Protein Engineering vol.6 no.8 pp.801–810, 1993

Regularities in interaction patterns of globular proteins

Adam Godzik[1], Jeffrey Skolnick and Andrzej Kolinski

Department of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, CA 92037, USA

[1]To whom correspondence should be addressed

The description of protein structure in the language of side chain contact maps is shown to offer many advantages over more traditional approaches. Because it focuses on side chain interactions, it aids in the discovery, study and classification of similarities between interactions defining particular protein folds and offers new insights into the rules of protein structure. For example, there is a small number of characteristic patterns of interactions between protein supersecondary structural fragments, which can be seen in various non-related proteins. Furthermore, the overlap of the side chain contact maps of two proteins provides a new measure of protein structure similarity. As shown in several examples, alignments based on contact map overlaps are a powerful alternative to other structure-based alignments.

Key words: contact maps/distance matrices/globular proteins/ Monte Carlo simulation

Introduction

Although protein structures have been studied for more than 30 years, we are still far from understanding them. For example, there are groups of proteins with substantial structural similarity and little or no sequence homology. There are arguments suggesting that the number of possible protein topologies is limited (Finkelstein and Ptitsyn, 1987) and the existence of such groups is often cited as an argument in favor of this hypothesis. Still almost every new example comes as a complete surprise. Newly refined tools for identifying structural similarity (Arytmiuk *et al.*, 1990; Nussinov and Wolfson, 1991; Vriend and Sander, 1991) have substantially increased the number of known examples of such 'structural twins', but do not answer fundamental questions concerning their existence. Why do two different sequences fold to very similar structures? What is so special about their common structure, that it can accommodate two vastly different sequences? Can we learn anything from such examples of unexpected structural similarity? Perhaps there are other sequences which would also fold into this structure? If so, can they be predicted? In this paper, we will attempt to address these questions by presenting a new way of looking at protein structures. By viewing a protein structure as a set of interactions, rather than as a set of points in space, one can discover a number of unexpected similarities between proteins, both on the global and on the local level. This perspective provides a deeper understanding of protein similarity than the vague notion of topological equivalence, and might eventually explain the existence of structural twins, or might even predict their existence before the structure of both becomes known.

There are strong arguments that the 'interaction'-based picture of protein structure does indeed introduce a new quality into our view of proteins. In a recently formulated inverse folding algorithm (Godzik and Skolnick, 1992; Godzik *et al.*, 1992), protein structure was described as a 'topology fingerprint', consisting of the interaction pattern for the whole protein and the buried/exposed classification for every residue. An extremely simple, binary definition of interactions, based on the minimal distance between side chain heavy atoms, was used in defining topological fingerprints. Simplified energy calculations, based on the fingerprint of one protein, can correctly recognize the protein's own sequence and in many cases, it can also recognize unrelated proteins which share the same topology. Several examples include phycocyanin-globin, several TIM barrels, the two domains of rhodanese, Leu/Val/Ile and glucose/galactose binding proteins, interleukin-2 and interleukin-4, granulocyte macrophage colony stimulating factor and many others.

Further progress in this direction depends to a large extent on our understanding of the rules and regularities in the interaction patterns in proteins. A number of important questions can be raised. For instance, a number of other inverse folding algorithms were formulated which use other definitions of interactions to estimate the protein energy: residue environment (Bowie *et al.*, 1991; Luethy *et al.*, 1992), binary interactions based on the minimum distance between backbone or Cβ atoms (Maiorov and Crippen, 1992) or continuous interactions based on the distances between Cβ positions of amino acid side chains (Jones *et al.*, 1992; Sippl and Weitckus, 1992; Bryant and Lawrence, 1993). The inverse folding approach is based on the assumption that the interactions in similar structures are similar. But how do the structure similarities as defined by various distance-based measures compare to each other? Or indeed are they any different from the standard root mean square deviation after optimal superposition (Kabsch, 1978) measure?

Protein folding simulations on simplified models provide another motivation for the need to better understand the interaction patterns in proteins. The *de novo* prediction of the structure of simple motif proteins at the level of 2.25–4.5 Å becomes possible after the introduction of additional energy terms favoring native-like contact patterns (Kolinski *et al.*, 1993). Even the fine differences between the molten globule and the native-like protein structure can be captured in these simulations (Kolinski *et al.*, 1993) and rely on the presence of a subset of templates favoring native-like packing. Again, there are important questions that have to be answered before further progress can be achieved in such simulations. How regular are local interaction patterns and can they and should they be reproduced in folding simulations? Do they depend on the protein structural class or are they specific to a particular protein? These and other questions will be addressed here.

The way of describing the protein structure as a 2-D matrix, where the element (i,j) carries information about the interactions of residues i and j, is quite old (Philips, 1970). Most often, the matrix elements are distances between Cα atoms (Philips, 1970; Ooi and Nikishawa, 1973; Rossman and Lilas, 1974; Kuntz, 1975; Liebman *et al.*, 1985; Richards and Kundrot, 1988) and the matrix is appropriately called a distance matrix. Because the 2-D information can be displayed in a diagram resembling a map,

it was also called a distance map or a distance plot. Another possibility is to store in the matrix the simplified yes/no (Godzik and Sander, 1989) or complex (including van der Waals energy and many other contributions) (Scharf, 1989) information about residue interactions, which is then called a contact map. Interaction or distance-based protein description was introduced to identify secondary structure elements and regions likely to act as folding nuclei (Philips, 1970) and later supersecondary structure elements (Kuntz, 1975; Richards and Kundrot, 1988) and protein domains (Rossman and Lilas, 1974). Both the similarity and differences between protein structures are easily studied using contact maps; examples include identifying changes in structure upon mutations (Liebmar *et al.*, 1985) or during molecular dynamics simulations (Nikishawa and Ooi, 1974) and similarities of the distance maps of homologous proteins (Kikuchi, 1992). The prediction of distance maps has been attempted for protein domains (Kikuchi *et al.*, 1988) and whole proteins (Ycas, 1990). Other applications include studying mutations in a family of related proteins as a function of constraints imposed by contacts (Godzik and Sander, 1989) and assessing the quality of the predicted protein structure as measured by conforming to contact preferences (Sander and Vriend, 1990). At present, these methods are not nearly as popular as other methods of describing protein structure.

In this paper, we would like to examine whether new regularities can be discovered by using a contact map description of protein structure. This indeed turns out to be the case and several examples will be presented in the first part of the paper, together with a discussion of their connection with well known secondary and supersecondary structure elements. The second part will describe several new tools developed in our group that use contact maps and their applications to the analysis of protein structure. Contact map-based alignments will be described along with a detailed comparison with more traditional structural alignments.

Methods

Contact maps and distance matrices

Simplified residue−residue contact maps can be built from the full protein coordinates, as deposited in the PDB (Bernstein *et al.*, 1977; PDB, 1992). We use here one of the simplest variants of the contact map (the term interaction pattern will be used interchangeably) where the C^A_{ij} matrix element is set to 1 if residues i and j in protein A are in contact and is 0 otherwise. In the definition adopted here, two amino acid residues are defined as being in contact when any two heavy side chain atoms, one from the first, the other from the second side chain, are closer than 5 Å. Such a map carries only yes/no information about the residue−residue interactions and any interactions over distances larger than 5 Å are ignored. An interaction definition based on the side chain atom distances, rather than the Cα distance, was adopted to emphasize side chain interactions, rather than backbone positions. It will be shown later in the paper that the side chain interactions tend to be preserved much more than the backbone positions. The chosen contact cut-off distance represents the best trade-off between the specificity (increasing with smaller cut-off) and the number of contacts (increasing with larger cut-off) (Godzik and Sander, 1989). It is possible that some other choices of parameters will yield a different variant of a contact map, which will be better suited for some purposes. As will be discussed in the Conclusions, several extensions of the present model are planned. However, most of the qualitative results of the present paper are unlikely to be changed by any change of definitions.

In the distance matrix, matrix element R^A_{ij} is equal to the $C\alpha_i - C\alpha_j$ distance. The Cα distance matrix is more closely related to the r.m.s. measure of structure similarity (Kabsch, 1978), since it also puts the emphasis on the backbone positions. Presentation of the results obtained with the Cα distance map is included here for comparison with other published work.

Both the contact maps and the distance matrices can be visualized as 2-D maps or plots, either in black and white (simplified maps; see Figure 1a) or in shades of grey (full Cα distance matrix, as seen in Figure 1b). Distinct protein topologies have their specific contact maps and with some experience, both can be analyzed and studied, yielding the same and in many cases more information than other representations. The best illustration can be provided by comparing contact maps or distance matrices of two similar proteins. The high similarity between Figure 1a (1b) of sperm whale myoglobin and Figure 1c (1d) of β-chain of human hemoglobin respectively, is obvious even after casual inspection and indeed the corresponding structures are very similar.

Measuring similarity between contact maps

At this point, it is necessary to introduce a measure that quantifies the similarity of two protein structures. Traditionally, the r.m.s. distance between Cα positions after optimal superposition (Kabsch, 1978) was used for such purposes. The overlap of the two contact maps, as defined by the number of contacts between equivalenced residues in two proteins that are simultaneously present in both structures can also be used. Here is an example of such a calculation. Let us start with the alignment between sequences A and B:

$$\begin{array}{cccccccccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & - & - & A_7 & A_8 & A_9 & (1) \\ B_1 & B_2 & B_3 & - & B_4 & B_5 & B_6 & B_7 & B_8 & - & B_9 & \end{array}$$

We shall describe this alignment by the function $AB(i) = k$, which maps residue i of sequence A, to residue k in sequence B. So, for instance in the example above, $AB(5) = 4$ and $AB(7) = 8$; unaligned residues are assigned the value of the last aligned residue from the other sequence, i.e. $AB(4) = 3$. This last choice is important for use in a Monte Carlo search protocol in alignment space (see next section). The contact map-based score for the alignment is calculated by equation (2), where the summation is taken over all aligned pairs.

$$S_c(AB) = \frac{1}{N_c} \sum_{i>j} C^A_{ij} C^B_{AB(i)AB(j)} \qquad (2)$$

where C^A_{kl} and C^B_{kl} are the binary contact maps of proteins A and B respectively, $C^B_{AB(i)AB(j)}$ is a contact map of a protein B, but numbered according to the alignment AB. Finally, the normalization factor, N_c is the maximal possible overlap between the two maps (it is equal to the number of contacts in the smaller of the two proteins). Figure 2 shows the overlap between the contact maps of sperm whale myoglobin and human hemoglobin, chain β. The contacts specific to the sperm whale myoglobin are indicated by open boxes, the ones specific for the human hemoglobin are indicated by boxes with diagonals and those common to both structures are black [the number of black points is equal to the alignment score, $S_C(AB)$].

A slightly modified equation (2) can be used to calculate the alignment score based on the distance matrices, R^A_{kl}. In this case, the absolute value of the difference between the distances between residues i and j in protein A and between residues $AB(i)$ and $AB(j)$ in protein B is summed over all possible pairs of residues:

$$S_D(AB) = \frac{1}{N_{pairs}} \sum_{i>j} \left| R^A_{ij} - R^B_{AB(i)AB(j)} \right| \qquad (3)$$

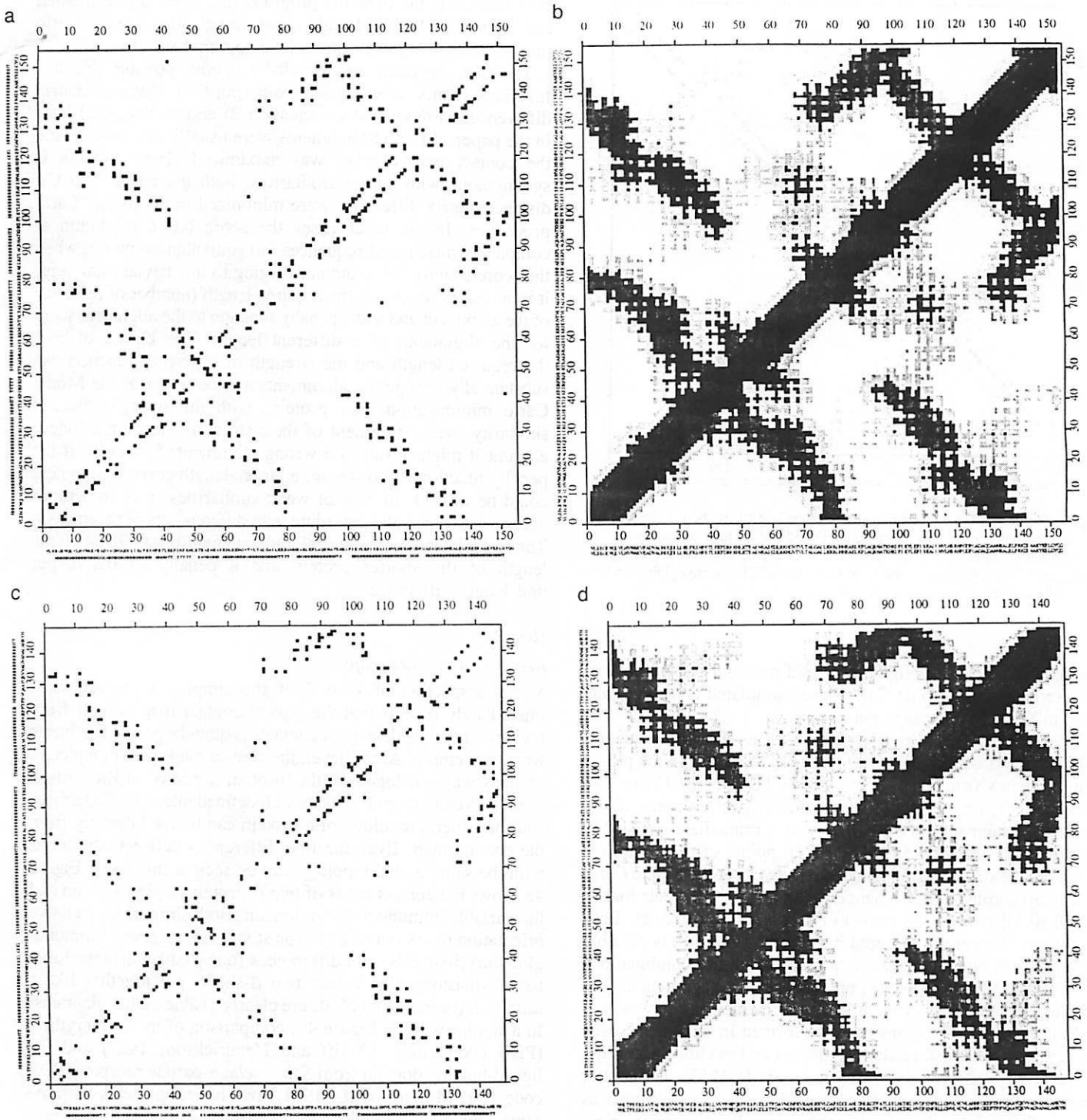


Fig. 1. An example of a contact map and distance matrix. (**a**) The contact map of the myoglobin (1mbc) structure; the black squares give information about the interaction between residues. (**b**) The distance matrix of the plastocyanin structure; the shades of grey are proportional to the distance between the Cα atoms (black, <5 Å; white, >20 Å). (**c**,**d**) Equivalent figures for the β chain of human hemoglobin (hhhb).

The notation used in equation (3) is analogous to that in equation (1), only the normalization factor N_{pairs} is now equal to the total number of pairs used in the calculations. Non-aligned residues are explicitly excluded from the calculations. To shift the emphasis in comparing full Cα−Cα distance matrices to the interacting residues, distances >15 Å are excluded from the calculations.

It is possible to assess the quality of an existing sequence alignment between two proteins by comparing their structures or, vice versa, structure comparisons can be used as a basis to build an alignment. Any of the measures of structure similarity can be used for this purpose, but we will concentrate on the contact map overlap, as defined by equation (2).

Contact map alignment algorithm

It is possible to build an alignment protocol that optimizes the contact map overlap. Since it is a highly non-local function of the alignment [it depends on the alignment in two different places $AB(i)$ and $AB(j)$; see equation (2)], standard dynamic programming

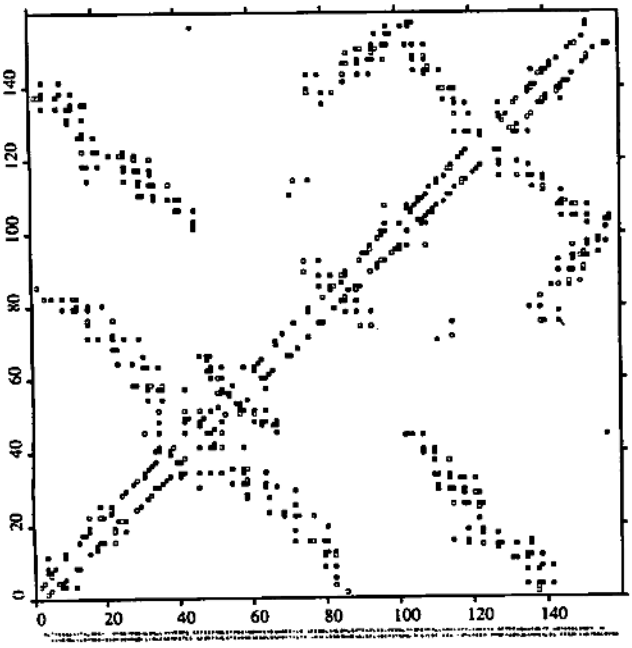


Fig. 2. Overlap of the contact maps, corresponding to the best alignment of the sperm whale myoglobin (open squares) and the β-chain of human hemoglobin (squares with diagonals). The overlapping contacts are in black. There are 159 overlapping contacts in total, out of 260 in myoglobin and 222 in hemoglobin structures (72% overlap).

methods for sequence alignment could not be used. Instead, in the present paper, Monte Carlo-based simulated annealing was used to find the best alignment.

The alignment between two sequences, as shown in equation (1), can be vizualized as a set of points $\{[i,ABH(i)]\}$ in the plane. Joining points from this set by vectors [1,0] or [0,1] one can obtain a path, such as shown in Figure 3a. Convex points along the path (combinations of the [0,1] vector immediately followed by the [1,0]) are the original $[i,AB(i)]$ points of the alignment (see Figure 3a). The set of elementary moves for the Monte Carlo calculations consists of exchange of the {[0,1],[1,0]} pair for the {[1,0],[0,1]} pair (down move) or vice versa (up move). Both elementary moves are presentd in Figure 3b and c. It is possible to implement other long range moves, which are combinations of several elementary moves; however, every alignment can be reached from any other by a series of elementary moves.

Monte Carlo simulations were performed in alignment space starting from an initial random alignment and modified according to the Metropolis algorithm (Metropolis *et al.*, 1953). The contact map overlap was used as a scoring function. The system was repeatedly heated and cooled to escape from local energy minima (i.e. a simulated annealing protocol was followed). In a typical application, it was necessary to perform 10^5-10^6 elementary moves in order for the score to converge. This represents 10–20 min of c.p.u. time on a SUN SPARC2+ workstation. This type of minimization does not guarantee that the global extremum would be found; therefore, all alignments presented in this paper could possibly be improved. To get a reasonable assurance of the stability of the final alignment, multiple minimizations from different starting points were always performed. In the cases presented here, the same solution was repeatedly found. Also, for test purposes, the same Monte Carlo alignment was used to obtain sequence-based alignments, where the exact solution can be found using the dynamic programming method (Needleman and Wunsch, 1970). In all cases tested, the Monte Carlo procedure converged to the unique, optimal solution.

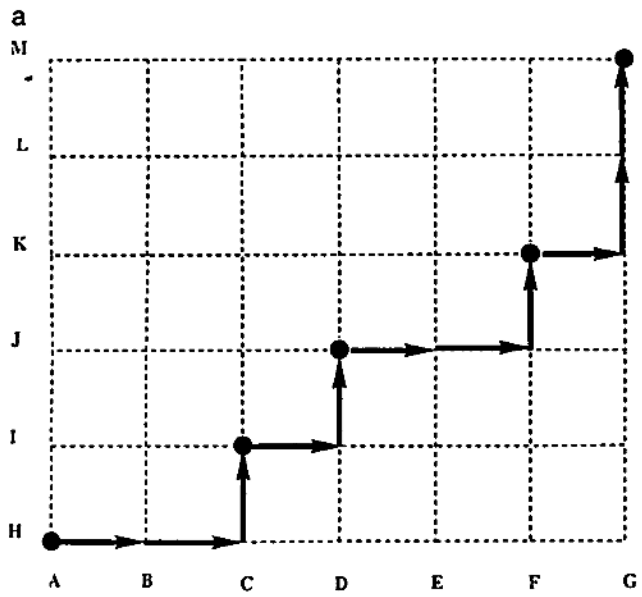
The same procedure was repeated with other possible objective functions (r.m.s. after optimal superposition, distance matrix difference, as described by equation 2) and as discussed later in the paper, different alignments were usually obtained. While the contact map overlap was maximized (large overlap is synonymous with strong similarity), both the r.m.s. and Cα distance matrix difference were minimized in the Monte Carlo procedure. In the latter cases the score has a minimum at completely misaligned sequences (no equivalenced pairs), when the score is zero. To avoid converging to this trivial minimum, it is necessary to specify the required length (number of residues) of the alignment and add a penalty function to the alignment score for the alignments of a different length. The choice of both the required length and the strength of a penalty function can substantially change the alignments and behavior of the Monte Carlo minimization. For proteins with the strong structural similarity over a fragment of their structure, forcing a longer alignment might result in a wrong alignment. Similarly, if the penalty function is too strong, a larger-length correct alignment could be missed. In case of weak similarities, it is sometimes necessary to repeat the procedure with different sets of parameters. The default choice for the alignment length was 70% of the length of the shorter protein and a penalty of 0.1 Å per unit length difference.

Results

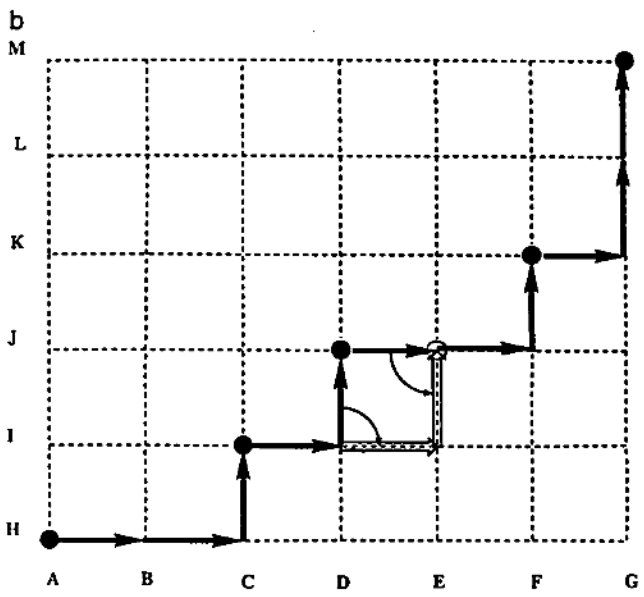
Regions of contact maps

Visual inspection of several of the simplified contact maps immediately reveals that the typical contact map is built from regions where the density of contacts is quite large and fragments which are empty. At this level, the contact maps give us a picture in many ways analogous to the 'protein cartoons' of Richardson (1981). Protein fragments have well defined interaction interfaces and the general topology of a protein can be read directly from the contact map. Even the fine differences between structures with the same global topology can be seen at this level. Figure 4a shows the contact maps of two β-proteins—plastocyanin and the variable immunoglobulin domain. Both similarities [relative orientation of six out of eight (plastocyanin) or seven (immunoglobulin) β-strands] and differences [one β-sheet arranged in a barrel (plastocyanin) versus two β-sheets put together like a sandwich (immunoglobulin)] are clearly visible on the diagrams. In a similar way, in Figure 4b, comparison of myohemerythrin (PDB code 2mhr; Sheriff and Hendrickson, 1987) and the ligand-binding domain from *Salmonella* aspartate receptor (PDB code 1lig; Milburn *et al.*, 1991), two helical proteins with the same four bundle topology (visible on both contact maps as a characteristic set of parallel strips), immediately shows the main difference in packing—cross interactions between helices 1 and 3 (in 2mhr) versus 2 and 4 (in 1lig).

A more careful analysis of the interactions within the 'dense regions' identifies a number of patterns, typically spanning seven to nine residues, that are repeated in various, unrelated proteins. Examples of such patterns are presented in Figure 5, where the parallel and antiparallel $\beta-\beta$ sheet pattern from several proteins is shown (Figure 5a and b), as well as the coiled-coil $\alpha-\alpha$ antiparallel pattern (Figure 5c). In all cases, the characteristic contact map pattern of one interaction area is repeated with an accuracy >80% in the structural database. For instance,



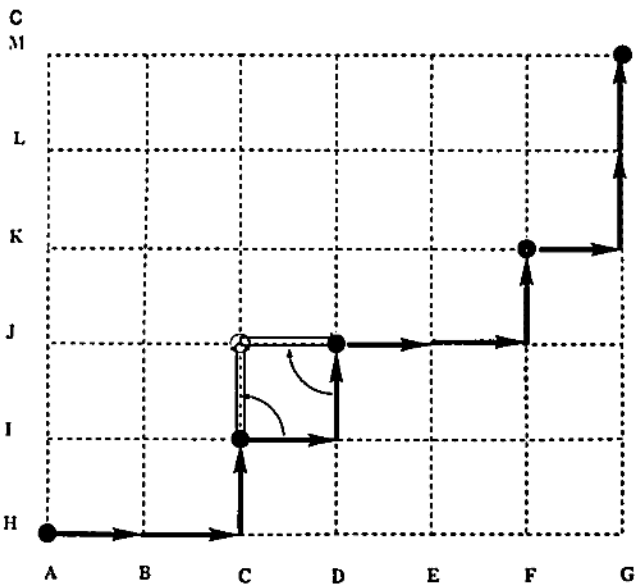
A B C D E F . G
H . I J . K L M



A B C D E F . G
H . I J . K L M

changes to

A B C D E F . G
H . I . J K L M



A B C D E F . G
H . I J . K L M

changes to

A B C D E F . G
H . I J . . K L M

Fig. 3. (a) The sequence alignment built from the basis vectors and a set of elementary moves for the Monte Carlo algorithm in the alignment space. (b) A 'down' move, when the convex point (D−J) is moved to a 'concave' point (I−E), to point. (c) The opposite move.

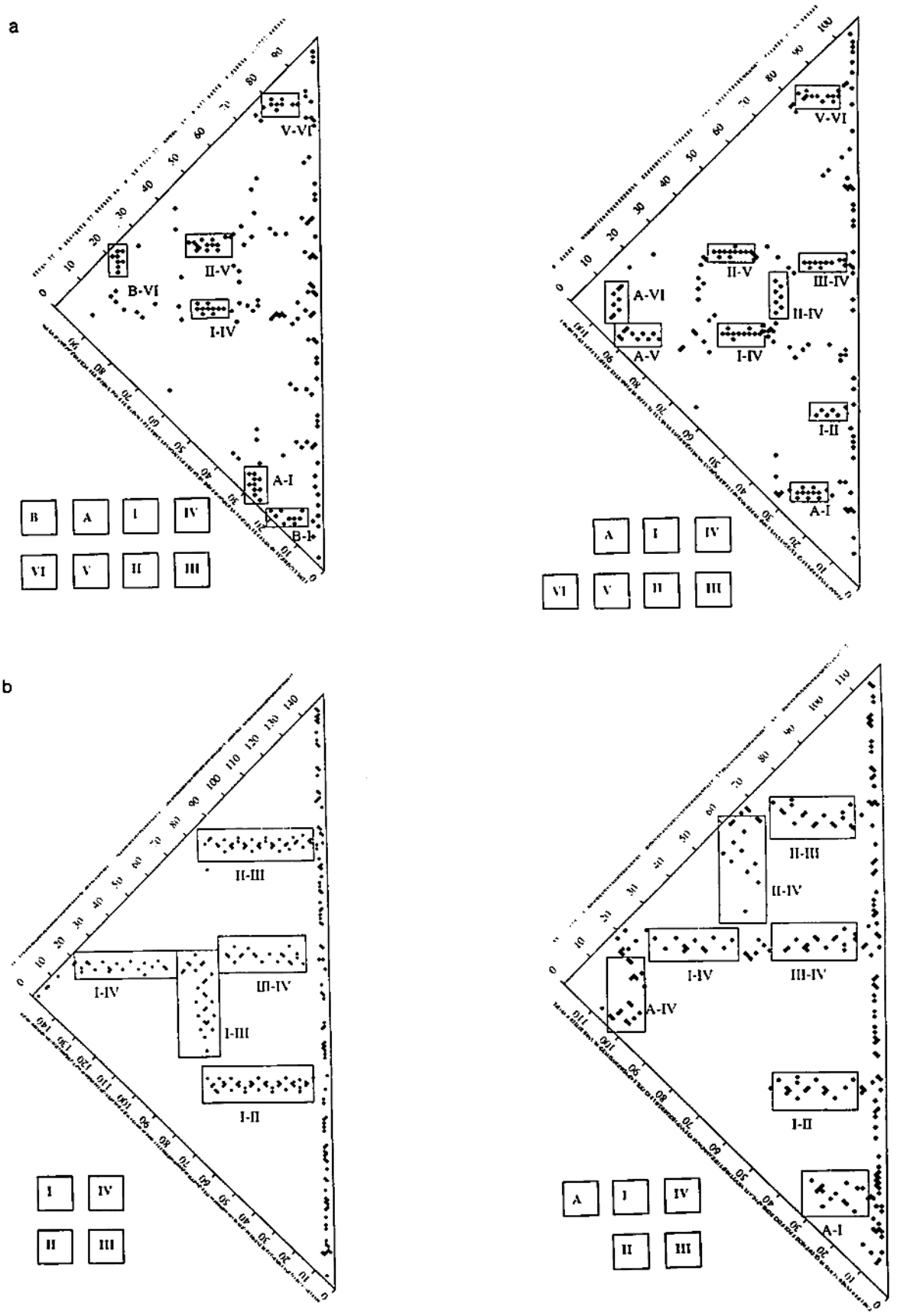


Fig. 4. Regions of interactions on protein contact maps. The contact regions are depicted as boxes and denoted by the number of interacting strands. (a) Two β-proteins, plastocyanin (1pcy) and variable immunoglobulin domain (2rhe). Strands in the immunoglobulin domain are numbered in such a way that the similarity to plastocyanin structure is more visible. Note the presence of additional contact regions. (b) Two four-helical bundle proteins: myohemorythrin (2mhr) and ligand-binding domain from *Salmonella* aspartate receptor (1lig). Again note the presence of the additional contact regions.

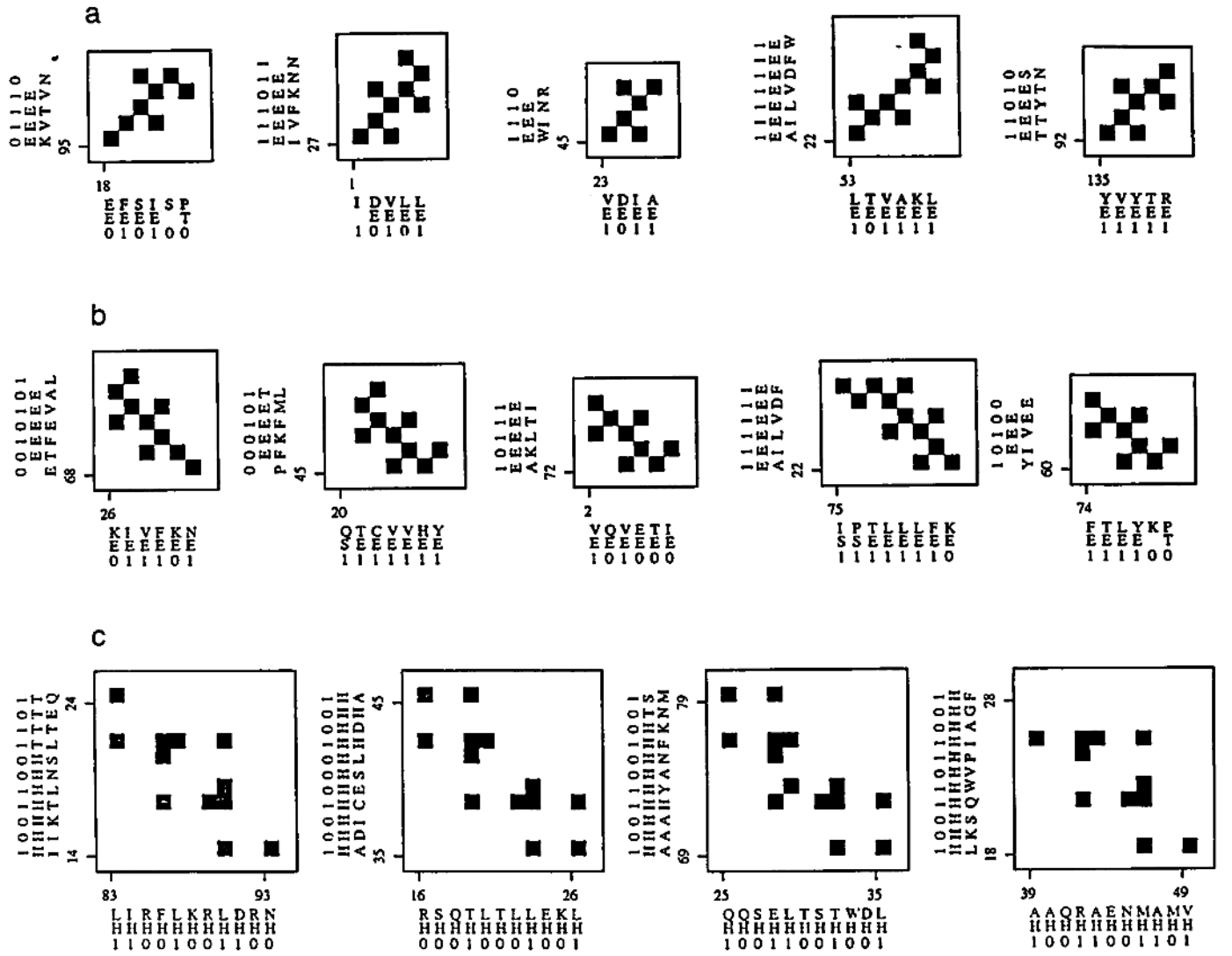


Fig. 5. Examples of the contact patterns of the different types of interaction regions from several proteins. The superset with (a) parallel $\beta-\beta$ sheet pattern, (b) antiparallel $\beta-\beta$ sheet, (c) coiled-coil $\alpha-\alpha$ antiparallel.

the parallel $\beta-\beta$ sheet pattern is repeated more than 400 times in the contact map database extracted from 258 high quality, unrelated proteins.

The examples presented here can be classified as characteristic patterns of interactions between secondary structure elements. However, there are systematic differences between the extent of secondary structure elements as defined on the basis of the local (H-bond patterns or Φ,Ψ values) criteria and the contact map-based definition. The characteristic interaction patterns tend to continue for longer than the locally defined secondary structure elements and also they tend to incorporate turn residues. Examples are seen in Figure 4a and b, where strong patterns extend beyond boxed regions into turns, which are identified by the Kabsch–Sander (Kabsch and Sander, 1983) secondary structure assignments which are also plotted along one of the axes. The full library of interaction patterns is under construction in our laboratory (M.Milik, A.Godzik, A.Kolinski and J.Skolnick, manuscript in preparation).

In total, ~50% of all contacts in the test database belong to one of the already classified interaction patterns. Another 20–30% are local interactions, defined as interactions between residues close (±5) in the sequence. It is an interesting coincidence that the sum of these two numbers is close to the level of contact similarity between identical or very closely related proteins (mean ~80%) and that the other 20% of contacts can be classified as a 'noise'. An interesting question, which will be investigated in the future, is whether an improved contact definition or an extension of the pattern library might increase the percentage of contacts which can be easily classified as being 'regular'.

Contact map-based alignments

As described in the Methods section, it is possible to compare two protein structures by building an alignment that maximizes the overlap between the contact maps. Allowing for the insertions/deletions in both sequences, it is possible to superimpose the contact maps of topologically related proteins at the level of 50–70% contact overlap. The first example is shown in Figure 2 and two more examples are presented in Figure 6a and b, that show the superposition between the contact maps of two β (plastocyanin–azurin) and two α (interleukin-2–interleukin-4) proteins, with each pair sharing the same topology. The sequence similarity expressed as the percentage of identical residues in the sequence-based alignment is close to random: 22.5% for the plastocyanin–azurin case, 21.8% for the Il-2–Il-4 alignment; both alignments were obtained with the BESTFIT program from the GCG package (Genetic Computer Group, 1991). The contact

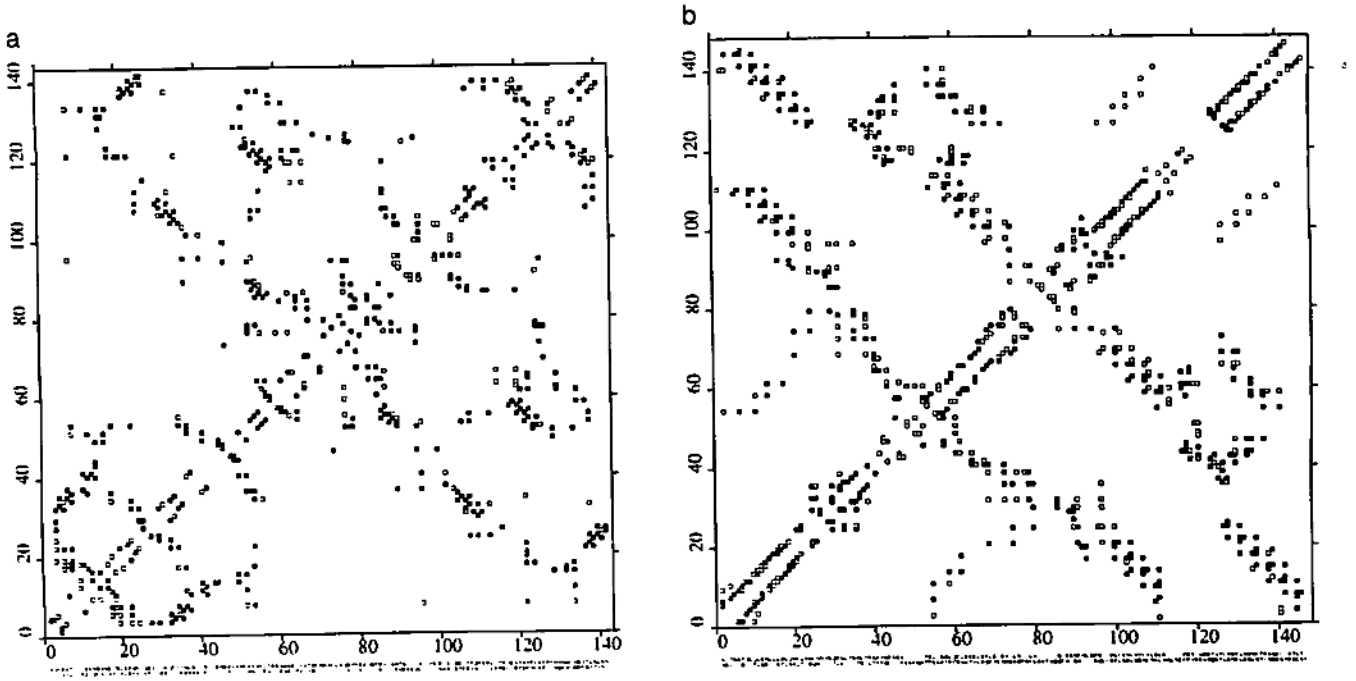


Fig. 6. Examples of the superposition of the contact maps. (a) Two β-proteins, plastocyanin and azurin, (b) two α-proteins, interleukin-2 and interleukin-4.

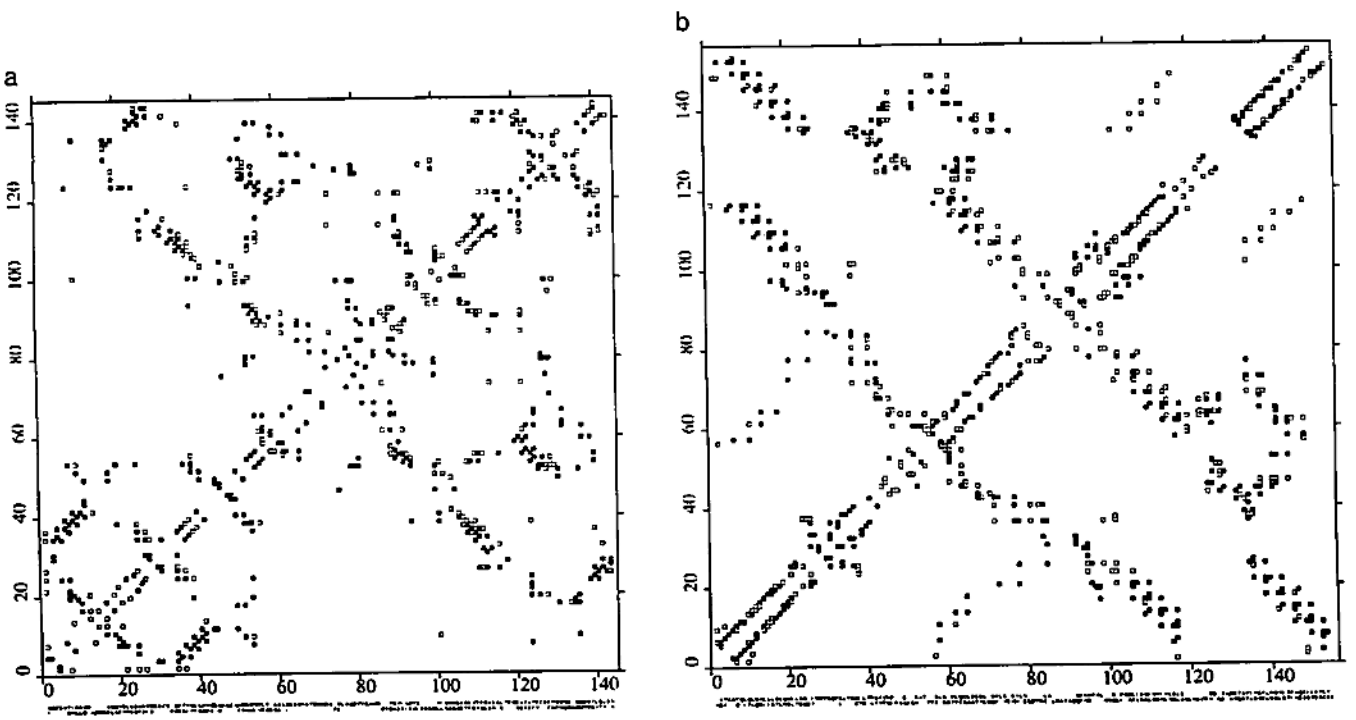


Fig. 7. Comparison between the r.m.s. and the contact map similarity. The same proteins as shown in Figure 6 are now aligned according to a r.m.s. after optimal superposition of Cα positions. As seen in these figures, important interaction regions are misaligned.

map overlap in both cases is substantial: 111 contacts overlap out of 242 in azurin and 172 in plastocyanin (65% overlap) and 126 out of 250 in interleukin-2 and 257 in interleukin-4 (50%). This should be compared to ~80% overlap between independently solved structures of highly homologous proteins (for example, there is 85% contact overlap between green alga and poplar plastocyanin and 65% overlap between sea hare and sperm whale myoglobin) and ~20–30% overlap between proteins with different topology.

Interestingly, the contact map overlap-based alignment is close to, but by no means identical with either the Cα, r.m.s. or the sequence-based alignment. For instance, in the above examples for alignments optimized for the contact map overlap, the r.m.s. for the plastocyanin azurin alignment is equal to 6.2 Å r.m.s. on 87 equivalenced positions and for the Il-2 and Il-4 alignment it is 2.9 Å for 104 positions. In both cases, it is possible to build different alignments, optimized for the r.m.s. measure of similarity (3.5 Å for 85 residues and 2.1 Å for 96 residues

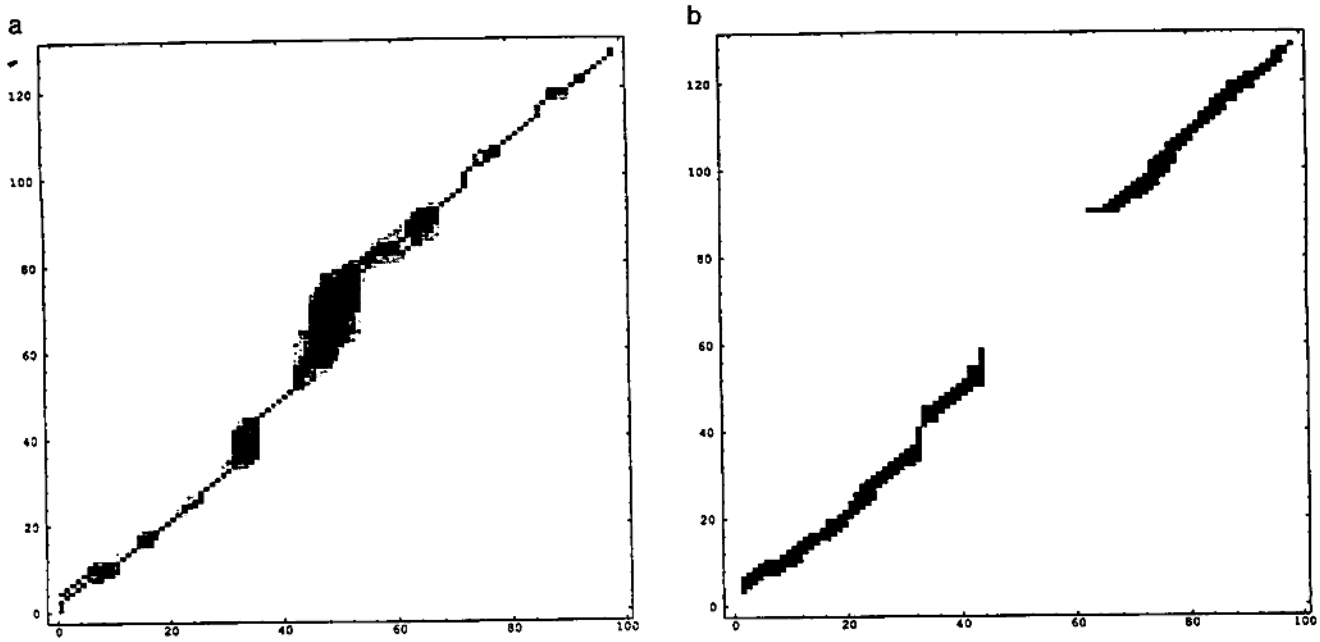


Fig. 8. Density of suboptimal alignments for (a) contacts map overlap- and (b) r.m.s.-based alignments.

respectively). On the other hand, r.m.s. optimized alignments show very little contact overlap (20 and 40% respectively for the examples above). The aligned contact maps for both r.m.s.-based alignments are shown in Figure 7a and b. In both cases, important interaction regions are misaligned, with shifts of up to two residues. Comparing different structure alignments derived on the basis of different measures of similarity, it is clear that such alignments are not unique, but strongly depend on the measure of structure similarity used to obtain them.

It is also interesting to study and compare the published model alignments, which are often used as a standard to verify sensitive sequence alignments. Plastocyanin and azurin were aligned by Adman (1985) and Lesk and Chothia (1982), phycocyanin and globins by Pastore and Lesk (1990), and immunoglobulin domains and superoxide dismutase by Richardson *et al.* (1976). As expected, it is possible to optimize each of these alignments with respect to different measures of structure similarity, obtaining alignments varying at almost all positions. Alignments based on the detailed analysis of packing inside the protein core (Lesk and Chothia, 1982; Pastore and Lesk, 1990) are closest to contact map-based alignments; the latter tend to equivalence larger fragments of both sequences extending the alignments into the loop regions.

The meaning of structure similarity

In the classical problem of protein homology, where the objective of an alignment is to study the evolutionary relation between proteins, we expect that for related proteins there is a unique solution. Here, we are interested in identifying the interactions responsible for a particular fold, so it is clear that by varying the definition of interactions one can arrive at different alignments.

All three types of structural alignment are essentially the same for very closely related proteins, i.e. when the sequence homology is >50%, no significant differences are observed. At <50% homology, the alignments start to differ, as seen by comparing examples in Figures 6a and 7a (6b and 7b) and it might be difficult to decide which alignments are better. However, the following example is very suggestive. Sequences of Il-2 and Il-4 were aligned based on all three structural methods (Cα r.m.s., Cα distances and contact map overlap) and each alignment was used to generate the sequence profile by the method of Gribskov *et al.* (1987). Each of the profiles was used to scan the SWISSPROT 20 protein sequence database. Only the contact-based profile was able to pick up other cytokines (interleukin-6, growth hormone, prolactin, granulocyte macrophage colony stimulating factor, interferon-β). All are very weakly related to both interleukins (Bazan, 1990) and have similar structure. This strongly suggests that the contact map-based alignment highlights the conservation of interactions within each protein and can identify these interactions in other proteins.

Another point, important for both sequence and structure-based alignments, is the uniqueness of the alignment. There are alignments for which the scores are lower, but close to the optimal one. Because structures contain errors and measures used to compare them depend on many arbitary decisions, it is necessary to understand the nature of the errors and precision of the alignments. For this purpose, long, equilibrium Monte Carlos were performed and statistics were collected for the best score that could be obtained for the alignment containing the equivalence between every residue pair. The resulting plot is analogous to the density of the suboptimal alignments discussed in the literature for the sequence alignments (Vingron and Argos, 1990), but obtained here using different methods. As seen in Figure 8, for the case of plastocyanin–azurin alignment, in some regions the alignments are well defined (the central peak is narrow, i.e. little or no shift is allowed) and in some others it is not (the central peak is wide, i.e. large shifts are allowed). As expected, the alignment in the core regions is usually well defined and indeed this method could be used as an automated definition of the protein core. Again, there is a distinct difference with both the r.m.s. or Cα distance-based alignment. In the latter, there is no difference in the width of suboptimal alignments between the core and outside regions.

Conclusions

Interactions between amino acid side chains and their characteristic patterns offer a new language for studying similarities between

protein structures. The structural similarity between more distantly related proteins is not uniquely defined. Depending on the similarity criteria used, alignments with relative shifts of three to four residues can be obtained. Contact map-based alignments are closest to the alignments based on the detailed analysis of packing inside the protein core (Lesk and Chothia, 1982; Pastore and Lesk, 1990) and they seem to capture the sequence similarity between weakly homologous proteins. Clearly, more work is necessary to understand the differences between various types of structural alignments.

There is a relatively small number of well defined contact patterns for interactions between protein fragments and the same patterns can be found in different proteins. The classification of local interaction patterns opens up the possibility of predicting the position (where it starts and where it ends) and type (for example, α or β) of interacting regions for proteins with unknown structure. Using empirical energy functions, it should be possible to calculate the energy of various contact patterns at a given position in the protein and find the one with the lowest energy. It should also be possible to build contact maps on the basis of a schematic topology (Godzik *et al.*, 1993), even for novel topologies for which no examples have so far been found. With better understanding of these patterns, they can be incorporated in the empirical energy function used in protein folding simulations. In the first examples of the latter approach, *de novo* folding simulations of a few small proteins with simple topologies led to predicted structures with accuracy at the level of 2.25–3 Å from experimentally determined structures (Skolnik *et al.*, 1993). Also a simulation of GCN4 leucine zipper was performed and has resulted in a predicted structure with an 0.6 Å r.m.s. from the native structure (M.Vieth, A.Kolinski, C.L.Brooks, III and J.Skolnick, submitted). In both applications, additional energy terms based on local interaction patterns proved crucial to achieve a high quality of predicted structures.

Inverse folding methods, based on the contact map description of protein structure (Godzik and Skolnik, 1992; Godzik *et al.*, 1992) can extend the application of sequence analysis methods. The applicability of these methods would be significantly improved with better understanding of interactions with proteins. Comparing contact maps of known structural twins and building multiple alignments of the contact map could help to solve this problem.

Acknowledgements

This research was supported in part by grant No. 2 PO1 GM38794 of the Division of General Medical Sciences of the National Institutes of Health.

References

Adman,E.T. (1985) *Metalloproteins*, **I**, 1–42.
Arytmiuk,P.J., Rice,D.W., Mitchell,E.M. and Willet,P. (1990) *Protein Engng*, **4**, 39–43.
Bazan,J.F. (1990) *Immunol. Today*, **11**, 350–354.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Simanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Bowie,J.U., Luethy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Bryant,S.H. and Lawrence,C.E. (1993) *Proteins*, **16**, 92–112.
Finkelstein,A.V. and Ptitsyn,O.B. (1987) *Prog. Biophys. Mol. Biol.*, **50**, 171–190.
Genetic Computer Group (1991) Program Manual for the GCG Package V7. Madison, WI.
Godzik,A. and Sander,C. (1989) *Protein Engng*, **2**, 589–596.
Godzik,A. and Skolnick,J. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 12098–12102.
Godzik,A., Skolnick,J. and Kolinski,A. (1992) *J. Mol. Biol.*, **227**, 227–238.
Godzik,A., Kolinski,A. and Skolnick,J. (1993) *J. Comput. Aided Mol.*, in press.
Gribskov,M., McLachlan,M. and Eisenberg,D.P. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) *Nature*, **358**, 86–89.
Kabsch,W. (1978) *Acta Crystallogr.*, **A34**, 827–828.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Kikuchi,T. (1992) *J. Protein Chem.*, **11**, 305–320.
Kikuchi,T., Nemethy,G. and Scheraga,H.A. (1988) *J. Protein Chem.*, **7**, 427–470.
Kolinski,A., Godzik,A. and Skolnick,J. (1993) *J. Chem. Phys.*, **98**, 7420–7433.
Kuntz,I.D. (1975) *J. Am. Chem. Soc.*, **97**, 4362–4366.
Lesk,A.M. and Chothia,C. (1982) *J. Mol. Biol.*, **160**, 309–324.
Liebman,M.N., Venanzi,C.A. and Weinstein,H. (1985) *Biopolymers*, **24**, 1722–1758.
Luethy,R., Bowie,J.U. and Eisenberg,D. (1992) *Nature*, **356**, 83–85.
Maiorov,V.N. and Crippen,G.M. (1992) *J. Mol. Biol.*, **277**, 876–888.
Metropolis,N.A., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) *J. Chem. Phys.*, **51**, 1087–1092.
Milburn,M.H., Prive,G.G., Milligan,D.L., Scott,W.G., Yeh,J., Jancarick,J., Koshland,D.E. and Kim,S.-H. (1991) *Science*, **254**, 1342–1347.
Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
Nikishawa,K. and Ooi,T. (1974) *J. Theor. Biol.*, **43**, 351–374.
Nussinov,R. and Wolfson,H.J. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 10495–10499.
Ooi,T. and Nikishawa,K. (1973) In Bergman,E.D. and Pullman,B. (eds), *Conformation of Biological Molecules and Polymers*. Academic Press, New York, pp. 171–187.
Pastore,A. and Lesk,A.M. (1990) *Proteins*, **8**, 133–155.
PDB (1992) *Newsletter*, No. 61.
Philips,D.C. (1970) *Biochem. Soc. Symp.*, **30**, 11–28.
Richards,F.M. and Kundrot,C.E. (1988) *Proteins*, **3**, 71–84.
Richardson,J. (1981) *Adv. Protein Chem.*, **34**, 167–339.
Richardson,J.S., Richardson,D.C. and Thomas,K.A. (1976) *J. Mol. Biol.*, **102**, 221–235.
Rossman,M.G. and Lilas,A. (1974) *J. Mol. Biol.*, **85**, 177–181.
Sander,C. and Vriend,G. (1990) *Protein Design on Computers*. EMBO Practical Course Manual, EMBL, Heidelberg.
Scharf,M. (1989) Diploma Thesis. University of Heidelberg, Heidelberg.
Sheriff,S. and Hendrickson,W.A. (1987) *J. Mol. Biol.*, **197**, 273–296.
Sippl,M.J. and Weitckus,S. (1992) *Proteins*, **13**, 258–271.
Skolnick,J., Kolinski,A., Brooks,C.L., III, Godzik,A. and Rey,A. (1993) *Curr. Biol.*, in press.
Vingron,M. and Argos,P. (1990) *Protein Engng*, **3**, 565–569.
Vriend,G. and Sander,C. (1991) *Proteins*, **11**, 52–58.
Ycas,M. (1990) *J. Protein Chem.*, **9**, 177–200.

Received April 5, 1993; revised June 29, 1993; accepted June 29, 1993