Decision support system for the evolutionary classification of protein structures*

Liisa Holm and Chris Sander

EMBL-EBI
Wellcome Trust Genome Campus, CB10 1SD Cambridge, U.K.
surname@embl-ebi.ac.uk

Abstract

The structures of nearly a thousand sequence-unique proteins represent only 300 different 3D shapes. Is structural resemblance between proteins with little sequence similarity the result of physical convergence to favourable folding patterns, or does it reflect a memory of common evolutionary history? Separating these two processes is important for organizing genome data in terms of protein families and for theoretical approaches to protein structure prediction by fold recognition techniques. Achieving separation requires a combination of structure, sequence and functional analysis of proteins. For this purpose, we are developing a decision support system that scans heterogeneous protein sequence and structure related databases, and collects or calculates characters indicative of common functional constraints. The criteria include sequence homology, analysis of 3D clusters of conserved residues, conservation of active sites, and keyword analysis of biological function. Even without extensive refinement, application of a combination of these criteria to a test set representing all currently known protein structures yields 87% coverage with 7 % false positives, compared to 53 % coverage by only 1D sequence criteria. Thus, the semiautomatic prototype system significantly enhances the efficiency of unifying families of functionally related proteins in spite of long evolutionary distances.

Introduction

Taxonomic classification has long traditions in biology. Classic work by Linné, Darwin, Wallace organized species of plants and animals in a hierarchy based on common morphological characters. Access to the genotype has allowed molecular phylogenies to be constructed not only of species of organisms but also within and between protein families. The concept of evolution in which gradual changes to protein phenotype (structure and function) result from amino acid replacements, has made searching databases for significant sequence similarities a standard technique of functional characterization of newly determined genes.

In constructing molecular phylogenies, the use of sequence information has two limitations. First, the accuracy of predicted biological function is different between orthologous (e.g., myoglobins in the muscle of whales and humans) and paralogous genes (e.g., myoglobin and leghemobin in the roots of plants). Second, protein folds appear to be compatible with a very wide range of amino acid substitutions, making detection of homology difficult at long evolutionary distances. Fortunately, comparison of 3D structures has led to the discovery of many distant evolutionary relationships that are not easily captured even by the most sophisticated 1D models of sequence evolution [2-3]. We have previously developed the Dali/FSSP structure alignment database and fold classification to automatically monitor where new structures map in fold space in terms of a geometrical similarity measure [4].

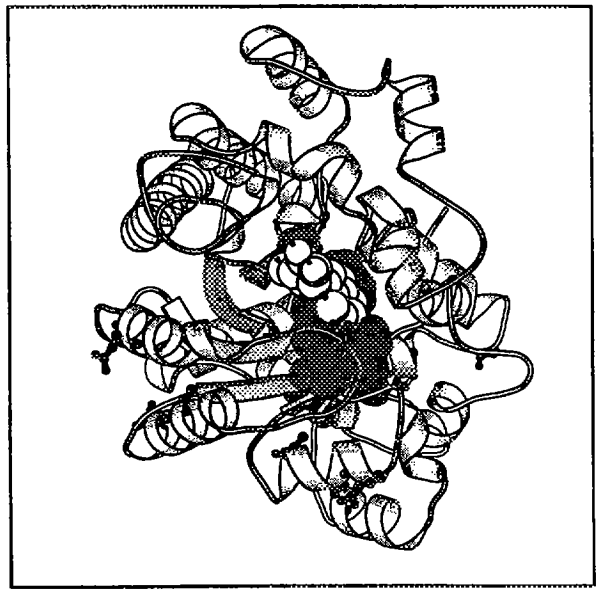


Figure 1. Functional residues are conserved and cluster in 3D.

Adenosine deaminase, phosphotriesterase and urease share a conserved active site with four invariant histidines and an aspartic acid supporting metal binding and a common biochemical mechanism [5]. The conserved residues were identified by structural alignment expanded by sequence homologs. Here, the clustering of the invariant residues (dark spacefilling representation) is shown mapped onto the structure of adenosine deaminase [18]. A purine nucleoside ligand is shown in light spacefilling representation. A number of peripheral residues are conserved between only two proteins of the triplet (ball-and-stick representation).

Figure 1 illustrates one case of remote evolutionary relatives discovered by structure comparison. Urease, phosphotriesterase and adenosine deaminase not only have a conserved structural core but support a conserved active site constrained to perform metal-assisted hydrolysis of amide bonds. The active site is made up of four histidines and an aspartic acid; these are the only residues which are invariantly conserved in all three families. Although the active site residues are widely dispersed along the polypeptide chain, they cluster together in the folded structure. They are invariantly conserved between the respective protein families, defining a sharp sequence signature for the superfamily that led to the evolutionary unification of a large set of distinct amidohydrolase families [5].

Here, we examine computational criteria for verifying hypotheses of functional homology between proteins with very low sequence similarity. Computationally, we work in a decision support system framework, using intelligent agents to access heterogeneous databases (Figure 2). We calibrate the selectivity and sensitivity of quantitative criteria related to biological function against a balanced, comprehensive test set. We discuss the potential of a composite criterion as the basis for an automatic expert system, assisting or replacing human experts [6-7], that tackles the problem of evolutionary classification of protein structures based on more than mere sequence and structure criteria.

Test set

The notion of a biologically significant relationship between proteins is intuitive and therefore rather imprecise. In this work, we formulate a number of quantitative criteria and test how well they correspond to human intuition. The test set was composed of 458 evolutionarily unrelated and 482 related pairs (classified manually by L.H.). The test set is available electronically over the Internet at the URL http://www2.embl-ebi.ac.uk/dali/testset.

Our basic assumption in constructing a test set was that remote homologs are detected in structure comparison and are positioned in structure space as near neighbours. The test set consisted of 941 proteins which have known 3D structures and less than 25 % mutual sequence identity. In order to obtain a balanced sample of pairs in terms of fold types and protein families, a minimal spanning tree with 940 links was constructed. Linkage was by Z-scores reported in the FSSP database. Each protein is linked to its closest structural neighbour, and the whole set is connected. As the FSSP database only reports pairs with Z-scores higher than 2, 103 arbitrary links were created to merge isolated branches into one connected set. One section of the tree is graphically presented in Figure 3.

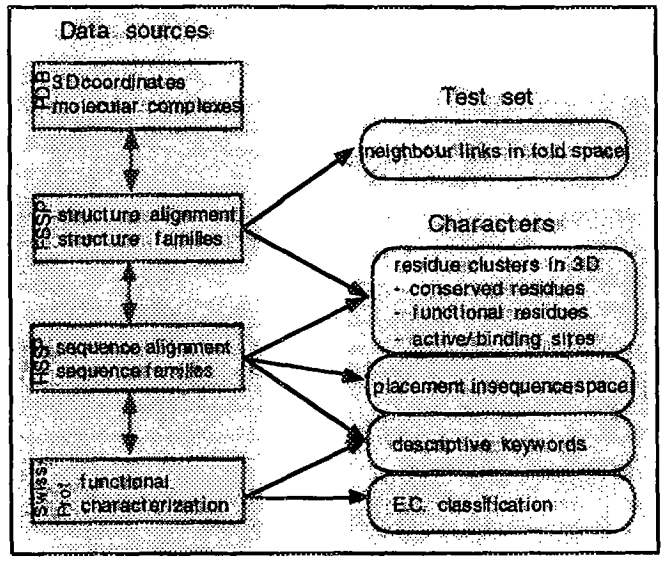


Figure 2. Flowchart.

The decision support system addresses the question whether two proteins are functionally homologous by collecting and combining information from heterogeneous databases. The query proteins are first mapped to their respective families, defined using a 25 % sequence identity cutoff [2], so that the computation of 'characters' can make use of all information attached to any family member and of family properties such as residue conservation. Primary databases for structures and sequences are the Protein Data Bank (PDB [19]) and Swissprot [20], respectively. The FSSP [21] and HSSP [2] databases contain derived multiple alignments of structures and sequences, respectively, and provide equivalence links between different entries in the primary databases at the protein level (families), at the residue level (columns of multiple alignments), and at the 3D site level (clusters of residues in spatial proximity).

Conservative characters

This section explains the biological background and computational details of six criteria related to evolutionary constraints on protein families.

1. Structure similarity

Protein structure is conserved over much longer evolutionary distances than amino acid sequences, in terms of being distinguishable from database background. The quantitative criterion of overall structure similarity between two proteins was the statistical significance (Z-scores) by the Dali method of distance matrix alignment [4].

2. Sequence family overlap

The alphabet of 20 natural amino acids generates a vast sequence space. Empirically, sequence identity above 25 %

between two proteins is a reliable indicator of common evolutionary descent [2]. This threshold is used in the HSSP database in listing members of the protein family centred around a protein of known 3D structure. We define that there is overlap between two protein families centred around proteins *A* and *C*, if there exists a bridging sequence *B* that is listed both in the HSSP dataset for protein *A* and in that for protein *C* , even though *A* and *C* may be less than 25 % identical.

3. Enzyme class

The biochemical reactions catalyzed by enzymes are codified in the E.C. numbers [8]. At the top level, there are six broad groups of oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. We used identity of the first numbers of the E.C. codes as a criterion of functional similarity.

It should be noted that even though biochemical function tends to be conserved in the evolution of protein families, the E.C. numbers are not a phylogenetic classification. Unrelated protein families may catalyze the same biochemical reactions, and members of one and the same protein family may be assigned several E.C. numbers. For example, the family known as class-III aminotransferases uses covalently bound pyridoxal-phosphate as a cofactor to catalyze reactions with E.C. numbers 2.6.1.(11,13,18,19,62), 4.1.1.64, and 5.4.3.8 [9].

4. Common functional sites (experimental)

We used two sources of experimental information about functional sites in proteins: sequence annotations in the Swissprot database, and crystal structures of protein-ligand complexes.

From Swissprot feature fields, we used sites spanning a single residue (excluding e.g. DNA-binding regions) and containing the words THIOLEST, METAL, *_BIND, ZN_FING, ACT_SITE or MUTAGEN but not ALLELE or NO EFFECT. Annotations of homologous family members were translated via sequence alignment to the sequence coordinates of the HSSP master sequence (i.e., the structurally known protein).

To extract information from crystal structures in the Protein Data Bank, we defined ligands as molecules given in HETATM records excluding sulfate, water, bromine, chloride, sodium, beta-merkaptoethanol, methyl, ethanol, acetic acid, nitrate, potassium, acetyl groups, methanol, and ammonium. The FSSP database linked 6377 Protein Data Bank entries to the 941 representatives by unambiguous sequence homology (>25 % sequence identity). To simplify contact calculations, the 3D superimposition of the protein chains was used to project the positions of known ligands in the crystal structure of any homologous protein onto the representative structure.

The requirements of common functional residues in a pair of proteins were evolutionarily conservation in both protein families (HSSP variability<10) and identity of amino acid type. A pair of proteins was defined to share a functional site if identically conserved residues in the structural alignment either (i) included any that had active site annotation in Swissprot, or (ii) included at least two residues in contact with a ligand molecule (at less than 4.0 Å atom-atom distance).

5. Common functional sites (predicted)

In cases where sequence annotation or structures of ligand complexes is not available, functional sites may be predicted based on sequence conservation and clustering of conserved residues in the 3D structure. The difficulty is that there are two types of conservation. Conserved hydrophobic residues typically have a structural role in the solvent-inaccessible core of the protein, whereas the active site is typically made up of conserved polar residues.

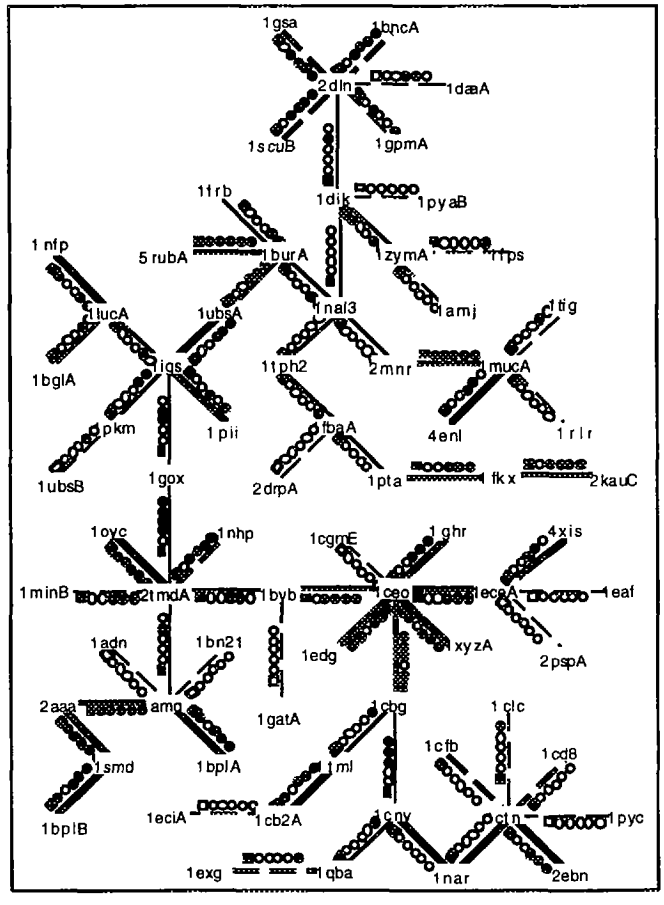


Figure 3. Resolution of clusters of functionally related protein families in the region of (βα)8 barrel folds.

Superfamilies are marked by thick links. Dotted lines denote links to non-(βα)8-barrel proteins. The bit-patterns (square: z-score, circles:sequence family, E.C., sites, function preference, keywords) next to each link depict the 6 criteria (black: above; white: below cutoff defined in Table I). The layout was created manually.

These prejudices were cast in numerical terms through the derivation of the following preference parameters for each amino acid type *r*. Let us define:

(1a) $N_{tot} = \Sigma N_r$ = total number of residues,

(1b) $C_{tot} = \Sigma C_r$ = total number of conserved residues,

(1c) $F_{tot} = \Sigma F_r$ = total number of functional residues,

where the summation is over all residues in the test set of 941 sequence-unique structures. The data set had a total of N_{tot} = 190978 residues, C_{tot} = 70791 conserved residues (HSSP variability < 10), and F_{tot} = 11834 functional residues within the conserved subset which are in contact with ligands (<4 Å atomic distance). Factoring out the general preference of being conserved from the preference of being functional, the log-odds ratio of observed over expected counts gives the preferences (Figure 4):

(2a) Conservation preference(r)=$\log[(C_r/N_r)/(C_{tot}/N_{tot})]$,

(2b) Functional preference(r) = $\log[(F_r/C_r)/(F_{tot}/C_{tot})]$.

In principle similar though more elaborate functional residue preferences have been derived earlier by Ouzounis, Sander and Valencia (*pers. comm.)* . The preferences should be useful in predicting active sites from multiple alignments of protein families which do not yet have a known structure. For illustration, we here apply the preferences to the same set of proteins from which the parameters were derived which, strictly speaking, is circular (but we do not think overlearning is severe in this particular case).

As with criterion 4, putative functional residues must have low variability in both protein families and identical sequence. We identified clusters of conserved residues and evaluated the functional potential of each cluster. Clusters of conserved residues were defined by single linkage clustering of residues in contact (atom-atom distances less than 4 Å). The functional preferences were summed over the residues in each cluster, and the highest preference score retained.

For example, comparison of adenosine deaminase and phosphotriesterase resulted in seven clusters with the compositions L, HHHGHD, N, A, G, E and G; comparison of adenosine deaminase and urease resulted in eight clusters with the compositions HHHHD, IN, G, E, D, A, E and A (see Figure 1). In both cases, the histidine-rich cluster at the active site has the highest functional preference.

6. Keyword overlap

The most fuzzy criterion is based on the subjective keyword annotation of the sequence entries in Swissprot. In this work, we restricted ourselves to keyword identity and crude elimination of noninformative keywords (3D-structure, acetylation, alternative initiation, alternative splicing, amidation, chloroplast, disease mutation, duplication, fusion protein, glycoprotein, hypothetical protein, membrane, mitochondrion, multigene family, nuclear protein, plasmid, polyprotein, polymorphism, repeat, signal, structural protein, transit peptide, transmembrane) which convey structural or genetic rather than functional information. A protein was represented by a vector in keyword space. The magnitude of each component was the relative frequency of a keyword in the family. The relative frequency of a keyword is the number of times it occurs with a homolog sequence listed in the family (HSSP dataset), divided by the number of sequences in the family. Keyword overlap between two proteins was quantified as the dot product of their keyword vectors.

Implementation

The current prototype of the decision support system is based on Perl scripts [11], using relational tables for storing intermediate results and HTML-viewers for display. Perl scripts parse information from the heterogeneous databases (Figure 2), do simple data manipulations, and keep track of equivalence links between residues in multiple sequence and structure alignments. The atomic contact calculations were programmed in Fortran.

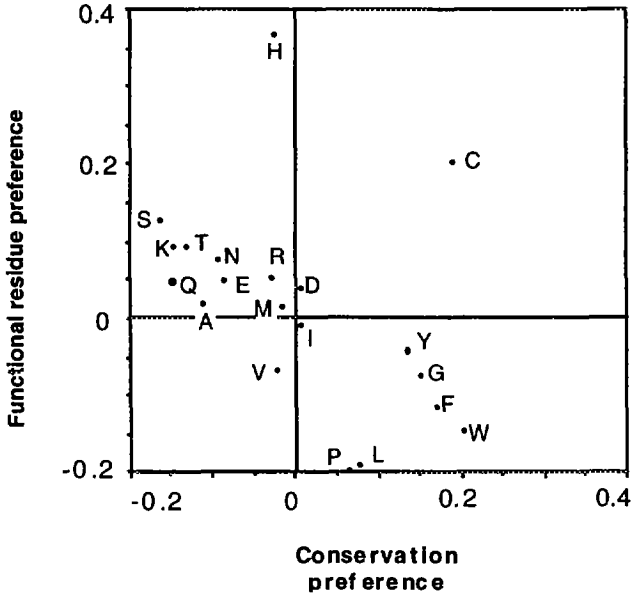
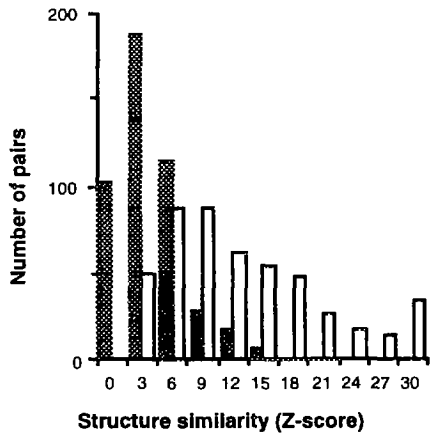
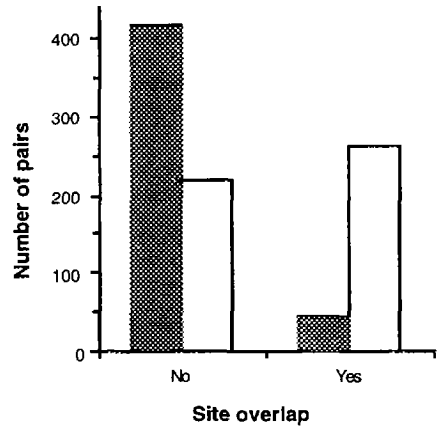
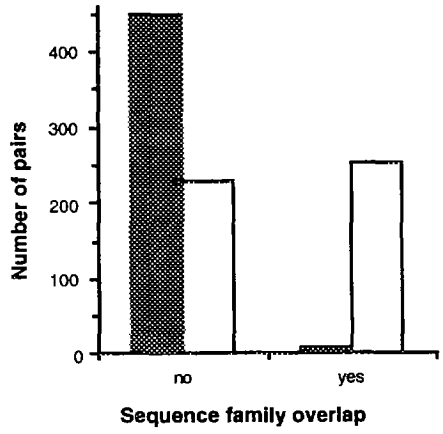


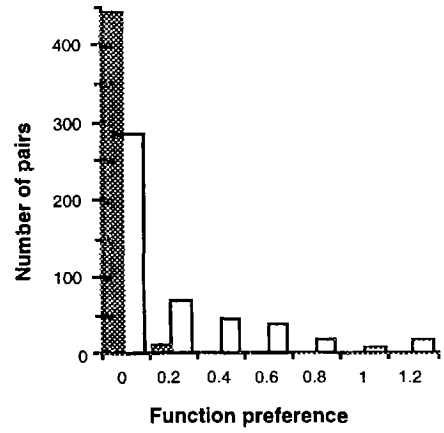
Figure 4: Functional residue preferences by amino acid type

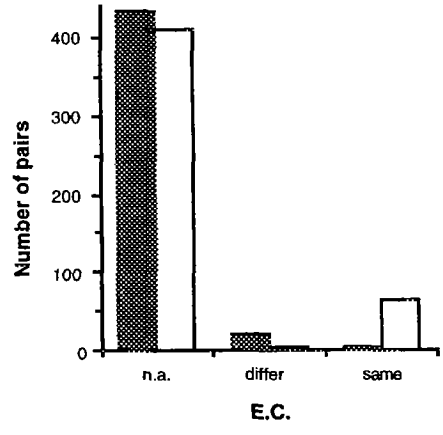
Statistical preferences show that histidine is the favourite functional residue, followed by cysteine and serine. The large aromatic residues and glycine tend to be conserved, but for structural and not functional reasons. In general, the tendency to be conserved (hydrophobic residues) is opposite to the tendency to be functional (polar residues). The exception is cysteine, which has strong preference both for structural conservation (disulphide bridges) and for functional conservation (metal binding).











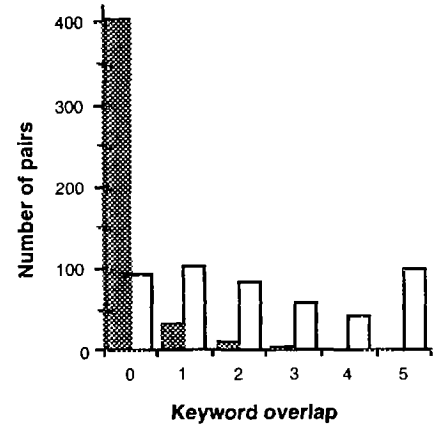


Figure 5. How well do the criteria discriminate?

The histograms show the selectivity and coverage of each of six computational criteria (functionally related pairs are white and unrelated gray).

Results

Figure 5 shows the calibration of the criteria against the test set of 940 pair relationships.

The distribution of structural similarities shows a broad range of overlap between divergently and convergently related proteins. On the one hand, there are popular fold classes such as parallel α/β domains and $(\beta\alpha)_8$ barrels which contain many superfamilies. On the other hand, structural divergence within a superfamily can proceed surprisingly far, as exemplified by glycogen phosphorylase and beta-glucosyltransferase [12] or different types of lysozymes [13].

Sequence family overlap proves to be a remarkably efficient criterion. The test set consists of proteins that have low mutual sequence similarity, but more than half of the pairs of proteins which are related by evolution have bridging sequences between them which are above the HSSP threshold to both test set proteins.

The Enzyme Classification and function preferences are quite selective but have low coverage. The present implementation of criteria involving overlap of functional sites in 3D has a relatively high number of false positives, which should be reduced by stricter comparison, for example requiring a match of the chemical types of the ligand compounds.

A number of false positives with the Swissprot keyword criterion were due to nonspecific keywords such as 'transcription regulation' and 'DNA binding'. Weighting keywords inversely to their frequency in the database would be an obvious remedy to this problem.

No single criterion has both high coverage and high selectivity. Encouragingly, a combined criterion that aims to minimize the sum of the numbers of false positives and false negatives yields reasonable coverage and selectivity (Table I).

Table I: Coverage and selectivity

criterion	coverage[1]	false positives[2]
structure similarity $Z \geq 4.5$	90 %	28 %
sequence family overlap	53 %	3 %
same E.C. class	13 %	6 %
site overlap	54 %	14 %
function preference ≥ 0.16	44 %	8 %
keyword overlap ≥ 0.9	82 %	13 %
bit-score ≥ 3 [3]	87 %	7 %

[1] *Percentage of true-and-positive of all true pairs defined in the test set.*

[2] *Percentage of false-but-positive of all positive pairs identified by the criterion.*

[3] *The bit-score is a linear combination of the six criteria at the top, with weights 1, 2, 1, 1, 1 and 2, respectively.*

Discussion

Conserved functional sites are useful indicators of common evolutionary ancestry between proteins with little sequence similarity but similar 3D structure. Conservation was analyzed both at the level of sequences (columns in multiple alignment) and structural equivalence (3D superimposition). Conceptually, the application of uniform criteria to a large test set leads to a calibration of the weight for different characters in assessing hypotheses of homology (for controversial views regarding one case, see [14] and [15]).

Technically, the prototype decision support system solves mainly syntactic problems associated with reading heterogeneous databases and bookkeeping of homology links. Future improvements must add more semantic understanding to the system, preferably by unsupervised machine learning techniques. For example, there is a small number of false positive homologs at the bottom of HSSP datasets, which should be detected and eliminated. Information theoretical approaches could be used to investigate whether a proposed family is better described by separate models for subfamilies. Furthermore, most larger proteins are composed of functionally distinct domains, which causes problems in associating keywords to entire sequences.

The present analysis centred on identifying common conserved biochemical functions. More detailed analysis of superfamilies and multiple alignments can account for adaptations of specificity in different subfamilies [16]. Let us take two examples. The active site of plant endochitinase was identified by structural homology to lysozymes from animals and phage [13]. In this case, there is hardly any overall sequence similarity, yet both are functionally related enzymes. On the other hand, alpha-lactalbumin has recently diverged from mammalian lysozymes which is evident from sequence identities around 30 % but the active site residues are not conserved.

The rapid increase in structure and sequence data will allow testing the criteria developed here on independent new data. The long-term goal is to refine the criteria introduced here for an expert system that would automatically resolve functional protein superfamilies in the Protein Data Bank. For example, a neural network could be trained on bit patterns (Figure 4) or on real-valued criteria (Figure 3). Automatic classification would not only give molecular biologists an overview of the evolution of protein families, but would, as a byproduct, also provide the fold recognition ("threading") [17] community with a clean test set of physically convergent protein structures.

References

1. Overington J.P., Zhu Z.Y., Sali A., Johnson, M.S., Sowdhamini, R., Louie, G.V. & 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.

2. Sander C. & Schneider R. (1991) Homology-derived secondary structure of proteins and the structural meaning of sequence homology. Proteins 9:56-68.

3. Krogh A., Brown M., Mian I.S., Sjölander K. & Haussler D. (1994) Hidden Markov models in computational biology. Applications to protein modelling. J. Mol. Biol. 235:1501-1531.

4. Holm L. & Sander C. (1997) Mapping the protein universe. Science 273:595-602.

5. Holm L. & Sander C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. Proteins, in press.

6. 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.

7. Orengo, C.A., Flores, T.P, Taylor, W.R. & Thornton, J.M. (1993). Identification and classification of protein fold families. Protein Eng. 6, 485-500.

8. Bairoch A. (1993) The ENZYME data bank. Nucleic Acids Res 21:3155-3156.

9. Bairoch A. (1993) Prosite. Nucleic Acids Res 21:3097-3103.

10. Hooft R.W.W., Sander C., Scharf M. & Vriend G. PDBFINDER database, unpublished.

11. Wall L., Christiansen T. & Schwartz R.L. (1996) Programming Perl. O'Reilly & Associates, Inc., Sebastopol, California.

12. Holm L. & Sander C. (1995) Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme. EMBO J 14:1287-1293.

13. Holm L. & Sander C. (1997) New structure - novel fold? Structure 5:165-171.

14. Pastore A. & Lesk A.M. (1990) Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. Proteins 8:133-155.

15. Holm L. & Sander C. (1993) Structural alignment of globins, phycocyanins and colicin A. FEBS Lett 315:301-306.

16. Casari G., Sander C. & Valencia A. (1995) A method to predict functional residues in proteins. Nat Struct Biol 2:171-178.

17. Ouzounis C., Sander C., Scharf M. & Schneider R. (1993) Prediction of protein structure by evaluation of sequence-structures fitness. Aligning sequences to contact profiles derived from three-dimensional structures. J. Mol. Biol. 232:805-825.

18. Sideraki V., Mohamedali K.A., Wilson D.K., Chang, Z., Kellems R.E., Quiocho F.A. & Rudoplh F.B. (1996) Probing the role of two conserved active site aspartates in mouse adenosine deaminase. Biochemistry 35:7862-7872.

19. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyers, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535-542.

20. Bairoch A. & Apweiler R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res 24:21-25.

21. Holm L. & Sander C. (1996) The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res 24:206-210.