doi:10.1006/jmbi.2000.3975 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 301, 691–711

JMB

An Integrated Approach to the Analysis and Modeling of Protein Sequences and Structures. III. A Comparative Study of Sequence Conservation in Protein Structural Families using Multiple Structural Alignments

An-Suei Yang[1,2]* and Barry Honig[1,2]*

[1]*Department of Biochemistry and Molecular Biophysics*

[2]*Columbia Genome Center Columbia University, 630 West 168th Street, New York, NY 10032, USA*

The information required to generate a protein structure is contained in its amino acid sequence, but how three-dimensional information is mapped onto a linear sequence is still incompletely understood. Multiple structure alignments of similar protein structures have been used to investigate conserved sequence features but contradictory results have been obtained, due, in large part, to the absence of subjective criteria to be used in the construction of sequence profiles and in the quantitative comparison of alignment results. Here, we report a new procedure for multiple structure alignment and use it to construct structure-based sequence profiles for similar proteins. The definition of ''similar'' is based on the structural alignment procedure and on the protein structural distance (PSD) described in paper I of this series, which offers an objective measure for protein structure relationships. Our approach is tested in two well-studied groups of proteins; serine proteases and Ig-like proteins. It is demonstrated that the quality of a sequence profile generated by a multiple structure alignment is quite sensitive to the PSD used as a threshold for the inclusion of proteins in the alignment. Specifically, if the proteins included in the aligned set are too distant in structure from one another, there will be a dilution of information and patterns that are relevant to a subset of the proteins are likely to be lost.

In order to understand better how the same three-dimensional information can be encoded in seemingly unrelated sequences, structure-based sequence profiles are constructed for subsets of proteins belonging to nine superfolds. We identify patterns of relatively conserved residues in each subset of proteins. It is demonstrated that the most conserved residues are generally located in the regions where tertiary interactions occur and that are relatively conserved in structure. Nevertheless, the conservation patterns are relatively weak in all cases studied, indicating that structure-determining factors that do not require a particular sequential arrangement of amino acids, such as secondary structure propensities and hydrophobic interactions, are important in encoding protein fold information. In general, we find that similar structures can fold without having a set of highly conserved residue clusters or a well-conserved sequence profile; indeed, in some cases there is no apparent conservation pattern common to structures with the same fold. Thus, when a group of proteins exhibits a common and well-defined sequence pattern, it is more likely that these sequences have a close evolutionary relationship rather

Present address: A.-S. Yang, Department of Pharmacology, Columbia University, 630 W. 168th St., PH 7 W Room 318, New York, NY 10032, USA.

Abbreviations used: PrISM, protein information system for modeling; PSD, protein structural distance; seqID, sequence identity.

E-mail addresses of the corresponding authors: ay1@columbia.edu; bh6@columbia.edu

0022-2836/00/030691–21 $35.00/0

© 2000 Academic Press

than the similarities having arisen from the structural requirements of a given fold.

© 2000 Academic Press

Keywords: protein structural distance; multiple structural alignment; sequence profile; protein evolution; sequence-structure relationship

*Corresponding authors

Introduction

It is well known that the 3D structures of proteins are more conserved than their amino acid sequences; that is, many pairs of proteins have very similar structures or topologies while having no obvious sequence relationship. This observation raises the question as to how common 3D structure information is encoded onto apparently unrelated sequences. In an accompaning (paper, Yang & Honig, 2000b), we demonstrated that protein pairs with similar structure but little sequence identity have non-random although not necessarily detectable sequence relationships that are below the statistical noise level of pairwise sequence alignments. We also demonstrated that protein structure similarity increases with increasing sequence conservation for these protein pairs. Thus, although there appears to be some remnant of common sequence information in many pairs of proteins with similar structures, it is generally difficult to detect the relevant pattern using pairwise sequence alignments.

The fact that seemingly unrelated sequences can code for very similar structures can be interpreted in a number of ways. One possibility is that structural information is contained in only a small number of critical residues and that conserved features are so sparsely distributed in the sequence that they cannot be detected with pairwise sequence alignments. If this is correct, then it is important to identify these residues using a structure-based sequence profile and to understand how it is possible that a small fraction of the complete sequence of a protein can be so critical in determining its 3D structure. An alternative possibility is that 3D structure is encoded in different locations throughout the sequence and that critical residues need not be located in the same position in different proteins of the same fold. The number of such residues could be small or large but in either case they would be difficult to detect with multiple sequence alignments. Here, we provide a basis for addressing these issues through the use of a novel multiple structure alignment procedure to construct structure-based sequence profiles for sets of structurally similar proteins that share little sequence identity.

It is conceivable that there may exist many parallel ways in coding 3D structural information onto a protein sequence. One way of coding structural information is through secondary structure propensities. However, residues with similar secondary structure propensities are not generally detectable with substitution matrices and need not be aligned in the sequence. There just needs to be enough signal in a particular sequence segment to favor a particular secondary structural element. Thus when one uses an expression such as ''critical residues at some location in a sequence'', an appropriate frame of reference for ''location'' might well be specific secondary structure elements. This issue will be considered further below.

There has been considerable disagreement in the literature as to the existence of a small number of critical residues in protein folding. In an attempt to identify such residues, Ptitsyn (1998) studied sequence profiles in seven subfamilies of c-type cytochromes and found that other than the heme-binding residues, only four positions were conserved in all subfamilies. These residues form a network of conserved contacts between the N-terminal and C-terminal helices. The fact that the docking of these two helices had been identified as an early step in the folding pathway (Roder & Colon, 1997) led Ptitsyn (1998) to conclude that the folding of all *c*-type cytochromes involves the formation of the same conserved folding nucleus (Fersht, 1997; Shakhnovich, 1997). This would imply that, at least for *c*-type cytochromes, only a few residues that participate in the folding nucleus need to be conserved and that 3D structural information is encoded in these residues. Of course, the location of these ''critical'' residues is context-dependent, i.e. they are located on α-helices and, as mentioned above, structural information must also be contained in the helical propensities of significant number of other residues.

In an earlier study of 226 globins, Bashford *et al*. (1987) found a much larger number of conserved residues than did Ptitsyn (1998). Bashford *et al*. (1987) found that some of the most conserved residues in globins are exposed to solvent and are not directly involved in contacts between helices. Furthermore, the most conserved residues are distributed among all eight helices, not just within the folding intermediate formed by helices A, G and H (Jennings & Wright, 1993). This suggests that residues other than those clearly identified with a folding nucleus can be important determinants of 3D structure.

The profiles constructed by Bashford *et al*. (1987) and by Ptitsyn (1998) were for large sequence families and, as a result, it is difficult to know if the conservation patterns that were found reflect important structural determinants for a particular fold or are simply the consequence of divergent evolution. Moreover, it is possible that the apparently contradictory conclusions that were reached

reflect differing levels of sequence identity within the two families that were studied. In an attempt to identify the sequence determinants for a particular fold while reducing any biases resulting from closely related sequences, Bork *et al.* (1994) constructed an immunoglobulin-like (Ig-like) profile from Ig-like domains that are not related by more than 25% pairwise sequence identity. Moreover, as opposed to the studies of the cytochromes *c* (Ptitsyn, 1998) and globins (Bashford *et al.*, 1987) the set of proteins studied by Bork *et al.* (1994) were not in a single sequence family. Given the difficulties in aligning sequences with low levels of identity, Bork *et al.* (1994) relied on structure-based alignments taken from Dali (Holm & Sander, 1995). A striking finding in their work was the absence of a common sequence pattern or localized set of interactions that could be identified as a crucial determinant of the Ig-like fold; even the number and location of disulfide bridges was variable. This led Bork *et al.* (1994) to suggest that all proteins with an Ig-like fold do not necessarily share the same folding pathway. Halaby *et al.* (1999) have also concluded that there is no common sequence signature for an Ig-like fold and that, despite the importance of the hydrophobic core, there is no strict conservation of the location of individual hydrophobic contacts.

In a study of 607 pairs of proteins with similar folds, Russell & Barton (1994) found that proteins with similar folds can use very different interactions to achieve this fold apart from a common scaffold of secondary structure elements. This finding is consistent with the conclusion reached by Bork *et al.* (1994), in that there do not appear to be a small number of crucial interactions that are characteristic of a particular fold. However, Mirny & Shakhnovich (1999) recently reached a very different conclusion. These authors studied the conservation of sequence and structure in five structural superfolds (Orengo *et al.*, 1994) and concluded that there are well-defined locations in a structure that are crucial for a particular fold even though the interactions at these key topological positions can be very different (e.g. a disulfide bond in one sequence family and a hydrophobic contact in another). However, it is not clear to what extent their conclusions are related to their choice of sequences for a particular family (for example, they did not use the same set of Ig-like domains that was used by Bork *et al.* (1994)) and to other ambiguities such as the threshold used to define a feature that occurs by chance (which varies by orders of magnitude from one superfold to another). Moreover, the sequence alignments they report are different from those used by Halaby *et al.* (1999), so that results based on these alignments must remain inconclusive. It is clearly extremely important to use objective standards as a basis of obtaining sequence and structure relationships, which is considered further in Discussion.

Here, we report a systematic study of the relationship between sequence and structure by constructing structure-based sequence profiles for members of the nine superfolds described by Orengo *et al.* (1994). Each of these superfolds contains a large number of members with varying levels of pairwise sequence identity. In order to avoid biases that can arise from a variety of sources, we used PrISM (protein information system for modeling) to construct a completely automatic protocol to generate structure-based sequence profiles and to identify conserved sequence features. This enables us to carry out comparative studies on a variety of protein folds using an identical protocol for all.

In this work, sequences included in all profiles of the nine superfolds were obtained from the PDB40D domain database of SCOP (version 1.37) (40% sequence identity or less) so as to avoid biases due to sequence redundancy. A structural distance (PSD) threshold (see the accompanying papers I and II, Yang & Honig, 2000a,b) is set so as to avoid a ''dilution'' effect on information content resulting from the inclusion in the profile of structures that are only marginally related. The quality of the structure-based profiles is examined in Results. In Discussion, we consider the importance of using a consistent set of criteria to evaluate sequence/structure relationships and describe the type of conservation patterns that are observed for the members of the nine superfolds.

Results

Comparison between multiple sequence alignments and multiple structural alignments of immunoglobulins

A sequence profile produced by a multiple structural alignment is expected to be superior, or at lease comparable, in quality to a sequence profile produced with a multiple sequence alignment. Immunoglobulins provide an ideal case for testing the alignment quality of sequence profiles, because this family contains a large number of diverse sequences and because the multiple alignments are complicated by highly variable regions in both sequence and structure (e.g. see Gerstein & Altman, 1995).

Figure 1(a) shows the structure-based sequence profile produced with PrISM for the immunoglobulin VH domains listed by Gerstein & Altman (1995). No sequence information was used in the construction of the profile. Each column lists residues that appear in equivalent topological positions in the 3D structure. The histogram above the sequence profile reflects the root-mean-square deviation (RMSD) of the C^{α} atoms from the geometric average of the C^{α} coordinates of the residues in each column of the profile (0.5 Å per $\wedge$). The integers listed directly below the histogram measure the consistency of the residues in each column and are obtained from the equation:

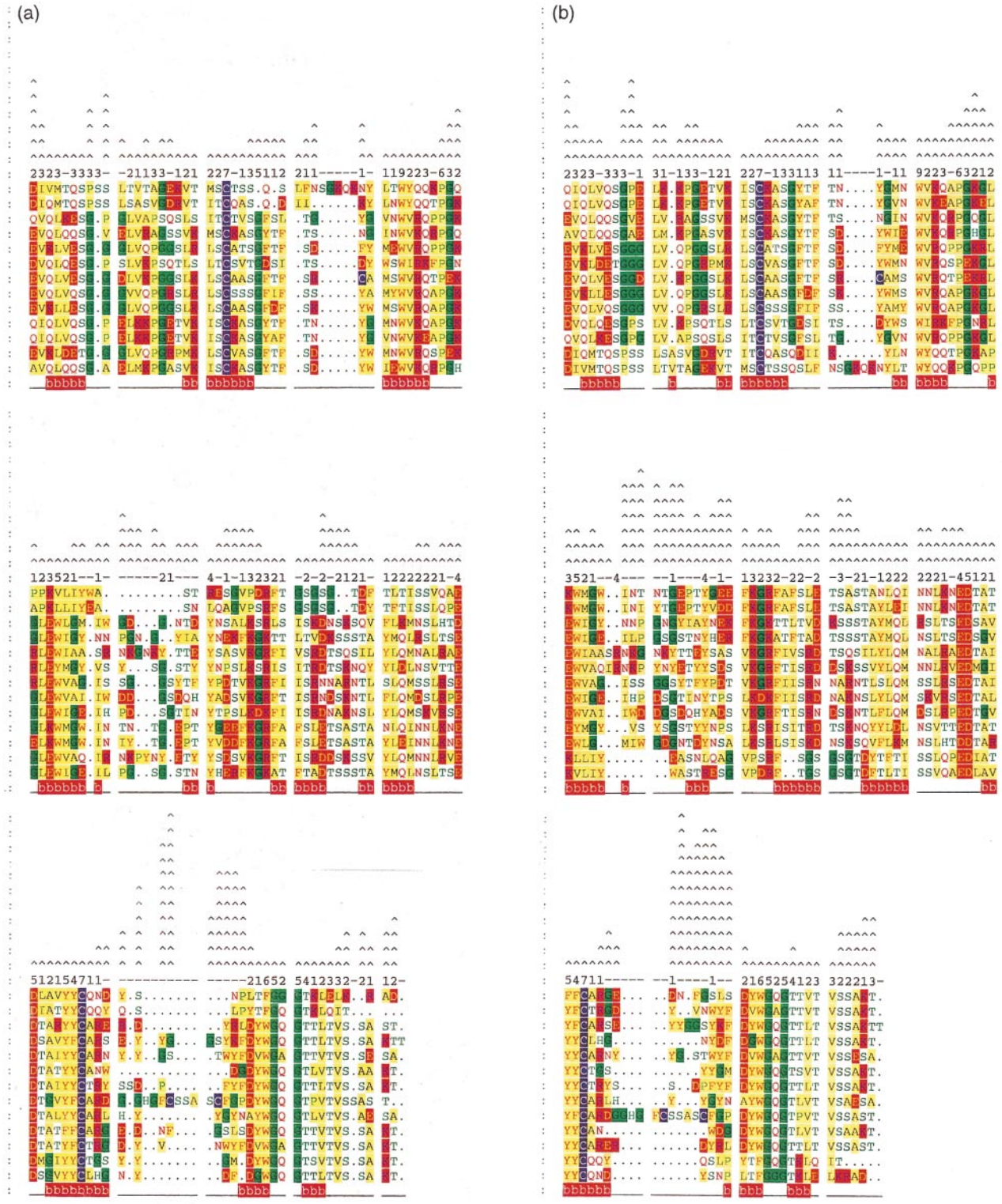


Figure 1. Immunoglobulin sequence profiles. (a) Profile produced with the PrISM multiple structure alignment procedure. From top row to bottom row, the sequences are 1hilA(1-110), 1reiA(1-107), 1fd1H(1-119), 6fabH(1-125), 1mcpH(1-125), 3hfmH(1-125), 1igfH(1-116), 2fb4H(1-121), 2fbjH(1-121), 1ncaH(1-116), 1dbaH(1-116), 4fabH(1-121), 3hflH(1-116). The first four letters are the PDB code, followed by chain name and the residue numbers for the first and the last residues. See the text for a detailed discussion. (b) The profile produced from the PrISM progressive multiple sequence alignment procedure. From top row to bottom row, the sequences are 1ncaH(1-116), 1dbaH(1-116), 6fabH(1-125), 3hflH(1-116), 1mcpH(1-125), 4fabH(1-121), 1igfH(1-116), 2fbjH(1-121), 2fb4H(1-121), 3hfmH(1-125), 1fd1H(1-119), 1reiA(1-107), 1hilA(1-110).

Amino acid consistency

$$= NINT\left[\frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} a(r_i, r_j)}{n} \times 0.8\right] \quad (1)$$

where, *NINT*[ ] takes the nearest integer of the value enclosed in the square brackets, N is the number of proteins in the profile, and $a(r_i,r_j)$ is the BLOSUM62 substitution matrix (Henikoff & Henikoff, 1993) element for replacing residue r_i by residue r_j. Residues r_i, $i = 1,N$ are residues in the same column of the profile. In the case of a deletion, where r_i or/and r_j is presented as a ''.'' in the profile, the substitution score $a(r_i,r_j) = 0$; n is the total number of terms in the summation not including deleted residues. The constant 0.8 on the right-hand side of equation (1) is used to reduce the integer to only one digit for display purposes. A similar but more sophisticated and normalized measure has been described by Landgraf *et al* (1999). In Figure 1, only amino acid consistency values that are greater than zero are displayed below the histogram. As can be seen from equation (1), the greater the consistency value for a particular column, the greater the degree of sequence conservation at that topological position in the structure. A consistency value of zero implies that the amino acid composition at that position is close to random, as defined by the BLOSUM62 substitution matrix. The bottom line of the profile shows the consensus secondary structure defined as the most common secondary structure assignment at that position in the structure as obtained from the DSSP program (Kabsch & Sander, 1983).

Figure 1(b) shows a multiple sequence alignment produced with only sequence information using the PrISM multiple sequence alignment procedure (similar to the progressive alignment procedure developed by Feng & Doolittle, 1990). The sequence and structure-based multiple alignments in Figure 1(a) and (b) are essentially identical in the β-strand regions. As indicated by the histogram, these regions are also the most conserved in structure. The alignment of the two profiles in loop regions, except for the two highly variable ones, is also quite similar, although both sequence and structure are less conserved in these regions. The largest discrepancies between the two profiles are located in the two hypervariable loops shown in the upper-left and upper-right corners of the superimposed structures in Figure 2. The sequence-based profiles tend be more compact, in that they have fewer insertions and deletions, but they exhibit greater structural variability than the structure-based sequence profile. There is no ''correct'' sequence alignment for the highly variable regions; indeed, the lack of obvious sequence similarity renders any sequence alignment meaningless. On the other hand, the structure-based sequence profile

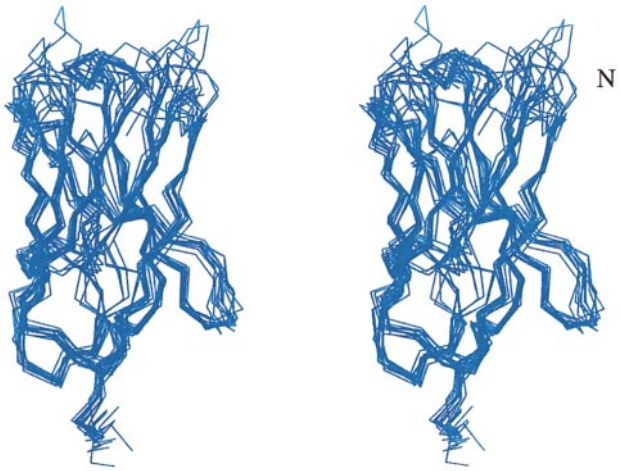


Figure 2. Multiple superimposition of the immunoglobulins shown in Figure 1 using the PrISM multiple structural alignment procedure. Only C^{α} traces of the structures are shown.

provides a somewhat more meaningful alignment, in that all the aligned residues are spatially close to one another in the superimposed structures shown in Figure 2.

Figure 3 shows that there is a significant correlation between sequence and local structural variability for immunoglobulin VH domains. Locations that have high amino acid consistency values always have low local structure deviations, so that the residues at these positions are expected to play an important structural and/or functional role. With some exceptions, the majority of the locations in the structure have low amino acid consistency values and small local structural deviations, consistent with the observation that structure is better conserved than sequence.

Structure-based sequence profile for serine proteases

The serine protease family has been a standard benchmark for testing multiple structure alignment algorithms (e.g. see Russell & Barton, 1992; Sali & Blundell, 1990). The ten serine proteases listed in Figure 4 are known to be difficult to align using sequence information alone (Russell & Barton, 1992). Russell & Barton (1992) devised an algorithm that uses structural information to improve multiple sequence alignments. The initial multiple alignment is generated from a sequence-based method and is then refined with a procedure that uses structural information and an iterative rigid-body superimposition method.

The procedure implemented in PrISM does not require a multiple sequence alignment procedure to generate the initial multiple alignment. This is particularly useful for a set of proteins with similar structure but undetectable sequence similarity. The alignment shown in Figure 4 is obtained from the

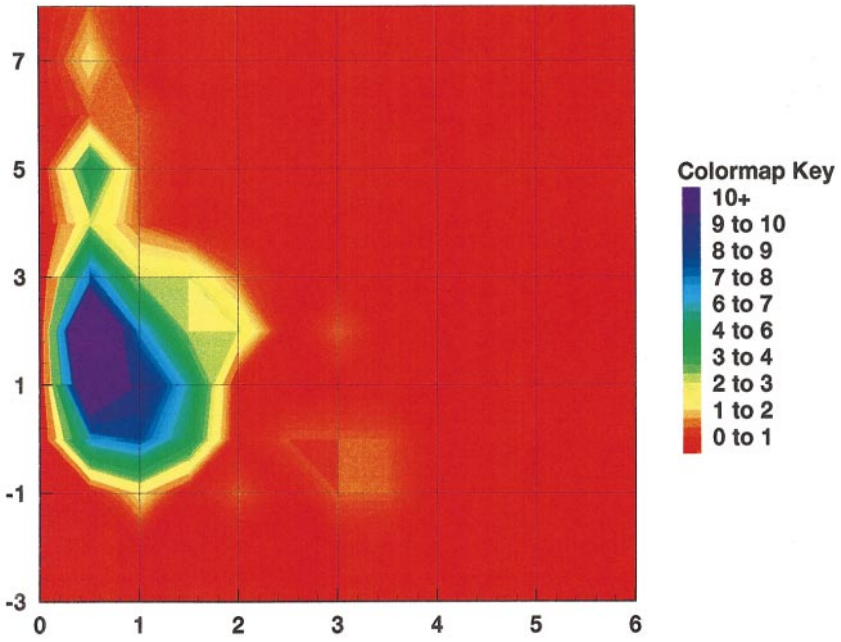


Figure 3. The correlation between sequence conservation and structure conservation for the set of immunoglobulines in Figures 1 and 2. The *y*-axis indicates the amino acid consistency value shown as defined by equation (1) (see also the numbers below the histogram in Figure 1(b)). The *x*-axis indicates the C^{α} RMSD from the averaged coordinate of a topological position in the optimally superimposed structures, as shown by the histogram above the sequence profile in Figure 1(b).

automatic multiple structural alignment procedure of PrISM. Only two of the 14 highly reliable regions (see Russell & Barton, 1992) of the profile contain alignments that are slightly different from those obtained by Sali & Blundell (1990) and by Russell & Barton (1992). In the region underlined with asterisks in Figure 4, PrISM identifies a β-bulge type of insertion of one residue into the second β-strand of the top three serine proteases in the profile that was not shown by the other two algorithms. The residues underlined by plus signs in the Figure correspond to a loop region and here the discrepancy between the different procedures involves the alignment of a four-residue segment in 1sgt to the rest of the proteases. Our alignment shifts this segment by one or two residues towards the N terminus as compared to the alignment made by Russell & Barton (1992) and Sali & Blundell (1990), respectively. The discrepancy results from structural variations in this loop that have been discussed elsewhere (Russell & Barton, 1992).

Figure 5 shows the multiple superposition of structures of the serine proteases listed in Figure 4 with the more conserved residues (amino acid consistency values greater than or equal to 3 and at least half of the column in the profile occupied) shown in red. More than half of these topological positions are located between the two β-barrels near the active site. Surprisingly, the remainder of the conserved residues are almost all located in the C-terminal β-barrel, suggesting that this domain did not mutate as actively as the N-terminal domain throughout the course of evolution. One possible explanation is that the C-terminal β-barrel needs to be more constant in structure so as to maintain a fixed conformation of the active-site serine residue and the oxy-anion hole.

The correlation between amino acid consistency and local structure variation of the serine proteases is shown in Figure 6. The proteases exhibit greater structural variability than the immunoglobulins listed in Figures 1-3. Nevertheless, the observations that positions that are well conserved in sequence are conserved in structure and that structure is better conserved than sequence holds for the serine proteases.

Structure-based clustering of immunoglobulin (Ig)-like structures

The conclusion reached by Bork *et al*. (1994) that Ig-like domains do not all have strictly conserved residues or a common folding pathway has important implications for the possibility of finding sequence-structure relationships for protein families. However, the conclusion depends, in part, on the definition of a family. Bork *et al*. (1994) obtained a non-redundant set (seqID < 25 %) of Ig-like proteins from a structural alignment search of the PDB using the Dali program (Holm & Sander, 1995). While the proteins that were identified share a common core structure consisting of four central β-strands, the peripheral strands do not all appear in the same location in the structure, a fact that has led to the definition of four structural subclasses of Ig-like domains; s, h, v, and c (Bork *et al*., 1994). The inclusion of proteins that belong to different sub-classes can have the effect of diluting the quality of a sequence profile through the presence of unrelated sequences and/or alignment errors for very different structures. Thus it is essential to use an objective structure similarity measure to decide what set of structures should be used to construct a sequence profile. With this goal in mind, we have devised a clustering procedure that relies on

the Protein Structural Distance (PSD) score defined in the accompanying paper I (see Methods).

Figure 7 illustrates the results of the clustering procedure applied to the set of Ig-like proteins listed by Bork *et al*. (1994). At a PSD threshold of 2 (which is a threshold for proteins with a similar fold, see the accompanying papers I and II), all the Ig-like protein domains are included in one cluster, and there is no clear sequence conservation pattern observed in the structure-based sequence profile (data not shown). This result is in agreement with the conclusion reached by Bork *et al*. (1994). However, at a PSD threshold of 1.5, the c-type and Ig-like protein domains form a separate cluster and sequence conservation, for example in the cystene residues that form a disulfide bridge, can be observed clearly (Figure 8(a)). Figure 8(b) shows a multiple structure alignment of the c-type domains with the highly conserved side-chains displayed in red (amino acid consistency values greater than or equal to 3 and at least half of the column in the profile occupied). The conserved residues appear in two distinct clusters located between the β-sheets; one includes the disulfide bond and the other is located in the lower half of the β-sandwich.

As seen in Figure 7, at a PSD threshold of 1, three clusters are formed, corresponding to s/h-type, v-type, and c-type domains, respectively. The relevant structure-based sequence profiles and multiply superimposed structures are shown in Figures 9 and 10 for v-type and s/h-type structures, respectively. While the v-type sequence profile has one cluster of conserved amino acids in the vicinity of the disulfide bond (see the side-chains in red in Figure 9(b)), it lacks the second cluster that is present in the c-type domains. The s/h-type structures have fewer conserved sequence patterns than the c-type and the v-type structures. Indeed, as shown in Figure 10(b), none of the topological positions in the s/h-type structures has a sequence feature that is conserved to the same extent as those seen in the c-type and v-type domains. The conserved disulfide bond shared by both v-type and c-type domains is missing from most of the s/h-type domains and its absence may contribute to the lack of other characteristic sequence features. Specifically, because the disulfide bond constrains the packing of the β-sheets, it should, in turn, dictate the sequence features in the core region of the protein in its vicinity (see also Halaby *et al*., 1999). Indeed, both c-type and v-type structures have a cluster of conserved residues around the disulfide bond, as shown in Figures 8(a) and 9(a), respectively.

In general, structures with more conserved sequence features also have more conserved local structure (Figures 7-10). Thus, the c-type structures are more conserved in sequence and structure than the v-type structures, and the s/h-type structures are the most diverse of the three classes of Ig-like domains. It is important to recognize that in some cases there may simply not be a distinct sequence profile for a group of proteins that appear structurally related, even when a PSD threshold of 1 is used. The s/h type Ig-like domains shown in Figure 10 offer one such example.

The importance of choosing an appropriate threshold for structural distance is evident from the analysis of Ig-like domains (Figures 7-10). If the PSD threshold is too small, the profile that is obtained from the structural alignment will generally be similar to that obtained from a sequence alignment. Indeed in the accompanying paper II we showed that protein pairs with PSD < 0.5 are likely to be related in sequence so that structure-based sequence profiles produced using this threshold will, for most cases, contain no more information than available from a sequence-based profile. On the other hand, as we have found for the Ig-like domains, a PSD score of greater than 1.5 can result in the inclusion of too many distantly related structures, which can dilute the structure-based sequence profiles, making it difficult to detect common sequence features. A PSD threshold of 1 is effective in clustering Ig-like domains into three structural families and was thus used in the analysis described below.

In the following section we consider sequence conservation patterns in the nine superfolds identified by Orengo *et al*. (1994). Superfolds correspond to the structural families with the largest number of members and contain approximately one-third of the total number of known structures in the PDB in 1994 (Orengo, 1994). One seed protein is selected from each superfold and a set of its structural neighbors is obtained using a PSD threshold of 1. All structure-based sequence profiles were constructed automatically with PrISM (see Methods for details).

It is worth noting that Figures 11-13 were generated using an approach somewhat different from that used to generate Figures 5, 8(b). 9(b) and 10(b) in that the previous Figures were obtained either from a predetermined set of proteins (Figure 5) or from clustering such a set into subfamilies (Figures 8-10). In contrast, Figures 11-13 were obtained by starting with a single seed protein. As can be seen in Figure 11, the cumulative correlation between amino acid consistency values and local structural variability for the nine sequence profiles is very similar to that observed for immunoglobulins (Figure 3) and serine proteases (Figure 6).

Due to the large number of superimposed structures, only the C^{α} traces for each of the nine seed structures are shown in Figures 12 and 13. Sequence and structure conservation information is mapped onto the seed structures. Consistent with Figures 5, 8(b), 9(b), and 10(b), the amino acid consistency threshold of 3 is used to identify the conserved residues, which are colored in magenta in Figure 12. The main-chains are color-coded for the local average structural deviation from the geometrically averaged positions. Backbone atoms colored in cyan are residues with average RMSD > 3 Å, while backbone atoms colored in

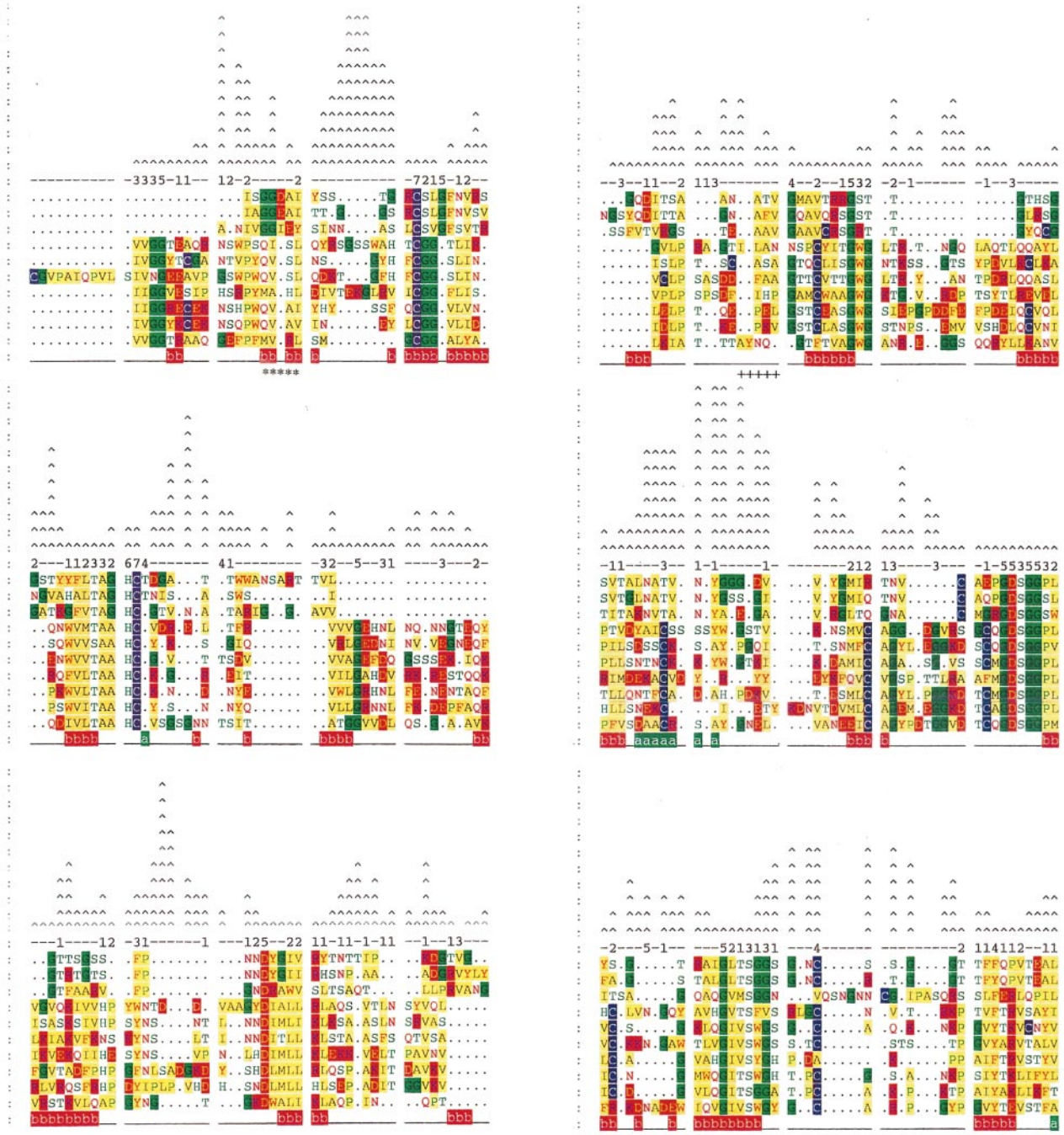


Figure 4 (*legend opposite*)

blue are residues with average RMSD $\leqslant 3$ Å. As discussed above, highly conserved residues are usually conserved in local structures. Overall, the conserved sequence features are sparsely distributed. As shown in Figure 12, many folds do not even contain conserved sequence feature. For folds that show conserved residues, these residues do not necessarily form clusters, in contrast with conclusions reported in recent work (Mirny & Shakhnovich, 1999; Ptitsyn, 1998).

It is possible that a more visible conservation pattern can be detected if a less stringent criterion is used to define conservation. Indeed, since there is no clear relationship between amino acid consistency value and evolutionary relatedness, an empirical amino acid consistency threshold should be chosen so as to reveal the clearest difference in sequence conservation patterns. Too large a threshold will result in no pattern at all in some protein families (see Figure 12), while too small a threshold produces conservation patterns with little meaning. For this reason, the threshold was relaxed so that any position with an amino acid consistency value of greater than or equal to 1 was

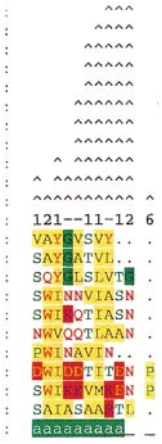


Figure 4. The structure-based sequence profile of a set of serine proteinases. From top row to bottom row, the sequences are 3sgbE(16-242), 2sga(16-242), 2alp(15A-245), 3est(16-245), 2ptn(16-245), 4chaA(1-245), 3rp2A(16-243), 2pkaB(16-246), 1ton(16-246), 1sgt(16-245).

marked in red (Figure 13). It is important that the same threshold is used to characterize all superfolds, so that the comparison among the superfold profiles is carried out at the same sequence conservation level. The threshold of one implies that any region shown in red in the Figure has some, albeit small, degree of sequence conservation. Figure 13 shows that, in general, the relatively conserved residues are clustered in the tertiary packing region, where the local structures are also relatively conserved.

Discussion

Here, we have described a protocol for automatic generation of multiple sequence alignments from the multiple structure alignment of closely related structures. Possible applications of these multiple alignments are the detection of conserved sequence features that can be used in fold recognition, homology modeling and in identifying the determinants of a particular fold. It is important to note that ''conserved sequence feature'' is a subjective term at best. The conserved residues in the sequence profiles reported here are determined by (1) the PSD threshold used to exclude dissimilar structures, (2) the sequence similarity threshold used to exclude redundant sequences, and (3) the amino acid consistency value threshold used to define ''conserved'' residues. None of these thresholds can be unambiguously determined based on a clearly defined physical or evolutionary model. The thresholds used in this work, although we believe them to be reasonable, are determined empirically so that the conserved sequence features are empirically derived as well. Nevertheless, as long as an unbiased and consistent set of criteria are used to define the structural alignment protocol and relevant thresholds, a comparative study of multiple structural alignments should be able to yield meaningful and objective conclusions. Similarly, it is difficult to evaluate the conclusions of studies where the choice of various criteria and the rationale for their use have not been clearly delineated.

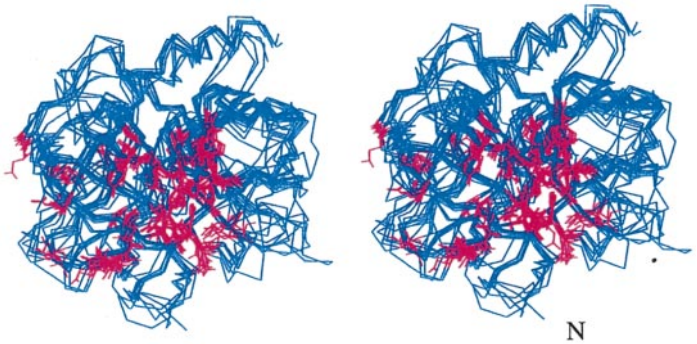


Figure 5. The superimposed structures of the serine proteases in Figure 4 using the PrISM multiple structural alignment procedure. Positions with an amino acid consistency value of greater than 3 are shown in red. The active site is in the center of the structure, facing front, and between the N-terminal domain (on the right of the active site) and the C-terminal domain (on the left of the active site).

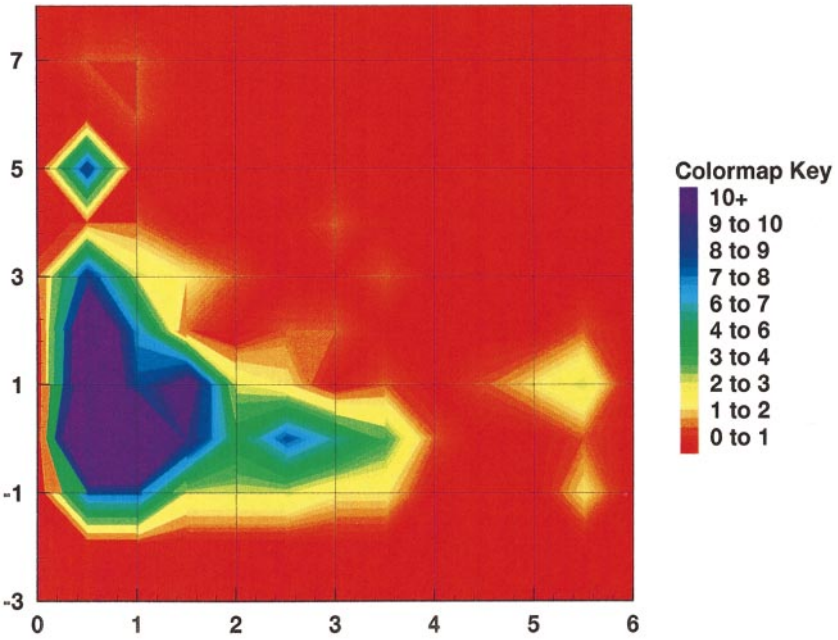


Figure 6. The correlation between sequence conservation and structure conservation for the set of serine proteases in Figure 5. See Figure 3 for details.

Analysis of nine superfolds

All-α proteins

The globin family (Figure 13(a)) has conserved residues clustered in the interface between helices A, G, and H and in the heme-binding pocket. The former consists of hydrophobic residues and is likely to be of structural importance because an early folding intermediate is formed by these three helices (Jennings & Wright, 1993). The cluster of residues that faces the heme-binding pocket is primarily involved in heme binding. A similar functionally related pattern of sequence conservation is observed in the up-down helix bundle in Figure 13(b). The group of proteins that was used in the alignment consists primarily of cytochromes that bind one heme group between the N and C-terminal helices. Except for a few conserved residues that are scattered throughout the structure, the main cluster of conserved residues are all close to the binding site and are expected to contribute to the binding of the heme, either directly or indirectly.

α/β Proteins

For small α/β proteins such as the split αβ sandwich (Figure 13(c)) and UB αβ roll (Figure 13(d)), the conserved residues are located primarily in the interface between the α-helix (helices) and the β-sheet. These residues are hydrophobic and primarily buried and clearly play structural and energetic roles. The conserved residues of the α/β doubly wound proteins (Figure 13(e)) are hydrophobic and are located in the central β-sheet, forming a two-sided hydrophobic plain that interacts with helices on either side of the β-sheet. The α-helices are not as conserved in sequence as is the β-sheet. In contrast, the majority of the conserved residues of the TIM-barrels (Figure 13(f)) are distributed in the peripheral α-helices rather than in the characteristic eight-stranded β-barrel. These conserved residues are also predominantly hydrophobic but their distribution is strikingly different from that in the α/β doubly wound proteins (Figure 13(e)). Both groups of proteins consist of a contiguous α-β-α structural motif, so that secondary structure propensities do not distinguish between the two types of folds. Rather, the distribution of amino acid polarity appears to be the decisive factor. This is in agreement with the conclusion that secondary structure formation can be viewed as providing the first tier of screening in the formation of protein structure, but it is the tertiary interactions that determine the final fold (Yang *et al.*, 1996).

All-β proteins

The majority of the conserved residues of Greek-key immunoglobulin (Figure 13(g)), jelly roll (Figure 13(h)), and trefoil (Figure 13(i)) folds are hydrophobic and are distributed in β-strands rather than in loops. The immunoglobulins have a relatively small number of conserved residues between the β-sheets. The structural neighbors of the seed immunoglobulin (PDB code 1tlk), which is defined as a v-type domain, include representatives of both v-type and s/h type domains, so that the sequence profile shows a lower degree of conservation than the v-type profile shown in Figure 9. Rather, the profile is more similar to that of the s/h-type domains shown in Figure 10.

The conserved residues in the jelly-roll fold are located in the central β-barrel and some loop regions. A similar pattern in the distribution of conserved residues is found in the trefoil fold. In contrast to the all-α and α/β proteins, it is difficult

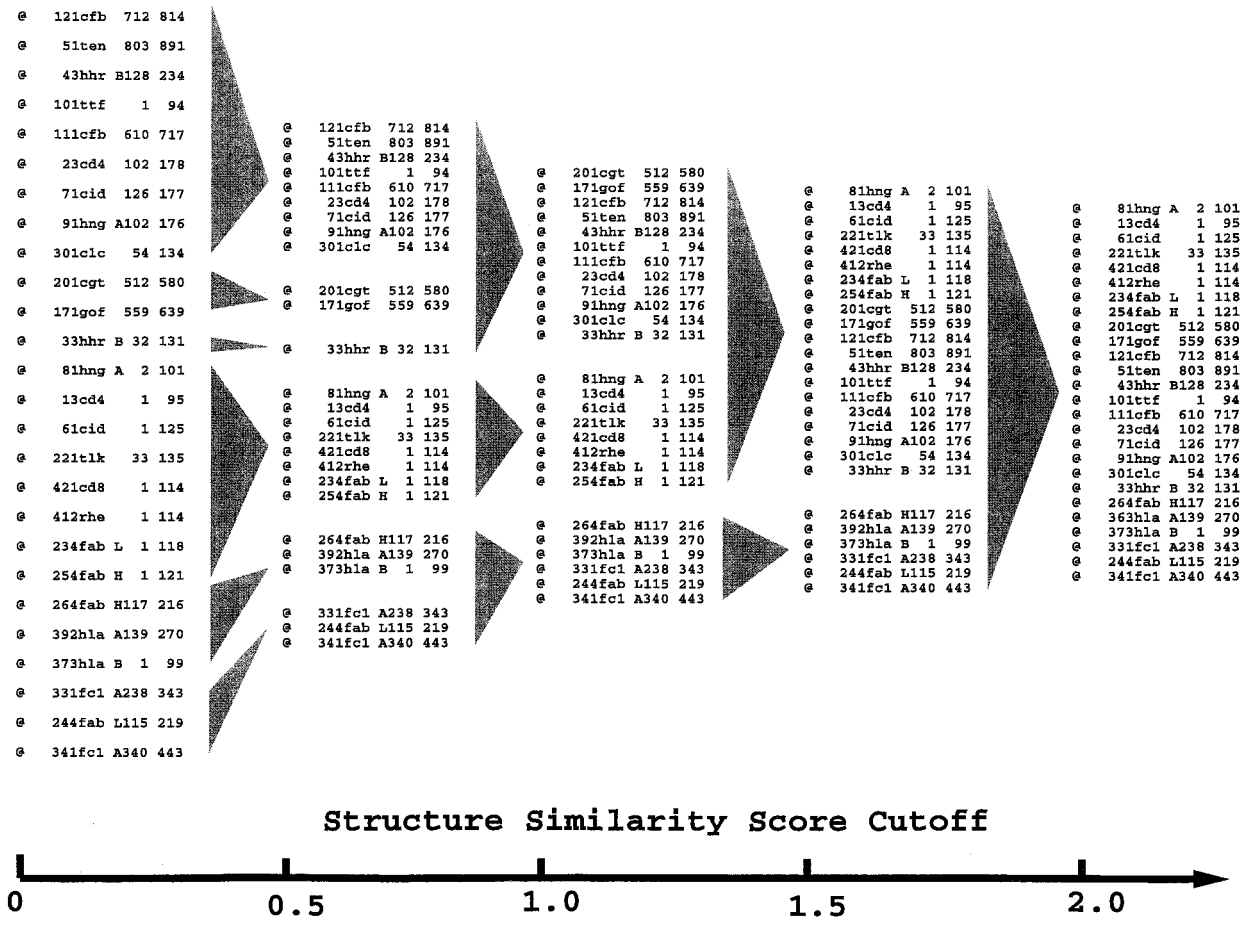


Figure 7. Clustering Ig-like proteins based on various protein structural distance (PSD) thresholds. The PSD threshold is shown along the *x*-axis.

to distinguish a clear pattern in the distribution of conserved residues in all-β proteins. This makes it difficult to distinguish among folds based on secondary structure propensities and residue polarity distribution alone, and accounts for the observation that fold recognition methods encounter their greatest difficulties in distinguishing between the fold topologies of all-β proteins. It is likely then, that there is another topological signal for this fold family that has not been identified.

How is 3D structure encoded on a sequence?

Figure 13 demonstrates that there is a distinct conservation pattern for each superfold, but the fact that an amino acid consistency threshold of 1 was used indicates that the pattern is weak. The fact that there exist residues with amino acid consistency values greater than 3 (see Figure 12) might argue that these correspond to the few critical locations that participate in a folding nucleus ((Mirny & Shakhnovich, 1999; Ptitsyn, 1998)). However, as shown in Figure 12, some folds simply do not present any conservation in sequence. The differences in conservation patterns and the extent of conservation between the v,c and s,h-type Ig-like domains certainly suggests that the same fold can be achieved with different sequence patterns and that these can be widely distributed throughout the structure.

Thus, although by no means conclusive, our results are consistent with the notion that, in general, structural information is encoded in many regions of an amino acid sequence. This is not unexpected, since structural information is contained partially in secondary structure propensities and this information does not require a unique sequential arrangement of amino acids. Similarly, the formation of a hydrophobic core is clearly a crucial feature of protein folding but the relevant residues need not appear in the same place in the sequence. Thus, when a group of proteins exhibits a common and clear sequence pattern, it appears likely that these proteins share a common evolutionary history. A crucial result of this work is that questions such as these must be addressed based on objective and well-characterized thresholds for sequence and structural distance.

Independent of their evolutionary history, the patterns revealed in Figures 12 and 13 may well reveal insights as to the features that produce a given fold. It is intriguing, for example, that different α/β fold families have different distributions of

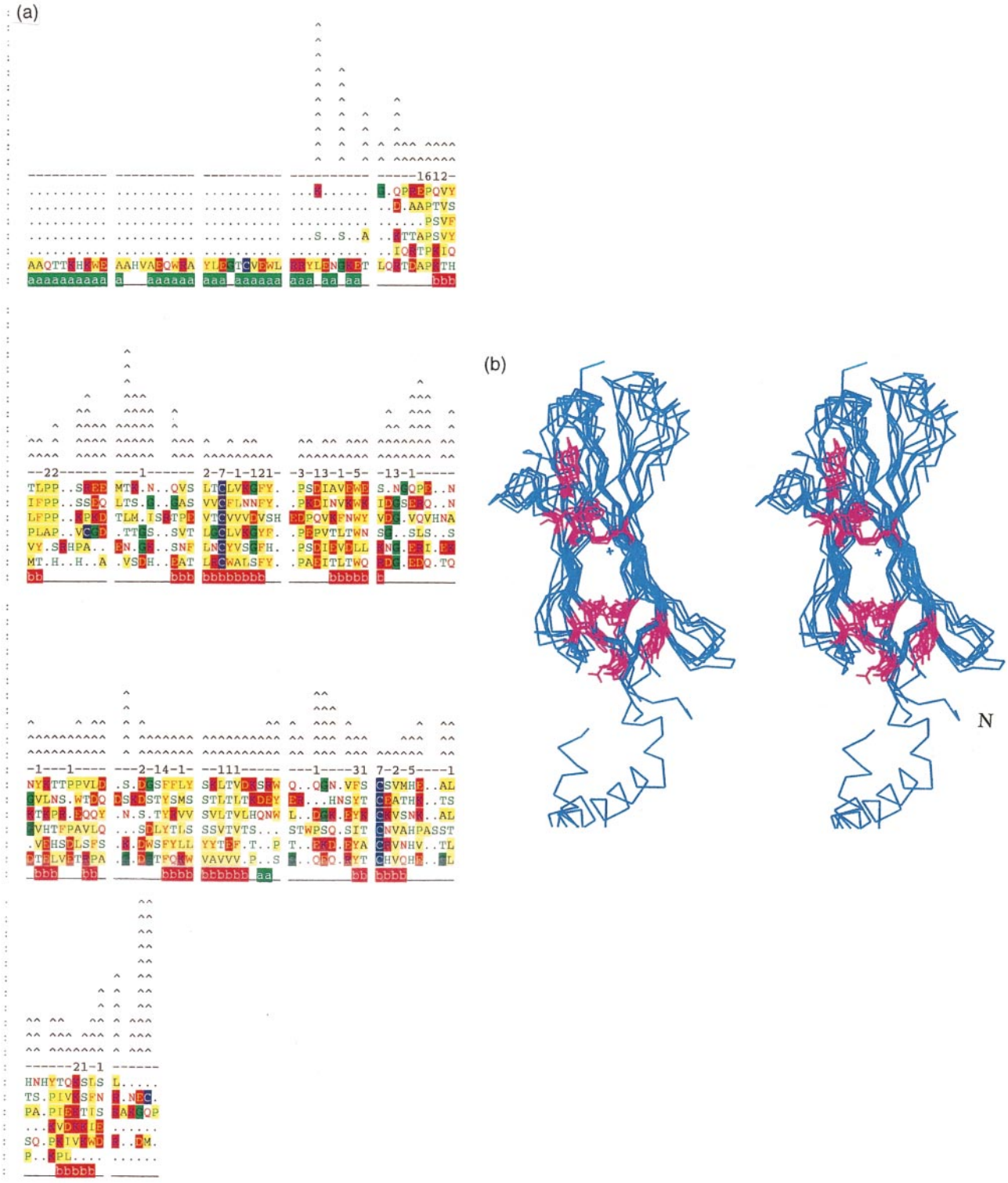


Figure 8. (a) The structure-based sequence profile and (b) the superimposed structures of c-type Ig-like domains. From top row to bottom row in (a), the sequences are 1fc1A(340-443), 4fabL(115-219), 1fc1A(238-343), 4fabH(117-216), 3hlaB(1-99), 2hlaA(139-270**).**

conserved residues; the TIM-barrels have more conserved residues in the α-helices than in the β-sheets, while the opposite is true for the α/β doubly wound proteins. This type of observation may contain hints that help us understand how sequence codes for topology although this will require a far more extensive analysis. Moreover, the existence of even weak conservation can be exploited in comparative modeling and fold recognition in cases where traditional sequence align-

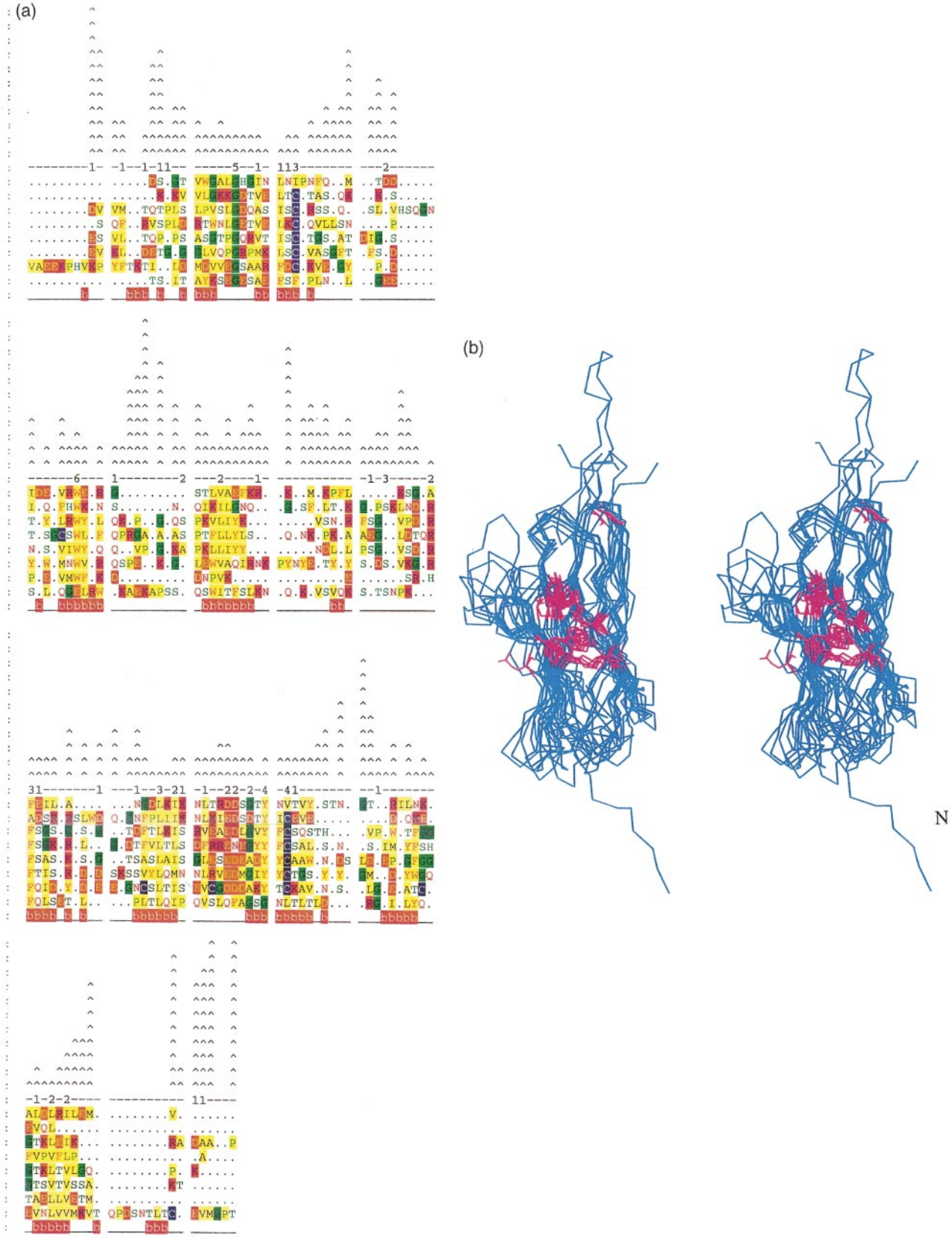


Figure 9. (a) The structure-based sequence profile and (b) the superimposed structures of v-type Ig-like domains. From top row to bottom row in (a), the sequences are 1hngA(2-101), 3cd4(1-95), 4fabL(1-118), 1cd8(1-114), 2rhe(1-114), 4fabH(1-121), 1tlk(33-135), 1cid(1-125).

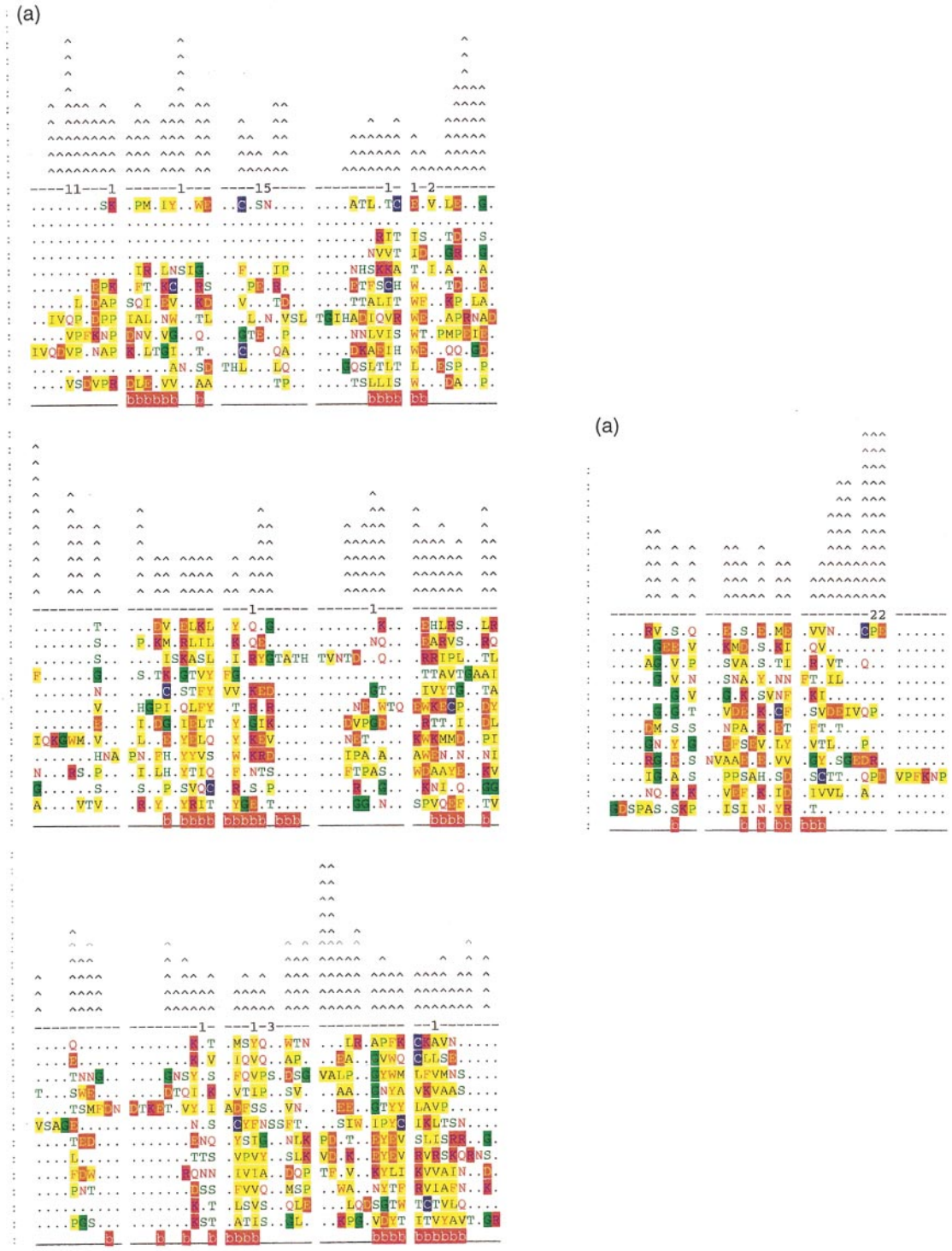


Figure 10 (*legend opposite*)

ment methods have failed due to low levels of sequence identity and extensive insertions and deletions. In future work we will carry out a more extensive study by using different seed proteins and varying PSD and amino acid consistency thresholds in order to determine if the patterns detected in Figures 12 and 13 are characteristic of proteins in general.

Classification or dynamic procedures

Although we have chosen a single member of each superfold to generate the alignments shown in Figures 12 and 13, the procedure we have described, which uses a single seed protein, does not depend on a pre-existing classification. In fact, one could in principle generate a set of neighbors

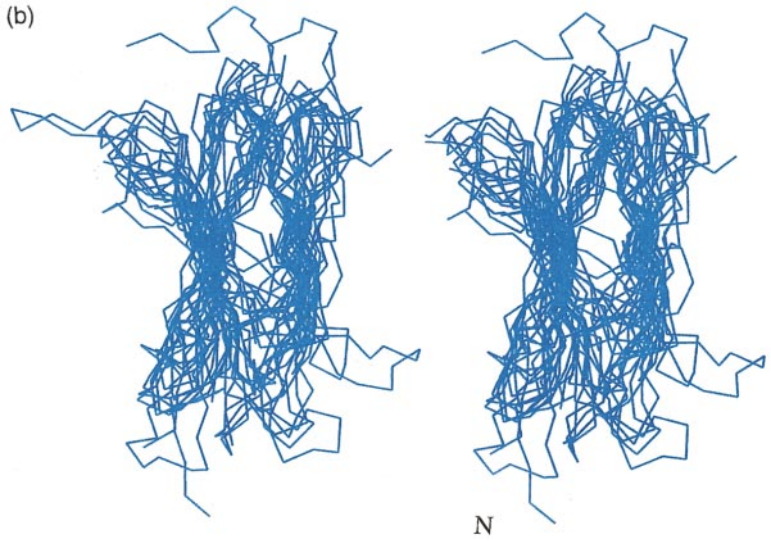


Figure 10. (a) The structure-based sequence profile and (b) the superimposed structures of s/h-type Ig-like domains. From top row to bottom row in (a), the sequences are 1hngA(102-176), 1cid(126-177), 1gof(559-639), 1cgt(52-580), 1clc(54-134), 3hhrB(32-131), 1ten(803-891), 3hhrB(128-234), 1cfb(712-814), 1cfb(610-717), 3cd4(102-178), 1ttf(1-94).

and alignments that would be distinct for each protein in the Protein Data Bank and avoid classification entirely.

There has been some interest in determining how best to select a set of proteins with the same fold so as to compare conserved sequence features and, in particular, to determine structures that represent a given fold (Orengo, 1999). However, this is a particularly difficult problem given the ambiguities involved in defining a fold (see e.g. our accompanying paper I). Our solution, described here, is to devise an automatic procedure (using a single set of thresholds) that is capable of creating a structure-based sequence profile for each member of a set of non-redundant proteins. The set of profiles obtained in this way can then be used in further applications, such as defining consensus sequences for a particular group of proteins. This can be done without a classification database and has the advantage of freeing the researcher from the reliance on a pre-existing classification scheme. The trade-off is that protein sequence features need to be discussed in the context of the seed protein and its immediate structural neighbors, rather than in terms of the features of complete fold as defined by a classification database. This is not necessarily a disadvantage given the uncertainty involved in the definition of some protein folds (see e.g. Orengo, 1999).

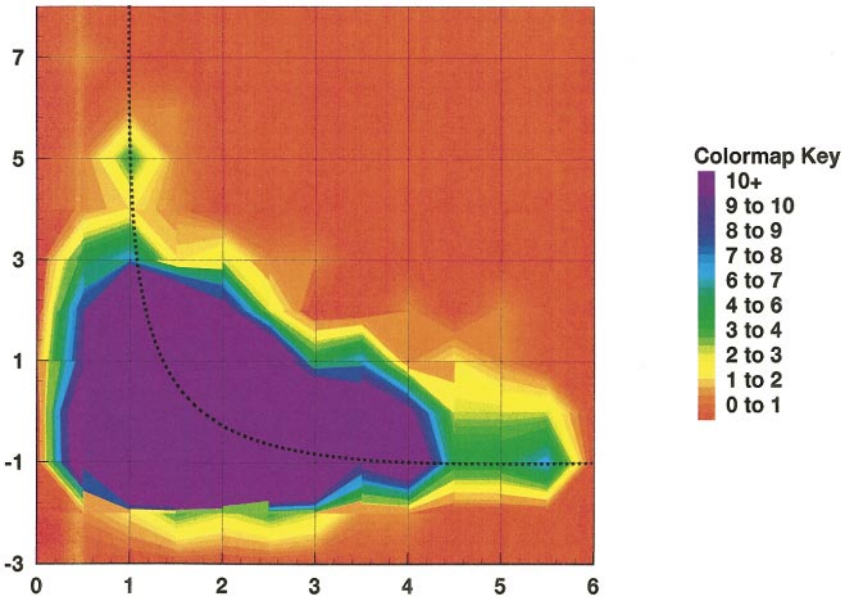


Figure 11. The cumulative correlation between sequence conservation and structure conservation for the nine superfolds. See Figure 3 and the text for details.

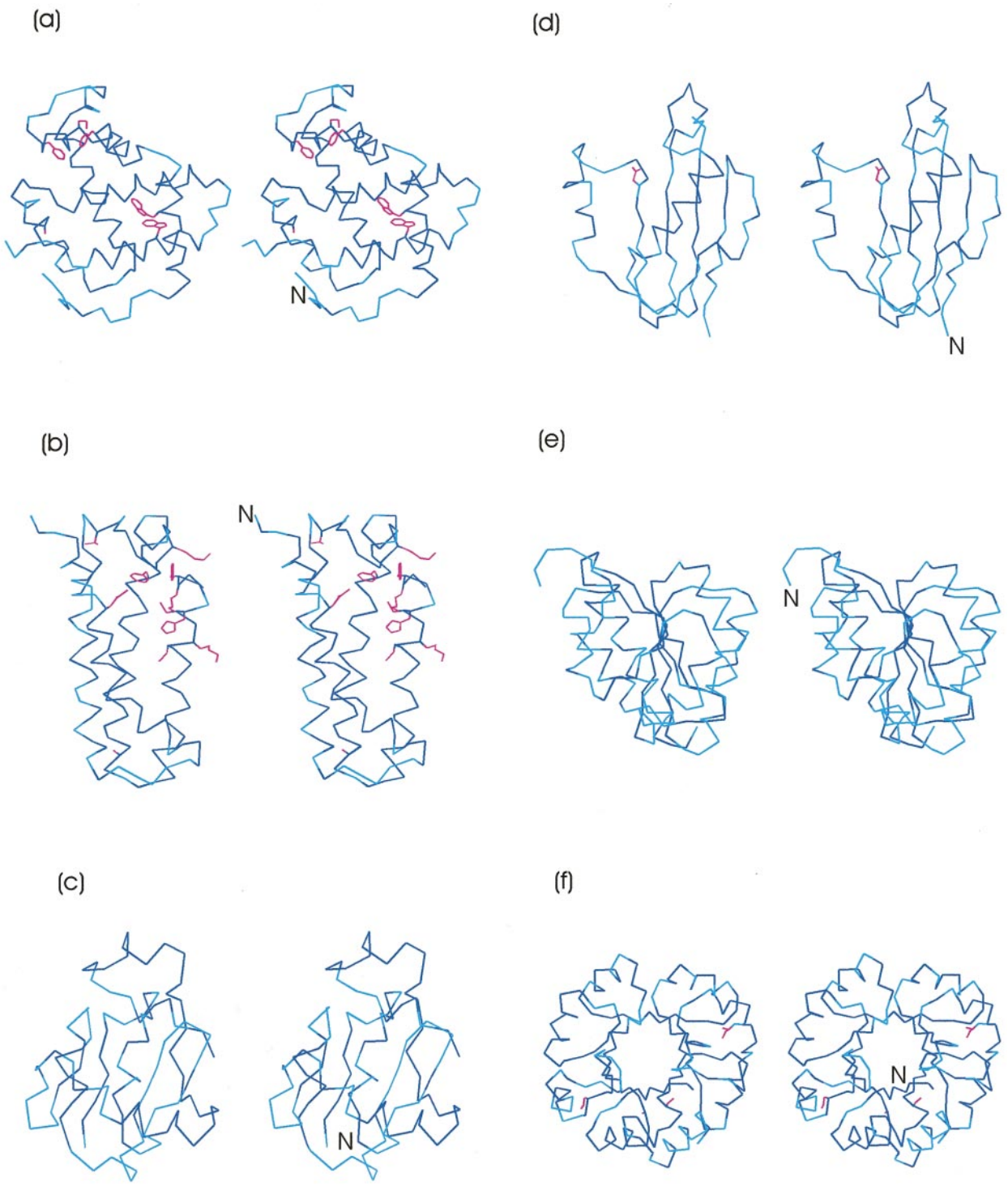


Figure 12 (*legend shown on page 708*)

Summary and Conclusions

We have designed an automatic multiple structure/sequence alignment procedure that can be applied in an objective and consistent fashion to any group of proteins or to a set generated from a single seed protein.

The quality of sequence alignments depends sensitively on the structural distance used to define the set included in the alignment. Too distant a set will dilute sequence profiles.

We have used our approach to identify conservation patterns for seed proteins taken from nine superfolds. Some conservation patterns were identified that may provide insights as to the factors

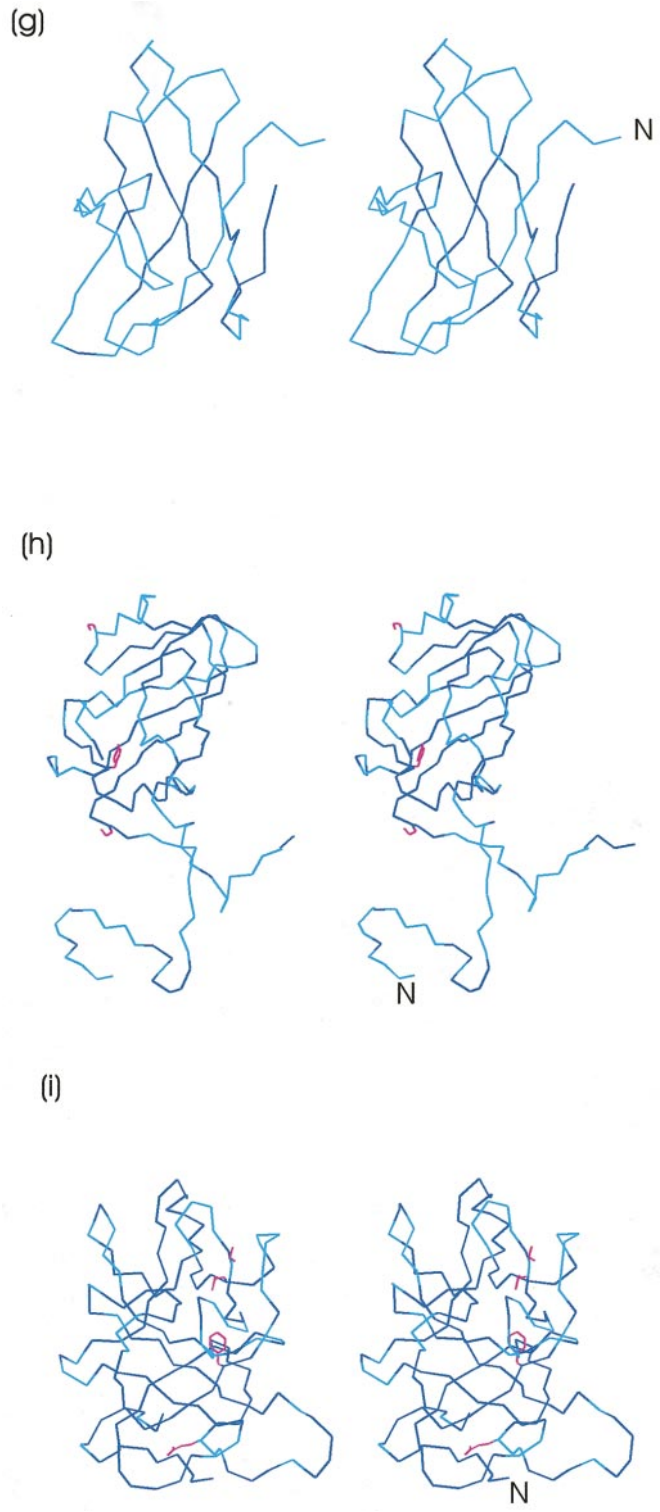


Figure 12 (*legend shown on page 708*)

that determine the topology of each group of proteins.

There does not appear to be a unique way to encode 3D structure on primary sequence. When a conserved sequence pattern is observed, it is likely to represent the consequences of divergent evolution rather than topological constraints associated with a particular fold.

We have presented an alternative view and methodology to those inherent in protein classification schemes. Rather than rely on pre-existing definitions, our approach, as embodied in PrISM,

has been to create a tool that makes it possible to carry out and evaluate alignments, and derive sequence and structure relationships that are based on objective criteria and parameters that can be varied and studied in a consistent fashion.

Methods

Multiple structure alignment

The PrISM multiple structural alignment procedure involves two stages. The initial multiple alignment involves a stepwise procedure. An all-against-all pairwise structure alignment is first carried out with the PrISM structure alignment algorithm (see our accompanying paper I) and a PSD score is obtained for each pair of proteins. The multiple alignment procedure starts with N groups of proteins, each of which initially contains one sequence. The structural distance between a pair of groups is simply the averaged pairwise PSD of the proteins involved. The pair of groups that are the closest in structure (i.e. the lowest average PSD) are then merged to generate a two-sequence profile based on the structural alignment for topologically equivalent residues of the pair of sequences. After the merge, the structural distances among the now $N-1$ groups are recalculated based on the averaged PSD as shown in equation (2):

$$\text{Structural distance between groups A and B} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} PSD(A_i, B_j) + PSD(B_j, A_i)}{2ab} \quad (2)$$

where protein A_i belongs to group A and protein B_j belongs to group B. Group A and B contain a and b sequences, respectively, and PSD is defined in our accumpanying paper I. Again, the pair of groups of proteins that are the closest in structure are merged to generate an $(a+b)$-sequence profile based on the pairwise structural alignment of the most similar pair of proteins, one from each of the groups. The merge cycle continues until all the N proteins merge into one group to generate an N-sequence profile, where residues that are located in equivalent topological positions in the 3D structure are aligned in the same column of the sequence profile. The initial structural multiple alignment is then followed by an iterative refinement procedure, as described in the following paragraph, to produce the final structure-based sequence profile.

The initial structure alignment procedure generates a structure-based sequence profile and multiple aligned 3D structures in the same coordinate system. The multiply aligned 3D structures are optimally superimposed, based on the same set of pairwise structural alignments that were used to generate the structure-based sequence profile. The refinement procedure proceeds by selecting sequentially one protein after the other. For each selected protein, a scoring matrix is calculated from the following equation:

$$S_{ij} = \sum_{k=1}^{n} \frac{p}{\max(DIST(C\alpha_i - C\alpha_j^k)^2, 1)} \quad (3)$$

where p is the constant defined by Yang & Honig (2000a) and a default value of 50 Å is used in this series of work. $DIST(C\alpha_i - C\alpha_j^k)$ is the C^{α}-C^{α} distance between residue i in the selected protein to each of the residues that are aligned in the same column j of the structure-based sequence profile. The C^{α}-C^{α} distances are calculated with the C^{α} coordinates of the superimposed structures produced by the initial multiple structural alignment. Note that not all the $N-1$ positions in the column j of the profile need to be filled with a residue, since some proteins might have a deletion in a particular position. The deletion contributes nothing to the element of the scoring matrix, i.e. n is less than or equal to $N-1$. A Needleman-Wunsch dynamic programming run follows the construction of the scoring matrix and the maximum match pathway of the scoring matrix aligns the sequence to the $(N-1)$-sequence profile to produce a new N-sequence profile. A single cycle involves repeating this procedure for each of the N proteins, and cycles are repeated until the structure-based sequence profile converges to a stable solution. In general, five cycles are usually enough to reach a converged profile. The multiple alignment of the 13 immunoglobulins shown in

Figure 12. The backbone traces of representative seed structures for the nine superfolds and the locations of residues with amino acid consistency values greater than 3 (shown in magenta). The backbone atoms are color-coded for the average local structure deviation from the geometrically averaged positions defined by the optimally superimposed structures. Backbone atoms colored in cyan are residues with average RMSD > 3 Å, while backbone atoms colored in blue are residues with average RMSD ⩽ 3 Å. (a) Globin. The seed structure shown is 1eca (1-136), and the structural neighbors are 2lhb (1-149), 1mbd (1-153), 1flp (1-142), 1mba (1-146), 1ash (0-146), 1babA (1-142), 1hbg (1-147), 1ithA (1-141), 3sdhA (2-146), 1hlb (1-157), 1lh1 (1-153). (b) α-Up-down helical bundle. The seed structure shown is 2ccyA (2-128). The structural neighbors are 1cgo (2-125), 1bbhA (1-131), 1cpq (1-129), 1cpq (1-129), 256bA (1-106), 1fapB (18-112). (c) Split αβ sandwich. The seed structure shown is 2fxb (1-81), and the structural neighbors are 1vjw (1-58), 1fxrA (1-64), 1fd2 (1-106), 1fca (1-55), 1xer (1-103), 1cnr (1-46), 1afi (1-72), 1pba (1-81), 1ctf (53-120). (d) UB αβ roll. The seed structure shown is 1ubi (1-76), and the structural neighbors are 1esfA (121-233), 1se4 (122-239), 1guaB (56-131), 1igd (1-61), 2ptl (1-78). (e) α/β Doubly wound. The seed structure shown is 3chy (2-129), and the structural neighbors are 2fx2 (2-148), 1rcf (1-169), 5nul (1-138), 1qrdA (1-273), 1rnl (5-142), 1srrA (3-121), 1reqA (561-728), 1bmtA (741-896), 1tadA (27-342), 1hurA (2-181). (f) TIM-barrel. The seed structure shown is 1pii (255-452), 1nsj (1-205), 1pii (1-254), 1igs (2-248), 1wsyA (1-265), 1byb (5-494). (g) Greek-key immunoglobulin. The seed structure shown is 1tlk (33-135), and the structural neighbors are 1fna (6-96), 1ten (803-891), 3hhrB (131-234), 4kbpA (9-120), 1ctn (24-132), 1cfb (610-709), 3hhrB (32-130), 1cd1A (186-279), 1dlhA (82-182), 1ggtA (628-728), 1ggtA (516-627), 3cd4 (98-178), 1ksr (1-100), 2hft (1-106), 1tiu (1-89), 2ncm (1-99), 1vcaA (1-90), 1wiu (1-93), 1zxq (1-86), 1tnm (1-91), 1nciA (2-100), 1who (1-94). (h) Jelly roll. The seed structure shown is 1bbt1 (1-208), and the structural neighbors are 2mev1 (1-268), 1tme1 (1-256), 1pvc1 (24-302), 1bbt3 (1-220), 1pvc3 (1-235), 1smvA (65-260). (i) Trefoil. The seed structure shown is 1i1b (3-153), and the structural neighbors are 1ilr1 (8-152), 1tie (1-170), 1wba (2-172).

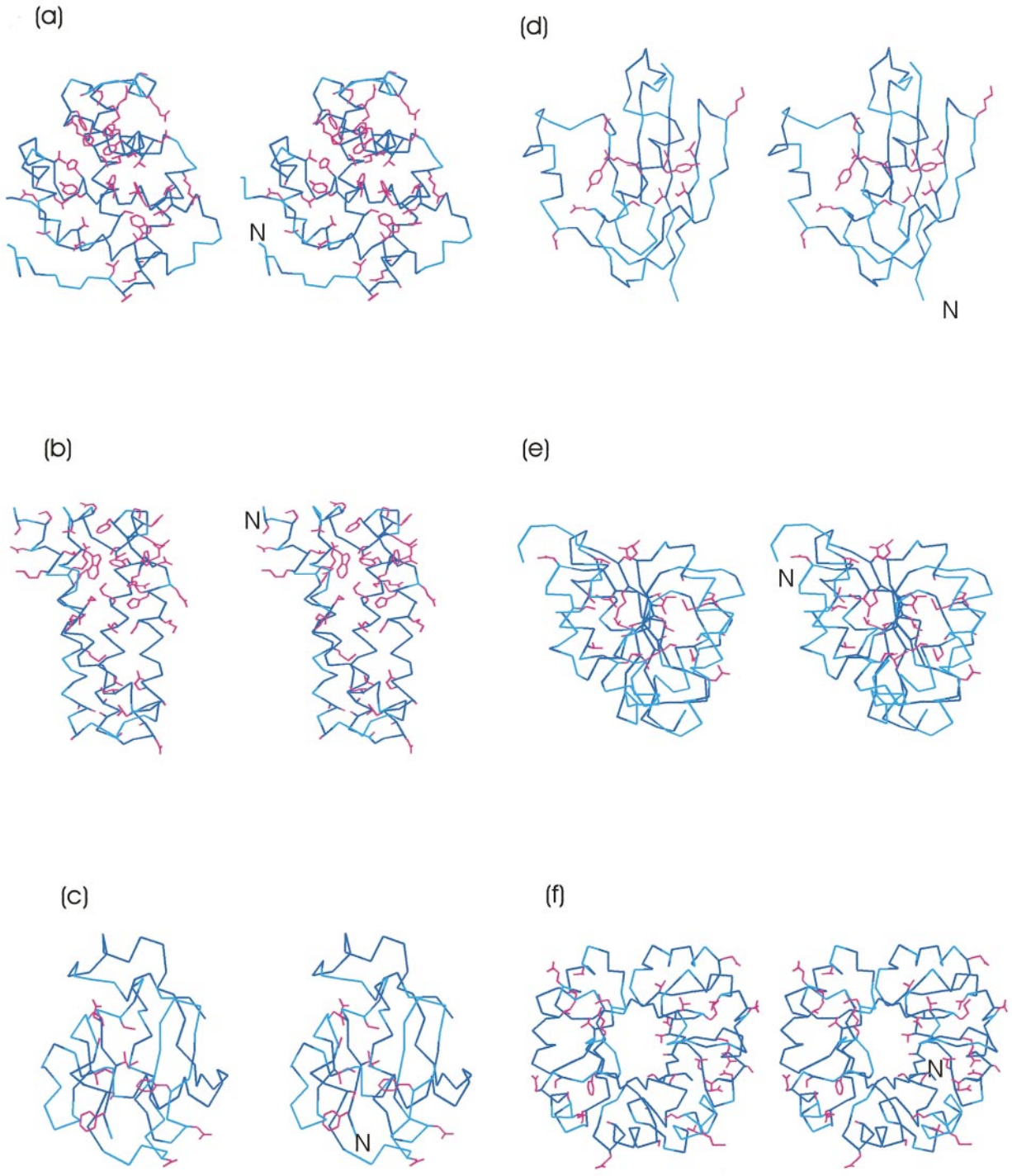


Figure 13 (*legend shown on opposite page*)

Figures 1 and 2, with about 120 residues per protein, takes less than a total of two minutes CPU time on an SGI R10000 work-station.

Protocol automatically to generate sequence profiles

All structure-based sequence profiles were constructed automatically with PrISM. The protocol starts with a search procedure in which a representative structure of a superfold family member is used to search for all of its structural neighbors with PSD scores less than some predetermined threshold (1 in this work). The structures used to represent superfolds were obtained from Orengo *et al*. (1994). All the structural neighbors are then clustered with a PSD threshold =1. A structure-based sequence profile is constructed using the PrISM multiple

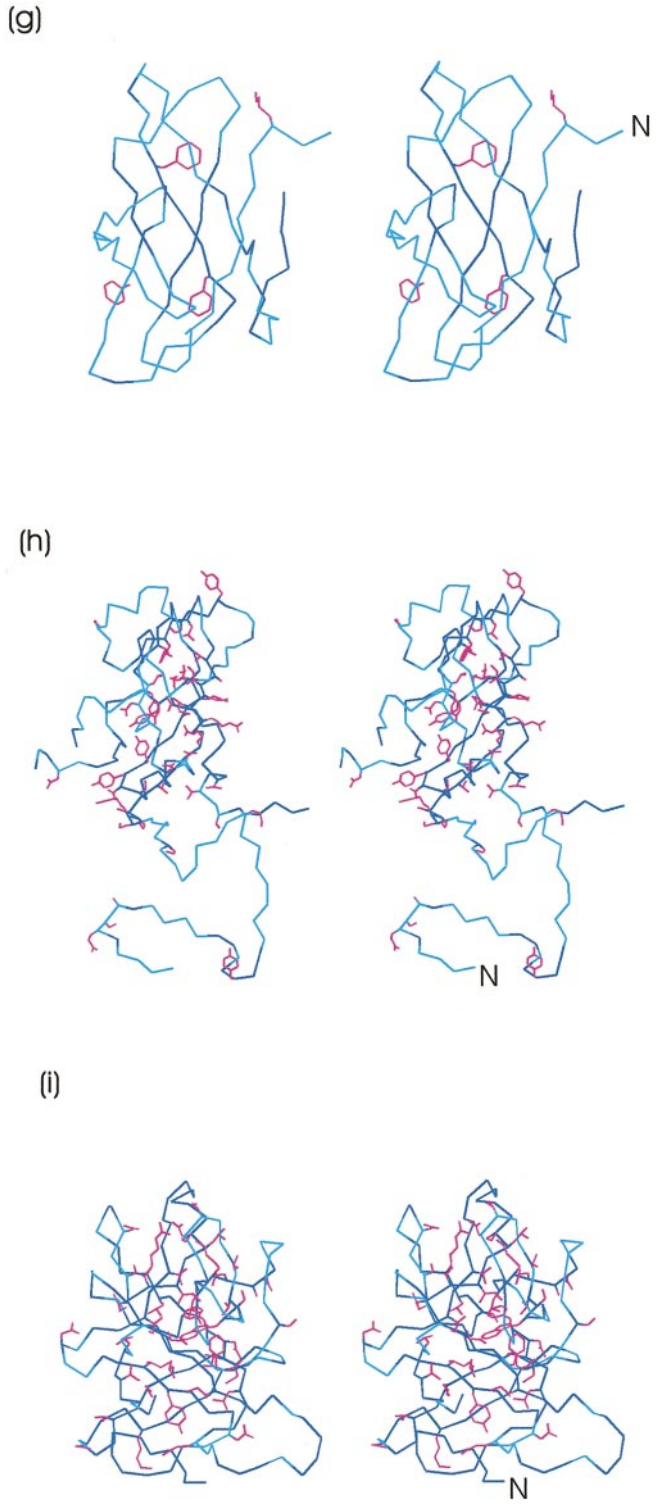


Figure 13. The backbone traces of representative seed structures for the nine superfolds and the locations of residues with amino acid consistency values greater than 1 (shown in magenta). See Figure 12 for details.

structure alignment procedure for the cluster that contains the original seed structure.

It is instructive to compare the PrISM multiple structural alignment procedure to that of CORA (Orengo, 1999). Both CORA and PrISM involve an initial stage that is used to select representatives from a database of protein structures using subjective measures for protein structure similarity. The CORA multiple structural alignment procedure is a simple successive alignment of selected members in the order of scores derived from pairwise structure alignment. A similar procedure is implemented in the PrISM first stage of multiple align-

ment. The PrISM second stage is unique in refining the initial sequence profile using an iterative structure-profile alignment procedure (see above). CORA was designed to work on protein sets defined in the CATH database, where proteins are classified hierarchically, while PrISM is designed to automatically produce multiple structural alignments without using any information derived from existing classification schemes. This is useful for proteins that have not been classified in databases such as CATH of SCOP, or for cases where the existing classification might not be appropriate and when one wishes to avoid classifications and to derive alignments based on objective criteria such as PSD threshold.

Software download

Information as to how to obtain the software tools described in this work can be obtained over the web from either http://www.columbia.edu/ ~ ay1/or honiglab.cpmc.columbia.edu

Acknowledgements

We thank Drs Burkhard Rost, Bissan Al-Lazikani and Diana Murray for their many useful comments and critical reading of the manuscript. This work was supported by DOE grant 96ER62265.

References

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199-216.

Bork, P., Holm, L. & Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* **242**, 309-320.

Doolittle, R. F. & Feng, D. F. (1990). Nearest neighbor procedure for relating progressively aligned amino acid sequences. *Methods Enzymol.* **183**, 659-669.

Feng, D. F. & Doolittle, R. F. (1990). Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol.* **183**, 375-387.

Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.

Gerstein, M. & Altman, R. B. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161-175.

Halaby, D. M., Poupon, A. & Mornon, J. (1999). The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng.* **12**, 563-571.

Henikoff, S. & Henikoff, J. (1993). Performance evaluation of amino acid substitution matrices. *Proteins: Struct. Funct. Genet.* **17**, 49-61.

Holm, L. & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478-480.

Jennings, P. A. & Wright, P. E. (1993). Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin [see comments]. *Science,* **262**, 892-896.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

Landgraf, R., Fischer, D. & Eisenberg, D. (1999). Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* **12**, 943-951.

Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stabiity, foldingkinetics and function. *J. Mol. Biol.* **291**, 177-196.

Orengo, C. (1994). Classification of protein folds. *Curr. Opin. Struct. Biol.* **4**, 429-440.

Orengo, C. A. (1999). CORA–topological fingerprints for protein structural families. *Protein Sci.* **8**, 699-715.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631-634.

Ptitsyn, O. B. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* **278**, 655-666.

Roder, H. & Colon, W. (1997). Kinetic role of early intermediates in protein folding. *Curr. Opin. Struct. Biol.* **7**, 15-28.

Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struct. Funct. Genet.* **14**, 309-323.

Russell, R. B. & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.* **244**, 332-350.

Sali, A. & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212**, 403-428.

Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.

Yang, A.-S. & Honig, B. (2000a). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* **301** 665-678.

Yang, A.-S. & Honig, B. (2000b). An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.* **301**, 679-690.

Yang, A.-S., Hitz, B. & Honig, B. (1996). Free energy determinants of secondary structure formation: III. Beta-turns and their role in protein folding. *J. Mol. Biol.* **259**, 873-882.

Edited by F. E. Cohen

(Received 23 February 2000; received in revised form 12 June 2000; accepted 12 June 2000)