Identification of homology in protein structure classification

Sabine Dietmann and Liisa Holm

Structural Genomics Group, EMBL-EBI, Cambridge CB10 1SD, UK.

Structural biology and structural genomics are expected to produce many three-dimensional protein structures in the near future. Each new structure raises questions about its function and evolution. Correct functional and evolutionary classification of a new structure is difficult for distantly related proteins and error-prone using simple statistical scores based on sequence or structure similarity. Here we present an accurate numerical method for the identification of evolutionary relationships (homology). The method is based on the principle that natural selection maintains structural and functional continuity within a diverging protein family. The problem of different rates of structural divergence between different families is solved by first using structural similarities to produce a global map of folds in protein space and then further subdividing fold neighborhoods into superfamilies based on functional similarities. In a validation test against a classification by human experts (SCOP), 77% of homologous pairs were identified with 92% reliability. The method is fully automated, allowing fast, self-consistent and complete classification of large numbers of protein structures. In particular, the discrimination between analogy and homology of close structural neighbors will lead to functional predictions while avoiding overprediction.

The organization of protein structure data in terms of evolutionary relationships is a field where subjective classification has prevailed and is considered by some to be impossible/inappropriate to address using numerical methods. A key difficulty is that the boundary between apparent homologs and analogs (convergent folds) corresponds to a very broad range of values in sequence, structural and functional similarity in different protein families. Finding attributes that would allow classification with both high coverage and a low error rate seems difficult[1–5]. Here we address homolog/analog discrimination using an underlying evolutionary model based on the concept of a protein space and natural selection[6]. The history of a superfamily starts from a common ancestor, and evolution occurs by gradual diffusion from this point source. The descendants of the common ancestor remain in the (structural) neighborhood of each other and inherit a similar spectrum of functional properties from the common ancestor. Thus, superfamilies can be delineated as continuous neighborhoods in the map of protein space (Fig. 1).

We have derived a representation of protein space by structure comparison. For computational convenience, we used hierarchical clustering (average linkage of DALI[7–9] Z-scores) to derive a fold dendrogram from structural similarities and limited the search for candidate superfamilies to branches of this tree. The weighted sum of intramolecular distance difference matrices from DALI captures the strong conservation of functionally constrained motifs — for example, in the surroundings of an active site — but simultaneously allows for structural deviations between more distant parts of the molecules[7]. We demonstrate, by comparison to SCOP[10], that DALI apparently successfully integrates over important aspects of the process of divergent structural evolution, preserving evolutionary neighbor relationships in protein space. In the resulting map of protein space, fold/superfamily boundaries are identified based on functional and sequence similarity clues of evolutionary relationships (such as shared sequence neighbors, conserved sites and functional annotations). A neural network is trained to sum up these heterogeneous inputs into a single number. This allows us to score any candidate superfamily in terms of functional similarities within the set and to select a 'best' partition of all proteins into homologous superfamilies in the light of available evidence.

Validating the map of fold space

Our key assumption is that members of a superfamily occupy monophyletic branches in the DALI fold dendrogram. To validate this assumption, the topology (that is, branching order) of the DALI fold dendrogram[9] was compared to the SCOP classification of Alexey Murzin[10]. SCOP classifies protein domains into a hierarchy of similarity levels (from the most general to the most specific): class, fold, superfamily, family and protein. The clustering score of Przytycka *et al.*[11] measures the extent to which members of a SCOP superfamily are grouped together in another hierarchical classification (here, the DALI fold dendrogram). For a given pair of structures belonging to the same SCOP superfamily, one first finds the smallest branch of the DALI fold dendrogram that contains both structures. One then examines the SCOP classification of all structures at the leaves of this branch. The clustering score is defined as the fraction of structures belonging to the same SCOP superfamily as the initial pair. The average clustering score of SCOP superfamilies in the DALI dendrogram was 0.76 evaluated on a set of 2,141 representative domains in SCOP classes 1–4. Strict monophyly (a clustering score of 1) was observed for 190 out of 330 (58%) of SCOP superfamilies, excluding singletons (Table 1). A clustering score <1 means that a SCOP superfamily is split (polyphyletic) in the DALI dendrogram. For example, SCOP superfamily 3.32.1, 'P-loop containing nucleotide triphosphate hydrolases', is divided on the fourth SCOP level based on β-sheet topologies, although topology is usually differentiated at the second SCOP level (the fourth level groups proteins with clear sequence similarity). In the fold dendrogram, most members of SCOP superfamily 3.32.1 occupy three separate monophyletic branches, leading to a clustering score of 0.32. The average clustering score was below 0.5 for 96 out of 140 polyphyletic SCOP superfamilies, suggesting that the decision to merge structures into superfamilies in SCOP was based on evidence other than structural similarity extending over an entire globular domain fold.

Pairwise classification accuracy

The boundaries of superfamilies in the fold dendrogram are identified based on functional similarity (Fig. 1*b*). A neural network was trained against the fold-to-superfamily transition in SCOP, using different functional attributes (Table 2) as input. The output of the neural network ranges from zero (analogous pairs) to one (homologous pairs) and is defined as our measure of functional similarity, ϕ. Although many homologous pairs are more functionally similar than most analogous pairs, the distributions of functional similarity (ϕ) values of homologs and analogs are broad and overlapping. For example, there is a strong (ϕ = 1.00) but false similarity between an α/β-hydrolase (cutinase 1cex, SCOP 3.17.8) and 2-hydroxyisocaproate dehydrogenase (1dxy, SCOP 3.17.11) due to a structurally equivalent pair of conserved His residues in the pentapeptide PH(I/L)AY, which is part of the catalytic triad and the NAD-binding site,

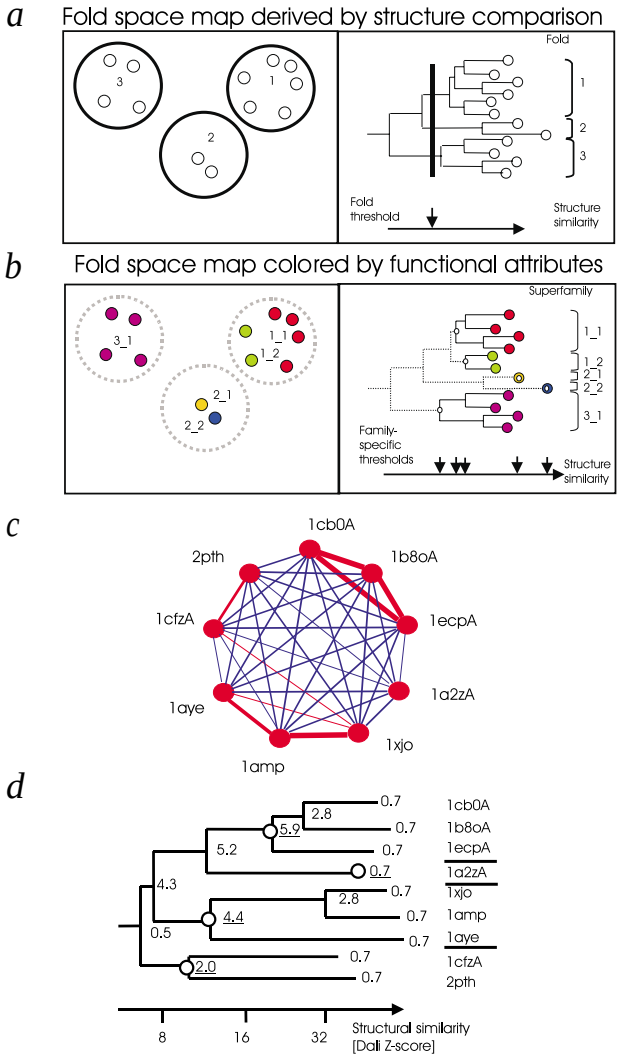

Fig. 1 Partitioning protein space into homologous families. ***a***, All-against-all structure alignment by DALI reveals a hierarchical organization of fold space. The method is sensitive enough to recognize similarities of general folding pattern — for example, the β-sandwich topology of superoxide dismutase and immunoglobulin domains — and selective enough to give higher scores to pairs of structures with more closely superimposable Cα traces — for example, any two globins score higher than any globin–phycocyanin pair. Structure similarity alone yields an operational definition of 'folds'. The thick circles denoting folds (left) are defined using a uniform radius for clusters of structural neighbors. The vertical bar (right) denotes cutting the fold dendrogram at a uniform value of structural similarity. However, the level of structural similarity, or degree of structural divergence, varies between different families, and we need other criteria to delineate superfamilies. ***b***, Divergent evolution from a common ancestor retains not only the fold but also many functional features. This means that homologs remain in a structural neighborhood and can be delineated by similar functional attributes (marked here by similar color) in the map of fold space. Functional convergence (from independent evolutionary origins) would appear as blotches of similar color in disconnected regions of the map of fold space and in disjoint branches of the fold dendrogram. Partitioning the fold dendrogram in terms of functional similarities yields family-specific thresholds in terms of structural similarity (nodes that partition the fold dendrogram into functionally conserved superfamilies are circled on the right). This combination of structural and functional similarity measures results in an automatically generated hierarchical classification m_n at the fold (m) and superfamily (n) levels. ***c***, The principles are illustrated on a branch of the fold dendrogram consisting of aminopeptidases (1xjo and 1amp), carboxypeptidase (1aye), purine nucleoside phosphorylases (1b8oA, 1cb0A and 1ecpA), pyrrolidone carboxyl peptidase (1a2zA), peptidyl–tRNA hydrolase (2pth) and hydrogenase maturating endopeptidase (1cfzA). The functional similarity between all pairs of structures is evaluated using a neural network with output ϕ in the range 0 (analogous)–1 (homologous) — for example, ϕ(1cb0A, 1b8oA) = 0.91, ϕ(1amp, 1aye) = 0.74, ϕ(1cfzA, 2pth) = 0.59, ϕ(1xjo, 1amp) = 0.30 and ϕ(1a2zA, 2pth) = 0.13. Here, line thickness indicates the magnitude of the term $\phi(i,j) - \theta$ (Eq. 1; see Methods) with color-coding for positive (red) or negative (blue) values. The threshold parameter θ was arbitrarily set to 0.30 in this numerical example. ***d***, The protein set is partitioned into superfamilies in the context of the fold dendrogram. Node scores s(C) are computed for each node (Eq. 1), with θ = 0.30. For example, each structure is homologous to itself; therefore, leaf nodes get a score s(leaf) = 1.00 – θ = 0.70, whereas s(1cfzA, 2pth) = (1.00 + 1.00 + (2 × 0.59)) / 4 – θ = 1.98. The optimal partition (circled nodes) maximizes the sum of node scores over selected nodes (underlined scores). This optimal partition is stable for threshold values 0.09 < θ < 0.53.

respectively. On the other hand, there is no significant functional similarity (ϕ = 0.07) between two fibronectin-like domains (SCOP 2.1.2) from interleukin-4 receptor (1iarB) and human fibronectin (1fnhA). In both cases, assessing the pairwise functional similarities in the context of the fold dendrogram leads to superfamily classification in agreement with SCOP. Overall, the global partitioning strategy yields a classification accuracy that is close to the upper limit determined by the monophyly of SCOP superfamilies in our fold dendrogram. Global partitioning leads to the identification of 77% of homologous pairs from SCOP with 92% reliability (Fig. 2). For comparison, a Markov transition model[2] of structural evolution was reported to recognize 48% of pairs from the same SCOP superfamily with 80% reliability (in a similar but not identical test set).

Calibrating the superfamily threshold parameter θ

We applied our automatic method to classify all structures in the Protein Data Bank (PDB) using a neural network trained on

Table 1 Mapping between automatically (DALI) and manually (SCOP) defined superfamilies

Query set Q	Total number of query sets Q	Reference sets R	$Q \subseteq R$[1]	$Q \not\subseteq R$
DALI singleton superfamily	738	SCOP superfamilies	738 (336)	–
DALI multimember superfamily	324	SCOP superfamilies	299 (108)	25[2]
All DALI superfamilies	1,062	SCOP superfamilies	1,037 (444)	25[2]
SCOP singleton superfamily	370	DALI superfamilies	370 (336)	–
SCOP monophyletic multimember superfamily	190	DALI superfamilies	132 (108)	58[3]
SCOP polyphyletic superfamily	140	DALI superfamilies	11 (0)	129[3]
All SCOP superfamilies	700	DALI superfamilies	513 (444)	187[3]

[1]The number in parentheses is the number of cases where Q = R.
[2]Overunification under automatic classification using θ = 0.33.
[3]Oversplitting under automatic classification using θ = 0.33.

SCOP 1.53. The θ threshold defines the level of functional similarity that constitutes 'compelling evidence' for merging structural neighbors into a superfamily (Eq. 1; see Methods). Partitions derived at different values of θ produce a hierarchical classification, where a functionally diverse superfamily splits into subfamilies at higher values of θ — for example, retroviral proteases *versus* classical aspartic proteases. In common practice, however, there is one particular value of θ which will be of most interest: the threshold value that most closely mirrors the superfamily/fold boundary defined in SCOP. A broad range of θ values yields very similar partitions. The sum of overunification and oversplitting errors is 215 at θ = 0.30, 212 at θ = 0.33 (Table 1), 216 at θ = 0.40, 219 at θ = 0.60 and 247 at θ = 0.90. Oversplitting means that the automatic classification fails to group all members of a SCOP superfamily in one DALI superfamily. Overunification means that a DALI superfamily contains members from two or more SCOP superfamilies.

The reliability of homology detection is very high: excluding singletons, all members of 299 out of 324 DALI superfamilies (92%) are members of one SCOP superfamily (Table 1). Borderline functional similarities can lead to overunification (example in Table 2). Unaware that 1cfzA is a metallopeptidase while 2pth has been shown to be metal-independent[12], our automatic classifier unifies these two enzymes into one superfamily but SCOP does not. In general, DALI defines superfamilies more conservatively than SCOP. Remarkably, however, a majority (444 / 700 = 63%) of SCOP superfamilies are recovered exactly — that is, an identical set of domains is grouped into one superfamily under our automated classification as in SCOP. Almost one-third of the monophyletic SCOP superfamilies suffer from oversplitting under the automatic classification (Table 1). Typically only one or a few members of a large SCOP superfamily remain unmerged at the outer fringe of a branch in the DALI fold dendrogram. Inspection indicates that the lack of evidence (undefined functional features) is a limiting factor. New information may support further mergers of superfamilies in the future. Adding more functional attributes to the feature vectors will also be technically straightforward. In particular, data from functional genomics provide new ways of quantifying functional similarity — for example, in terms of the similarity of transcriptional profiles of two genes in a large number of experimental conditions.

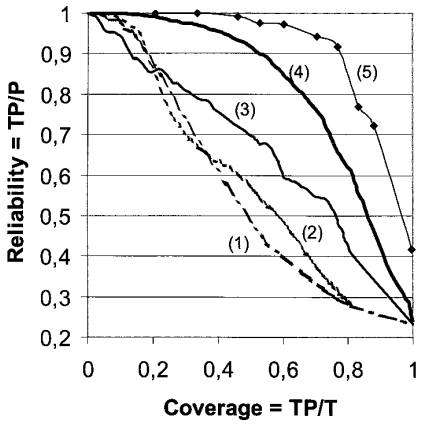

Fig. 2 Jack-knife evaluation of the prediction accuracy of the neural network. Neural networks were trained to weigh the relative contributions of the heterogeneous inputs (Table 2) in order to discriminate between related and unrelated pairs of structures as defined by the SCOP fold/superfamily classification. The training set consisted of 11,907 unrelated pairs (same SCOP fold but different superfamily) and 3,635 related pairs (same SCOP superfamily) from a representative set of single-domain PDB structures. N = 15,542 training runs were made, each using (N – 1) examples for training and testing the example left out. The prediction accuracy is summarized by coverage and reliability. The tested examples are rank-ordered according to the neural network output, with the strongest predictions at the top. Let us consider the P (positive) highest scoring examples. In this set, reliability is defined as TP / P, where TP (true positive) is the number of positive examples that are correctly identified as related. Coverage is defined as TP / T, where T (true) is the number of related pairs in the whole test set. P varies from 1 to N along each curve. A perfect classifier would rank all true pairs above the first false pair, driving the curve to the top-right corner of the graph. Ranking by (1) sequence identity alone; (2) structure similarity alone; (3) keyword similarity alone; and (4) neural network output. The dots on curve (5) correspond to optimal partitions obtained at different values of θ in Eq. 1 (see Methods) — that is, using the context in the fold dendrogram as a noise filter on neural network predictions. Filtering based on the DALI fold dendrogram was superior to clustering based on neural network outputs (data not shown).

Classifying proteins of unknown function

Functional annotation from literature is unlikely to be available for most proteins targeted in structural genomics. We simulated the classification of hypothetical proteins using only the first five features in Table 2 for a test protein. Each simulation perturbed the neural network inputs of all pairs involving the test protein, and the automatic procedure for superfamily classification was repeated. Classification of the PDB using θ = 0.33 and the full feature description yielded 456 superfamilies with at least two members and comprising 1,900 representative domains. The sparser feature vectors affected the result of unification in only 105 out of 1,900 simulations (6%). For example, restriction

Table 2 Feature vector for the (2pth, 1cfzA) pair

Feature	Evidence	Value
Z-score	DALI structural alignment	10.2
Sequence identity	DALI structural alignment	14%
Sequence family overlap	No common BLAST sequence neighbors	'no'
Identical conserved residues in contact to ligand	No ligand information for 2pth	'unknown'
Functional preference	2pth G7 & G23 and 1cfzA G7 & G19 are structurally equivalent, conserved and in spatial proximity[1]	0.75
Keyword similarity	2pth: 1 / 13 'sporulation', 12 / 13 'hydrolase'; 1cfzA: 12 / 12 'protease'and 12 / 12 'hydrolase'	0.93
Common E.C. numbers	2pth: E.C.3.1.1.29, 1cfzA: E.C.3.4.–.–	1
Overlap of annotated sites	2pth G100D is a temperature-sensitive mutation and structurally equivalent to 1cfzA G72; both residues are conserved	'yes'

[1]2pth N10 / 1cfzA N10 are also conserved, yielding a functional preference of 0.38 but only the highest scoring cluster is used.

Table 3 Classification of recently solved structural genomics targets

Protein	PDB	Authors' functional classification	Evolutionary classification[1] (this work)
Mth1615	1eijA	DNA-binding, putative transcription factor[2]	New fold
Mth1184	1gh9A	Putative metal-binding protein[2]	New fold
Hi1434	1dbuA	Putative nucleotide or oligonucleotide binding domain	New fold
E. coli YrdC	1hruA	Putative dsRNA binding protein	New fold
Mth538	1eiwA	Unknown[2]	New family
Mth1175	1eo1A	Unknown[2]	New family
Clostridium CipC	1ehxA	Scaffolding protein and the first prokaryotic member of the I set of the immunoglobulin superfamily[3]	New family
B. subtilis maf	1ex2A	Nucleotide binding, putative NTPase	Same superfamily as with Mj0226 pyrophosphatase (1b87A)
Mth152	1ejeA	FMN- and nickel-binding protein[2]	Same superfamily as ferric reductase (1i0rA)
Mj0541	1f9aA	NMN adenylyltransferase	Same superfamily as two nucleotidylyl transferases (1b6tA, 1cozA)
Mouse doppel	1i17A	Paralog of the cellular prion protein but with a distinct physiological role and distinct pathology	Same superfamily as prion proteins (1b10A, 1qlzA)
Yeast Ure2	1g6wa	Prion protein, lacks GST activity[4]	Same superfamily as 6 glutathione S-transferases (GSTs)
Mj0882	1dusA	Unpublished	Same superfamily as 16 methyltransferases
E. coli CyaY	1ew4A	Belongs to the frataxin family which is linked to the neurodegenerative disease Friedreich ataxia	Same superfamily as frataxin (1dlxA)
Mth649	1i81A	Belongs to the SnRNP Sm protein family; 37% sequence identity to 1d3bA	Same superfamily as small nuclear ribonucleo-proteins (SnRNPs)

[1]Superfamilies defined at $\theta = 0.33$. Folds defined by cutting the fold dendrogram at $Z = 2$.
[2]Ref. 13.
[3]Ref. 19.
[4]Ref. 20.

endonucleases are typical of a superfamily that in the automatic classification is held together by common keywords. All other features are weak (such as low average Z-score of 4.7, low sequence identities, small and nonoverlapping sequence families, and no ligands in the crystal structures) so that simulations on four out of six members in the endonuclease family resulted in a family break-up. However, the simulations suggest that manual functional annotation requiring human expertise (keywords, enzyme classification numbers and site annotation) is usually redundant with functional attributes extracted automatically from sequence and structure conservation. Moreover, the strategy of averaging over candidate superfamilies, defined by the fold dendrogram, confers general robustness to the evolutionary classifier with respect to a lack of functional information for one or a few proteins in a selected set. This is true as long as evidence for homology in neighboring pairs is strong.

Automated classification uses a limited set of generic functional attributes to determine superfamily membership. Once superfamily membership is established, detailed functional predictions can be based on judicious carry-over of the complete functional description from experimentally characterized members within a superfamily. Of a digest of 15 recently solved structural genomics targets, four were without structural neighbors (Table 3). Three structures had structural neighbors but insufficient functional similarity for grouping them into a superfamily. Biochemical experiments to test for functional similarity to the closest structural neighbors of *Methanobacterium thermoautotrophicum* proteins Mth538 and Mth175 were inconclusive[13], indicating that classification into a new family is probably a correct decision. Eight structures joined existing and emerging superfamilies, leading to experimentally testable hypotheses about biochemical function. In particular, we predict an NADP binding site in the 'FMN- and nickel binding protein' Mth152 (ref. 13) based on its remote homology to ferric reductase from *Archaeoglobus fulgidus* and the presence of a putative His marker for FMN:NADP oxidoreductases (His 126 in ferric reductase[14]/His 144 in Mth152). Also, a sulphate ion in Mth152 is bound in a structurally equivalent position to the diphosphates of NADP in the crystal structure of ferric reductase.

Conclusion

We have proposed a numerical taxonomy leading to robust automatic evolutionary classification of protein structures. The topology of protein space is probed using structural similarity. Searching for clusters of structural neighbors where the members consistently share many functional attributes leads to an optimal partitioning of protein space. This clustering corresponds well to the analog/homolog boundaries drawn by biologists, with applications in the generation of functional hypotheses in nonhypothesis-driven structural genomics efforts.

Methods

Data sets. All data sets, feature tables, neural network training and test sets, and results are available electronically from *http://www.ebi.ac.uk/dali/domain/3.1beta/*

Feature vectors. The input to the neural network is a feature vector (Table 2). The features are derived from structural conservation, sequence conservation and sequence annotation, exploiting sequence alignments and structure superimposition to transfer position-specific information from sequence-homologs to the query structures. 'Keyword similarity' is the dot product of vectors representing the frequency of occurrence of SWISSPROT keywords in sequence homologs of either query structure. Noninformative keywords such as '3D-structure' are excluded. 'Functional prefer-

ence' is defined per amino acid type and is summed over all residues in a three-dimenional cluster of conserved residues. Feature computation is described at depth in ref. 4.

Data normalization. The DALI Z-score, functional preference and keyword similarity are open scales of similarity and were linearly rescaled to zero mean and a standard deviation of one. The problem of missing values in specific components of the input vectors is severe in our case. For example, the similarity of enzyme classification codes is a strong feature but is defined only for 20% of the pair examples in the training set. There are various heuristics in the literature for dealing with missing data in classification problems. Here the missing values for the enzyme classification codes were filled with the mean value for all known pairs. Similarly, ligand information is unavailable or incomplete for many structures in the PDB. The feature 'identical conserved residues in contact to a ligand' was, therefore, encoded as 'yes', 'no' or 'unknown'.

Neural networks. Layered feed-forward neural networks[15–17] were optimized by a back-propagation algorithm[18]. Networks of widely different architectures were tested using one layer of hidden units, where the number of hidden units was initially set to 2 × (number of input units) + 1 and reduced until an optimum was reached. The final architecture had nine input units, 10 units in the hidden layer and one output unit leading to a total of (9 × 10) + (10 × 1) = 100 adjustable weights. All weights in the neural network were randomly initialized to a value from the interval [–1, 1] prior to network training. The early stopping technique was used to prevent overfitting of the free parameters of the network[15]. During training, the error function (difference between desired and obtained outputs) of the training set falls continuously until it converges on some value. An independent validation set consists of examples not in the training set. The error function of the validation set is usually higher than that of the training set; it falls initially but then rises again as overfitting sets in. Training of the neural network is stopped at the minimum.

Optimal partitioning of protein space. Our goal is to partition the fold dendrogram so that the observed functional similarities (strong neural network predictions) concentrated as much as possible within the selected clusters (branches, superfamilies). The objective function should balance the size of the clusters against their 'quality' as a homologous set. This is achieved with a sum-of-pairs formulation similar to that used to delineate the common structural core in distance matrix alignment by DALI[7]. The neural network outputs are thresholded so that there is a gain from including 'similar' pairs and a penalty for including 'dissimilar' pairs in a cluster. More formally, the optimal partitioning of the fold dendrogram results in a set of superfamily-ancestor nodes (C), over which the sum of node scores s(C) is maximal:

$$s(C) = \Sigma_{i\in C} \Sigma_{j\in C} (\phi(i,j) - \theta) = N_c^2 (<\phi>_C - \theta) \quad (1)$$

The node score s(C) is summed over all pairs of leaves (structures i,j) under a superfamily ancestor node C. $\phi(i,j)$ is the output from the neural network for a protein pair (i,j), $<\phi>_C = (\Sigma\Sigma\ \phi(i,j)) / N_c^2$ is the cluster average, N_c is the number of members in the cluster and θ is the threshold parameter. A branch of the fold dendrogram has a high score if the members have a high average neural network prediction for being homologous.

The merging of two branches of the fold dendrogram is favored if their average connection strength is above θ. Algebraically, if a cluster C consists of two subsets (branches) A and B, then s(C) = s(A) + s(B) + 2s(AB), where s(AB) denotes the sum over pairs where one structure belongs to subset A and the other belongs to subset B. A condition for merging A and B is s(C) > s(A) + s(B), which clearly holds only if s(AB) > 0. A straightforward tree traversal algorithm yields the optimal partition, where no subdivision or merger increases the sum of node scores s over the selected set of nodes. Fig. 1*c*,*d* gives a worked example of the partitioning procedure.

Received 23 March, 2001; accepted 27 September, 2001.

1. Wood, T.C. & Pearson, W.R. *J. Mol. Biol.* **291**, 977–995 (1999).
2. Kawabata, T. & Nishikawa, K. *Proteins* **41**, 108–122 (2000).
3. Matsuo, Y. & Bryant, S.H. *Proteins* **35**, 70–79 (1999).
4. Holm, L. & Sander, C. *ISMB* **5**, 140–146 (1997).
5. Russell, R.B., Saqi, M.A., Sayle, R.A., Bates, P.A. & Sternberg, M.J. *J. Mol. Biol.* **269**, 423–439 (1997).
6. Smith, J.M. *Nature* **225**, 563–564 (1970).
7. Holm, L. & Sander, C. *Science* **273**, 595–602 (1996).
8. Holm, L. & Sander, C. *Proteins* **333**, 88–96 (1998).
9. Dietmann, S. *et al. Nucleic Acids Res.* **29**, 55–57 (2001).
10. LoConte, L. *et al. Nucleic Acids Res.* **28**, 257–259 (2000).
11. Przytycka, T., Aurora, R. & Rose, G.D. *Nature Struct. Biol.* **6**, 672–682 (1999).
12. Fritsche, E., Paschos, A., Beisel, H.-G., Bock, A. & Huber, R. *J. Mol. Biol.* **288**, 989–998 (1999).
13. Christendat, D. *et al. Nature Struct. Biol.* **7**, 903–909 (2000).
14. Mosbah, A. *et al. J. Mol. Biol.* **304**, 201–217 (2000).
15. Bishop, C.M. *Neural networks for pattern recognition* (Oxford University Press, Oxford;1995).
16. Baldi, P. & Brunak S. *Bioinformatics: the machine learning approach* (MIT Press, London;1998).
17. Theodoridis, S. & Koutroumbas, K. *Pattern recognition* (Academic Press, San Diego; 1999).
18. Fahlmann, S.E. & Lebiere, C. In *Advances in neural information processing systems* (ed. Touretzky, D.) 524–532 (Morgan Kaufmann, San Mateo;1990).
19. Chiu, H.-J., Johnson, E., Schroder, I. & Rees, D.C. *Structure* **9**, 311–319 (2001).
20. Bousset, L., Belrhali, H., Janin, J., Melki, R. & Morera, S. *Structure* **9**, 39–46 (2001).