Protein Science (1995), *4*:93–102. Cambridge University Press. Printed in the USA.
Copyright © 1995 The Protein Society

A procedure for the automatic determination of hydrophobic cores in protein structures

MARK B. SWINDELLS

Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan

(RECEIVED May 18, 1994; ACCEPTED November 9, 1994)

Abstract

An algorithm is described for automatically detecting hydrophobic cores in proteins of known structure. Three pieces of information are considered in order to achieve this goal. These are: secondary structure, side-chain accessibility, and side-chain–side-chain contacts. Residues are considered to contribute to a core when they occur in regular secondary structure and have buried side chains that form predominantly nonpolar contacts with one another. This paper describes the algorithm's application to families of proteins with conserved topologies but low sequence similarities. The aim of this investigation is to determine the efficacy of the algorithm as well as to study the extent to which similar cores are identified within a common topology.

Keywords: hydrophobic cores; protein structure; structural analysis

Anyone familiar with structure analysis will know that, although globular proteins adopt many different topologies, two simple structural properties are always observed. Firstly, there is a strong preference for the main chain to adopt α-helical and β-strand conformations, and secondly, a large proportion of the nonpolar side chains pack together to form a buried hydrophobic core. In addition, although loops can vary extensively between homologous proteins, the tertiary arrangements of helices and strands can remain conserved, even in the absence of sequence similarity.

Algorithms for detecting such features constitute important research tools for the structural biologist and, through the work of previous researchers, it is now possible to assign secondary structure (Kabsch & Sander, 1983; Richards & Kundrot, 1988), calculate residue accessibilities (Lee & Richards, 1971), and search for recurrent topologies in a database of known structures (Mitchell et al., 1989; Taylor & Orengo, 1989; Alexandrov et al., 1992; Holm & Sander, 1993). It is surprising, therefore, that hydrophobic cores have received relatively little attention. One of the only algorithms available employs a combination of residue hydrophobicities and side-chain–side-chain distances for determining the core residues (Umezawa & Umeyama, 1988). However, this frequently leads to the inclusion of residues with unusually high accessibilities (>30% relative solvent accessibility) and contrasts sharply with most other analyses, where the highest structural conservation has been observed in regions with around 5–15% relative accessibility (Chothia, 1975; Hubbard & Blundell, 1987; Miller et al., 1987).

The algorithm described in this paper essentially formalizes the discoveries of many researchers who have studied a variety of well-known folds, such as the globins, immunoglobulins, and azurins (Lesk & Chothia, 1980, 1982; Chothia & Lesk, 1982; Bashford et al., 1987; Murzin et al., 1992). Their quantitative assessments have collectively shown that:

1. Many different sequences (frequently with no statistically significant similarity) are able to adopt the same topology.
2. Their least variable regions correspond to elements of regular secondary structure.
3. Within these elements, buried hydrophobic residues are the most highly conserved.

Thus, a suitable working definition for hydrophobic cores might be collections of residue sites with low solvent accessibilities, which are located in regions of regular secondary structure and whose nonpolar side-chain moieties interact with one another. Even with this rather severe definition, core sites may still vary between members of the same topology because global shifts in tertiary structure, resulting from the unique properties of each sequence, will ultimately determine the constituents of each core (Lesk & Chothia, 1980; Bashford et al., 1987; Swindells & Thornton, 1993). Nevertheless, this is the region where variation is least likely. The following algorithm, which requires full atom coordinates for its implementation, automatically assigns hydrophobic cores based on these observations.

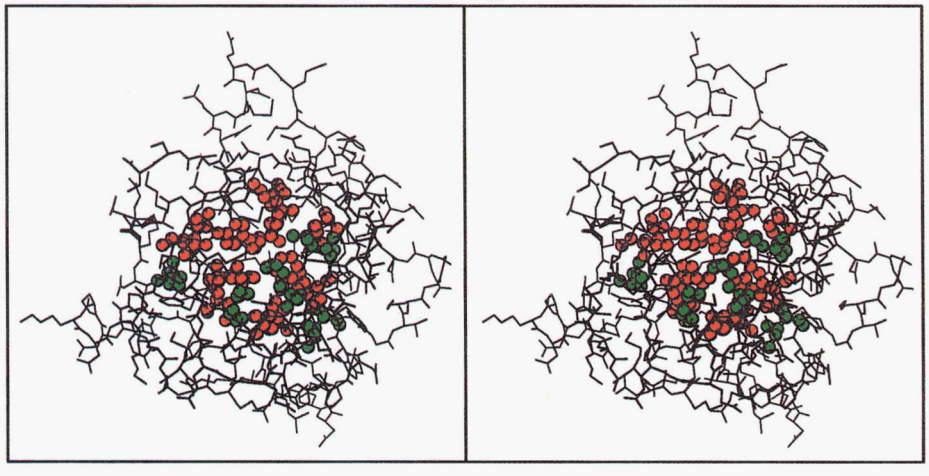
Reprint requests to Mark B. Swindells at his present address: Molecular Design Department, Yamanouchi Pharmaceutical Co. Ltd., 21 Miyukigaoka, Tsukuba, Ibaraki 305, Japan; e-mail: mark@yamanouchi.co.jp.

Results

In order to assess the efficacy of this algorithm, hydrophobic cores have been assigned to members of various structural superfamilies. Within each superfamily, proteins with low sequence identities were chosen. The aim of selecting such sequentially disparate proteins is to show that the algorithm generally identifies similar regions of a topology even when the specific structural features of each protein vary. In each of the following examples, protein cores are assigned using two values of accessibility. A 7% cutoff is chosen because it is in line with a previous study of buried residues (Hubbard & Blundell, 1987), whereas 15% (approximately twice 7%) is used to illustrate the concomitant increase in core size with accessibility.

Assignment of cores for the β-trefoil fold

The β-trefoil topology occurs in many proteins with no detectable sequence similarity (Murzin et al., 1992; Swindells & Thornton, 1993). Twelve strands constitute this fold; six form a barrel structure and the remainder are located at the open end of the barrel. Because β-barrels are less able to adapt to large structural changes, the topology remains highly conserved even in the absence of sequence similarity and therefore represents a good test case for the program. Figure 1 shows the core assignments for interleukin-1β (IL-1β), basic fibroblast growth factor, and *Erythrina* trypsin inhibitor. From this figure, it is clear that the assignments for all three structures agree well with one another, as well as with the 18 core sites (denoted by black circles) de-



```
1i1b    1 APVRSLNCTLRDSQQ---KSLVMSG------PYELKALHLQGQDMEQQVVFSMSFVQGEESN-
1tie    9 VVQNGGTYYLLPQVWAQGGGVQLAKTGEETCPLTVVQSPNELSD---GKPIRIESRLRSAFIP
2fgf   14 PGHFKDPKRLYCKNG--GFFLRIHP------DGRVDGVREKSDP---HIKLQLQAEE------

1i1b   54 DKIPVALGLKEK-----NLYLSCVLKDD-KPTLQLESVDPKNYPKKKMEKRFVFNKIEIN---
1tie   69 DDDKVRIGFAYAPKCAPSPWWTVVEDEQEGLSVKLSEDESTQF-------DYPFKFEQVSDQL
2fgf   60 -RGVVSIKGVCA-----NRYLAMKEDG----RLLASKC---------VTDECFFFERLESN--

1i1b  108 NKLEFESAQF---PNWYISTSQAEN--MPVFLGGTKGGQDI-TDFTMQFVSS
1tie  125 HSYKLLYCEGKHEKCASIGINRDQKGYRRLVVTEDYP-----LTVVLKKDESS
2fgf  102 NYNTYRSRKY---TSWYVALKRT----GQYKLGSKTGPGQKAILFLPMSAKS
```

Fig. 1. Structurally derived sequence alignment of interleukin-1β (1i1b), basic fibroblast growth factor (2fgf), and *Erythrina* trypsin inhibitor (1tie), with the hydrophobic core residues highlighted. Red residues identify the cores automatically assigned using a relative accessibility cutoff of 7%. Green residues show the corresponding increase in core size when this cutoff is set to 15%. Black circles indicate core sites identified by Murzin et al. (1992). β-Strands are indicated by arrows below each sequence. The stereo diagram shows the IL-1β hydrophobic core in a structural context.

scribed previously by Murzin et al. The results indicate that similar topological features are being identified in each structure, despite the different sequences and structural variations that are known to exist (particularly near the C-terminus, where no regular secondary structure is recorded in basic fibroblast growth factor).

The effects of limiting hydrophobic core assignments to regions of secondary structure can be assessed by repeating the calculation while discounting secondary structure assignments. When this is implemented, the following changes are observed. The IL-1β core increases by four residues (Val-47, Ala-59, Met-95, and Val-132), whereas the basic fibroblast growth factor core increases by only two residues (Val-116, Val-118). In contrast, the *Erythrina* trypsin inhibitor core decreases by one residue because, although Val-10 and Ile-67 are added, Leu-31, Tyr-127, and Val-164 are lost due to the additional interactions that are now considered. For researchers who are interested in studying the contribution of hydrophobic cores to the stability of a protein, implementing the algorithm with all residues may be preferable. However, in this paper, emphasis is given to detecting sites that are conserved over an entire superfamily. Because these will inevitably be restricted to regions of regular secondary structure, the following examples only show the results of limiting assignments to these regions.

Although the size of the core inevitably depends on the accessibility cutoff applied, this variation is less than might be expected. In fact, a graph showing the variation of core size with accessibility (Fig. 2) clearly shows that, even when accessibility is discounted (relative accessibility cutoff set at 100%), the number of residues identified does not exceed 20% of each structure.

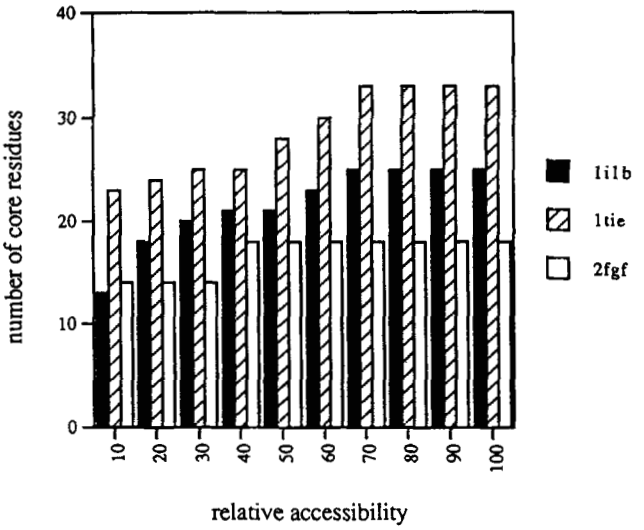


Fig. 2. Graph showing the dependence of core sizes on the relative accessibility cutoff for three structures; IL-1β (1il b), basic fibroblast growth factor (2fgf), and *Erythrina* trypsin inhibitor (1tie).

Assignment of cores for the globin fold

The second application is to members of the all-helical globin fold. In Figure 3, a structurally derived sequence alignment of four globins is shown together with cores assigned by the algorithm. Once again, there is good agreement between the cores assigned for all four structures. These results are also in line with the collective results of Lesk, Chothia, and coworkers, who have performed two distinct, but complementary analyses. In their first paper (Lesk & Chothia, 1980) a comparison of nine globin structures identified 31 buried sites (accessibility ≤ 10 Å^2) whose residues consistently made helix–helix interactions (black squares in Fig. 3). In the second study (Bashford et al., 1987), a comparison of 226 globin sequences led to the identification of 32 sites containing predominantly hydrophobic residues (black circles in Fig. 3). However, although more than 30 residue sites were identified in each case, only 25 sites were common to both, and of these, 20 sites are automatically detected by my algorithm in at least one of the four structures shown. As well as showing that there is a correlation between both automated and manual analyses, these results also show that core characteristics will vary over an entire superfamily.

Assignment of cores in immunoglobulin topologies

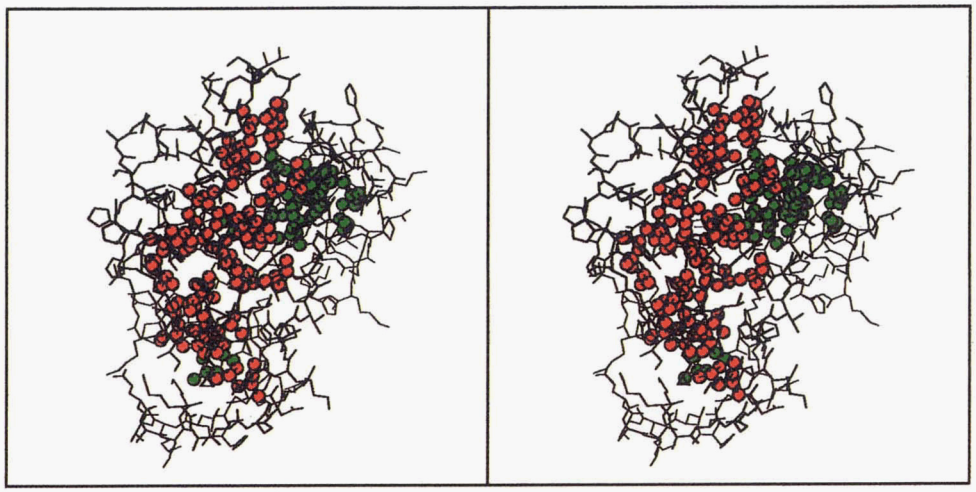
The cell surface glycoprotein CD4 consists of two domains, the first of which adopts a classical immunoglobulin fold. Despite this structural similarity, no significant sequence similarity is observed with the immunoglobulins. In Figure 4, assignments for the first domain of human CD4 are compared with the equivalent assignments for the Bence–Jones immunoglobulin variable domain (REI). Between 7% and 15% accessibility, 12 core residues are assigned to the first domain of CD4, whereas 11 are assigned to REI. Nine sites are common to both cores.

Assignment of cores for plastocyanin and azurin

Although both plastocyanin and azurin are β-sheet structures with low sequence similarities, the latter contains a large insertion between residues 56 and 88, and this leads to significant conformational differences between the two proteins. In addition, both proteins contain nonsuperimposable α-helices in this region and so it was of interest to determine whether any of their residues contribute to the core or whether, as anticipated (Chothia & Lesk, 1982), they are of minor importance. Figure 5 shows the cores assigned for these two structures at 7% and 15% relative accessibility. With a cutoff of 15% accessibility, 14 sites are identified in plastocyanin and 16 in azurin, with 12 being common to both structures. No core residues are assigned to the inserted regions, thus confirming their peripheral importance.

Assignment of cores in four-helix bundles

Although cytochrome *b*562 and hemerythrin both have four-helix bundles that can be aligned to give small RMS deviations (Orengo et al., 1993), their functions differ; cytochrome *b*562 is a heme-containing electron carrier, whereas hemerythrin is a non-heme oxygen-transport protein. Larger variations are observed between these core assignments and these can be attributed to the presence of different prosthetic groups as well as the long N-terminal peptide in hemerythrin (Fig. 6). This extra N-terminal region, which is stabilized by interactions with the four-helix bundle, leads to Phe-30 of hemerythrin being classified as a core residue, whereas its structurally aligned site in



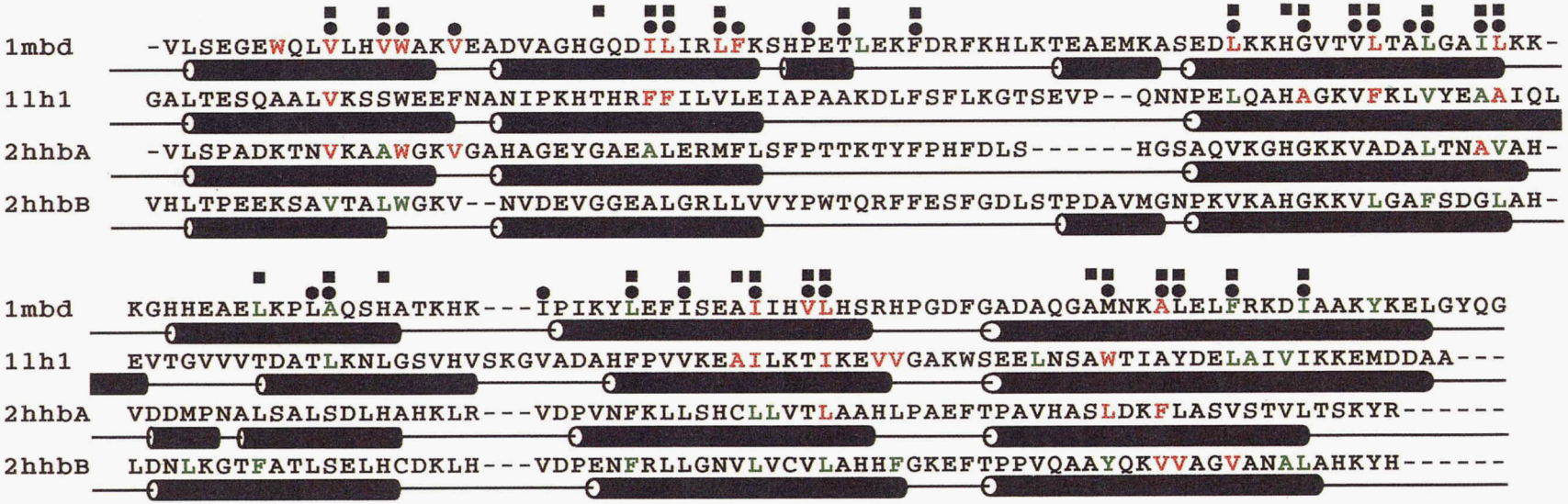


Fig. 3. Structurally derived sequence alignment of leghemoglobin (1lh1), myoglobin (1mbd), and hemoglobin chains A and B (2hhb), with the hydrophobic core residues highlighted. Red residues identify the cores automatically assigned using a relative accessibility cutoff of 7%. Green residues show the corresponding increase in core size when this cutoff is set to 15%. Black squares show the 31 buried sites identified through structural comparisons (Lesk & Chothia, 1980), whereas black circles show the 32 sites identified using homologous sequences and structures (Bashford et al., 1987). α-Helices are indicated by cylinders below each sequence. The automated structural alignment of leghemoglobin around helix F (residues 88–99) is shifted by one helical turn relative to the alignment published by Chothia and coworkers. However, this has a negligible effect on the comparisons described in this paper. A stereo diagram shows the myoglobin core in a structural context.

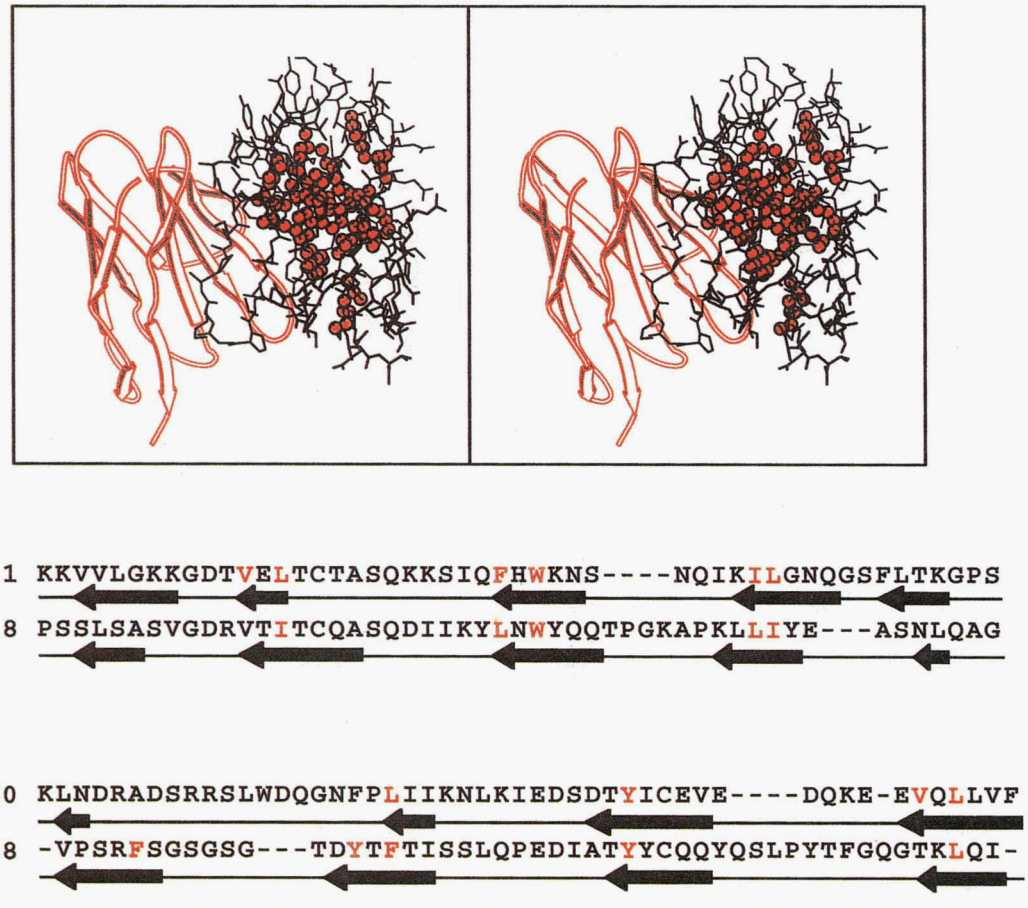


Fig. 4. Structurally derived sequence alignment of human CD4 protein (1cd4) and the Bence–Jones immunoglobulin variable domain (1rei). Layout of the figure follows previous examples. The stereo diagram shows the Bence–Jones variable domain dimer, with one indicating the hydrophobic core and another indicating the regions of secondary structure.

the cytochrome *b*562 is solvent accessible. More importantly, however, the structural environments of the prosthetic groups differ significantly, with the large heme group of cytochrome *b*562 slotting between two helices like a small wedge and the μ-oxo bridged iron group of hemerythrin locating more centrally within the four-helical bundle. These variations, coupled with different helix lengths, result in the different core assignments highlighted (Fig. 6).

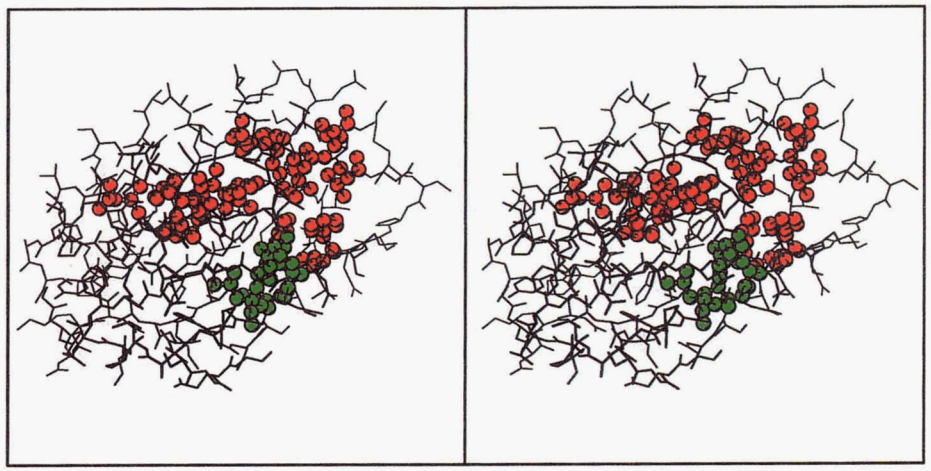
Assignment of cores in small proteins

The size of a hydrophobic core is inevitably linked to a protein's overall size. Below 50 residues, stable structures are generally limited to those with disulfide bridges and, even above this size, disulfide bonds are frequently recruited for extra stability. One such example is bovine pancreatic trypsin inhibitor (BPTI, 58 residues), which contains three disulfides but has no residues assigned to a hydrophobic core by this program. Closer analysis reveals that there are only four residues (excluding cysteine) with side-chain accessibilities of less than 10%, and even these do not cluster (Fig. 7A). In contrast, the slightly larger λ repressor (92 residues), which contains no disulfide bridges, has a seven-residue core clearly identified by the program (Fig. 7B).

Discussion

In this paper, I have described an algorithm for automatically assigning hydrophobic cores to proteins of known structure. One of the main difficulties encountered during its development was to show that the cores assigned by the algorithm were meaningful rather than merely being an inevitable product of the assignment process. In order to circumvent this problem, the program was applied to well-known, topologically similar, yet structurally unique proteins, because it was anticipated that the cores should remain similar even though the details of each structure were varying.

Overall, the cores identified appear to be in good agreement, both with one another and with previously published analyses. However, the following points should be noted. Cores identified in β-topologies appear to be more conserved than in α-proteins. Evolutionary distance between members of each superfamily may play a role, but this seems unlikely because many of the β-proteins compared (e.g., IL-1β and *Erythrina* trypsin inhibitor) have no statistically significant sequence identity. A more likely explanation is that long-range hydrogen bonds limit strand movements within each β-sheet. In this manner, β-barrel structures would be the most conserved. This is also in line with previous analyses (Hubbard & Blundell, 1987), which have shown



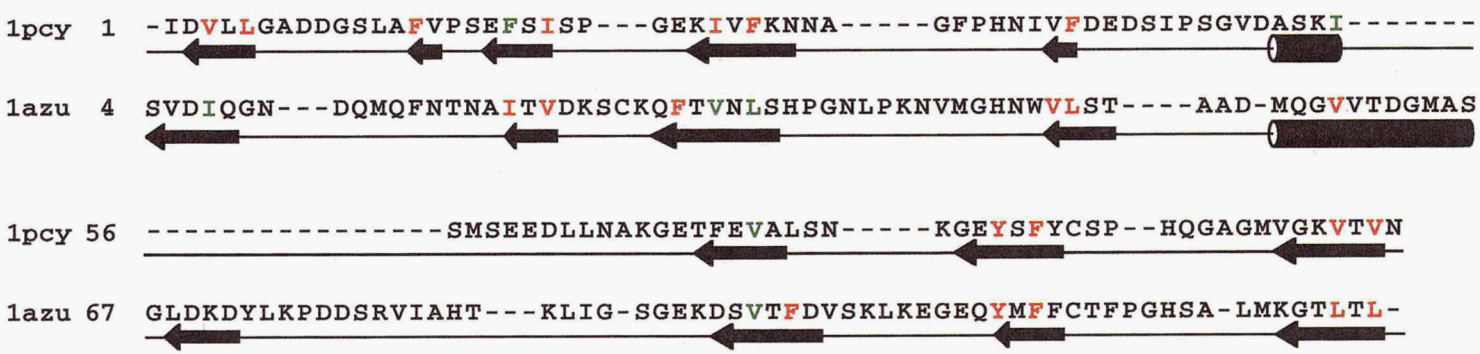
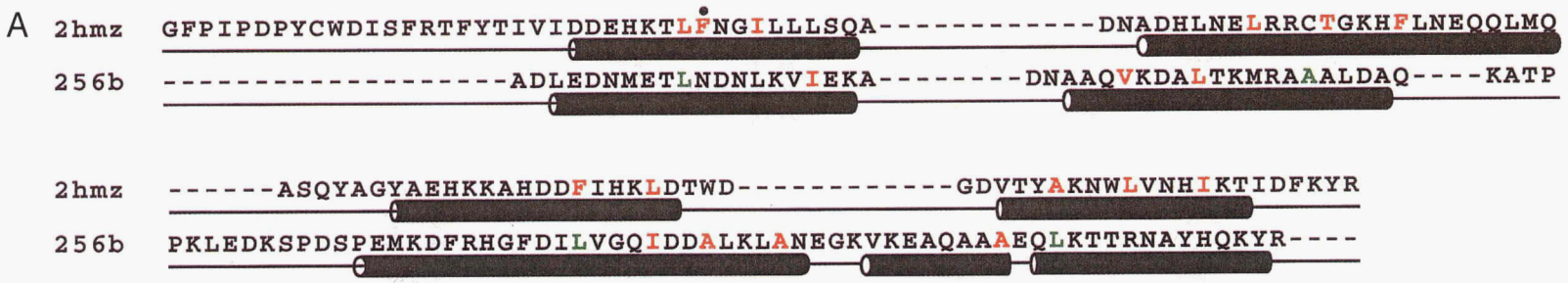


Fig. 5. Structurally derived sequence alignment of plastocyanin (1pcy) and azurin (1azu). Layout of the figure follows previous examples. The stereo diagram shows the azurin core in a structural context.



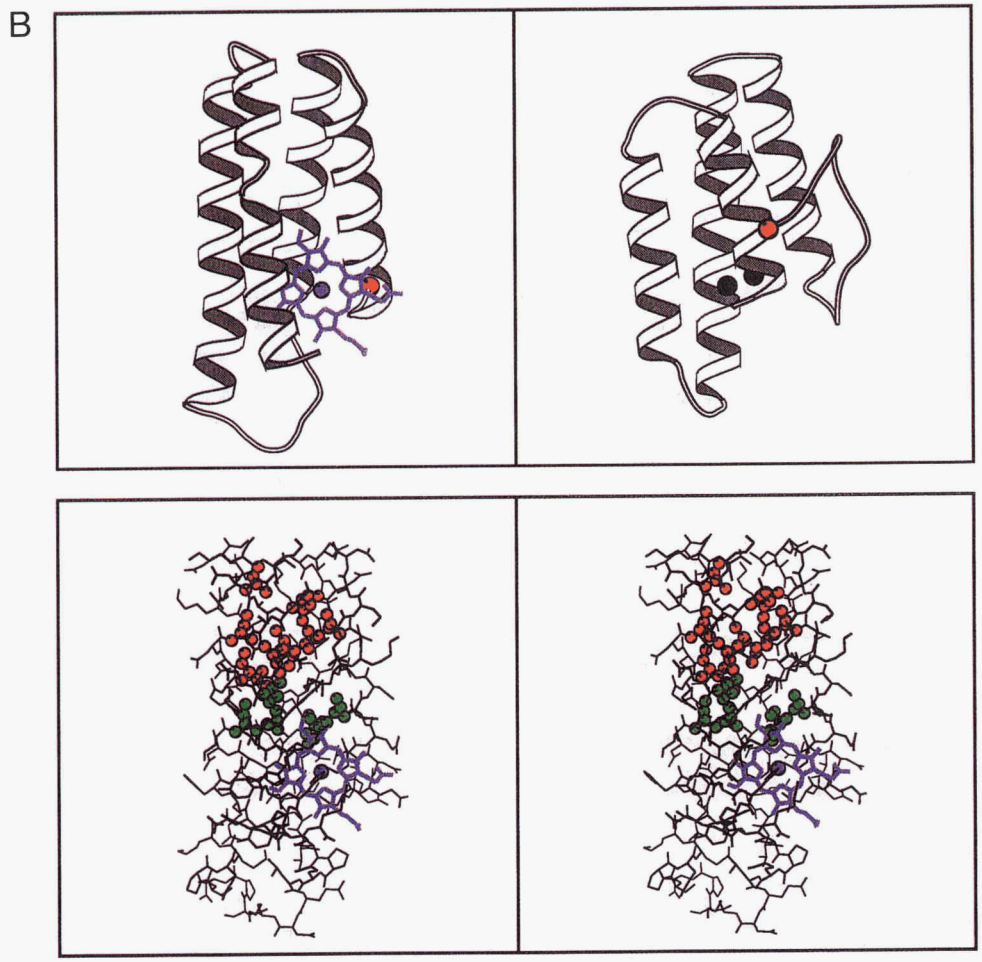


Fig. 6. A: Structurally derived sequence alignment of cytochrome *b*562 (256b) and hemerythrin (2hmz). Layout of the figure follows previous examples. Phe-30 of hemerythrin is marked with a black circle. **B:** Ribbon diagrams of cytochrome *b*562 (top left) and hemerythrin (top right). N-termini are indicated with red balls. The di-iron site in hemerythrin is indicated with black balls. Below is a stereo diagram of cytochrome *b*562 showing the core in a structural context.

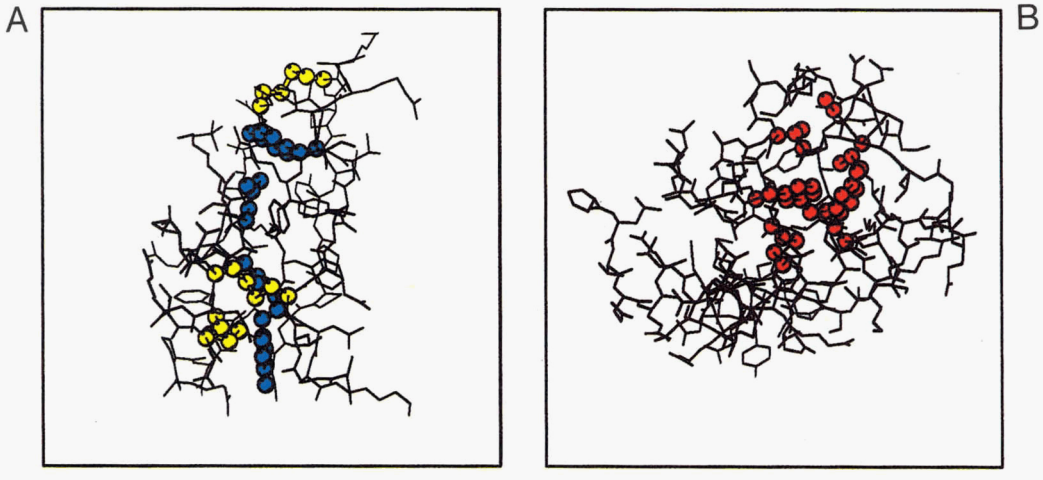


Fig. 7. A: Bovine pancreatic trypsin inhibitor, which has no core detected, is displayed with its disulfide bridges (yellow) and four buried non-glycine residues (Tyr-23, Tyr-35, Asn-43, Asn-44) highlighted in cyan. **B:** λ Repressor shown with its seven hydrophobic core residues (Leu-18, Tyr-22, Val-36, Ala-37, Val-47, Phe-51, Leu-65) colored red.

that superimposed α-proteins have higher RMS deviations than β-structures. Furthermore, this difference is even more pronounced when the comparisons are limited to residues with no more than 7% relative accessibility. Therefore, differences detected by this algorithm may reflect important structural variations within each superfamily.

Another point is that the cores assigned using this algorithm tend to be smaller than those identified by manual analysis. This can generally be attributed to the use of secondary structure in the calculation and the subsequent reliance on the Kabsch and Sander (1983) algorithm, which usually assigns shorter helices and strands than the equivalent author assignments (Morris et al., 1992). Although the reader may initially consider it surprising to limit core residues to regions of regular secondary structure, it is in fact a reasonable approximation for this analysis. Loop conformations vary substantially, even between proteins with high sequence identities. As a result, their contributions cannot be consistent, even though each loop may contribute to a protein's stability. Following this line of reasoning, common structural determinants will coincide with the helices and strands, which are generally conserved across the entire superfamily. This situation is clearly observed in the independent analyses of Chothia and coworkers, whose assignments have been used for comparison in this paper. Furthermore, implementing the algorithm without the requirement of regular secondary structure, as shown for the β-trefoil fold, still places the majority of core sites in the regions of secondary structure. Improvements may be gained by using a more lenient strategy for secondary structure assignment. One possibility is to include regions of a 3_{10}-helix in the calculation, whereas another possibility is to apply the Kabsch and Sander algorithm with a different energy threshold.

Applications of the algorithm

An immediate application of this algorithm is to fold recognition, where the structure of a novel sequence is predicted by assessing its compatibility with the structural profiles of known proteins. In these approaches, three-dimensional structural information is either reduced to a simpler list of residue environments, derived from accessibility and secondary structure data (Bowie et al., 1991), or used as a model, through which the sequence of unknown structure is threaded (Jones et al., 1992). Although fold recognition depends on these profiles describing all members of a common topology, this is in fact the Achilles heel, because at low sequence identities the accessibility profiles of structures with common topologies can vary significantly. For example, a comparison of the side-chain accessibilities for 88 structurally alignable positions in IL-1β and *Erythrina* trypsin inhibitor gives a rather poor correlation coefficient of 0.41. Thus, a site that is buried in one structure may not necessarily be buried in another. In contrast, sites belonging to the hydrophobic core are more conserved and have the potential to provide a "cleaner" signal during fold recognition. However, until now there has been no automated method for identifying these sites from a single structure. The algorithm described here provides a solution.

The second application is to domain recognition. From the application of this algorithm to multidomain proteins it has become apparent that the number of cores detected corresponds well with the number of domains previously identified by manual analyses. Because of this good correlation, it is possible (with certain modifications) to deduce domain locations from the residues constituting each core. As well as being more intuitive than previous methods, it also has the advantage of being able to detect continuous as well as discontinuous domains with equal efficacy. Details of the method for identifying domains are described in the companion paper (Swindells, 1995). It is intended to make a version of this algorithm available for public use. For more information, please contact the author by e-mail at mark@yamanouchi.co.jp.

Methods

Assigning a core

For the assignment of residues to a hydrophobic core, secondary structure, solvent accessibility, and side-chain interaction

data are required for each residue. These data were calculated in the following manner:

1. Side-chain solvent-accessible surface area (ASA) was calculated using an implementation of the Lee and Richards (1971) algorithm written by Simon Hubbard using a probe size of 1.4 Å and sphere slice of 0.05 Å. Atomic van der Waals radii were taken from (Chothia, 1976). Relative accessibilities were calculated as a residue's observed ASA relative to its ASA in an Ala-X-Ala peptide of extended conformation. An Ala-X-Ala peptide was chosen in preference to Gly-X-Gly because the former is more typical of the situation observed in protein structures.
2. Secondary structure assignments were calculated using the method of Kabsch and Sander (1983).
3. Side-chain–side-chain interaction data were calculated in the following manner. Atomic interactions were considered to exist when two atoms were closer than the sum of their van der Waals radii plus 1 Å. Van der Waals radii were taken from Chothia (1976). By considering all possible atomic interactions between two side chains (extending from the C^{β} atom), data were collected for each residue pair (i,j) and subdivided into how many of these interactions were hydrophobic (simplified as those between pairs of carbon atoms; $h_{(i,j)}$) and how many were nonhydrophobic (all others; $nh_{(i,j)}$). After calculating the interaction data for all residue pairs, the total numbers of hydrophobic, $H_{(i)}$, and nonhydrophobic, $NH_{(i)}$, atomic interactions made by each residue were calculated.

After these initial calculations were performed, the following assignment procedure was applied.

For each residue i recorded as:

1. buried (the cutoff applied for buried side chains is discussed in the Results section);
2. located in helix or strand;
3. having more than 75% of its atomic side-chain interactions classified as hydrophobic (i.e., $H_{(i)}/[H_{(i)} + NH_{(i)}] > 0.75$).

A search was made in order to determine which other residues j were also:

4. located in helix or strand;
5. with $(H_{(j)}/[H_{(j)} + NH_{(j)}] > 0.75)$;
6. a majority of nonhydrophobic atomic interactions between the two residues i and j under consideration $(h_{(i,j)} > nh_{(i,j)})$.

If these criteria were passed, further investigations were made in order to determine whether residue j was also buried (b) or accessible (a).

If residue j was also buried:

1. $b_{(i)}$ was incremented to record an interaction between residue i and another buried residue; and
2. a specific record of the interaction (i,j) was made in a new matrix $Z_{(i,j)}$.

If residue j was accessible, $a_{(i)}$ was incremented to record an interaction between residue i and an accessible residue.

After all relevant comparisons had been completed, only buried residues forming a majority of buried interactions with other residues $(b_{(i)} > a_{(i)})$ were retained for further consideration. These residues were subsequently clustered on the basis that interacting pairs (recorded in matrix Z) will belong to the same core (Fig. 8).

On occasions where two sequentially adjacent core residues belonged to different clusters, with either of the clusters accommodating at least five residues and the adjacent residues both containing more than one buried contact $(b > 1)$, clusters were merged. This procedure was adopted to ensure that clusters detected on either side of a β-sheet were considered as a single hydrophobic core. A similar procedure was also applied when residues located at i and $i + 2$ of a β-strand belonged to different clusters.

Of the clusters calculated, only those consisting of at least five residues were classified as hydrophobic cores for the protein under consideration. The choice of five residues was somewhat arbitrary, but during the search for a suitable cutoff, it was found that smaller values frequently led to spurious clusters being recorded.

Structural alignments and atomic data

Structural alignments of related proteins were made using the method of Taylor and Orengo (1989). Atomic coordinates were

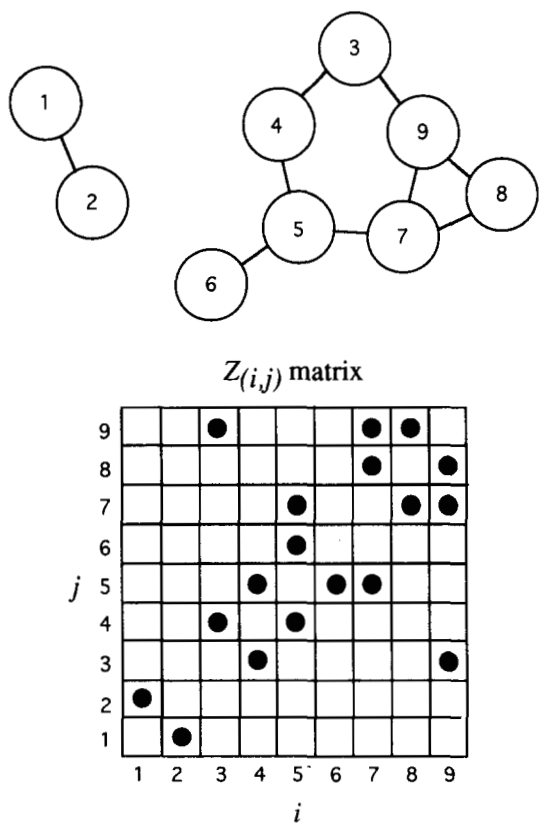


Fig. 8. Schematic diagram showing how residues are clustered. Each side chain is represented by a circle, and interactions between residue pairs are indicated by lines. Interaction data shown in the figure are summarized by the $Z_{(i,j)}$ matrix. Residues interacting with one another are assumed to belong to the same cluster. In this case, two clusters are identified, one including residues 1 and 2 and another encompassing residues 3–9. The cluster formed by residues 1 and 2 is too small to be classified as a core because this requires five residues.

taken from the Protein Data Bank (Bernstein et al., 1977). Files and references to the structures used in this paper are written as: protein name, PDB code (reference).

IL-1β, 1il1b (Finzel et al., 1989); *Erythrina* trypsin inhibitor, 1tie (Onesti et al., 1991); basic fibroblast growth factor, 2fgf (Zhang et al., 1991); myoglobin, 1mbd (Phillips & Schoeborn, 1981); leghemoglobin, 1lh1 (Vainshtein et al., 1977); hemoglobin, 2hhb (Fermi et al., 1984); human CD4 glycoprotein, 1cd4 (Ryu et al., 1990); Bence-Jones immunoglobulin, 1rei (Epp et al., 1974); plastocyanin, 1pcy (Colman et al., 1978); azurin, 1azu (Adman & Jensen, 1981); hemerythrin, 2hmz (Holmes & Stenkamp, 1991); cytochrome *b*562, 256b (Lederer et al., 1981); BPTI, 6pti (Wlodawer et al., 1987); λ repressor, 1lmb (Beamer & Pabo, 1992).

Computing details

This algorithm is written in standard Fortran77. Tests using IL-1β show that on a single R8000 chip of a Silicon Graphics Power Challenge, the complete calculation (including secondary structure and accessibility data) takes less than 30 s of central processor time. This makes it suitable for processing large amounts of structural data, such as that available from the Protein Data Bank.

References

Adman ET, Jensen LH. 1981. Structural features of azurin at 2.7 Å resolution. *Isr J Chem 21*:8-12.

Alexandrov NN, Takahashi K, Go N. 1992. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J Mol Biol 225*:5-9.

Bashford DW, Chothia C, Lesk AM. 1987. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J Mol Biol 196*: 199-216.

Beamer LJ, Pabo CO. 1992. Refined 1.8 Å crystal structure of the lambda repressor-operator complex. *J Mol Biol 227*:177-196.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer based archival file for macromolecular structures. *J Mol Biol 122*:535-542.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three dimensional structure. *Science 253*:164-169.

Chothia C. 1975. Structural invariants in protein folding. *Nature 254*: 304-308.

Chothia C. 1976. The nature of the accessible and buried surface in proteins. *J Mol Biol 105*:1-14.

Chothia C, Lesk AM. 1982. Evolution of proteins formed by beta sheets. I. Plastocyanin and azurin. *J Mol Biol 160*:309-323.

Colman PM, Freeman FC, Guss JM, Murata M, Norris VA, Ramshaw JAM, Venkatappa MP. 1978. X-ray crystal structure analysis of plastocyanin at 2.7 Å resolution. *Nature 272*:319-324.

Epp O, Colman P, Fehlhammer H, Bode W, Schiffer M, Huber R, Palm, W. 1974. Crystal and molecular structure of a dimer composed of the variable portions of the Bence-Jones protein. *Eur J Biochem 45*:513-524.

Fermi G, Perutz MF, Shaanan B, Fourme R. 1984. The crystal structure of deoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol 175*:159-174.

Finzel BC, Clancy LL, Holland DR, Muchmore SW, Watenpaugh KD, Einspahr HM. 1989. Crystal structure solution of IL-1β at 2.0 Å resolution. *J Mol Biol 209*:779-791.

Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol 233*:123-138.

Holmes MA, Stenkamp RE. 1991. Structures of Met and Azidomet hemerythrin at 1.66 Å resolution. *J Mol Biol 220*:723-737.

Hubbard TJP, Blundell TL. 1987. Comparison of solvent inaccessible cores of homologous proteins: Definitions useful for protein modelling. *Protein Eng 1*:159-171.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to fold recognition. *Nature 358*:86-89.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure. *Biopolymers 22*:2577-2637.

Lederer F, Glatigny A, Bethge PH, Bellamy HD, Mathews FS. 1981. Improvement of the 2.5 Å resolution model of cytochrome b562 by redetermining the primary structure and using molecular graphics. *J Mol Biol 148*:427-448.

Lee B, Richards FM. 1971. Interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379-400.

Lesk AM, Chothia C. 1980. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J Mol Biol 136*:225-270.

Lesk AM, Chothia C. 1982. Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. *J Mol Biol 160*:325-342.

Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. *J Mol Biol 196*:641-656.

Mitchell EM, Artymiuk PJ, Rice DW, Willett DW. 1989. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol 212*:151-166.

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. 1992. Stereochemical quality of protein structure coordinates. *Proteins Struct Funct Genet 12*:345-364.

Murzin AG, Lesk AM, Chothia C. 1992. β-Trefoil fold. Patterns of structure and sequence in the Kunitz inhibitors, interleukins 1β and 1α and fibroblast growth factors. *J Mol Biol 223*:531-543.

Onesti S, Brick P, Blow DM. 1991. Crystal structure of a Kunitz-type trypsin inhibitor from *Erythrina caffra* seeds. *J Mol Biol 217*:153-176.

Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485-500.

Phillips SEV, Schoeborn BP. 1981. Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin. *Nature 292*:81-82.

Richards FM, Kundrot CE. 1988. Identification of structural motifs from protein coordinate data: Secondary structure and first level supersecondary structure. *Proteins Struct Funct Genet 3*:71-84.

Ryu SE, Kwong PD, Truneh A, Porter TG, Arthos J, Rosenberg M, Dai X, Xuong N, Axel R, Sweet RW, Hendrickson WA. 1990. Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature 348*: 419-426.

Swindells MB. 1995. A procedure for detecting structural domains in proteins. *Protein Sci 4*:103-112.

Swindells MB, Thornton JM. 1993. A study of structural determinants in the interleukin-1 fold. *Protein Eng 6*:711-715.

Taylor WR, Orengo C. 1989. Protein structure alignment. *J Mol Biol 208*:1-22.

Umezawa Y, Umeyama H. 1988. Computer screening and visualisation of hydrophobic core of protein. *Chem Pharm Bull 36*:4652-4658.

Vainshtein BK, Arutyunyan EG, Kuranova IP, Borisov VV, Sosfenov NI, Pavlovskii AG, Grebenko AI, Konareva NV, Nekrasov YV. 1977. Three dimensional structure of lupine leghemoglobin with a resolution of 2.8 Å. *Dokl Akad Nauk SSSR 223*:238-241.

Wlodawer A, Nachman J, Gilliland GL, Gallagher W, Woodward C. 1987. Structure of form III crystals of bovine pancreatic trypsin inhibitor. *J Mol Biol 198*:469-480.

Zhang J, Cousens LS, Barr PJ, Sprang SR. 1991. Three dimensional structure of human basic fibroblast growth factor, a structural homologue of interleukin-1 beta. *Proc Natl Acad Sci USA 88*:3446-3450.