Identification of Homologous Core Structures

Yo Matsuo and Stephen H. Bryant*
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

ABSTRACT **Using a large database of protein structure–structure alignments, we test a new method for distinguishing homologous and "analogous" structural neighbors. The homologous neighbors included in the test set show no detectable sequence similarity, but they may be well superimposed and show functional similarity or other evidence of evolutionary relationship. Analogous neighbors also show no sequence similarity and may be well superimposed, but they have different functions and their structural similarity may be the result of convergent evolution. Confirming results of other analyses, we find that remote homologs and analogs are not well distinguished by measures of pairwise structural similarity, including the percentage of identical residues and root-mean-square (RMS) superposition residual. We show, however, that with structure–structure alignments of analogous neighbors rarely superimpose the particular substructure that is shared among homologous neighbors. We call this characteristic substructure the homologous core structure (HCS), and we show that a cross-validated test for presence of the HCS correctly identifies 75% of remote homologs with a false-positive rate of 16% analogs, significantly better than discrimination by RMS or other measures of pairwise similarity. The HCS describes conservation of spatial structure within a protein family in much the way that a sequence motif describes sequence conservation. We suggest that it may be used in the same way, to identify homologous neighbors at greater evolutionary distance than is possible by pairwise comparison. Proteins 1999;35:70–79.**

Key words: protein structure; comparative analysis; molecular evolution

INTRODUCTION

Experimental determination of protein structures has shown that three-dimensional structure is highly conserved during molecular evolution. Similarity of structure can therefore suggest that two proteins are homologous, even when their sequences are not detectably similar, and indeed there are now many examples of evolutionary relationships that have been recognized in exactly this way. However, not all proteins with some similarity of three-dimensional structure are necessarily homologous, the products of descent from a common ancestral gene. It is possible, for example, that certain polypeptide folds are strongly preferred for physical reasons, and thus likely to have been discovered independently in different evolutionary lineages.[1–4] These analogous structures, as they have been called, may be the result of convergent evolution.

The distinction between homologous and analogous structures has long been a question of academic interest.[5,6] More recently, with the growth of the known-structure database, it has also become a question of practical importance for molecular biologists who use collections of comparative analysis results. Automated methods for structure–structure comparison now list the structural neighbors of each chain or domain in the Protein Data Bank,[7,8] often very many of them.[9–12] But which of these neighbors are homologs, from which one might infer biological function or properties, and which are analogs, where no such inference is possible? Is there a way to make this distinction in the difficult cases, when sequences are not detectably similar, and to identify the structural neighbors that are most likely to be remote homologs?

Methods for automatic discrimination of homlogues vs analogs have been considered recently by several investigators. Russell and Barton[13] examined various measures of pairwise structural similarity, including root mean square (RMS) superposition residual and the proportion of conserved side-chain contacts. They found that these quantities follow similar distributions for homologous and analogous pairs, and they suggested that homologs cannot be distinguished on this basis. Russell et al.[14] found that the percentage of identical residues in structure–structure alignments was on average somewhat greater for a set of remote homologs, as compared to analogs, but distributions nonetheless overlapped substantially and discrimination was poor. Holm and Sander[15] similarly report poor discrimination on the basis of DALI score, a measure of pairwise structural similarity, and they suggest use of keyword overlap and other measures intended to detect functional similarity. On the whole, these investigations have suggested that measures of pairwise structural similarity cannot well distinguish remote homologs and analogs.

In the present work we ask whether homologs and analogs may be distinguished on the basis of information

Grant sponsor: National Institutes of Health Intramural Research Program.
*Correspondence to: Dr. Stephen H. Bryant, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894. E-mail: bryant@ncbi.nlm.nih.gov
Received 28 April 1998; Accepted 7 December 1998

derived from multiple structure–structure alignments with previously identified homologs. We define the homologous core structure (HCS) of a protein as the subset of Cα coordinates whose spatial locations are conserved across structure–structure alignments with previously identified homologs. We then examine structure–structure alignments of that protein with other structural neighbors, including other homologs and analogs, and we ask whether these alignments involve the HCS or instead an alternative or partially overlapping substructure. This approach to homologue vs analogue classification is analogous to definition of a sequence motif from multiple sequence alignments,[16] but it is based on conservation of the three-dimensional coordinates of a characteristic substructure rather than conservation of its sequence.

Considering a large set of sequence-dissimilar structural neighbors, we find that homologs and analogs are not well distinguished by RMS or other measures of pairwise structure–structure similarity, in agreement with earilier studies. We find, however, that analogous neighbors rarely conserve the HCS, the core structure characteristic of a given protein's evolutionary family. For this test set we show that homologue vs analogue discrimination on the basis of HCS overlap is clearly superior to discrimination by RMS, the percentage of identical residues, or structure–structure alignment length as a fraction of domain length. We therefore suggest that remote homologs in the lists of structural neighbors generated by automated structure–structure comparison methods may be better identified by conservation of the HCS, as compared to measures of pairwise structural similarity, and that it may be useful to sort lists of structural neighbors on this basis.

METHODS

Sequence-Dissimilar Domain Structures

Three-dimensional structure data were taken from the PDB (Protein Data Bank), a set of 7,150 protein structures available as of May 27, 1998, containing 13,162 polypeptide chains. Definitions of domains within each chain were taken from MMDB, the three dimensional structure database distributed with Entrez (http://www.ncbi.nlm.nih.gov/Structure).[17–19] These definitions are based on quantitative compactness criteria.[10] They yielded a total of 18,993 domains, including chains not subdivided into smaller units. We excluded domains with fewer than 80 residues, reducing the total to 14,528 domains.

The amino acid sequences of each domain were compared to one another using the BLAST algorithm, a sensitive indicator of pairwise sequence similarity.[20] Calculations were performed with a database size parameter[21] of 5×10^5, using routines from the NCBI toolkit (ftp://ncbi.nlm.nih.gov/toolbox/). Domains were then clustered into a total of 1,611 sequence-similar groups by a neighbor joining procedure, in which a domain was merged into a preexisting group if it showed a BLAST P value of 10^{-7} or less with any member of that group. A threshold of 10^{-7} was necessary in single-linkage clustering to avoid any chance of false merging of dissimilar groups. A similar threshold was recently suggested as optimal for identifying significant similarity among PDB-derived sequences.[22]

Representative structures for each of the 1,611 sequence-similar groups were chosen automatically according to the precision and completeness of structural data and a requirement that domain definitions agree with those of SCOP (see below). Within each group domains were ranked according to the percentage of unassigned residue types, the percentage of residues with incomplete or missing coordinates, crystallographic resolution, the number of chains in the structure, the number and types of nonpolymer ligands, and the total number of residues. The highest ranked member whose domain definition agreed with SCOP was chosen as a representative of its group. A list of the selected domains is available from the authors.

Structural Neighbors From Entrez

We identified structural neighbors among the sequence-dissimilar domains in this set by reference to the database of structure–structure alignments provided by Entrez (http://www.ncbi.nlm.nih.gov/Entrez/Structure).[23] This database contains the results of automated all-against-all comparison of domain structures from the PDB, using the MMDB domain definitions that we have employed in selecting sequence-dissimilar domains. We used Entrez data available as of July 30, 1998, which was up to date with respect to the 1,611 sequence-dissimilar domains in the set. Entrez provided a total of 19,516 pairwise structure–structure alignments involving these domains. Note that counts of structure–structure alignments include separately the superposition of a protein onto its neighbor, and of that neighbor onto the protein.

Structure–structure neighbors in Entrez are identified by the VAST algorithm.[10,11] In comparing two structures VAST searches for similar orientations and chain-connectivity of the axial vectors of secondary structure elements. VAST then refines a rigid-body superposition for any substructure similarity crossing its significance threshold. Refinement employs a distance-matrix Monte Carlo procedure to align only those Cα coordinates that well superimpose, and the resulting alignments thus exclude insertions or loop regions of dissimilar conformation. We selected structural neighbors using the significance threshold employed in Entrez, corresponding to $P < 0.05$ for search of a database containing 500 unrelated structures.

Homologous Neighbors From SCOP

To test alternative methods of homologue vs analogue discrimination we employ the classifications provided by the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop/) as an independent standard of truth.[24] We also use SCOP to identify the subsets of homologous neighbors required for definition of the HCS. SCOP provides a hierarchical classification that places protein domains within the same superfamily only when the authors of SCOP have identified evidence of descent from a common ancestral gene. Evidence of homology is derived from comparison of sequences, manual inspection of three-dimensional struc-

tures, and survey of relevant scientific literature.[24] For sequence-dissimilar domains evidence of homology typically consists of a correspondence of function-associated features, for example, a common ligand or ligand-binding site.

Among the structural neighbors identified by Entrez, we identified a subset of homologous neighbors using SCOP release 1.37, the latest available as of July 30, 1998. This release of SCOP provided classifications for 1,532 of the 1,611 domains in the sequence-dissimilar set we consider. SCOP and MMDB domain definitions were equivalent for 987 of these, according to criteria described below, and it is these 987 domains that form the test set used in the analysis of homologue/analogue discrimination. Of the 15,632 Entrez structure–structure alignments involving these domains, 2,879 involve domains within the same SCOP superfamily, which are thus identified by SCOP as homologous neighbors. The remaining 12,753 structure–structure alignments are assumed to involve analogous neighbors.

We have included in the analysis only those Entrez structure–structure alignments where we could verify that the SCOP homology classification referred to the same similarity. We could not directly verify that structural alignments from Entrez and SCOP are equivalent, since SCOP does not provide alignments, but we could check that domain boundaries defined by MMDB and SCOP agree, and thus that the regions identified as similar must necessarily overlap. The specific criterion we adopted is that 80% of an MMDB domain be contained within a SCOP domain and simultaneously that 40% of the SCOP domain be contained within the MMDB domain. This criterion allowed for the different nature of domain definition in SCOP vs. MMDB, based on observed recombination[24] vs compactness[10,11] such that MMDB may split a chain when SCOP does not. MMDB divides the arabinose-binding protein (PDB code 8ABP) into two domains, for example, where SCOP considers these regions one domain. We note that most equivalent domains also agreed at a more stringent criterion of 80% mutual overlap (681 of 987).

Characterization of the Structural Neighbor Set

We have included in the analysis only structural neighbors that are identified by direct structure–structure comparison of sequence-dissimilar domains using the VAST algorithm. This choice is consistent with our purpose, which is to examine methods for identification of homologs among the lists of structural neighbors generated by automated comparison methods. This choice implies, however, that we have excluded from the analysis structural neighbors that might be identified indirectly, in particular by hierarchical clustering procedures. The set of 2,879 homologous neighbors we consider thus includes only 61.6% of the pairs (involving the 987 sequence-dissimilar domains in the test set) that are identified as belonging to the same superfamily cluster in SCOP. For domains in this set the intrinsic sensitivity of VAST vis-à-vis SCOP is relatively high, 89.2%, since an analogous clustering procedure based on the VAST *P* value threshold we employ places 89.2% of SCOP superfamily pairs into the same groups. But our analysis excludes the roughly one third of homologous pairs that are recognizable by neighbors-of-neighbors clustering or by SCOP's consideration of functional similarity. It is restricted to a subset where structural similarity is extensive enough to be recognized by automatic pairwise structure–structure comparison.

The analysis also includes a large number of analogous structural neighbors, as are typically identified by automated comparison methods such as DALI[9,15] or VAST.[10,11] Of the 12,753 analogous pairs identified by VAST only 25.8% are identified as having similar folds in the SCOP classification, presumably because the authors of SCOP employ a stringent similarity threshold for analogous neighbors. We include all these neighbors, however, since our purpose is to consider identification of homologs among such automatically generated lists. We also note that objective measures of pairwise structure–structure similarity, such as RMS, show the analogs in this set to be about as similar to one another as are the homologs (see below). Nonetheless, to examine the effect of test set size, we have computed rates of false positive identification as homologs (at 75% sensitivity, as in Table II) for subsets containing 2,879 analogous pairs, a number equal to the number of homologous pairs in our test set. For subsets containing the 2,879 analogous pairs with the lowest RMS and the greatest percentage of identical residues the false positive rates by the HCS-overlap method are 11.9% and 15.5%, as compared to 16.4% for the complete set of 12,753 analogous pairs. This control experiment indicates that the size of the analogue test set is not favorably biasing the discrimination results we report.

RESULTS

Pairwise Structural Similarity of Homologs and Analogs

To determine the extent to which structure–structure alignments are comparable for homologs and analogs we plot in Figure 1 the distributions of some commonly used similarity measures. These are the RMS Cα superposition residual, the percentage of aligned residue pairs where amino acid types are identical, and the length of structure–structure alignments as a fraction of domain length. One may see that distributions for homologs and analogs overlap substantially, in agreement with previous findings for similar test sets.[13–15] Discrimination involving linear combinations of these measures is no more successful (not shown). It would appear that pairwise structure–structure alignments are largely comparable for the analogs and remote homologs considered here, and that it is difficult to tell them apart on this basis.

We note that structure–structure similarity results shown for homologs in Figure 1 reflect our selection of only sequence-dissimilar domains. If we include homologous domains at closer evolutionary distances, with greater sequence similarity, then discrimination by alignment length, RMS or the percentage of identical residues all improve as one might expect (not shown). We also note

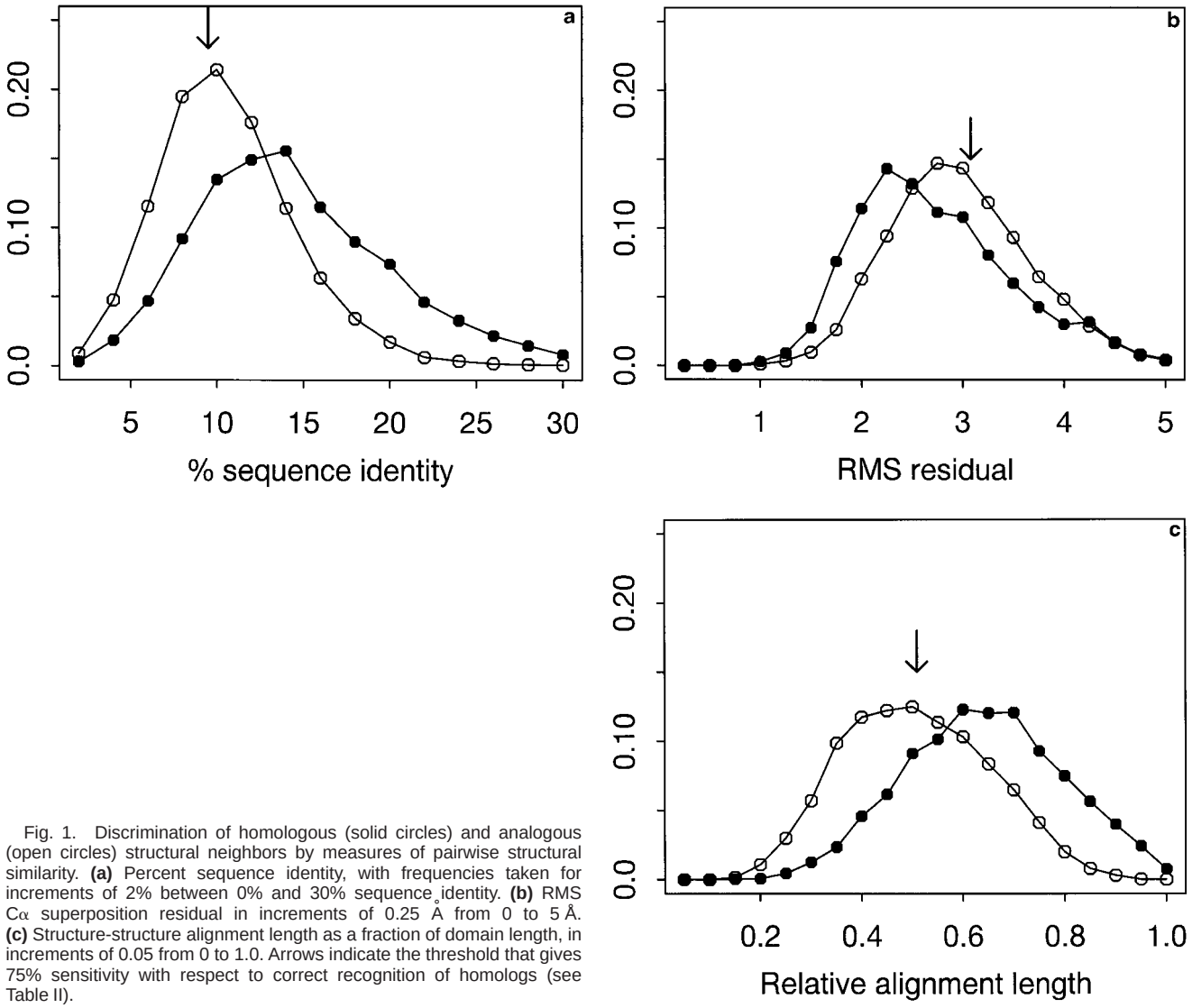


Fig. 1. Discrimination of homologous (solid circles) and analogous (open circles) structural neighbors by measures of pairwise structural similarity. **(a)** Percent sequence identity, with frequencies taken for increments of 2% between 0% and 30% sequence identity. **(b)** RMS Cα superposition residual in increments of 0.25 Å from 0 to 5 Å. **(c)** Structure-structure alignment length as a fraction of domain length, in increments of 0.05 from 0 to 1.0. Arrows indicate the threshold that gives 75% sensitivity with respect to correct recognition of homologs (see Table II).

that the measures of pairwise structure–structure similarity shown in Figure 1 reflect our choice of the VAST structure–structure alignment algorithm. One may see, for example, that discrimination of homologs and analogs is somewhat more successful by alignment length than by the percentage of identical residues, an observation not reported in previous investigations.[13–15] VAST rejects from the alignment residue sites that disproportionately increase the rigid-body RMS of the superimposable substructure, and it is probably this algorithmic feature that leaves structure–structure alignment length as the measure of pairwise structure–structure similarity that best indicates homology.

Definition of the Homologous Core Structure

We define the HCS of a domain as the subset of residues included in structure–structure alignments with 80% or more of its homologous structural neighbors. This definition is illustrated in Figure 2 for a particular example, the N-terminal domain of NADH peroxidase (PDB code 1NHP).[25] In this example there are 10 homologous neighbors, and a total of 71 of 160 domain residues (44%) are included in at least 8 of these ($8/10 \geq 80\%$). The HCS includes Cα coordinates surrounding the active site of NADH peroxidase and a structural scaffold of secondary structure elements extending some distance from the active site. This pattern of conservation is not surprising for remote homologs, and it seems likely that the HCS corresponds to the family-specific substructure identified manually by the authors of SCOP.[24]

We can define the HCS for 614 out of the 987 domains in our test set, those for which SCOP identifies at least one sequence-dissimilar homologous neighbor. On average the HCS includes 53% (standard deviation 18%) of domain residues. As a control we have applied the same definition to analogs, identifying the subset of domain residues present in 80% or more of that domain's analogous (but not homologous) neighbors. These analogous core structures

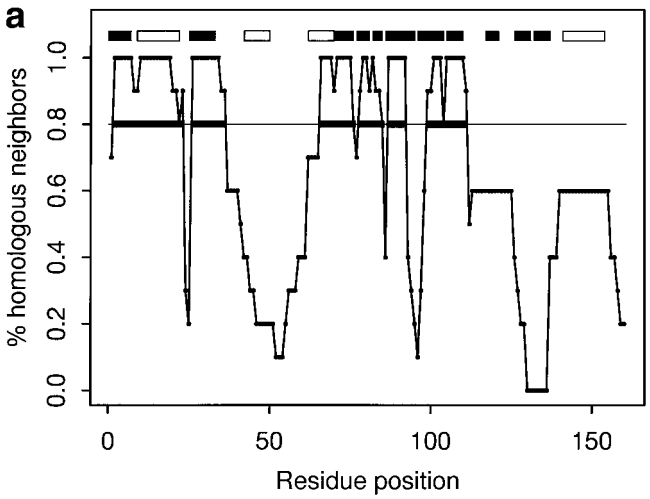


Fig. 2. The HCS of 1NHP_1, NADH peroxidase from *Enterococcus faecalis*. **(a)** Percentage of structure neighbors that had structurally equivalent residues at each residue position of 1NHP_1, the N-terminal domain, residues 1–111 and 270–318, of 1NHP. The positions of helical (open rectangle) and strand (filled rectangle) residues are shown in the upper region of the graph. If 80% or more of homologous neighbors had equivalent residues at a given position that residue position was included in the HCS of 1NHP_1. The 80% threshold is shown by the thin solid line in the graph, and thick segments indicate the positions of HCS residues. **(b)** Structure of 1NHP_1 with its HCS in red. Also shown are the homologous structure neighbors 1AOGA4, 1FCDA1, 1GAL_1, 1AOGA1, 1IUT_1, 1TDE_1, 1NHP_3, 2TMDA3, 1XAN_1, and 1TDE_2 (see Table I). 1NHP_1 is rendered as a thick tube, while its homologous structural neighbors are rendered in wireframe. The cofactors FAD or NAD bound to each of the domains (except for 1AOGA4, 1NHP_3, 2TMDA3 and 1TDE_2) are shown in cyan.

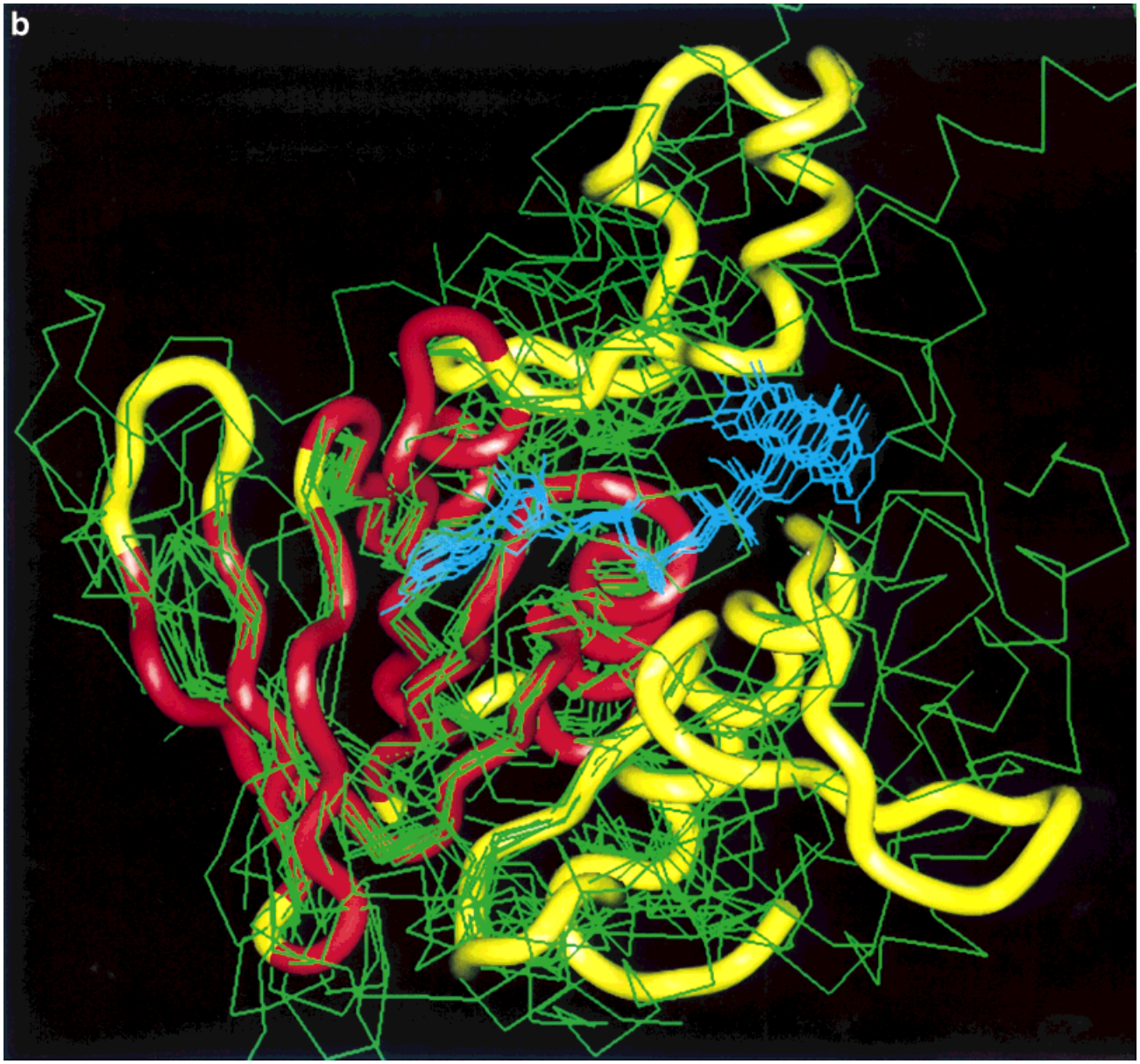


TABLE I. Structural Neighbors of the N-Terminal Domain of NADH Peroxidase (1NHP_1, 1NHP Residues 1–111 and 270–318), Sorted by HCS Overlap Score

Code[a]	SCOP superfamily name[b]	H[c]	S[d]	RMS[e]	% id[f]	L[g]
1AOGA4	FAD/NAD(P)-binding domain	h	1.00	1.6	12.0	82
1FCDA1	FAD/NAD(P)-binding domain	h	0.986	2.5	18.2	142
1GAL_1	FAD/NAD(P)-binding domain	h	0.986	2.6	9.7	115
1AOGA1	FAD/NAD(P)-binding domain	h	0.986	2.5	18.6	126
1IUT_1	FAD/NAD(P)-binding domain	h	0.972	2.3	18.3	113
1TDE_1	FAD/NAD(P)-binding domain	h	0.972	2.3	22.0	114
1NHP_3	FAD/NAD(P)-binding domain	h	0.958	1.7	18.8	78
2TMDA3	FAD/NAD(P)-binding domain	h	0.944	2.4	12.2	80
1XAN_1	FAD/NAD(P)-binding domain	h	0.944	2.5	16.6	129
1TDE_2	FAD/NAD(P)-binding domain	h	0.930	1.6	17.1	76
2TMDA2	A nucleotide-binding domain	a	0.859	1.7	20.8	94
1AA8A1	A nucleotide-binding domain	a	0.817	1.9	16.5	89
1BMDA1	NAD(P)-binding Rossmann-fold domains	a	0.676	2.1	17.6	62
8ATCA2	Aspartate/ornithine carbamoyltransferase, catalytic chain	a	0.662	2.3	10.2	56
3BTOA2	NAD(P)-binding Rossmann-fold domains	a	0.662	1.7	16.7	59
1LDND1	NAD(P)-binding Rossman-fold domains	a	0.634	1.5	22.4	56
1BNCA1	Biotin carboxylase N-terminal domain-like	a	0.634	1.9	4.0	50
1ORTA3	Aspartate/ornithine carbamoyltransferase, catalytic-chain	a	0.634	2.5	8.3	55
1QUE_2	Ferredoxin reductase-like, C-terminal NADP-linked domain	a	0.577	1.6	9.6	52
2PIA_2	Ferredoxin reductase-like, C-terminal NADP-linked domain	a	0.563	1.4	10.4	48
1AHPA2	β-Glucosyltransferase & glycogen phosphorylase	a	0.563	2.8	10.4	66
1DNPA1	N-terminal domain of DNA photolyase	a	0.535	1.8	8.0	50
1RABA2	Aspartate/ornithine carbamoyltransferase, catalytic-chain	a	0.521	2.0	11.1	45

[a]Domain identification code. The first four characters indicate the PDB entry code; the fifth, the PDB chain name or "_" if blank; and the last character is the domain number as assigned in the MMDB database. The names of the proteins and the residues covered by the domains are as follows: 1AOGA4, trypanothione reductase from *Trypanosoma cruzi*, residues 179–286; 1FCDA1, flavocytochrome c sulfide dehydrogenase, flavin-binding subunit from *Chromatium vinosum*, residues 1–104 and 282–327; 1GAL_1, glucose oxidase from *Aspergillus niger*, residues 1–143, 228–322, and 531–583; 1AOGA1, trypanothione reductase from *Trypanosoma cruzi*, residues 3–38, 119–161, and 316–356; 1IUT_1, *p*-hydroxybenzoate hydroxylase from *Pseudomonas aeruginosa*, residues 1–39, 102–181, and 268–354; 1TDE_1, thioredoxin reductase from *Escherichia coli*, residues 1–112 and 265–316; 1NHP_3, NADH peroxidase from *Enterococcus faecalis*, residues 122–242; 2TMDA3, trimethylamine dehydrogenase C-terminal domain from *Escherichia coli*, residues 488–648; 1XAN_1, glutathione reductase from human, residues 18-156 and 310-360; 1TDE_2, thioredoxin reductase from *Escherichia coli*, residues 120–243; 2TMDA2, trimethylamine dehydrogenase, middle, ADP-binding domain from *Escherichia coli*, residues 374–487 and 649–729; 1AA8A1, pig D-amino acid oxidase, N-terminal domain, residues 1–82, 141–188, and 288–340; 1BMDA1, malate dehydrogenase from *Thermus flavus*, residues 0–155; 8ATCA2, aspartate carbamoyltransferase from *Escherichia coli*, residues 151–275; 3BTOA2, horse liver alcohol dehydrogenase, residues 168–322; 1LDND1, lactate dehydrogenase from *Bacillus stearothermophilus*, residues 15–162; 1BNCA1, biotin carboxylase subunit of acetyl-CoA carboxylase from *Escherichia coli*, residues 1–104; 1ORTA3, Ornithine transcarbamoylase from *Pseudomonas aeruginosa*, residues 153–277; 1QUE_2, ferredoxin reductase from Anabaena pcc 7119, residues 141–303; 2PIA_2, Phthalate dioxygenase reductase from *Pseudomonas cepacia* db01, residues 109–229; 1AHPA2, maltodextrin phosphorylase from *Escherichia coli*, residues 456–777; 1DNPA1, N-terminal domain of DNA photolyase from *Escherichia coli*, residues 1–140; 1RABA2, aspartate carbamoyltransferase from *Escherichia coli*, residues 38–133.
[b]The name of the superfamily to which the domain belongs, as defined in SCOP.
[c]"h" if the domain is homologous to 1NHP_1, according to SCOP.
[d]HCS overlap score.
[e]Root-mean-square Cα superposition residual in angstroms.
[f]Percentage of identical residues.
[g]Length of structure–structure alignment.

include on average only 29% (standard deviation 13%) of domain residues, roughly half the size of the HCS. This comparison shows that consistency of structure–structure alignments is greater for members of a homologous family than among structural neighbors in general, an observation that prompts our naming this substructure the homologous core structure.

Presence of the HCS in Homologs and Analogs

We may classify the structural neighbors of domains in our test set according to the extent to which they conserve the HCS. To do so we calculate the fraction of the HCS that is included in the pairwise structure–structure alignment of that domain with its structural neighbors, referring to this quantity as the HCS overlap score. The HCS overlap score is 1 when all domain residues forming the HCS are included in the pairwise structure–structure alignment with a neighbor. The minimum possible value is zero, obtained if the structure–structure alignment with the neighbor were to involve a different region altogether, not overlapping the HCS at all.

As an example we list in Table I the structural neighbors of the N-terminal domain of NADH peroxidase, sorted according to HCS overlap score. Table I also indicates whether each neighbor has been identified as homologous by the authors of SCOP.[24] One may see that homologous

structural neighbors appear at the top of the list and that HCS overlap is lower for all of the analogous structural neighbors. While these analogs all show surprising structural similarity to NADH peroxidase, the corresponding structure–structure alignments include a smaller fraction of the HCS.

We note that HCS overlap scores for homologs shown in Table I have been calculated in a cross-validated manner, such that values are representative of what can be expected when this calculation is applied to new or as yet unclassified structural neighbors. In calculating HCS overlap score for a particular homologous neighbor we exclude structure–structure alignment data for that neighbor from the definition of the HCS, so that the calculated score reflects presence of a substructure common among other homologous neighbors, excluding the domain in question. Cross-validated HCS overlap scores may be calculated for 422 domains in our test set, those for which at least two homologous neighbors are identified by SCOP. The analysis of homologue vs. analogue discrimination below similarly employs cross-validated HCS overlap scores and considers structural neighbors of only this subset of 422 domains.

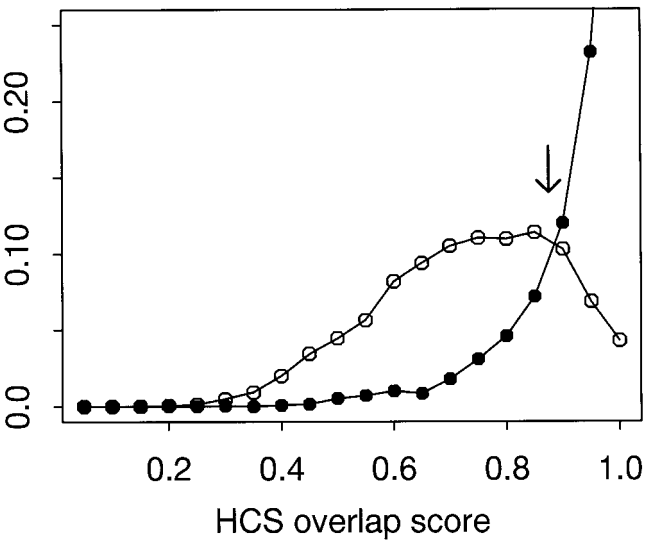


Fig. 3. Discrimination of homologous and analogous structural neighbors by HCS overlap score. Frequencies are taken for increments of 0.05 over the interval from 0 to 1.0. The arrow indicates the threshold that gives 75% sensitivity with respect to correct identification of homologs (see Table II).

Discrimination of Homologs and Analogs by HCS Overlap

Discrimination of homologs and analogs in our test set by the HCS-overlap method is shown in Figure 3. One may see that homologs have high HCS overlap score, generally above 0.90. The distribution for analogs, on the other hand, is centered at about 0.70. At an HCS overlap threshold of 0.88, 75.0% of homologs are correctly identified as such, with a false-positive rate of 16.4% analogs. As shown in Table II, discrimination of homologs and analogs by HCS overlap score is significantly better than discrimination by the percentage of identical residues, RMS, or the length of structure–structure alignments. We conclude that homologs are better identified not by these measures of pairwise structural similarity, but instead by the presence of the HCS, the substructure characteristic of an evolutionary family.

To verify that the discrimination we observe is a unique property of the HCS we have conducted a control experiment in which we used the analogous core structure in an attempt to discriminate analogs and homologs. We define the analogous core structure for each domain based on structure–structure alignments with analogs (but not homologs), calculate overlap scores for all neighbors, and then examine the discrimination obtained. We find that 68.5% of analogs have analogous core structure overlap score of 0.88 or greater. However, we also find that 80.7% of homologs have scores this high and that it is impossible to distinguish homologs and analogs on this basis. This result shows that discrimination by HCS overlap score is greater than can be expected from multiple structure–structure alignments for an arbitrary set of structural neighbors. On the contrary, it seems that the consistency of structure–structure alignments of homologous neighbors is unique and essential to the success of this method in distinguishing homologs.

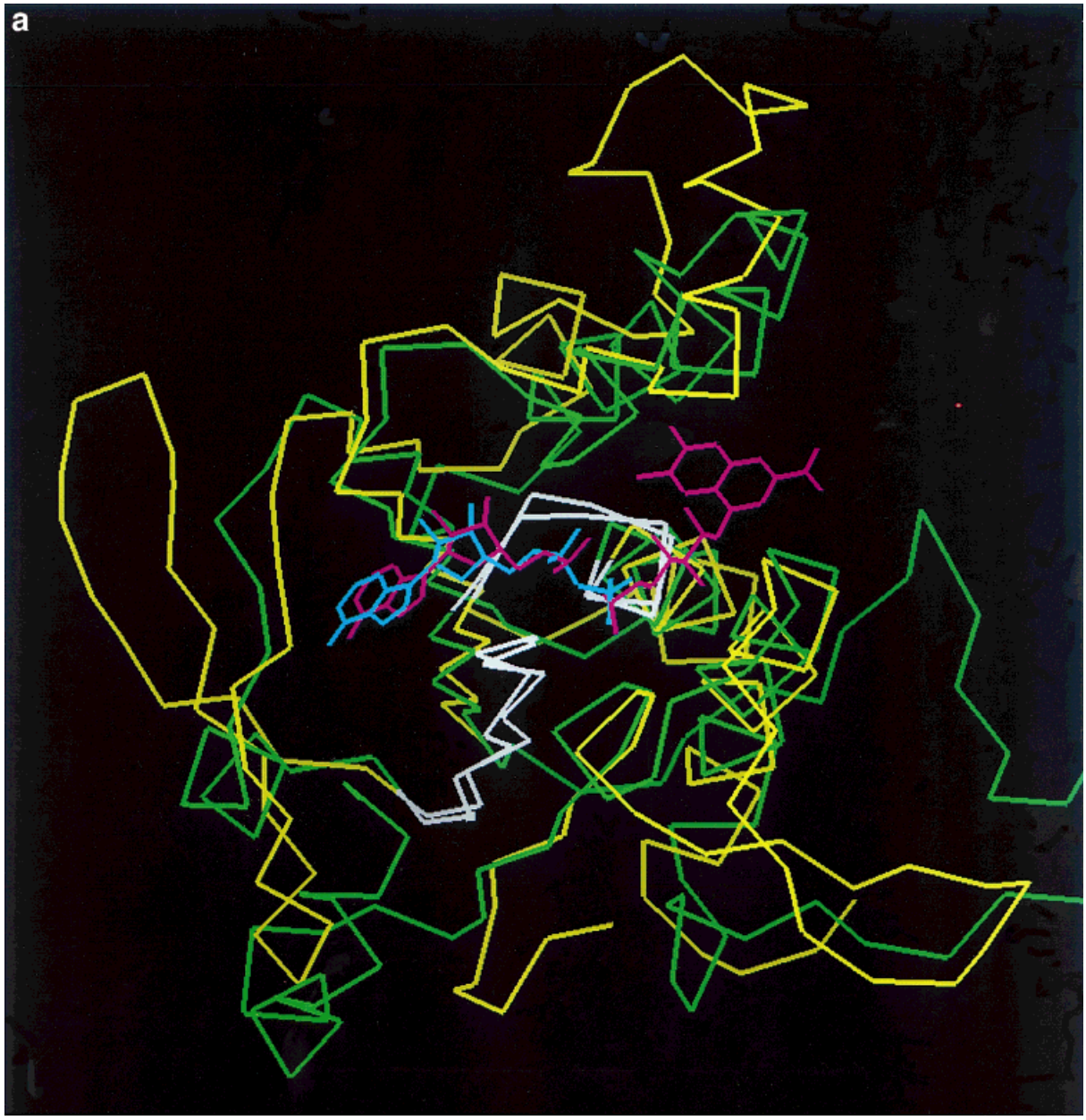
TABLE II. Discrimination of Homologs and Analogs by Different Similarity Measures

Similarity measure (threshold)	Homologs (%)	Analogs (%)
HCS overlap score (>0.88)	75.0	16.4
Structural alignment length (>0.51)	75.0	42.3
Percent identical residues (>9.5%)	75.0	48.2
RMS (<3.1 Å)	75.0	65.7

Does HCS Overlap Indicate Evolutionary Distance?

We see in Table I that two analogous neighbors of 1NHP_1 (NADH peroxidase) have relatively high HCS overlap scores, 0.859 for 2TMDA2 (trimethylamine dehydrogenase) and 0.817 for 1AA8A1 (D-amino acid oxidase). 1NHP_1 belongs to the FAD/NAD(P)-binding domain superfamily, according to the authors of SCOP,[24] while the two analogous neighbors belong to the nucleotide-binding domain superfamily. Other pairs of domains from these two superfamilies also show high HCS overlap scores.

Fig. 4. Structural similarity of the N-terminal domain of NADH peroxidase (1NHP_1) and trimethylamine dehydrogenase ADP-binding domain (2TMDA2). **(a)** shows the 3D superposition of 1NHP_1 (yellow) and 2TMDA2 (green) as determined by VAST. The FAD (magenta) bound to 1NHP_1 and the ADP (cyan) bound to 2TMDA2 are also shown. Two regions with conserved sequence motifs are indicated in white. **(b)** shows the VAST structural alignment of 1NHP_1 (and a homologue, 1TDE_1, Thioredoxin reductase) with 2TMDA2, in a sequence-only view. The two well-conserved regions are indicated by boxes. Strictly conserved residues are indicated by an asterisk and locations of secondary structure elements are indicated by segments labeled a for helix, and b for strand.



b

```
           <--b-->  <------a------>       <--b--->
1NHP_1     1 MKVIVLGSSHGGYEAVEELLNLH (2) AEIQWYEKGDFIS    38
1TDE_1     7  KLLILGSGPAGYTAAVYAARAN (0) LQPVLITGMEKGGQLT 44
2TMDA2   390 DSVLIVGAGPSGSEAARVLMESG (0) YTVHLTDTAEKIGGHL 428
           <--b-->  <------a------>       <--b--->

                          <-a->      <-----b----->        <---b----> <-b->
1NHP_1    66              MESRG (0) VNVFSNTEITAIQ  (17) VENYDKLIISPGAVPFEL 118
1TDE_1    59 LTGPLLMERMHEHATKFE (1)   EIIFDHINKVDL (12) EYTCDALIIATGASARYL 119
2TMDA2   438 GEWSYHRDYRETQITKLL (5) SQLALGQKPMTADD ( 2) QYGADKVIIATGARWNTD 494
               <-----a-----> <-b-->    <a>       a>  <-b--> <-b-> b

                                    <-b->              <-----a------>
1NHP_1   275                        DVFAVGDAT    (13) ALATNARKQGRFAVKNLEEP  316
1TDE_1   240 FVAIGHSPNTAIFEGQL (22) PGVFAAGDVMDH ( 2) RQAITSAGTGCMAALDAER  311
2TMDA2   668 VLVTGRHSECTLWNELK (11) KGIYLIGDAEAP ( 0) RLIADATFTGHRVAREIEEA  698
           <---b---> <--a-->       <-b-->              <-----a------>
```

Figure 4.

1TDE_1 (thioredoxin reductase) of the FAD/NAD(P)-binding domain superfamily, for example, has an HCS overlap score of 0.866 against 1AA8A1. Since the list in Table I resembles a ranking by evolutionary distance, with known homologs at the top, it is interesting to examine those analogous neighbors, as classified by SCOP, that have relatively high HCS overlap scores.

To further consider the example of 2TMDA2, we show in Figure 4 VAST structure–structure alignments of 2TMDA2 with 1NHP_1 and 1TDE_1.One may see that the mononucleotide ligand of 2TMDA2 (ADP) binds in the same region as the dinucleotide ligand of 1NHP_1 (FAD). There are also two conserved sequence features located in close proximity to the ligands. One of them contains the same pattern as the glycine-rich dinucleotide-binding motif GXGXXG seen in many Rossmann-fold domains, even though 2TMDA2 binds a mononucleotide. These proteins also have similar domain organization. They include two domains, the first divided into discrete regions by insertion of the second in topologically equivalent positions, between the fourth and fifth strands of the β sheet. These observations suggest that 2TMDA2 may in fact be homologous to 1NHP_1 and 1TDE_1. The authors of SCOP are unambiguous in placing 2TMDA2 in a different superfamily, and in fact in a different fold category. But they may also have noticed these suggestive similarities, since they comment in SCOP that 2TMDA2 (and 1AA8A1) provide "a link between the Rossmann-fold NAD(P)-binding and FAD/NAD(P)-binding domains."

We emphasize that there is no reason to believe that a high HCS overlap score is a sufficient condition to identify a structural neighbor as homologous, since structural similarity that is the result of convergent evolution might certainly involve the HCS region by chance. In the case of NADH peroxidase, however, HCS overlap does identify interesting candidates (2TMDA2 and 1AA8A1), that were also noted as such by human-expert analysis. The same is true for a several other examples we have examined, among those analogs with high HCS overlap scores (not shown). We can thus suggest that HCS overlap may be a useful criterion for sorting the lists of structural neighbors generated by automated structure–structure comparison algorithms. This may help biologists using these automated resources to identify the structural neighbors that are most likely to be remote homologs.

DISCUSSION

Distinguishing homologs and analogs among sequence-dissimilar structural neighbors is a difficult problem. Structural biologists have solved it primarily by detailed comparison of conformation and binding interactions, to identify features that seem unlikely to have arisen by convergence, much in the way that taxonomists consider particular anatomical details in classification of different organisms. Indeed, this is largely the method used by the authors of SCOP to produce their encyclopedic classification of homology groups among known structures.[24] This approach is laborious, however, since it relies on expert identification of functionally important sites and/or interactions, information that is difficult to encode in a computer and that is in any case missing from current macromolecular structure databases. Identification of homologous features can also be problematical, just as in taxonomic classification, and different experts may not always identify the same set of structures as homologous.[26]

Definition of the HCS, on the other hand, relies neither on the fine details of protein–ligand interaction nor on the presence of functional annotation. The HCS is simply the subset of Cα coordinates that may be superimposed onto those of known homologous neighbors, the backbone-structure scaffold that is conserved within a protein family. With respect to classification of new structural neighbors a test for HCS overlap may certainly be less accurate than classification based on detailed comparison of active sites, since it uses less information. But HCS overlap nonetheless correctly identifies 75% of the sequence-dissimilar homologs that are identified by SCOP, with a false-positive rate of only 16% analogs, significantly better than is achieved by RMS or other measures of pairwise structural similarity. It would appear that conservation of this structural scaffold is strongly associated with conservation of functionally important sites and that one may use this property for detection of remote homologs.

It is interesting to note that the HCS may be considered the three-dimensional structure equivalent of a sequence motif. Sequence motifs identify a particular region that is well conserved in sequence alignments with homologs, and they describe characteristic residue preferences within that region.[16] The HCS is the conserved region similarly identified from structure–structure alignments with homologs, the core structure characteristic of that homologous family. The utility of sequence motifs is in identification of homologous sequences not detected by pairwise sequence comparison alone.[16] The HCS would appear to be useful for the same purpose, in as much as it can identify remote homologs among the structural neighbors of a protein domain with better specificity than RMS or other measures of pairwise structure–structure similarity. The method to improve specificity is in both cases the same: Focus the comparison on the substructure that has been conserved in protein evolution, as defined by multiple alignments with previously identified homologs.

Holm and Sander[15] recently proposed an alternative method for automated discrimination of homologs and analogs among structural neighbors. Homologs may be identified, they suggest, by a combination of measures intended to detect functional similarity, those with greatest weight being sequence family overlap and keyword overlap. The former indicates that neighbors-of-neighbors clustering based on sequence similarity creates a link between structural neighbors, and the latter detects similar terminology in the text annotation of sequences similar to those of either structure, using a technique similar to that used in Entrez[23] for Medline® neighboring. These authors do not base discrimination on presence of a conserved substructure, and in that sense their method is quite different from HCS overlap. But these authors

similarly rely on previously recognized homologs, using them to define sequence and keyword profiles for a protein family, much in the way that we use previously recognized homologs to define a characteristic structural motif. Obviously, these methods are not mutually exclusive and might be combined.

For purposes of illustration we have adopted in the current analysis very simple definitions of the HCS and HCS overlap, based on empirical thresholds, but we note that more rigorous probabilistic definitions are possible. One may, for example, define the relative variability of HCS coordinates, and calculate HCS overlap scores in a way that takes this information into account. In the current analysis we have also chosen to use SCOP's identification of homologous neighbors, so as to test the ability of HCS overlap calculations to reproduce this independent classification. It seems likely, however, that homologs identified by sequence motif analysis will provide a useful definition of the HCS, and that it may be possible to construct a comprehensive HCS database automatically, without the need for manual classifications. In this case one might use HCS overlap scores systematically, as one choice for ranking the lists of structural neighbors identified by automated structure–structure comparison methods. We also note that a HCS database might be useful in other computations, such as threading[10,27,28] or homology modeling,[29] which similarly require definitions of the core structure likely to be conserved in protein evolution.

ACKNOWLEDGMENTS

We thank E. Koonin, D. Lipman, T. Madej, and A. Marchler-Bauer for valuable discussions, and T. Madej for assistance with use of VAST alignment data.

REFERENCES

1. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372:631–634.
2. Mizuguchi K, Go N. Seeking significance in three-dimensional protein structure comparisons. Curr Opin Struct Biol 1995;5:377–382.
3. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
4. Brenner SE, Chothia C, Hubbard TJ. Population statistics of protein structures: lessons from structural classifications. Curr Opin Struct Biol 1997;7:369–376.
5. Jukes TH. Comparisons of the polypeptide chains of globins. J Mol Evol 1971;1:46–62.
6. Richardson JS, Richardson DC, Thomas KA, Silverton EW, Davies DR. Similarity of three-dimensional structure between the immunoglobulin domain and the copper, zinc superoxide dismutase subunit. J Mol Biol 1976;102:221–235.
7. Bernstein FC, Koetzle TF, Williams GJB, Meyer EFJr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
8. Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, editors. Crystallographic Databases: Information Content, Software Systems, Scientific Applications. Bonn: Data Commission of the International Union of Crystallography, 1987:107–132.
9. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. A database of protein structure families with common folding motifs. Protein Sci 1992;1:1691–1698.
10. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23:356–369.
11. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
12. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH: a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
13. Russell RB, Barton GJ. Structural features can be unconserved in proteins with similar folds: an analysis of side-chain to side-chain contacts, secondary structure and accessibility. J Mol Biol 1994;244: 332–350.
14. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 1997;269:423–439.
15. Holm L, Sander C. Decision support systems for the evolutionary classification of protein structures. ISMB 1997;5:140–146.
16. Bork P, Koonin EV. Protein sequence motifs. Curr Opin Struct Biol 1996;6:366–376.
17. Ohkawa H, Ostell J, Bryant S. MMDB: an ASN.1 specification for macromolecular structure. ISMB 1995;3:259–267.
18. Hogue CW, Ohkawa H, Bryant SH. A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. Trends Biochem Sci 1996;21:226–229.
19. Marchler-Bauer A, Addess KJ, Chappey C, Geer L, Madej T, Matsuo Y, Wang Y, Bryant SH. MMDB: Entrez's 3d structure database. Nucleic Acids Res 1999;27:240–243.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
21. Altschul SF, Boguski MS, Gish W, Wootton JC. Issues in searching molecular sequence databases. Nature Genet 1994;6:119–129.
22. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. PNAS 1998;95:6073–6078.
23. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. Methods Enzymol . 1996; 266:141–162
24. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
25. Stehle T, Ahmed SA, Claiborne A, Schulz GE. Structure of NADH peroxidase from *Streptococcus faecalis* 10C1 refined at 2.16 Å resolution. J Mol Biol 1991;221:1325–1344.
26. Artymiuk PJ, Poirrette AR, Rice DW, Willett P. A polymerase I palm in adenylyl cyclase? Nature 1997;388:33–34.
27. Bryant SH. Evaluation of threading specificity and accuracy. Proteins 1996;26:172–185.
28. Marchler-Bauer A, Bryant SH. A measure of success in fold recognition. Trends Biochem Sci 1997;22:236–240.
29. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.