDIX B. As shown in Figures 2 and 4, the systems found in lysozyme can be grouped in 5 blocks. In this appendix we give the residue composition (sequence number) of each system and their division into blocks.

Block 1	Block 2	Block 3	Block 4	Block 5
4 6 7 8	19 21 22 23 24	1 39 40 41	89 91 92 93	103 105 106 107
5 6 7 8	20 16 17 18	39 40 41 42	90 92 93 94	104 105 106 107
6 8 9 10	20 21 22 23	40 84 85 86	91 92 94 95	107 105 106 107
7 3 4 5	23 19 20 21 22	43 51 52 53	92 88 89 90	100 105 106 107
7 9 10 11	24 26 27 28	45 49 50 51	92 93 94 95 96	108 105 106 107
8 3 5 6	25 27 28 29	46 47 48 49 50 51 52	93 89 90 91 92	108 110 111 112
8 10 11 12	26 28 29 30	50 59 60 61	93 95 96 97	109 110 111 112 113
9 5 6 7	27 23 24 25	51 43 44 45 46	94 90 91 92	110 112 113 114 115 116
9 11 12 13	27 28 29 30 31	52 57 58 59	94 96 97 98	112 108 109 110 111
10 6 7 8	28 23 24 25 26	53 57 58 59 60	95 91 92 93	113 109 110 111 112
10 12 13 14	28 29 30 31 32	54 55 56 57	95 97 98 99	114 110 111 112 113
11 7 8 9	29 25 26 27 28	55 38 39 40	96 92 93 94	119 120 121 122
11 13 14 15	29 31 32 33	57 42 43 44	96 98 99 100	121 123 124 125
12 8 9 10	30 26 27 28	57 52 53 54 55	97 93 94 95	122 123 124 125
13 9 10 11	30 32 33 34	59 50 51 52 53	98 94 95 96 97	123 120 121 122
13 15 16 17 18	31 27 28 29	59 60 61 62 63 64	98 99 100 101	124 121 122 123
14 10 11 12	31 33 34 35	60 62 63 64	99 95 96 97 98	125 121 122 123
15 11 12 13 14	32 28 29 30	61 71 72 73	101 97 98 99 100	
38 2 3 4 5	32 34 35 36	62 59 60 61		
	33 29 30 31	62 73 74 75		
	33 34 35 36 37 38	63 58 59 60 61 62		
	34 30 31 32 32	63 74 75 76		
	35 31 32 33	64 58 59 60		
		64 78 79 80		
		65 78 79 80		
		69 70 71 72		
		74 62 63 64 65		
		74 75 76 77 78		
		80 64 65 66		
		82 78 79 80		
		82 83 84 85		
		83 80 81 82		
		84 40 41 42 43		
		88 90 91 92		

pectation value of P given by Eq. (5), S_p is the sample SD of P, and n is the total number of sequence windows of length N considering regions with at least one strong interaction. It should be noted that a proper choice of the value N (the window length) is crucial for a correct statistical evaluation. We choose to accept only windows of length N with at least one strong interaction such that we do not consider sequence regions with a zero probability of having strong interactions with the ith residue (regions not available for nonbonded interactions with the ith residue).

Applying this analysis to the set of proteins that we have studied, and considering all strong interactions, we found that there is no way to have both χ^2 and t values consistent with a random distribution. Only if we choose to reject all strong interactions belonging to segments with $\alpha \geq 3$ and setting $N \cong 20$ we obtain, on the contrary, χ^2 and t values really consistent with the discussed random distribution (probability $>5\%$).

From these results it follows that strong interactions of a residue with three or more amino acids in sequence ($\alpha \geq 3$) are absolutely not generated by a random distribution while the other strong interactions seem to be really randomly distributed in the $N \cong 20$ residues sequence windows.

Following this statistical analysis which brought into evidence a nonrandom pattern of strong interactions, we decided to identify and study the sets of residues in a protein structure behaving according to the pattern itself (see Scheme 2).