CAMPASS: a database of structurally aligned protein superfamilies

R Sowdhamini[†], David F Burke, Jing-fei Huang[‡], Kenji Mizuguchi, Hampapathalu A Nagarajaram, N Srinivasan[§], Robert E Steward and Tom L Blundell*

Address: Department of Biochemistry, University of Cambridge,
80 Tennis Court Road, Cambridge CB2 1GA, UK.

Present addresses: [†]National Centre for Biological Sciences, TIFR Centre, PO Box 1234, Indian Institute of Science Campus, Bangalore 560012, India, [‡]Kunming Institute of Zoology, The Chinese Academy of Sciences, Eastern Jiaochang Road, Kunming, Yunnan 650223, PR China and [§]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India.

*Corresponding author.
E-mail: tom@cryst.bioc.cam.ac.uk

Structure 15 September 1998, **6:**1087–1094
http://biomednet.com/elecref/0969212600601087

Introduction

Homologous proteins resemble each other in sequence, three-dimensional structure and usually function; they are related through evolutionary divergence [1–7]. Divergent relationships undoubtedly also occur beyond the ‘nuclear’ family [8,9] but they can be difficult to identify on the basis of sequence alone [10–16] and are easily confused with proteins that have evolved convergently.

Divergent evolution has resulted in families of homologous proteins with similar sequences, three-dimensional structures and usually functions. Evidence is now accumulating that divergent evolution has also led to the existence of superfamilies with very low sequence identities, but similar topologies and often related functions. Sequences of such superfamilies can best be recognised and aligned if the three-dimensional structure of one or more members is known. For example, mammalian relaxin [17–19] and silkworm bombyxin [20] are not easily recognised from sequence analysis as members of the extended insulin/insulin-like growth factor (IGF) superfamily [21]. Their membership of the family was recognised by a careful analysis of the sequences with respect to the insulin fold and has more recently been confirmed experimentally [22,23]. Members of the insulin superfamily are all hormones, growth factors or neurotransmitters, being synthesised in one cell and binding receptors on another. Other superfamilies show similar divergence in sequence but retention of function. For example, members of the well studied pepsin-like/retroviral aspartic proteinase superfamily have similar catalytic sites, but can exist either as dimers or as single chains, with sequence identities around 15% and differing in length by a factor of two in equivalent domains/subunits [24,25]. A further example is provided by the cystine-knot superfamily, which includes nerve growth factor, transforming growth factor-β2 and platelet-derived growth factor. All of these proteins bind to cell-surface receptors but have no significant sequence identity [26]. Similarly, the lectin superfamily includes legume lectins and mammalian pentraxins that adopt an elaborated jelly-roll fold implicated in sugar binding [27,28], but have sequence identities of less than 10%. The observation of such superfamilies is becoming increasingly common [10,29,30] with the determination and availability of many thousands of protein structures in the Brookhaven Protein Data Bank (PDB) [31].

Superfamilies may have evolved by divergent evolution, although this is difficult to establish unequivocally. Analyses of genomes have shown that 40–60% of new sequences belong to known homologous families, however [32,33]; in such instances, the presence of functional sites can usually be predicted on the basis of sequence alignment [34]. Many of the remaining sequences are likely to be members of superfamilies that include previously identified members of known function and even proteins of known three-dimensional structure. If membership of a superfamily can be established, then this may give clues as to the function of the protein encoded by a new sequence (e.g. for a review see [35]).

It is possible that two proteins share a similar three-dimensional structure but do not perform similar functions. For example, the C-terminal domain of hepatocyte growth factor and its homologues have high sequence similarity with the classical serine proteinases, but two of the three residues in the ‘catalytic triad’ are substituted and lack the characteristic activity. Similarly, haptoglobin is also a member of the serine proteinase family but does not cleave peptide bonds. Such examples occur relatively rarely in closely related homologues and can be identified when key catalytic or binding residues are absent, although those residues stabilising the structure are conserved.

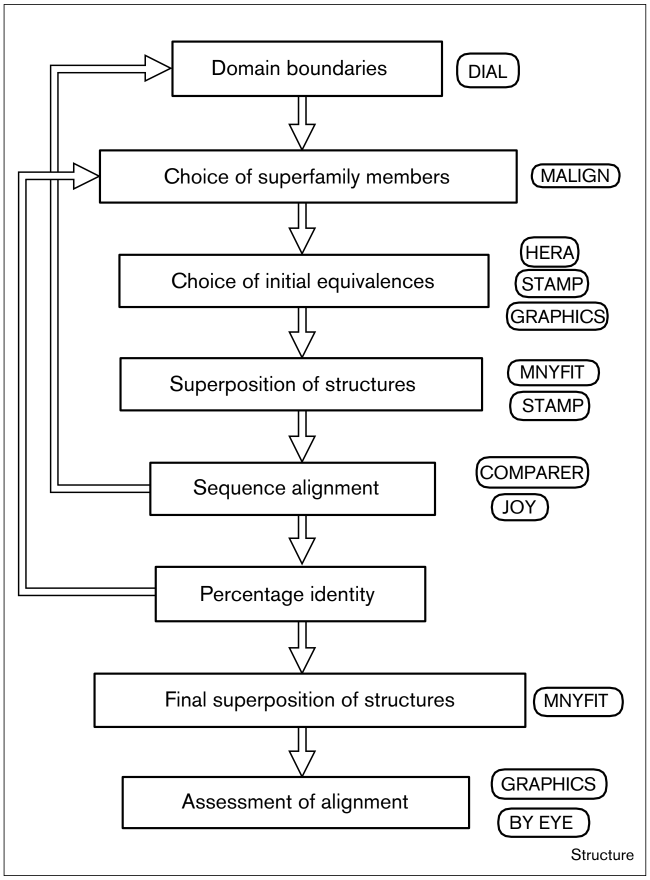
Derived databases are now available that classify protein structures deposited in the PDB into homologous families, superfamilies and folds [14–16,36,37]. Together with databases of sequence motifs [38], these are useful tools for fold prediction and for suggesting functions for new sequences. The recognition of distant analogies can often be facilitated if sequence alignments for the relevant superfamily are available. Such analyses have been addressed previously but have usually been restricted to particular systems of immediate interest to the authors.

We have aligned sequences of protein domains belonging to superfamilies on the basis of the conservation of local three-dimensional structural features, relationships and functional sites. We have considered 69 superfamilies, consisting of 288 protein domains representing 713 homologous proteins. We report the compilation of a database of superfamily alignments (CAMPASS, CAMbridge database of Protein Alignments organised as Structural Superfamilies) available on the World Wide Web (http://www-cryst.bioc.cam.ac.uk/~campass).

Structure-based sequence alignment and compilation of the database

For most proteins, the program DIAL [39] was used to define domain boundaries on the basis of clustering of secondary structure distances. The definition of domains was often refined on the basis of structural comparisons of the superfamily; large insertions corresponding to compact clusters of secondary structures that occur in only one superfamily member were omitted. Differences in the definition of domain boundaries within a superfamily also resulted from rigid-body movements.

Figure 1



Flow-chart indicating the various steps involved in the structure-based sequence alignment of proteins belonging to superfamilies. The tools or programs used to perform a particular analysis are shown on the right in ellipsoid boxes.

Superfamilies were defined as families of proteins where not only the three-dimensional structures were similar, but where there was also similarity in function. Superfamilies were chosen using the results of an earlier analysis on the clustering of structural domains based on structural similarity [16] and also by referring to the SCOP database [36] for functional and gross structural similarity. We restricted our analysis to a dataset where no two proteins shared more than 25% sequence identity (i.e. the ratio between the number of identical residues and the number of aligned, non-gap positions as a product of 100) [16]. Out of several homologous proteins, the protein with the highest resolution structure was usually chosen as the representative member of a superfamily. In cases where there was more than one protein of equally high resolution proteins were chosen where with a ligand or a cofactor, was available. In earlier studies [16], domains smaller than seven secondary structure units were not considered, as this often led to spurious matches of substructures in automatic procedures. In the present alignment database, however, we have intervened in the automatic classification to include superfamilies comprising a smaller number of secondary structures, such as the cytochromes. A simple sequence-based alignment (MALIGN [40]) was used to identify and to eliminate clear homologues with a sequence identity of > 25%. Members within superfamilies have high structural similarity (usually SEA score values less than 0.55 [16]) but this varies in fold space making the automatic choice of superfamilies still a difficult task. For well-defined superfamilies with a consistent assignment of domain boundaries, the final alignment was obtained automatically (see below and Figure 1 for subsequent steps). Other features such as the provision of links to the homologous alignment database [6,7] and inclusion of 'single-member superfamilies' are now being improved.

The three-dimensional coordinates were superposed using the programs MNYFIT [41] or STAMP [42] to obtain fitted coordinates for all possible pairs within a superfamily. The initial equivalences cannot usually be identified from sequence-based alignments. Initial equivalences required by MNYFIT were selected on the basis of common structural or functional features. Common structural features included residues in a buried strand or helix identified from JOY [4,5] or STAMP [42] or residues in secondary structures that display similar patterns of hydrogen bonding identified by HERA [43]. Common functional features used to define initial equivalences included residues that were involved in catalysis, cofactor binding, etcetera.

The superposed coordinates were used to seed the alignment using COMPARER [44,45], which also exploits accessory files containing information on hydrogen bonding (HBOND, JP Overington and TLB, unpublished results), backbone secondary structural assignment (DSSP [46]; SSTRUC, DK Smith, unpublished results) and solvent accessibility (PSA, A Sali and TLB, unpublished results, based on the algorithm by Lee and Richards [47]) of individual proteins. Gaps were assigned to retain maximum conservation of secondary structure and structural environments such that the solvent-buried nature and hydrogen bonding patterns were conserved at an alignment position rather than the amino acid itself [45]. The optimal alignment was performed using dynamic programming and simulated annealing.

As the structure-based alignment is usually different from the preliminary alignment obtained by simple amino acid matches, the pairwise percentage identity values were recalculated for the final multiple alignment derived from COMPARER. If there were values higher than 25%, proteins giving rise to such high sequence identity were eliminated such that there was a minimum loss in the number of proteins. For example, if the case arose where one protein shared less than 25% identity with two proteins

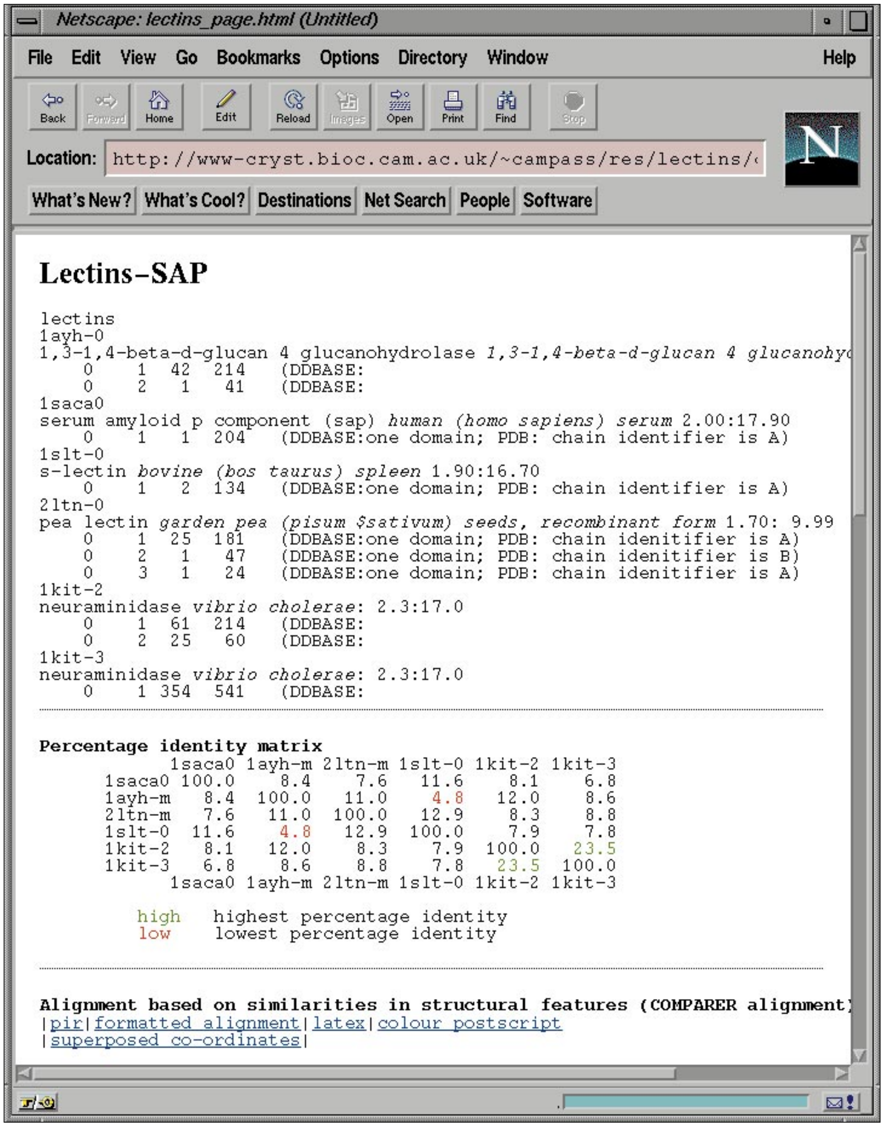
Table 1

Superfamilies in the CAMPASS database.

Superfamily code*	Superfamily name†	Superfamily code*	Superfamily name†
4helud (3)	Cytochromes	lectins (6)	ConA-like lectins/glucanases
FAD-binding-like (13)	FAD/NAD(P)-binding domain	lipocalin (5)	Lipocalins
FMN_typeI (2)	FMN-linked, oxidoreductases	methyltransferases (5)	*S*-adenosyl-L-methionine-dependent methyltransferases
PH (3)	PH-domain-like		
SH3 (2)	SH3 domain	muconate_lactonising (3)	Enolase and muconate-lactonising enzyme, C-domain
ab5_toxins (5)	Bacterial enterotoxins		
ab_hydrolases (8)	α/β-Hydrolases	muconate_ndomain (3)	N-terminal domain of enolase and muconate-lactonizing enzyme
actinIA (3)	Actin-like ATPase domain		
actinIIA (3)	Actin-like ATPase domain	nip (3)	P-loop containing nucleotide hydrolases
actin_binding (2)	Actin depolymerizing proteins	p450 (4)	Cytochrome P450
adk (2)	Nucleotide and nucleoside kinases	p450 (4)	Cytochrome P450
adp (4)	ADP ribosylation	pbgd1 (4)	Periplasmic binding II (domain 1)
anticodon_binding (2)	Bacteriophage ssDNA-binding (family)	pbgd2 (4)	Periplasmic binding II (domain 2)
asp_hiv (3)	Acid proteases	periplasmic_binding_I1 (4)	C-domain of periplasmic binding type I
bacteriophage (2)	Bacteriophage ssDNA-binding (family)	periplasmic_binding_I2 (6)	N-domain of periplasmic binding type I
β-γ-crystallin_like (3)	Crystallins/protein S	phospholipase (2)	Phospholipase A2
bgt-gpb (2)	β-Glucosyltransferase and glycosyltransferase	plp1 (4)	PLP-dependent transferases
		plq (2)	PLP-dependent transferases
cbp (7)	EF-hand	porins (3)	Porins
ccperoxy (3)	Heme-dependent peroxidases	ppase1 (3)	Sugar phosphatases
creatinase (2)	Creatinase/methionine aminopeptidase	ppase2 (3)	Sugar phosphatases
ctt (2)	Cytidine deaminase	propeller (3)	7/8-bladed propeller (fold)
cys (2)	Papain-like	ras (4)	G proteins (family)
cystineknot (6)	Cystine-knot cytokines	repressor_like (4)	λ repressor-like DNA-binding
cytc (3)	Monodomain cytochrome *c* (family)	ribonucleaseh_like (5)	Ribonuclease H-like
cytokine (2)	Cytokine	rubredoxins (3)	Rubredoxin-like (fold)
exopeptidase (3)	Zn-dependent exopeptidases	serineproteases1 (5)	Trypsin-like serine proteases
ferredoxin_reductases (3)	Ferredoxin reductase-like, C-domain	serineproteases2 (4)	Trypsin-like serine proteases
flav (7)	Flavodoxin-like (fold)	sial_neur (3)	Sialidases (neuraminidases)
globins (7)	Globin-like	sslipid (2)	Bifunctional inhibitor/lipid-transfer protein
glucoamylase_like (3)	Glycosyltransferases of the superhelical fold		
		strep (2)	Avidin/streptavidin
glucosyltransferases (18)	Glycosyltransferases	superantigen_toxins (2)	Superantigen toxins, N-domain (family)
gshase_2 (4)	Glutathione synthetase ATP-binding	thiamin_binding (6)	Thiamin-binding
gshase_3 (5)	Glutathione synthetase ATP-binding	thioredoxin (6)	Thioredoxin-like
ig (12)	Immunoglobulins	trp-biosynthesis (3)	Tryptophan biosynthesis enzymes
il8_like (2)	Interleukin-8-like chemokines	tyrosine_phosphatases (3)	Phosphotyrosine (protein) phosphatases
kinases (3)	Protein kinases (PKs), ca. core	viral_coats (13)	Viral coat and capsid proteins

*The number of members in the superfamily is given in parentheses. †Superfamily name as defined in SCOP [36]. In a few cases where there is considerable functional similarity, a broader class of proteins were considered under one superfamily (marked as fold). In a few other cases, the choice of superfamily members was restricted to a group of proteins, defined as a family in SCOP (marked as family), to permit reliable structural superposition and structure-based sequence alignment.

Figure 2


Netscape: lectins_page.html (Untitled)
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Edit Reload Images Open Print Find Stop
Location: http://www-cryst.bioc.cam.ac.uk/~campass/res/lectins/
What's New? What's Cool? Destinations Net Search People Software

Lectins–SAP

```
lectins
1ayh-0
1,3-1,4-beta-d-glucan 4 glucanohydrolase 1,3-1,4-beta-d-glucan 4 glucanohy
    0    1  42  214   (DDBASE:
    0    2   1   41   (DDBASE:
1saca0
serum amyloid p component (sap) human (homo sapiens) serum 2.00:17.90
    0    1   1  204   (DDBASE:one domain; PDB: chain identifier is A)
1slt-0
s-lectin bovine (bos taurus) spleen 1.90:16.70
    0    1   2  134   (DDBASE:one domain; PDB: chain identifier is A)
2ltn-0
pea lectin garden pea (pisum $sativum) seeds, recombinant form 1.70: 9.99
    0    1  25  181   (DDBASE:one domain; PDB: chain idenitifier is A)
    0    2   1   47   (DDBASE:one domain; PDB: chain idenitifier is B)
    0    3   1   24   (DDBASE:one domain; PDB: chain idenitifier is A)
1kit-2
neuraminidase vibrio cholerae: 2.3:17.0
    0    1  61  214   (DDBASE:
    0    2  25   60   (DDBASE:
1kit-3
neuraminidase vibrio cholerae: 2.3:17.0
    0    1 354  541   (DDBASE:
```

Percentage identity matrix

```
            1saca0 1ayh-m 2ltn-m 1slt-0 1kit-2 1kit-3
     1saca0 100.0    8.4    7.6   11.6    8.1    6.8
     1ayh-m   8.4  100.0   11.0    4.8   12.0    8.6
     2ltn-m   7.6   11.0  100.0   12.9    8.3    8.8
     1slt-0  11.6    4.8   12.9  100.0    7.9    7.8
     1kit-2   8.1   12.0    8.3    7.9  100.0   23.5
     1kit-3   6.8    8.6    8.8    7.8   23.5  100.0
            1saca0 1ayh-m 2ltn-m 1slt-0 1kit-2 1kit-3

        high   highest percentage identity
        low    lowest percentage identity
```

Alignment based on similarities in structural features (COMPARER alignment)
|pir|formatted alignment|latex|colour postscript
|superposed co-ordinates|

World Wide Web page for the lectin superfamily. The page gives the name in the title and provides PDB information for the domains (see text for details). The percentage identity matrix corresponds to the final structure-based alignment: the lowest identity is in red and the highest in green.

that had a higher than 25% similarity with each other, one of the higher-similarity proteins was eliminated.

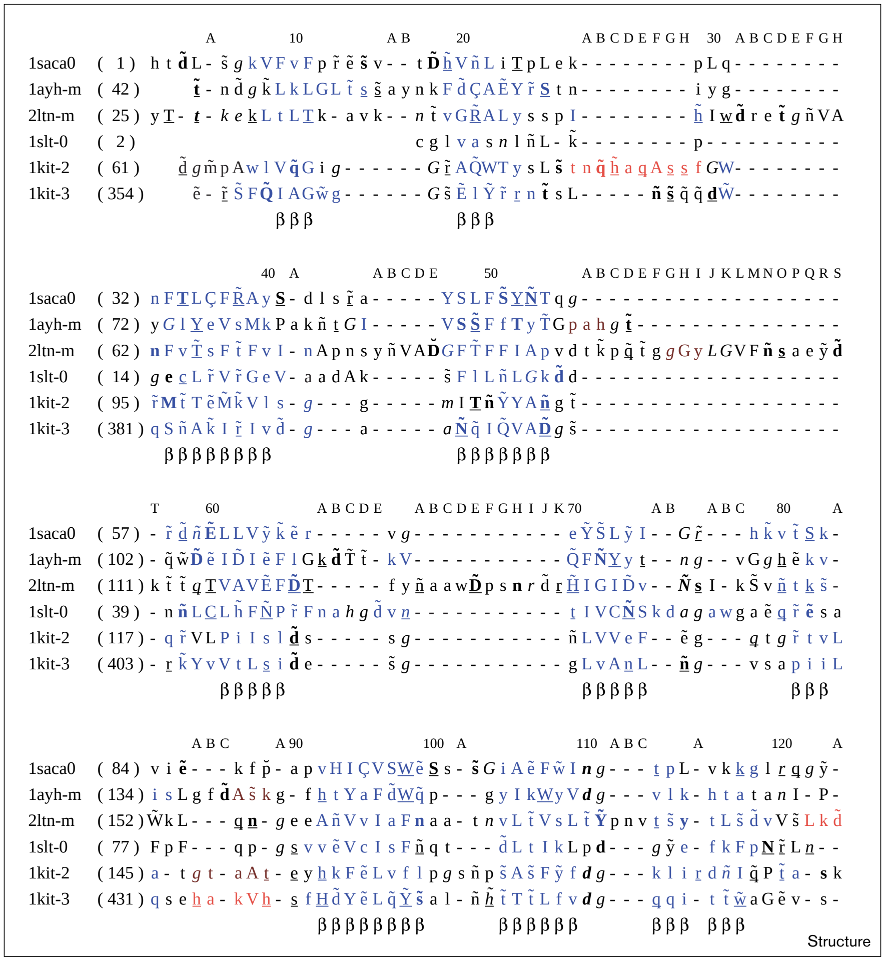
Segments corresponding to non-gap positions in the final sequence alignment of members in a superfamily were used as initial equivalences to superpose structures using MNYFIT [41] without the update of the equivalences supplied. This set of multiply superposed structures can be viewed on the World Wide Web using the RASMOL graphics interface [40]. Large structural variations are observed in the loop regions and even in the structural core; insertions of a few secondary structure elements are also seen.

Description of the database

Table 1 lists the superfamilies for which structure-based sequence alignments are available in CAMPASS. A complete list along with members in the superfamilies can be accessed from the World Wide Web site. Each individual superfamily member can represent a set of homologous distinct proteins and several protein entries of the PDB. Indeed, the 69 superfamilies described involve 288 representative domains, 142 families, 713 distinct homologous proteins (from different species) and 2466 entries in the PDB. Although SCOP [36] suggests the existence of around 453 superfamilies, most (357) [48] include either a single representative or a single homologous family.

We have accommodated significant insertions/deletions in many alignments. The porin superfamily includes representatives from porins and maltoporins, both of which are multistranded, β-strand-rich membrane proteins forming a closed barrel. The extra strands in maltoporins (18, 22; n, S

Figure 3



```
                       A          10           A B      20              A B C D E F G H   30   A B C D E F G H
1saca0   (   1 )  h t d̃ L - s̃ g k V F v F p r̃ ẽ s̃ v - - t D̃ h̃ V ñ L i T p L e k - - - - - - - - p L q - - - - - - - -
1ayh-m   (  42 )        t̃ - n d̃ g k̃ L k L G L t̃ s s̃ a y n k F d̃ Ç A Ẽ Y r̃ S t n - - - - - - - - i y g - - - - - - - -
2ltn-m   (  25 )  y T - t - k e k L t L T k - a v k - - n t̃ v G R̃ A L y s s p I - - - - - - - - h̃ I w d̃ r e t̃ g ñ V A
1slt-0   (   2 )                                    c g l v a s n l ñ L - k̃ - - - - - - - - p - - - - - - - - - -
1kit-2   (  61 )    d̃ g m̃ p A w l V q̃ G i g - - - - - - G r̃ A Q̃ W T y s L s̃ t n q̃ h̃ a q A s s f G W - - - - - - - -
1kit-3   ( 354 )      ẽ - r̃ S̃ F Q̃ I A G w̃ g - - - - - - G s̃ Ẽ l Ỹ r̃ r n t̃ s L - - - - - ñ s̃ q̃ q̃ d W̃ - - - - - - - -
                            β β β                     β β β

                             40  A           A B C D E      50              A B C D E F G H I J K L M N O P Q R S
1saca0   (  32 )  n F T L Ç F R̃ A y S - d l s r̃ a - - - - - Y S L F S̃ Y Ñ T q g - - - - - - - - - - - - - - - - - - -
1ayh-m   (  72 )  y G l Y e V s M k P a k ñ t G I - - - - - V S S̃ F f T y T̃ G p a h g t̃ - - - - - - - - - - - - - - - -
2ltn-m   (  62 )  n F v T̃ s F t̃ F v I - n A p n s y ñ V A Ď G F T̃ F F I A p v d t k̃ p q̃ t̃ g g G y L G V F ñ s a e ỹ d̃
1slt-0   (  14 )  g e c L r̃ V r̃ G e V - a a d A k - - - - - s̃ F l L ñ L G k d̃ d - - - - - - - - - - - - - - - - - - -
1kit-2   (  95 )  r̃ M t̃ T ẽ M̃ k̃ V l s - g - - - g - - - - - m I T ñ Ỹ Y A ñ g t̃ - - - - - - - - - - - - - - - - - - -
1kit-3   ( 381 )  q S ñ A k̃ I r̃ I v d̃ - g - - - a - - - - - a Ñ q̃ I Q̃ V A D̃ g s̃ - - - - - - - - - - - - - - - - - - -
                  β β β β β β β β                      β β β β β β β

                  T      60             A B C D E      A B C D E F G H I J K 70         A B    A B C    80     A
1saca0   (  57 )  - r̃ d̃ ñ Ẽ L L V ỹ k̃ ẽ r - - - - - v g - - - - - - - - - - - e Ỹ S̃ L ỹ I - - G r̃ - - - h k̃ v t̃ S k -
1ayh-m   ( 102 )  - q̃ w̃ D̃ ẽ I D̃ I ẽ F l G k d̃ T̃ t̃ - k V - - - - - - - - - - - Q̃ F Ñ Y y t - - n g - - v G g h ẽ k v -
2ltn-m   ( 111 )  k t̃ t̃ q T V A V Ẽ F D̃ T - - - - - f y ñ a a w D̃ p s n r d̃ r H̃ I G I D̃ v - - Ñ s I - k S̃ v ñ t k s̃ -
1slt-0   (  39 )  - n ñ L C L h̃ F Ñ P r̃ F n a h g d̃ v n - - - - - - - - - - - t I V C Ñ S k d a g a w g a ẽ q r̃ ẽ s a
1kit-2   ( 117 )  - q r̃ V L P i I s l d̃ s - - - - - s g - - - - - - - - - - - ñ L V V e F - - ẽ g - - - q t g r̃ t v L
1kit-3   ( 403 )  - r k̃ Y v V t L s i d̃ e - - - - - s̃ g - - - - - - - - - - - g L v A n L - - ñ g - - - v s a p i i L
                        β β β β β                                   β β β β β                β β β

                       A B C      A 90               100  A              110  A B C     A       120    A
1saca0   (  84 )  v i ẽ - - - k f p̆ - a p v H I Ç V S W ẽ S s - s̃ G i A ẽ F w̃ I n g - - - t p L - v k k g l r q g ỹ -
1ayh-m   ( 134 )  i s L g f d̃ A s̃ k g - f h t Y a F d̃ W q̃ p - - - g y I k W y V d g - - - v l k - h t a t a n I - P -
2ltn-m   ( 152 )  W̃ k L - - - q n - g e e A ñ V v I a F n a a - t n v L t̃ V s L t̃ Ỹ p n v t s̃ y - t L s̃ d̃ v V s̃ L k d̃
1slt-0   (  77 )  F p F - - - q p - g s v v ẽ V c I s F ñ q t - - - d̃ L t I k L p d - - - g ỹ e - f k F p N r̃ L n - -
1kit-2   ( 145 )  a - t g t - a A t - e y h k F ẽ L v f l p g s ñ p s̃ A s̃ F ỹ f d g - - - k l i r d ñ I q̃ P t̃ a - s k
1kit-3   ( 431 )  q s e h a - k V h - s f H d̃ Y ẽ L q̃ Ỹ s̃ a l - ñ h̃ t̃ T t̃ L f v d g - - - q q i - t t̃ w̃ a G ẽ v - s -
                                     β β β β β β β β        β β β β β β        β β β   β β β
```

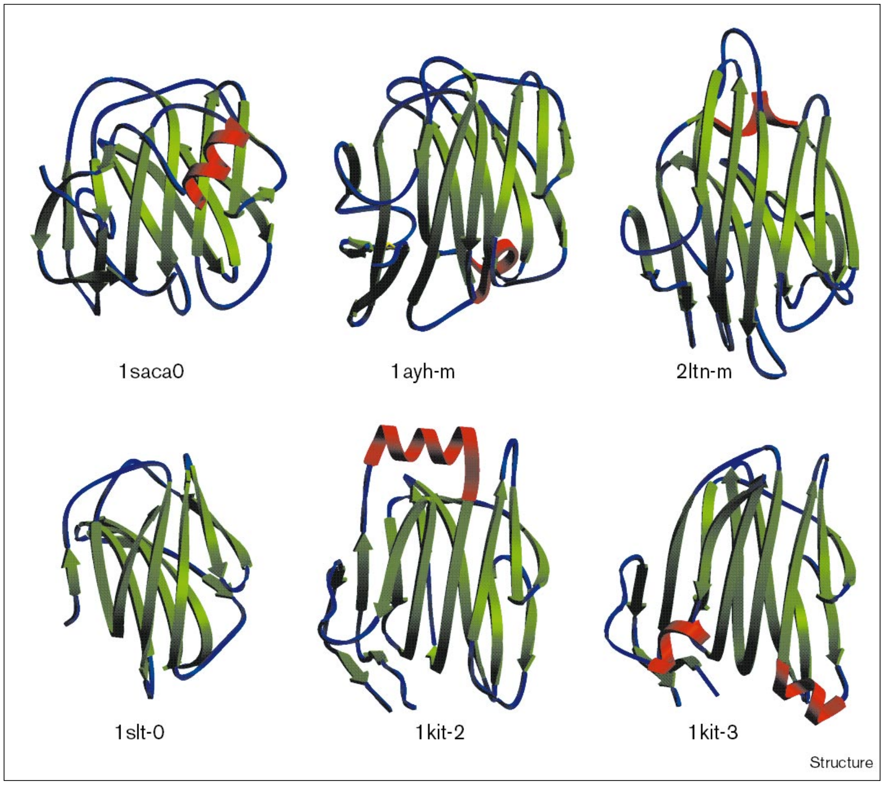
Structure-based sequence alignment of the lectin superfamily compiled using the program COMPARER [44] and structure-annotated using JOY [4,5]. Solvent-accessible and solvent-inaccessible residues are shown in lower case and upper case, respectively. Residues with a postive phi are indicated in italics; residues with a *cis* peptide in the backbone or disulphide bonds are indicated by the presence of a breve (e.g. š) or cedilla (e.g. ç), respectively. Hydrogen bonds formed to the sidechains, mainchain amides and mainchain carbonyls of other residues are indicated by the presence of a tilde on top, boldface or underline, respectively. Residues in α helices, β strands or 3_{10} helices are shown in red, blue or brown, respectively. (Above the alignment, residue numbers given are for 1saca0 – where there are insertions with respect to 1saca0, the upper-case letters indicate insertion codes.)

system, where n denotes the number of strands and S denotes the shear number, a measure of the stagger of the strands in the β-sheet) compared to the porins (16, 20 system; see the SCOP [36] entry of the porin superfamily) were easily accommodated as a C-terminal overhang in the structure-based alignment (figure not shown; see the CAMPASS Web page for a superposition). In the superfamily of salt-sensitive (magnesium/lithium) phosphatases, despite long insertions, the salt-coordination sites that were not used as initial equivalences have also been aligned well (e.g. Asp54, Glu79 and Glu80 of inositol polyphosphate 1-phosphatase (PDB code 1inp); see block 2 in the alignment of ppase in CAMPASS).

Because of large structural variations, structure-based sequence alignment has often proved difficult in proteins that have previously been considered as superfamilies. For example, α/β-hydrolases consist of proteins with diverse specificity and exhibit large variations and insertions in their loop regions. Superposition and structure-based alignment were possible, however, due to higher levels of similarity in the core regions. The final alignments of all the superfamilies have been examined for the correct positioning of secondary structure elements, functionally important residues and the conservation of structural environments. In case of any misalignments, the process was sometimes repeated after a better choice of initial equivalences or members (Figure 1). In other superfamilies we have performed simulated annealing (see [44]) in order to obtain reasonable alignments.

Figure 2 shows the CAMPASS Web page of a typical superfamily. It contains the name in the title and provides PDB information for the domains, such as protein name, source, resolution and R factor. The nomenclature follows DDBASE [16], where each domain is referred to by a six-character code: the first four characters refer to the PDB code, the fifth character to the chain identifier and the last character to the domain number as defined in the domain database. A comparison of domain boundaries with the original domain database definition has been provided (see above for reasons for differences). The percentage

Figure 4



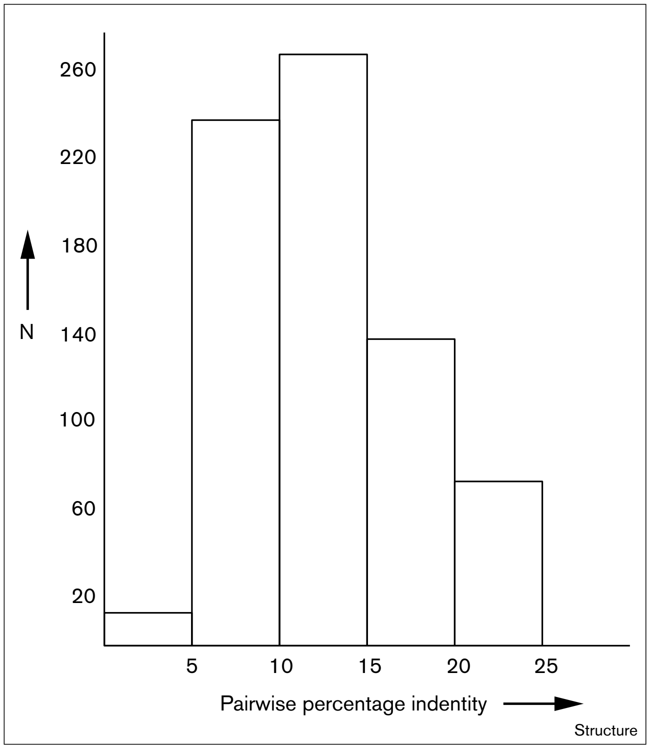
Members of the lectin superfamily. The structures of six members of the lectin superfamily were aligned by best-fit superposition using MNYFIT [41] and are represented using SETOR [50]. α Helices are shown in red and β strands in green. The name of each protein was assigned according to DDBASE [16] and corresponds to those in Figures 2 and 3.

identity matrix corresponds to the final structure-based superfamily alignment; the lowest percentage identity is indicated in red and the highest (always ≤ 25%) in green. On the Web page for each superfamily in CAMPASS, the alignment is also shown in an annotated format produced by JOY [4,5], such that structural features (e.g. solvent burial or solvent accessibility and hydrogen bonding) are represented by modification of the characters in the single-letter code for residue types. Secondary structure elements are coloured using the recent version of JOY [5]. The annotated alignment allows users to consider the conservation of secondary structure and particular structural features within a superfamily, even where there is poor sequence identity. The superfamily alignment itself is available for extraction as PostScript or LATEX files of JOY annotated colour-enhanced form, as plain files or as plain formatted alignments. The structures of superfamily members, superposed by considering equivalent residues corresponding to the final alignment, can be viewed using RASMOL [49] (see the CAMPASS Web page for more details).

In order to illustrate the information on the database, the structure-based sequence alignment of a superfamily, the lectins, is shown in Figure 3. All of these proteins bind sugars at the concave face of the tertiary structure and adopt a jelly-roll fold. Although amino acid residues are not conserved amongst members, several structural features are conserved. For example, the solvent accessibility pattern in strands 3–6 (starting from residue 62 of pea lectin (PDB accession code 2ltn; see Figure 4) are very similar amongst the proteins. The solvent-buried residues in these four strands are major contributors to the hydrophobic core. Figure 4 shows the structures in the lectin superfamily after best superposition. The percentage sequence identity between the members in this superfamily is quite low (see Figure 4); for example, human serum amyloid P component (SAP; 1saca) and pea lectin (2ltn-m) share sequence identity of 7.6% and the root mean square deviation (rmsd) of the superposed structures is 3.3 Å. In contrast, the two jelly-roll fold domains in sialidase (1kit-2 and 1kit-3) share 23.5% sequence identity and the final rmsd of the superposed structures is 1.3 Å.

The distribution of pairwise percentage identities within each of the superfamilies is shown in Figure 5. Of the 747 protein pairs, 514 have a percentage identity between 5 and 15%. The percentage sequence identity between some of the superfamily members is very low, however: the lowest sequence identity (2.5%) in the database is between one of the domains of bean pod mottle virus (1bmv22) and canine parvovirus (2cas1m) of the viral coat protein superfamily. Incidentally, the observed range of sequence identity

Figure 5



Distribution of pairwise percentage identities of members within superfamilies. N represents the number of pairs. This analysis includes 69 superfamilies and values correspond to the final structure-based alignment.

between a large number of computer-generated random sequences, with a bias for amino acid composition as in the globular proteins, is between 2 and 9%; the average sequence identity for such a set is 5.9% and the standard deviation is 2.4% (NS, RS and TLB, unpublished results).

Conclusions

The CAMPASS sequence alignments provide a means of understanding the structural and functional similarities in protein superfamilies and interpreting additional information when structures of new members of a superfamily are determined. CAMPASS can also be used to construct amino acid substitution tables [4] and templates [40] of protein superfamilies, which can assist in the assignment of a previously known fold to a new sequence in cases of poor overall sequence similarity.

Other databases, such as SCOP [36], depict structural hierarchies amongst protein folds and consider the evolution of structure and function amongst proteins in order to classify them. SCOP does not involve automatic methods or sequence alignments. CATH [14,15] and FSSP [12,37] do employ automatic methods and scoring schemes for structural classification, but have less emphasis on structure-based sequence alignments and structural annotations. CAMPASS does not consider fold families as in other databases.

Comparative modelling methods have proved useful in extrapolating the available information for known proteins to the three-dimensional structures and functions of new sequences. Where the new sequence has no known homologue, it may still belong to an established superfamily. Tools such as CAMPASS, which can assist in the recognition of such similarities, are useful for predicting the fold and function of new proteins identified in genome sequencing studies. Structure-based alignments of superfamilies confirm that identities and/or conservative variation in sequence are usually associated with structural determinants (key packing relationships or hydrogen bonds) or functional requirements, common to the superfamily. Structure-based alignments, therefore, provide a firm basis for understanding and predicting amino acid substitutions in superfamilies and for developing methods of fold recognition. These should be of value in proteomics — understanding the functions of proteins identified in genome sequences.

Acknowledgements

We thank Andrej Sali, John Overington and Mark Johnson for use of the programs COMPARER, JOY and MALIGN, respectively. We thank the anonymous referees for their useful comments. For financial support, RS thanks the Imperial Cancer Research Fund; DFB and HAN Oxford Molecular Ltd; KM HFSP; NS The Wellcome Trust; and RES the MRC.

References

1. Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**, 167-339.
2. Rossmann, M.G. & Argos, P. (1977). The taxonomy of protein structure. *J. Mol. Biol.* **109**, 99-129.
3. Chothia, C. (1984). Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **53**, 537-572.
4. Overington, J.P., Johnson, M.S., Sali, A. & Blundell, T.L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. R. Soc. (Lond.), B* **241**, 132-145.
5. Mizuguchi, K., Deane, C.M., Johnson, M.S., Blundell, T.L. & Overington, J.P. (1998) JOY: protein sequence structure representation and analysis. *Bioinformatics*, in press.
6. Overington, J.P., *et al.*, & Blundell, T.L. (1993). Molecular recognition in protein families: a database of aligned three-dimensional structures of related proteins. *Biochem. Soc. Trans.* **21**, 597-604.
7. Mizuguchi, K., Deane, C.M., Blundell, T.L. & Overington, J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, in press.
8. Rossmann, M.G., Moras, D. & Olsen, K.W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature* **250**, 194-199.
9. Matthews, B.W. & Rossmann, M.G. (1985). Comparison of protein structures. *Methods Enzymol.* **115**, 397-420.
10. Murzin, A.G. & Chothia, C. (1992). Protein architecture: new superfamilies. *Curr. Opin. Struct. Biol.* **2**, 895-903.
11. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein-structure families with common folding motifs. *Protein Sci.* **1**, 1691-1698.
12. Holm, L. & Sander, C. (1998). Touring fold space with Dali/FSSP. *Nucleic Acids Res.* **26**, 316-319.
13. Alexandrov, N.N. & Go, N. (1994). Biological meaning, statistical significance, and classification of local spatial similarities in non-homologous proteins. *Protein Sci.* **3**, 866-875.
14. Orengo, C.A., Jones, D.T. & Thornton, J.M. (1994). Protein superfamilies and domain superfolds. *Nature* **372**, 631-634.

15. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997). CATH – a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108.
16. Sowdhamini, R., Rufino, S.D. & Blundell, T.L. (1996). A database of globular protein structural domains: clustering of representative family members into similar folds. *Fold. Des.* **1**, 209-220.
17. Schwabe, C., Mcdonald, J.K. & Steinetz, B.G. (1976). Primary structure of the A-chain of porcine relaxin. *Biophys. Biochem. Res. Commun.* **70**, 397-405.
18. Schwabe, C., Mcdonald, J.K. & Steinetz, B.G. (1977). Primary structure of the B-chain of porcine relaxin. *Biophys. Biochem. Res. Commun.* **75**, 503-510.
19. James. R.H., Niall, S., Kwok, G. & Bryant-Greenwood, G. (1977). Primary structure of porcine relaxin: homology with insulin and related growth factors. *Nature* **267**, 544-546.
20. Jhoti, H., McLeod, A.N., Blundell, T.L., Ishizaki, H., Nagasawa, H. & Suzuki, A. (1987). Prothoracicotropic hormone has an insulin-like tertiary structure. *FEBS Lett.* **219**, 419-425.
21. Blundell, T.L. & Humbel, R.E. (1980). Hormone families: pancreatic hormones and homologous growth factors. *Nature* **287**, 781-787.
22. Eigenbrot, C., *et al.*, & Kossiakoff, A.A. (1991). X-ray structure of human relaxin at 1.5 Å – comparison to insulin and implications for receptor-binding determinants. *J. Mol. Biol.* **221**, 15-21.
23. Nagata, K., *et al.*, & Inagaki, F. (1995). Identification of the receptor-recognition surface of bombyxin-II, an insulin-like peptide of the silkworm *Bombyx mori* – critical importance of the B-chain central part. *J. Mol. Biol.* **253**, 749-758.
24. Lapatto, P., *et al.*, & Hobart, P.M. (1989). X-ray analysis of HIV-1 proteinase at 2.7 Å resolution confirms structural homology among retroviral enzymes. *Nature* **342**, 299-302.
25. Miller, M., Jaskolski, M., Rao, J.K.M., Leis, J. & Wlodawer, A. (1989). Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature* **337**, 576-579.
26. Murray-Rust, J., *et al.*, & Bradshaw, R.A. (1993). Topological similarities in TGF-β2, PDGF-BB and NGF define a superfamily of polypeptide growth-factors. *Structure* **1**, 153-159.
27. Emsley, J., *et al.*, & Wood, S.P. (1994). Structure of pentameric human serum amyloid-P component. *Nature* **367**, 338-345.
28. Srinivasan, N., Rufino, S.D., Pepys, M.B., Wood, S.P. & Blundell, T.L. (1994). A superfamily of proteins with the lectin fold. *Chemtracts* **6** 149-164.
29. Holm, L. & Sander, C. (1995). Evolutionary link between glycogen-phosphorylase and a DNA modifying enzyme. *EMBO J.* **14**, 1287-1293.
30. Murzin, A.G. (1996). Structural classification of proteins – new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386-394.
31. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
32. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992). What's in a genome? *Nature* **358**, 287.
33. Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393-403.
34. Zvelebil, M.J., Barton, G.J., Taylor, W.R. & Sternberg, M.J.E. (1987). Prediction of protein secondary structure and active-sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957-961.
35. May, A.C.W., *et al.*, & Blundell, T.L. (1994). The recognition of protein-structure and function from sequence – adding value to genome data. *Phil. Trans. Roy. Soc. Lond. B* **344**, 373-381.
36. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP – a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
37. Holm, L. & Sander, C. (1994). Structural similarity of plant cattiness and lysozymes from animals and phage – an evolutionary connection. *FEBS Lett.* **340**, 129-132.
38. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
39. Sowdhamini, R. & Blundell, T.L. (1995). An automatic method involving cluster-analysis of secondary structures for the identification of domains in proteins. *Protein Sci.* **4**, 506-520.
40. Johnson, M.S., Overington, J.P. & Blundell, T.L. (1993). A structural basis for sequence comparisons: an evaluation of scoring methodologies. *J. Mol. Biol.* **233**, 716-738.
41. Sutcliffe, M.J., Haneef, I., Carney, D. & Blundell T.L. (1987). Knowledge-based modelling of homologous proteins. Part I: Three-dimensional frameworks derived from simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377-384.
42. Russell, R.B. & Barton, G.J. (1992). Multiple protein sequence alignment from tertiary structure comparison – assignment of global and residue confidence levels. *Proteins* **14**, 309-323.
43. Hutchinson, E.G. & Thornton, J.M. (1990). HERA – a program to draw schematic diagrams of protein secondary structures. *Proteins* **8**, 203-212.
44. Sali, A. & Blundell, T.L. (1990). The definition of topological equivalence in homologous and analogous structures: a procedure involving a comparison of local properties and relationships. *J. Mol. Biol.* **212**, 403-428.
45. Zhu, Z.-Y., Sali, A. & Blundell, T.L. (1992). A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng.* **5**, 43-51.
46. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure – pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.
47. Lee, B. & Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.
48. Brenner, S.E., Chothia, C. & Hubbard, T.J.P. (1997). Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7**, 369-376.
49. Sayle, R.A. & Milner-White, E.J. (1995). RASMOL – biomolecular graphics for all. *Trends. Biochem. Sci.* **20**, 374-376.
50. Evans, S.V. (1993). SETOR – hardware-lighted 3-dimensional solid model representations of macromolecules. *J. Mol. Graph.* **11**, 134-138.