BIOINFORMATICS

Vol. 16 no. 12 2000
Pages 1111–1119

Analysis of conservation and substitutions of secondary structure elements within protein superfamilies

Kenji Mizuguchi* and Tom L. Blundell

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Old Addenbrookes Site, Cambridge CB2 1GA, UK

Received on May 30, 2000; revised on June 30, 2000; accepted on July 25, 2000

Abstract

Motivation: *Structural alignments of superfamily members often exhibit insertions and deletions of secondary structure elements (SSEs), yet conserved subsets of SSEs appear to be important for maintaining the fold and facilitating common functionalities.*

Results: *A database of aligned SSEs was constructed from the structure-based alignments of protein superfamily members in the CAMPASS database. SSEs were classified into several types on the basis of their length and solvent accessibility and counts were made for the replacements of SSEs in different types at structurally aligned positions. The results, summarized as log-odds substitution matrices, can be used for two types of comparisons: (1) structure against structure, both with secondary structure assignments; and (2) structure against sequence with predicted secondary structures. The conservation of SSEs at each alignment position was defined as the deviation of observed SSE frequencies from the uniform distribution. This offers a useful resource to define and examine the core of superfamily folds. Even when the structure of only a single member of a superfamily is known, the extended method can be used to predict the conservation of SSEs. Such information will be useful when modelling the structure of other members of a superfamily or identifying structurally and functionally important positions in the fold.*

Availability: *The database is available on the world wide web at http://www-cryst.bioc.cam.ac.uk/~kenji/ssdb/. The conservation of SSEs was translated into colour values and can be visualized through the web interface.*

Contact: *kenji@cryst.bioc.cam.ac.uk*

Introduction

Many protein domains have been identified which have similar three-dimensional (3D) structures and functions despite low percentages of sequence identities (see, e.g. the SCOP (Murzin *et al.*, 1995), CATH (Orengo *et al.*, 1997), FSSP (Holm and Sander, 1996) and CAMPASS (Sowdhamini *et al.*, 1998) databases). In these domain superfamilies, detailed 3D structures can be diverse: insertions and deletions of secondary structure elements (SSEs) and relative shifts in the orientation of SSEs are often observed. Structure-based alignments of the sequences of superfamily members have been compiled in several databases including CAMPASS (Sowdhamini *et al.*, 1998) and more recently CATH-DHS (Bray *et al.*, 2000). CAMPASS displays the sequence alignments in an annotated format (Mizuguchi *et al.*, 1998a) that represents the structural environment, such as solvent accessibility, hydrogen bonding and secondary structure state, of each amino acid residue. In CAMPASS, superfamily members may have low sequence identities but are probably divergently evolved with similar structures and broadly similar functions. For example, members of the cystine-knot superfamily are all secreted proteins, most of which are dimeric growth factors/hormones that bind to specific cell-surface receptors, even though the signals transmitted by these proteins are diverse.

Despite low percentages of sequence identities, alignments of superfamily members do reveal conserved amino acid residues that are key to the fold and/or function. However, unlike the alignments of homologous proteins (as shown in the HOMSTRAD database, Mizuguchi *et al.*, 1998b), these conserved residues are often scattered over a wide range of sequence, and conventional sequence motif searches may fail to identify some or all members of the superfamily. In fact the common functionality of each superfamily can often be better explained in terms of the general arrangements of SSEs rather than detailed sequence motifs. For example, the sugar binding region of the lectin superfamily is at the concave face of a β-sheet where each member has different specific residues for sugar binding. Members of the superfamily of aspartate-α-decarboxylase and formate dehydrogenase contain a six-stranded β-barrel, with the active site located on one

*To whom correspondence should be addressed.

of the Ψ loops (Castillo *et al.*, 1999). These examples suggest the importance of comparative analyses of protein domains at the level of secondary structures, where we can define the core of a superfamily fold and derive rules for substitutions of SSEs.

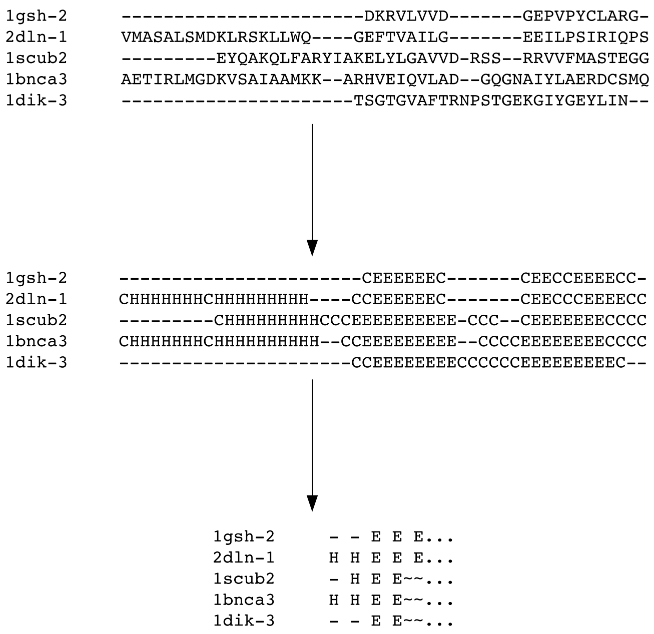
Previously, patterns of amino acid substitutions in a particular secondary structure type have been analyzed. Overington *et al.* (1990) constructed amino acid substitution tables for different structural environments defined by sidechain accessibility, hydrogen bonds and secondary structure. Secondary structure-dependent (or more generally, environment-dependent) amino acid substitution tables have been used for fold recognition (Luthy *et al.*, 1991; Johnson *et al.*, 1993) and secondary structure predictions (Wako and Blundell, 1994; Mehta *et al.*, 1995). Other fold recognition methods used related but different types of scoring matrices, which were designed for evaluating matches between the predicted secondary structure state for the probe and the known or predicted states for the target at each residue position (Russell *et al.*, 1998; Aurora and Rose, 1998).

Here, we describe a database of aligned secondary structures for the superfamilies in CAMPASS. Conserved features of SSEs have been investigated by analyzing this database, and substitution matrices have been derived to evaluate the matching of two structures, or a structure and a sequence with predicted SSEs. These substitution matrices are calculated for SSEs rather than individual amino acids (or secondary structure assignments for individual residues) as in the previous work mentioned above. We demonstrate that the SSE substitution matrices can be used for: (1) rapid structure comparisons (e.g. SSAP, Orengo *et al.*, 1992; DALI, Holm and Sander, 1998; VAST, Gibrat *et al.*, 1996; SEA, Rufino and Blundell, 1994; and COSEC, Mizuguchi and Go, 1995); (2) fold recognition (Rice and Eisenberg, 1997); and (3) identification of the core of a fold.

Results and discussion

Alignments of secondary structure elements

Residue-by-residue basis alignments in CAMPASS were automatically converted to aligned secondary structures (Figure 1). We consider 3_{10}-helices (G), α-helices (H) and β-strands (E) as SSEs. In our conversion algorithm, the coil state (C) appears only when a region of secondary structure is aligned with a non-secondary structure region in the original residue-by-residue basis alignment and therefore, coil is not regarded as a proper SSE. Similarly, if an entire segment of secondary structure has no equivalent region (either secondary structure or coil) in the sequence alignment, a gap (-) is introduced. Multiple alignments of SSEs were thus generated for all the superfamilies in CAMPASS.



```
1gsh-2    -----------------------DKRVLVVD-------GEPVPYCLARG-
2dln-1    VMASALSMDKLRSKLLWQ----GEFTVAILG-------EEILPSIRIQPS
1scub2    ---------EYQAKQLFARYIAKELYLGAVVD-RSS--RRVVFMASTEGG
1bnca3    AETIRLMGDKVSAIAAMKK--ARHVEIQVLAD--GQGNAIYLAERDCSMQ
1dik-3    ----------------------TSGTGVAFTRNPSTGEKGIYGEYLIN--
```

```
1gsh-2    -----------------------CEEEEEEC-------CEECCEEEECC-
2dln-1    CHHHHHHHCHHHHHHHHH----CCEEEEEEC-------CEECCCEEEECC
1scub2    ---------CHHHHHHHHHCCEEEEEEEEEEE-CCC--CEEEEEEECCCC
1bnca3    CHHHHHHHCHHHHHHHHHH--CCEEEEEEEEE--CCCCEEEEEEEECCCC
1dik-3    ----------------------CCEEEEEEEECCCCCEEEEEEEEEEC--
```

```
1gsh-2      - - E E E...
2dln-1      H H E E E...
1scub2      - H E E~~...
1bnca3      H H E E~~...
1dik-3      - - E E~~...
```

Fig. 1. Conversion of a residue-by-residue alignment to an alignment of SSEs. Top section shows an alignment of amino acid sequences for the members of the superfamily of glutathione synthetase, domain 3 (CAMPASS (Sowdhamini *et al.*, 1998) entry gshase_3). Middle section shows the secondary structure assignment by DSSP (Kabsch and Sander, 1983) for individual residues. H, G, E and C denote α-helix, 3_{10}-helix, β-strand and coil, respectively. Bottom section shows a derived alignment of SSEs. H, G, E and C indicate a segment of α-helix, 3_{10}-helix, β-strand and coil, respectively, rather than secondary structure assignment for individual residues. Tilde (∼) indicates that the preceding SSE is aligned with more than one SSEs.

Classification of secondary structure elements

We used two variables, length and solvent accessibility, for characterizing and classifying SSEs. The length of an SSE was defined as the number of amino acid residues in it. This naturally led to a classification of SSEs into *m* categories, each differing by one amino acid residue. The accessibility measure (see **Methods** for details) was categorized so that SSEs were also classified into *n* categories according to their accessibility values. Then counts were made for the joint occurrences of these two variables, the results being summarized in a ($m \times n$) contingency table. We produced two separate contingency tables, for α-helices and for β-strands. 3_{10}-helices were excluded from this analysis and treated as a separate single entity. In the following discussion, helices refer to both α- and 3_{10}- helices.

Contingency tables can be visualized by using correspondence analysis (Greenacre, 1984). The resultant scatter diagram reveals not only characteristic patterns of the distribution of each of the two variables, but also the association of these two. Correspondence analysis of the contingency table for β-strands shows that there is a distinct pattern relating length and accessibility (Figure 2a). Strands are clustered into two groups in terms of their accessibility: buried (accessibility categories 1 and 2; see **Methods**) and accessible (accessibility categories 3 or higher). The former is characteristically associated with the medium range of the length categories (3–7). Shorter strands (length categories 1 and 2) are associated with the medium range of accessibility. Based on these observations, we classified the strands into six types (Figure 2b): short buried (E1), short exposed (E2), medium buried (E3), medium exposed (E4), long buried (E5) and long exposed (E6).

No clear association was observed between the length and accessibility of α-helices. The only discernible pattern from correspondence analysis with various parameters was that large accessibility (accessibility categories 1, 2 and 3) is correlated with short length (length categories 1–4) (data not shown). We, therefore, classified the α-helices into four types (Figure 2c): short exposed (H1), short buried (H2), long exposed (H3) and long buried (H4).

Substitution frequencies of helices and strands

By applying the classification described above, the alignments of SSEs show matches and mismatches of each type of SSE. We took alignment positions that contain no gaps, and counted the number of matches and mismatches of each type of SSE in an analogous way to that used in the construction of the BLOSUM matrices (Henikoff and Henikoff, 1992). The raw substitution counts and a log-odds matrix computed from them are shown in Table 1.

To visualize the characteristics of the log-odds substitution matrix, Euclidean distances were calculated between all types of SSEs (Johnson and Overington, 1993). A dendrogram was constructed from the distance matrix (Figure 3). For helices, the short exposed (H1) and 3_{10}-helices (G0) form a cluster, which are then clustered with the short buried (H2). The long accessible (H3) and the long buried (H4) types form another cluster. For strands, the medium exposed (E4) and the long exposed (E6) are first clustered, and they are then clustered with the short exposed (E2). This cluster is combined successively with buried strands (E1, E3 and E5), although the topology of these larger clusters is unstable.

The substitution matrix described above can be used for comparing two known structures. While it cannot be directly applied to the comparison of a structure with a sequence whose 3D structure is unknown, it is often possible to predict the secondary structure of a sequence with reasonable accuracy (Rost and Sander, 1993). A detailed classification of SSEs (as in Figures 2b and c) would be difficult for predicted secondary structures

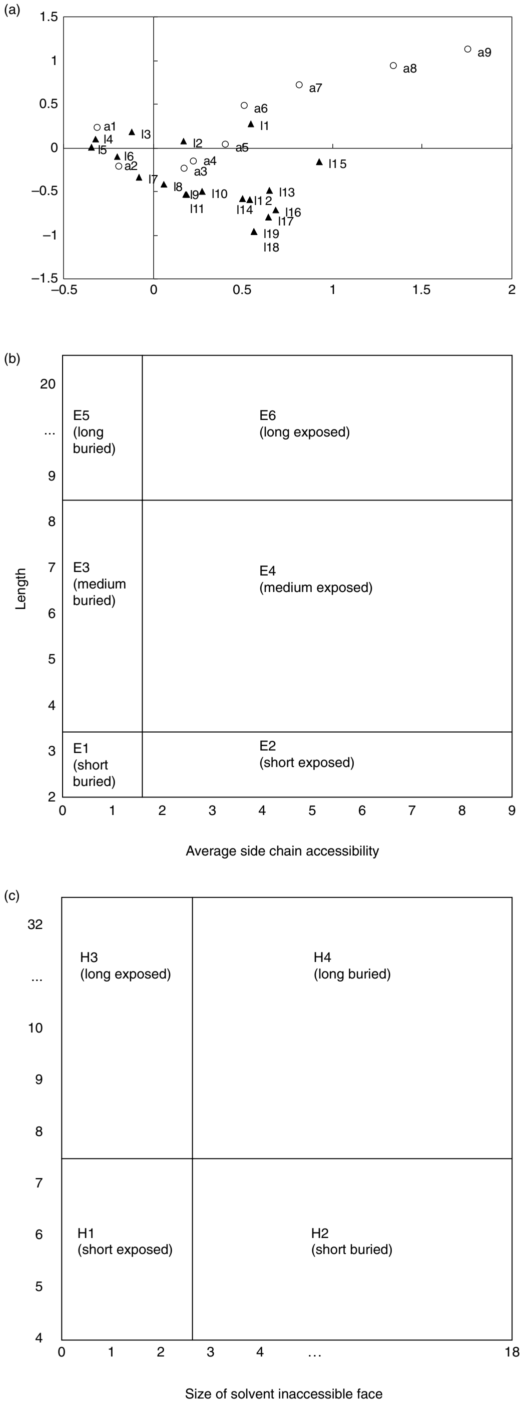


Fig. 2. Classification of secondary structure elements. (a) Correspondence analysis for β-strands. The scatter diagram shows the best two-dimensional representation of the contingency table relating length (categories l1–19 shown as filled triangles; see **Methods**) and accessibility (categories a1 to a9 shown as open circles; see **Methods**). (b) Six types of β-strands, defined on the basis of the length and average sidechain accessibility. (c) Four types of α-helices, defined on the basis of the length and size of the solvent inaccessible face.

Table 1. Observed substitution counts (lower triangle) and a log-odds matrix (upper triangle) for 67 superfamilies of aligned secondary structure elements. The log-odds values are expressed in 1/3 bits unit

	G0	H1	H2	H3	H4	E1	E2	E3	E4	E5	E6	
	8	6	5	3	2	−15	−15	−17	−14	−7	−16	G0
		9	4	5	1	−14	−15	−13	−13	−8	−15	H1
G0	46		10	−2	2	−15	−16	−18	−17	−15	−16	H2
H1	45	28		8	2	−8	−9	−20	−16	−16	−18	H3
H2	42	26	73		5	−14	−15	−24	−19	−15	−20	H4
H3	56	65	16	157		9	2	2	0	−1	−1	E1
H4	145	94	187	276	1072		8	−3	5	−3	2	E2
E1	0	0	0	3	2	55		4	−2	2	−2	E3
E2	0	0	0	3	2	27	69		6	0	3	E4
E3	0	2	0	0	0	166	61	1213		8	5	E5
E4	1	1	0	1	2	52	188	248	343		7	E6
E5	2	1	0	0	1	9	6	163	38	37		
E6	0	0	0	0	0	17	48	125	205	80	144	

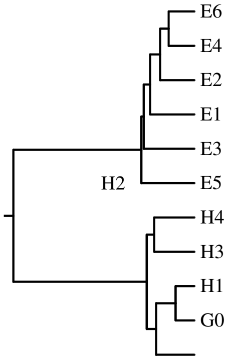


Fig. 3. Dendrogram showing the similarities between SSE types. Euclidean distances were calculated between each pair of SSE types from the log-odds matrix in Table 1. Hierarchical clustering was performed using KITSCH (Felsenstein, 1985).

but it will still be possible to assign a broader class to each predicted SSE. For example, by merging 3_{10}- and α-helices and ignoring the accessibility, five different types of SSEs can be defined (Hs, short helix; Hl, long helix; Es, short strand; Em, medium strand and El, long strand). Another set of substitution counts were then made between the original 11 SSE types (G0, H1-H4 and E1-E6) and the five broader states just mentioned. The log-odds matrix was calculated (Table 2) in an analogous way to the environment-specific amino acid substitution matrices (Overington *et al.*, 1990, K.Mizuguchi, unpublished).

This new type of substitution matrix can be used for the comparison of a known structure with a sequence with predicted secondary structures. Matches and mismatches between observed and predicted SSEs can be evaluated not uniformly, but in a position-specific manner. For example, a predicted short strand should produce a high score when aligned with a short buried strand (E1), but an alignment of the same type of SSE with a predicted long strand should be penalized.

In actual applications, it is necessary to construct a dynamic programming matrix whose (i, j) element contains the score of the best alignment between the first i residues of one protein and the first j residues of another. When these substrings terminate in the middle of an SSE, it is not trivial to evaluate the contributions from SSE matching. Various other parameters will be also required, such as relative weights for amino acid and SSE matching and gap penalties for SSEs. While these details are currently being examined in the development of our profile alignment method FUGUE (http://www-cryst.bioc.cam.ac.uk/~fugue/; J.Shi, T.L.B. and K.M., submitted), the use of the SSE substitution matrix for evaluating sequence alignments is demonstrated in Figure 4. Five new superfamilies have been added to CAMPASS after the collection of the statistics for the present analysis. We used these alignments to see whether the SSE substitution matrix can discriminate good sequence alignments from bad ones without knowing the detailed probe structure. From each of the five structure-based alignments, we first extracted all pairwise alignments. (For example, if the superfamily includes three members, A, B and C, three alignments AB, AC and BC were extracted.) These alignments were regarded as a reference set. We also realigned each of these alignments using the sequence comparison method CLUSTALW (Higgins *et al.*, 1996) and regarded them as a test set. Because of the low percentage identities and the lack of structural information, the test set includes both correct and incorrect alignments. SSE alignment scores were then calculated using Table 2 for the two sets, with the full classification

Table 2. Log-odds substitution scores from the full 11 SSE types to the reduced five SSE types (Hs, short helix; Hl, long helix; Es, short strand; Em, medium strand; El, long strand). The log-odds values are expressed in 1/3 bits unit

	G0	H1	H2	H3	H4	E1	E2	E3	E4	E5	E6
Hs	7	6	7	3	2	−18	−19	−20	−17	−10	−21
Hl	2	2	2	4	4	−12	−13	−27	−19	−16	−21
Es	−18	−17	−19	−9	−16	7	7	0	3	−2	0
Em	−17	−13	−19	−19	−24	1	1	3	2	1	0

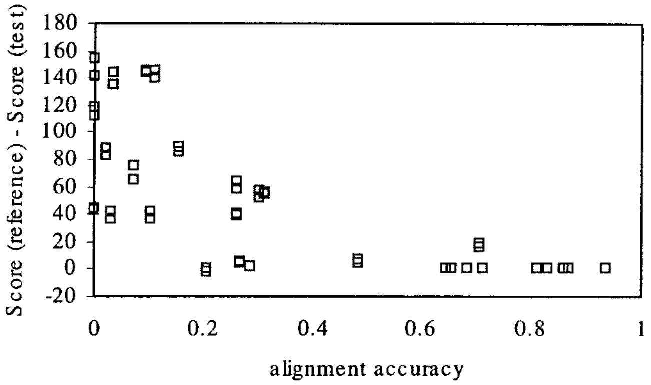


Fig. 4. Scores for SSE alignments calculated using Table 2. Score for the correct (reference) alignment minus that for the test alignment is plotted against the alignment accuracy (0–1) of the test alignment. The reference alignments were generated from the following five CAMPASS entries (SH2, muconate_ndomain, periplasmic_binding_I1, periplasmic_binding_I2 and propellor). The test alignments were obtained by realigning these sequences with CLUSTALW (Higgins *et al.*, 1996).

applied to one structure of a pair (target) and the broader classification applied to the other (probe). Figure 4 shows the score for the correct (reference) alignment minus that for the test alignment plotted against the alignment accuracy of the test alignment. It can be seen that the correct alignments produced better scores in 59 of the 60 cases and the difference is particularly pronounced when the alignment accuracy is low. This result suggests that the substitution matrix in Table 2 would help discriminate good and bad alignments.

Deletion and substitution to coil

The analysis in the previous section concerned only substitutions between different types of SSEs and did not include substitution to coil or deletion. This was because coil and gap were not proper SSEs (attributes depending on comparison, rather than each sample) and the normalization for obtaining the log-odds matrix could not be applied to them. The existence of coil has no counterpart in the derivation of amino acid substitution matrices but it cannot be simply ignored as, for example, shorter helices are expected to be replaced by coil more often than longer ones. Substitutions to coil or gap (i.e. deletion) are, therefore, expected to reveal additional features of different types of SSEs, not expressed in the substitution matrix.

To investigate this, separate counts were made from the alignments of SSEs for matches of each type of SSE and a gap or coil. Edge gaps and edge coils were excluded to account for uncertainties in the definition of domain boundaries. The results, summarized in Table 3, show similar patterns both for substitutions to coil and deletions. In general, length appears to be the major factor for determining the probability of deletion and substitution to coil; short SSEs (3_{10}-helices, helix types H1 and H2 and strand types E1 and E2) show higher tendency of deletion or substitution to coil, than longer ones. This result may in part be due to the difficulty in defining short SSEs because of their structural variability. For short and medium strands, however, the buried types (E1 and E3) are much less likely to be substituted than the accessible ones and this may indicate that buried short strands, often found in the middle of a parallel sheet, are difficult to substitute.

Conservation of secondary structure at each position in a domain fold

The multiple alignment of SSEs can be used to calculate the degree of conservation at each position in a domain fold. Both substitutions between different types of SSEs (Table 1) and coil substitutions and deletions (Table 3) should be taken into account. We adopted the structure logo formula (Gorodkin *et al.*, 1997) to define the information content $I(i)$ of position i:

$$I(i) = \sum_{k \in A} p_k(i) \log \frac{p_k(i)}{q_k(i)}.$$

This quantifies the contrast between the observed probabilities $p(i)$ and a reference probability distribution $q(i)$, where A is the set of all SSE types (G1, H1,..., E6) plus '-' (coil substitution/deletion). For the reference probability distribution, we use a flat distribution ($q_k = 1/11$), and we set q equal to one for k = '-', following Gorodkin *et al.*

Table 3. Probabilities of substitution to coil and deletion

	Del*	Coil†	Match#	$P_{\mathrm{del}}^{\S}$	$P_{\mathrm{coil}}^{\P}$	$P_{\mathrm{dc}}^{\|}$
G0	1311	725	1414	0.38	0.21	0.59
H1	409	149	655	0.34	0.12	0.46
H2	406	282	1076	0.23	0.16	0.39
H3	164	75	1314	0.11	0.05	0.15
H4	838	460	6548	0.11	0.06	0.17
E1	204	104	753	0.19	0.10	0.29
E2	605	394	821	0.33	0.22	0.55
E3	348	250	4483	0.07	0.05	0.12
E4	258	219	1870	0.11	0.09	0.20
E5	59	78	535	0.09	0.12	0.20
E6	21	50	893	0.02	0.05	0.07

*Observed counts of deletion excluding edge gaps.
†Observed counts of substitution to coil excluding N- and C-terminal coils.
#Observed counts of substitution to other SSE types.
§Del/(Del + Coil + match).
¶Coil/(Del + Coil + match).
‖ $P_{\mathrm{del}} + P_{\mathrm{coil}}$.

(1997). This choice ensures high values of I for alignment positions that are dominated by a certain type of SSE and with few coil substitutions/deletions. Alignment positions with high information content can be regarded as conservative and thus form the 'core' of the domain fold.

The number of structures in each CAMPASS alignment is generally small and $p(i)$ is not reliable if obtained simply by counting observed SSE types. Instead, we used the substitution table (Table 1) to estimate the observed probabilities. This will also produce a high value of information content even if the position is not dominated by a single SSE type but is occupied by similar, inter-changeable SSEs. To estimate $p(i)$, the smoothing procedure was employed, which had been used in deriving environment-specific amino acid substitution tables (Topham *et al.*, 1993):

$$p_k(i) = \omega_1 A_k(i) + \omega_2 W_k(i),$$

where $A_k(i)$ is obtained by averaging the substitution probabilities to type k from each of the observed SSEs at position i and $W_k(i)$ is the observed frequency of type k. The weights (ω_1, ω_2) are

$$\omega_1 = 1/(1 + N_k/(11\sigma))$$

where N_k is the number of structures adopting the kth type of SSE and σ is a constant (for the present work, $\sigma = 0.5$ is used).

The information content was converted to colour values (blue conservative, red variable) and mapped onto each structure in the CAMPASS database. An example is shown in Figure 5a, where a superposition of the members of the globin superfamily (CAMPASS entry globins) is displayed. Eight helices, labeled A–H, are common to most structures but some members lack the small D helix and also helix C can be either an α- or 3_{10}-helix. These helices, shown in yellow and orange, have small values of information content, indicating that they are variable. Of the rest of the core helices, helix B appears to be the most conservative (shown in dark blue), whereas helix F appears the least conservative (shown in green), consistent with the previous observation that in some members of the superfamily, helix F can be broken into two parts (called F′ and F) (Lesk and Chothia, 1980). There are a few additional helices (such as the N-terminal helix of 3sdha0), which are highly variable.

As an example of the α/β fold, a flavodoxin-like domain (CAMPASS entry flav) of methionine synthetase is shown in Figure 5b. It can be seen that the strands forming the central β-sheet are well conserved while some of the peripheral helices are variable. Among the β-strands, middle strands are better conserved than the edge strands. Other α/β proteins including α–β hydrolases and the superfamily of thiamin-binding-like domains show similar tendencies. The α/β barrel of the glycosyltranferase superfamily also shows that the barrel strands are better conserved than the surrounding helices.

As an example of the all-β fold, the structure of the PH domain is shown in Figure 5c. The fold of PH domains is a partly opened β-barrel capped by an α-helix. Strand 4 is at the partly opened barrel edge and is less conserved than the other strands; its length varies among the members of the superfamily.

The information content for all the structures in CAMPASS can be visualized through the RasMol (Sayle and Milner-White, 1995) interface, which is available at http://www-cryst.bioc.cam.ac.uk/~kenji/ssdb/. Normalized values of the information content were used in FUGUE (http://www-cryst.bioc.cam.ac.uk/~fugue/) to modulate the position-specific gap penalties (the more conserved, the higher the gap penalties) and this procedure was shown to improve the fold recognition performance (J.Shi, T.L.Blundel and K.Mizuguchi, unpublished).

Conserved segments from a single structure

For the majority of superfamilies currently known, only a single representative structure is available (at present, these superfamilies are not included in CAMPASS and thus not used for the present analysis). One advantage of the use of the smoothing technique is that the above formula for the conservation of SSEs can be also applied to the single-member superfamilies. This will provide useful information for modelling the structure of other members of the superfamily based on the known structure.

Because of the present choice of the smoothing parameter σ, contributions from the observed probabilities become significant only when the number of structures is

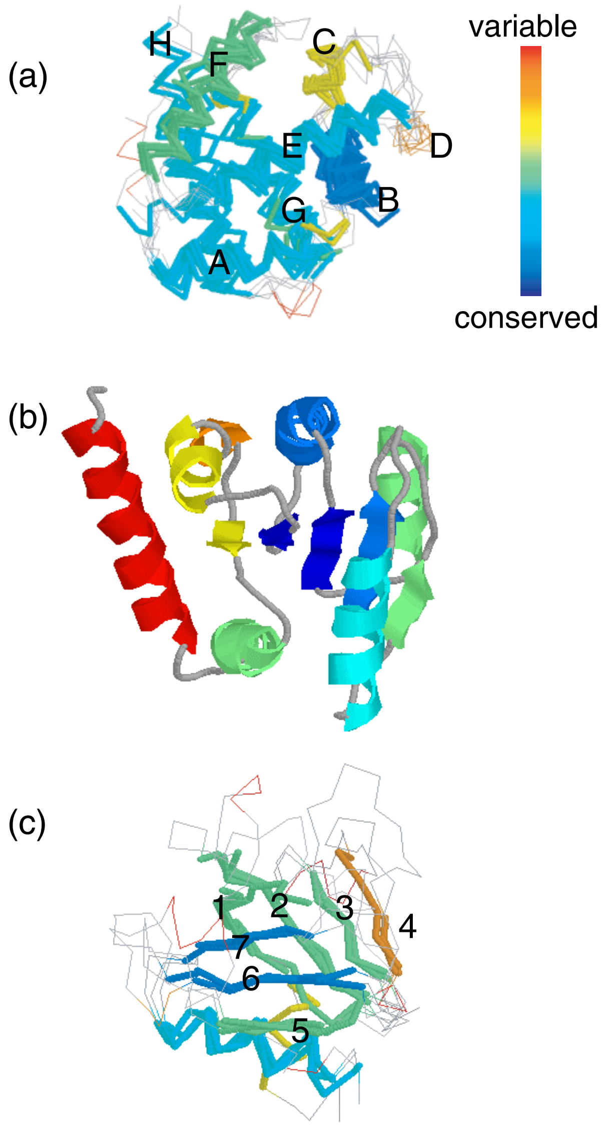


Fig. 5. Conservation of SSEs. The information content (degree of conservation) is translated into colour values (ranging from −0.7, red to 2.75, blue). (a) Superposed structures of the globin superfamily (CAMPASS entry globins) shown as Cα traces. Loops and least conserved SSEs (with information content less than 0.2) are shown in thin lines. (b) Cartoon representation of the structure of methionine synthetase (1bmta1), as a representative of the flavodoxin-like superfamily (CAMPASS entry flav). (c) Superposed structures of the PH domain superfamily (CAMPASS entry PH), shown as in (a). Figure drawn with RasMol (Sayle and Milner-White, 1995).

greater than five. There are not enough new superfamilies of this size, which had not been used to collect the statistics and therefore, a rigorous jackknife test is difficult to perform at this stage. Instead, we show one example to demonstrate the usefulness of this procedure. Figure 6a shows the superposed structures of the members of the superfamily of class I periplasmic binding proteins, domain 2 (CAMPASS entry periplasmic_binding_I2), colour-coded as in Figure 5. One member, AmiC (1pea-1), was then taken and the information content for each position was calculated only from this structure. Figure 6b shows the structure of AmiC using the same colour scheme. The agreement of these two is generally good, with the exception of helix V308-L315 (top right in Figure 6a and b). This helix was classified as long buried (H4) and predicted to be fairly conservative (shown in light blue in Figure 6b). Some members of this superfamily, however, lack this helix and from the observed probabilities in the alignment,

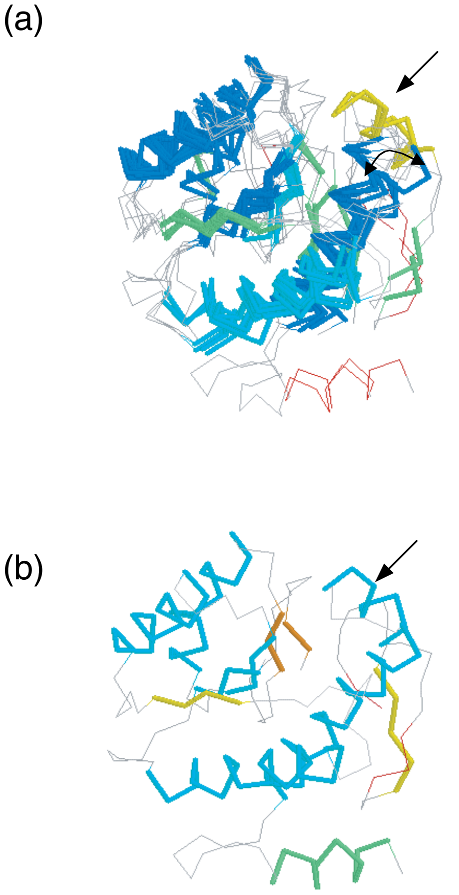


Fig. 6. Conserved core of a fold defined from a multiple structural alignment (a) and a single structure (b). (a) Superposed structures of the superfamily of class I periplasmic binding proteins, domain 2 (CAMPASS entry periplasmic_binding_I2), shown as in Figure 5. (b) Structure of AmiC (1pea-1) colour-coded with the information content calculated from this structure alone. Arrow indicates the helix V308-L315. The movement of the preceding helix is also shown. Figure drawn with RasMol (Sayle and Milner-White, 1995).

the helix was regarded as variable (shown in yellow in Figure 6a). In the members without this helix, a large shift of the preceding helix (equivalent to helix A284-A303 of AmiC) is observed and this appears to compensate the loss of this relatively long and buried helix.

Conclusion

The conservation of SSEs in protein superfamilies has been analyzed and substitution matrices for comparing two structures or a structure with a sequence with predicted secondary structures have been derived. The first matrix can be used as part of a structure comparison method. The second matrix can be used for the construction of 'secondary structure profiles'. For each SSE position of a domain fold, SSE replacement scores and gap-penalty information can be extracted from the substitution tables. Use of secondary structure prediction was shown to improve the sensitivity of fold recognition methods (Rice and Eisenberg, 1997; Hargbo and Elofsson, 1999) and the secondary structure profiles discussed here can be integrated into a new fold recognition algorithm. The analysis of the conservation of SSEs has also been used to define the core of a fold, either from a multiple structural alignment or a single structure. The core of a fold from SSE is similar in concept but distinct from the Cα-based methods for core definitions (Gerstein and Altman, 1995; Matsuo and Bryant, 1999; Orengo, 1999) and will be useful in identifying new members of the superfamily or modelling their structures.

Methods

Structure-based alignments of protein sequences were taken from the July 1998 version of the CAMPASS database, which contains 67 superfamilies and 272 structural domains (a complete list of the superfamilies used for analysis can be found at http://www-cryst.bioc.cam.ac.uk/~kenji/ssdb/.

Secondary structures were assigned by SSTRUC, an implementation of the Kabsch and Sander algorithm (Kabsch and Sander, 1983). For α-helices, the accessibility was defined as the size of the inaccessible face. The inaccessible face of a helix is the arc of a helical wheel, which is made by inaccessible residues in the helix (Blundell and Zhu, 1995). The size is measured by the number of residues in the arc and ranges between 0 and 18 (denoted here categories 1–19). Residues with relative sidechain accessibility $<7\%$ were regarded as inaccessible (Hubbard and Blundell, 1987). For β-strands, the accessibility was defined as relative sidechain accessibility, averaged over all the residues in a strand. The average accessibility values, expressed in percent, were classified into ten bins (categories 1–9; e.g. strands with average accessibility $<10\%$ are in category 1, 10–20% in category 2). Solvent accessibility was calculated by PSA, which is an implementation of the method of Richmond (1984).

After the classification of SSEs, we calculated the frequencies A_{ab} of observing type a (G0, H1, E1, etc.) aligned against type b in the alignments of SSEs. From the A_{ab}, the probability $P_{a\to b}$ of substituting type a with type b was calculated as $P_{a\to b} = A_{ba}\Big/\sum_b A_{ba}$. The frequency q_a was estimated as $q_a = \sum_b A_{ab}/\sum_{ab} A_{ab}$ and the log-odds values were derived by $s_{ab} = \log\frac{P_{a\to b}}{q_b}$. Elements of the log-odds matrix were multiplied by a scaling factor of 3/log2 and rounded to the nearest integer value (Table 1).

Euclidean distances were calculated between pairs of columns in the log-odds matrix. The distance between columns i and j was defined as $D_{i,j} = \sqrt{\sum_k (X_{k,i} - X_{k,j})^2}$, where $X_{k,j}$ is the (k, j) element of the log-odds matrix. The distance matrix $D_{i,j}$ was used as the input to KITSCH (Felsenstein, 1985) to perform hierarchical clustering and produce a dendrogram.

Acknowledgement

We thank SmithKline Beecham Pharmaceuticals for financial support.

References

Aurora,R. and Rose,G.D. (1998) Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons. *Proc. Natl. Acad. Sci. USA*, **95**, 2818–2823.

Blundell,T.L. and Zhu,Z.Y. (1995) The alpha-helix as seen from the protein tertiary structure: a 3D structural classification. *Biophys. Chem.*, **55**, 167–184.

Bray,J.E., Todd,A.E., Pearl,F.M., Thornton,J.M. and Orengo,C.A. (2000) The CATH dictionary of homologous superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.*, **13**, 153–165.

Castillo,R.M., Mizuguchi,K., Dhanaraj,V., Albert,A., Blundell,T.L. and Murzin,A.G. (1999) A six-stranded double-psi beta barrel is shared by several protein superfamilies. *Structure*, **7**, 227–236.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

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

Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Greenacre,M. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, Orlando, Fl.

Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci.*, **13**, 583–586.

Hargbo,J. and Elofsson,A. (1999) Hidden Markov models that use predicted secondary structures for fold recognition. *Proteins*, **36**, 68–76.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10 915–10 919.

Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1996) Using CLUSTAL for multiple sequence alignments. *Meth. Enzymol.*, **266**, 383–402.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.

Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.

Hubbard,T.J. and Blundell,T.L. (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.*, **1**, 159–171.

Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.*, **233**, 716–738.

Johnson,M.S., Overington,J.P. and Blundell,T.L. (1993) Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.*, **231**, 735–752.

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

Lesk,A.M. and 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.

Luthy,R., McLachlan,A.D. and Eisenberg,D. (1991) Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, **10**, 229–239.

Matsuo,Y. and Bryant,S.H. (1999) Identification of homologous core structures. *Proteins*, **35**, 70–79.

Mehta,P.K., Heringa,J. and Argos,P. (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci.*, **4**, 2517–2525.

Mizuguchi,K. and Go,N. (1995) Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng.*, **8**, 353–362.

Mizuguchi,K., Deane,C.M., Johnson,M.S., Blundell,T.L. and Overington,J.P. (1998a) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.

Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998b) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

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

Orengo,C.A., Brown,N.P. and Taylor,W.R. (1992) Fast structure alignment for protein databank searching. *Proteins*, **14**, 139–167.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—A hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Overington,J., Johnson,M.S., Sali,A. and Blundell,T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. R. Soc. Lond. B. Biol. Sci.*, **241**, 132–145.

Rice,D.W. and Eisenberg,D. (1997) A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.*, **267**, 1026–1038.

Richmond,T.J. (1984) Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.*, **178**, 63–89.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rufino,S.D. and Blundell,T.L. (1994) Structure-based identification and clustering of protein families and superfamilies. *J. Comput. Aided. Mol. Des.*, **8**, 5–27.

Russell,R.B., Saqi,M.A., Bates,P.A., Sayle,R.A. and Sternberg,M.J. (1998) Recognition of analogous and homologous protein folds—assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng.*, **11**, 1–9.

Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.

Sowdhamini,R., Burke,D.F., Huang,J.F., Mizuguchi,K., Nagarajaram,H.A., Srinivasan,N., Steward,R.E. and Blundell,T.L. (1998) CAMPASS: a database of structurally aligned protein superfamilies. *Structure*, **6**, 1087–1094.

Topham,C.M., McLeod,A., Eisenmenger,F., Overington,J.P., Johnson,M.S. and Blundell,T.L. (1993) Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.*, **229**, 194–220.

Wako,H. and Blundell,T.L. (1994) Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *J. Mol. Biol.*, **238**, 693–708.