Sensitivity and selectivity in protein structure comparison

MICHAEL L. SIERK AND WILLIAM R. PEARSON

Department of Biochemistry and Molecular Genetics, University of Virginia Health System, Charlottesville, Virginia 22908, USA

(RECEIVED July 23, 2003; FINAL REVISION November 26, 2003; ACCEPTED November 28, 2003)

Abstract

Seven protein structure comparison methods and two sequence comparison programs were evaluated on their ability to detect either protein homologs or domains with the same topology (fold) as defined by the CATH structure database. The structure alignment programs Dali, Structal, Combinatorial Extension (CE), VAST, and Matras were tested along with SGM and PRIDE, which calculate a structural distance between two domains without aligning them. We also tested two sequence alignment programs, SSEARCH and PSI-BLAST. Depending upon the level of selectivity and error model, structure alignment programs can detect roughly twice as many homologous domains in CATH as sequence alignment programs. Dali finds the most homologs, 321–533 of 1120 possible true positives (28.7%–45.7%), at an error rate of 0.1 errors per query (EPQ), whereas PSI-BLAST finds 365 true positives (32.6%), regardless of the error model. At an EPQ of 1.0, Dali finds 42%–70% of possible homologs, whereas Matras finds 49%–57%; PSI-BLAST finds 36.9%. However, Dali achieves >84% coverage before the first error for half of the families tested. Dali and PSI-BLAST find 9.2% and 5.2%, respectively, of the 7056 possible topology pairs at an EPQ of 0.1 and 19.5, and 5.9% at an EPQ of 1.0. Most statistical significance estimates reported by the structural alignment programs overestimate the significance of an alignment by orders of magnitude when compared with the actual distribution of errors. These results help quantify the statistical distinction between analogous and homologous structures, and provide a benchmark for structure comparison statistics.

Keywords: structure alignment; database search; statistical significance; CATH database

Pairwise comparison of protein molecules is fundamental to modern biological research. If two proteins share significantly more similarity with regard to some characteristic (e.g., amino acid sequence) than is expected by chance, the most parsimonious explanation is that they descended from a common precursor (they are homologous) and thus are likely to share other similarities as well. The most commonly compared features are a protein's sequence and its three-dimensional structure, although other features, such as function, ligand(s), the location of specific conserved residues, the location of specific cofactors or posttranslational modifications, subcellular localization, phyletic profile (i.e., the pattern of organisms in which the protein is present or absent), or the expression profile of the protein can also be compared.

There are three general steps in protein comparison: (1) deciding what feature to compare, (2) deciding how to compare the chosen feature, and (3) determining whether the feature exhibits an excess of similarity compared to chance. In the case of sequence comparison, these are all fairly straightforward decisions. The simplest feature to compare is the linear sequence of the 20 naturally occurring amino acids. Sequence comparison typically involves sequence alignment (Smith and Waterman 1981; Pearson and Lipman 1988; Altschul et al. 1997), attempting to maximize the correspondence of identical residues while minimizing the number of gaps inserted, although other methods have been described (Wu et al. 1996). There is a well-established theo-

Reprint requests to: William R. Pearson, Department of Biochemistry and Molecular Genetics, University of Virginia Health System, P.O. Box 800733, Charlottesville, VA 22908, USA; e-mail: wrp@virginia.edu; fax: (434) 924-5069.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.03328504.

retical (Karlin and Altschul 1990) and empirical (Mott 1992; Altschul and Gish 1996; Brenner et al. 1998; Pearson 1998) basis for estimating the probability of a local alignment score occurring by chance. In contrast, the comparison process is more complicated with three-dimensional protein structures. One can compare the coordinates of Cα atoms, secondary structure elements (SSEs), or internal mappings such as a contact map or distance matrix. There are multiple algorithms for comparing each of these features, and one can combine different features in a single comparison algorithm. The structures being compared are models that best fit experimental data, and therefore have varying degrees of precision, accuracy, and dependence upon the experimental conditions (e.g., salt, pH, crystal packing, etc.) Finally, there is no clear statistical definition of what constitutes an excessive amount of similarity. This is due largely to three circumstances: (1) The range of protein structures appears far more constrained by chemical and physical forces than the range of sequences, (2) there is no definition of an optimal three-dimensional alignment, and (3) it is difficult to specify a "random" protein structure, and it is difficult to compare very different protein structures (e.g., all-α versus all-β). The uncertainty regarding the range of possible protein structures and the lack of a mathematical definition for "random" structures inhibits a theoretical understanding of how likely it is for two proteins to independently end up with similar structures (which is extremely unlikely in the case of sequences), whereas the lack of an optimal alignment definition inhibits efforts to empirically determine the distribution of structural similarity scores and to compare different methods (for reviews, see Eidhammer et al. 1999, Koehl 2001).

The above complications notwithstanding, comparison of protein structures has been of great interest since the first two atomic resolution structures of myoglobin and hemoglobin were determined (Kendrew et al. 1960; Perutz et al. 1960). Protein structures offer a great deal of information about protein function, evolution, and the fascinating problem of how the three-dimensional structure is imbedded in the one-dimensional sequence (the protein folding problem). It is only recently, however, that enough protein structures have become available to enable large-scale comparisons of protein structures. In the future, high-throughput methods will produce even larger numbers of protein structures, many of which will have unknown functions and/or evolutionary relationships (Brenner and Levitt 2000). Thus, a better understanding of protein tertiary structure alignment statistics should allow more accurate classification of proteins, and may provide more quantitative insights into the balance of homology (descent from a common ancestor) and analogy (chance similarity due to convergence). In addition, because the process of obtaining the three-dimensional structure of a protein via X-ray crystallography or NMR is so laborious and expensive compared to computer-based structure predictions, structure prediction will become increasingly important as the predictions become more accurate (Marti-Renom et al. 2000; Vitkup et al. 2001). Comparison of protein structures is crucial for evaluating the accuracy of predictions, and by extension for increasing their performance (Moult et al. 2001). Finally, tertiary structure comparison is essential to elucidating the relationships between protein sequence change and the resulting changes in structure and function throughout evolution.

Both sequence and structure comparison methods are confounded by protein domains. Because many proteins are made up of multiple unrelated domains that have been spliced together into one polypeptide, biologists usually break proteins into their constituent domains. This can be done based upon sequence, structure, or both. Several domain-based structural databases were created in recent years, which organize the data from the Protein Data Bank (PDB; Berman et al. 2000) in various ways. The most popular of these secondary databases are Structural Classification of Proteins (SCOP; Murzin et al. 1995), CATH (which stands for Class, Architecture, Topology, and Homology; Orengo et al. 1997), and FSSP (which stands for Families of Structurally Similar Proteins; Holm and Sander 1998). The first two are hierarchical databases, which first break the protein structures in the PDB into domains, then classify the domains into groups with common secondary structure components (Class in SCOP and CATH), common arrangement (Architecture in CATH) and topology of secondary structure elements (fold in SCOP or Topology in CATH), and homologous superfamilies (superfamily in SCOP and Homologous family in CATH) and sequence families (both). FSSP consists of comparisons between all the protein chains in the PDB, but no specific hierarchy is assigned. We chose CATH as a standard because it is more automated in its classification procedure than SCOP, and it has explicitly defined sequence and structure-based criteria for assigning homology, which both SCOP and FSSP lack.

Here we evaluate various structure comparison algorithms and scoring schemes in their abilities to detect Homologs and Topologs (domains with the same Topology, or T, designation) as defined by CATH (we use capital letters to distinguish the specific Homologs/Topologs as defined by CATH from the general reference to truly homologous proteins, which may or may not coincide with the CATH definitions). The performance of the methods varies widely. For structure alignment, statistical measures of structural similarity calculated by the programs moderately outperform root-mean-square distance (RMSD)-based scores. Our statistical analysis seeks to quantify the difficulty of distinguishing homology and analogy based upon pairwise structural comparison alone. Most importantly for the issue of reliably identifying distant homologs, with the exception of statistical estimates derived from Dali Z-scores, the statis-

tics provided by structure comparison programs greatly overstate the significance of structural alignments.

Results

We created a nonredundant subset of CATH (see Materials and Methods), and selected a single member from each of 86 families to serve as a query. This was done instead of doing an all-versus-all comparison, both to reduce computational requirements and to mitigate the high redundancy found in the PDB: We did not want the results to be dominated by the structural family members that are most overrepresented in the PDB, such as the globins or immunoglobulins. Each query was compared to each member of the library using five different structural superposition programs (Dali, Holm and Sander 1996; Structal/LSQMAN, Kleywegt 1996 and Levitt and Gerstein 1998; CE, Shindyalov and Bourne 1998; VAST, Madej et al. 1995; and Matras, Kawabata and Nishikawa 2000). For comparison, we also used two sequence comparison programs (SSEARCH, Pearson 1991; PSI-BLAST, Altschul et al. 1997). In addition, we tested two measures of protein structural similarity that calculate a structural distance between the two domains but do not align them (SGM, Rogen and Fain 2003; PRIDE, Carugo and Pongor 2002).

Two characteristics of a search algorithm are important when searching a database: sensitivity and selectivity. A more sensitive algorithm will find a larger percentage of the total number of true positives, or homologs in the database, at a given threshold of statistical significance or false positives. A more selective algorithm will find a smaller number of false positives, or nonhomologs that receive high similarity scores to the query, at a given threshold of coverage. Generally there is a trade-off between these two characteristics, such that improving the performance of one degrades the performance of the other.

Errors per Query versus Coverage

To present sensitivity and selectivity simultaneously, we plotted Errors per Query (EPQ) versus Coverage curves, which are similar to receiver operating characteristic (ROC) curves (Gribskov and Robinson 1996), as described in the Materials and Methods section. The Errors versus Coverage curves show how much coverage is obtained at a given error level, or the sensitivity (the number of true positives detected) at a given level of selectivity (the number of false positives detected). Figure 1A shows the EPQ versus Coverage for seven of the methods (the PRIDE score performed worse than the SGM score and was left off of the plots) in detecting CATH Homologs. At an error level of 0.1 EPQ, which corresponds to ~8 false positives, PSI-BLAST shows the best coverage, identifying 32.6% of the 1120 possible hits. Dali is the next best, with 28.7% coverage. The SGM algorithm, which does not align the protein domains, has the lowest coverage at all error rates. At an EPQ of 1.0, Matras is the best performer, with 49% coverage, slightly ahead of Dali and Structal at ~45%; PSI-BLAST has 36.9% coverage.

EPQ or ROC curves can be misleading if some homologous proteins are misidentified as nonhomologous, and thus counted as false positives. CATH has fairly stringent criteria for assigning homologous relationships (Orengo et al. 1997), and some structural alignments that were labeled false positives, and thus contributed to the "error" axis, may be homologs that CATH had placed into a different H classification. To reduce the effect of misclassification of homologs, we also plotted EPQ versus Coverage when hits to members of the same Topology classification were not counted as errors (Fig. 1B). This changes the absolute amount of coverage, more homologs are found with fewer errors, but it only moderately changes the relative performance of the different structure comparison methods (Dali performs better relative to Matras and Structal, and VAST improves with respect to the other methods). Dali now finds the most homologs at 0.1 EPQ (45.7%), whereas PSI-BLAST finds 32.8%. At an EPQ of 1, the structure/sequence difference is even more pronounced, with Dali finding 70% of the possible homologs whereas PSI-BLAST finds 36.5%.

We also looked at the ability of the programs to detect members of the same CATH Topology (T) group, which we call Topologs (Fig. 1C). This avoids the problem of homolog misclassification; we expect that non-Topologs are much more likely to be unrelated than non-Homologs in the CATH classification. Dali finds the most Topologs at an EPQ of 0.1, finding 9.2% of the 7056 total Topolog pairs, whereas PSI-BLAST identified only about 5% of Topologs. All of the structural alignment methods improve relative to PSI-BLAST at low EPQ; VAST in particular improves dramatically, going from one of the worst methods at detecting CATH Homologs to one of the best at detecting Topologs (although VAST and PSI-BLAST have very similar performance at 0.1 EPQ). CE and SGM still perform worse than PSI-BLAST until EPQ >1. Most of the improvement in performance reflects reduced errors per query early in the EPQ plot, which are due to high-scoring alignments to structures in a different Homology class. When both Homology matches and Homology errors are excluded from the EPQ calculation (Fig. 1D), showing the ability to detect Topologs that are not Homologs, none of the methods do very well, but VAST and Dali perform best.

In their paper on the Structal method, Levitt and Gerstein (1998) reported a coverage of 29.8% (627/2107) at an error rate of 1%, or an EPQ of 0.01. They used a library of 941 domains from the SCOP database and performed an all-versus-all comparison. They used pairs sharing the same

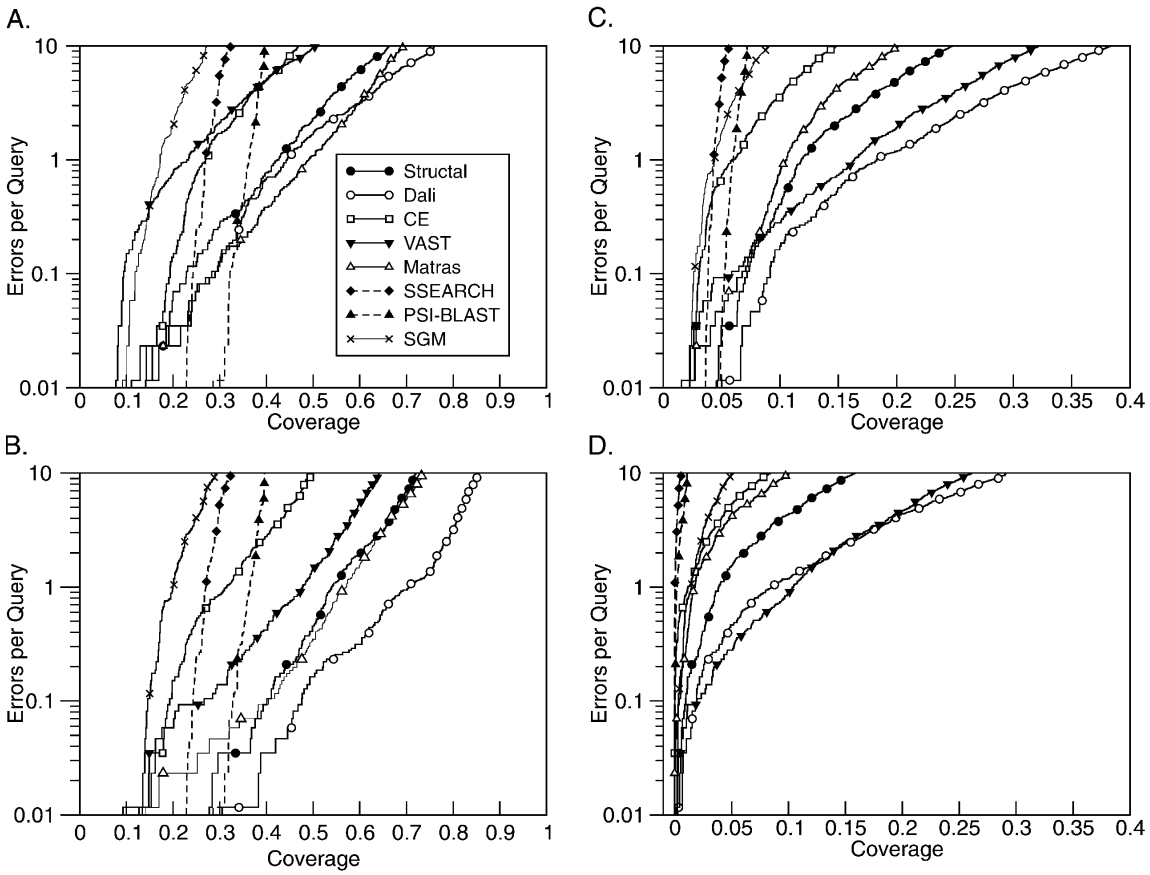

Figure 1. Errors per Query vs. Coverage plots for eight of the nine methods tested (PRIDE data not shown). (*A*) CATH Homolog set of true positives. (*B*) CATH Homolog set of true positives, but only non-Topologs are false positives. (*C*) CATH Topolog (same Topology) set of true positives, non-Topolog false positives. (*D*) Non-Homolog CATH Topolog set of true positives, non-Topolog false positives. The sequence alignment programs are shown with dashed lines; the structural comparison programs, with solid lines. Programs using Z-scores as the scoring criterion have open symbols; those using E()-values have filled symbols. Symbols are shown at every 200th point.

SCOP superfamily as their set of true positives (roughly analogous to Homolog pairs in CATH), but counted as errors only pairs not in the same Fold category (roughly the same as the Topology classification in CATH). This is analogous to our Figure 1B, and we see similar performance—27.8% coverage (311/1120) at an EPQ of 0.01.

EPQ versus Coverage curves can be distorted due to poor performance by one or two queries, if those queries produce many errors at low coverage levels. Thus we examined the performance of the methods with individual queries. Figure 2, A and C show the level of coverage generated by the median query (43 queries performed better, 43 worse) at the 1st, 3rd, 10th, 30th, and 100th false positive for Homologs and Topologs. To compare these results with those shown in Figure 1, we also grouped the queries into nine sets by query length (shortest to longest) and plotted the coverage obtained by the median set at 0.1, 0.3, and 1.0 EPQ (see Materials and Methods). Figure 2, B and D show the same results for the 25th percentile (i.e., 21 of the queries have better coverage and 65 have worse coverage). (For EPQ < 1, the 25th percentile corresponds to the second-lowest level of coverage of the nine groupings.)

Comparison of Figures 1 and 2 shows that the relative performance of the different methods does not change appreciably. However, absolute performance does change significantly for the structural comparison programs when each family is considered independently. For Dali, half of the queries obtain a coverage of ~85% or better before generating the first false positive, and half reach 100% coverage before the third false positive. The discontinuity between the upper and lower parts of Figure 2A shows that grouping different families together by length of query greatly affects the median amount of coverage. This implies that there is no clear correspondence between family length and selectivity (i.e., the amount of coverage generated at one false positive). There is also no clear correspondence between selectivity and the number of family members, nor between the rankings of the different methods (data not shown).

Table 1. *Errors by CATH class*

	α (21%)		β (26%)		α/β (50%)	
	H	T	H	T	H	T
Structal (100)	9	36	8	6	83	58
Dali (100)	6	44	6	12	88	44
CE (95)	64	90	4	1	32	9
VAST (100)	0	5	11	4	89	91
Matras (93)	25	31	39	32	36	37
PSI-BLAST (53)	22	24	17	22	61	54
SSEARCH (37)	25	26	35	34	40	40

The CATH class of the query domain is shown for the top 100 errors in detecting Homologs (H) and Topologs (T) for seven of the tested methods. The percentages in the column headers are the percentage of the library in the given class. The number of errors in which the query and target are in the same CATH class (when detecting Homologs) is shown in parentheses after the name of the program.

Evaluation of errors

We investigated the nature of the first 100 errors made by the different algorithms. For 93 of the first 100 errors that Dali makes in finding Homologs, the query and target have the same Topology; Structal and VAST have 85 and 92 errors with the same Topology, respectively, whereas Matras has 67 and CE has 43. Thus, the homology "errors" generated by the structural alignment programs are topologically similar, and may, in fact, be misclassified homologs. In contrast, SSEARCH has seven errors with the same topology, similar to the number (three) one would expect if the errors were randomly distributed among topologies. PSI-BLAST has 22 errors with the same topology, suggesting that PSI-BLAST may also be recognizing misclassified Homologs.

The programs show distinct differences in their misclassifications according to structural class. In our library of 2771 domains, there are 586 (21%) all α-helix (CATH class 1) domains, 720 (26%) all β-sheet domains (class 2), 1384 (50%) mixed α/β (class 3) domains, and 82 (3%) little secondary structure (class 4) domains; the queries follow a similar distribution—23%, 29%, 48%, and 0% for classes 1, 2, 3, and 4, respectively. Table 1 shows the CATH Class of the query for the top 100 errors for six of the eight methods tested. In detecting Homologs, the PSI-BLAST and SSEARCH errors approximate the expected distribution if the errors were randomly distributed among the three Classes, but errors in four of the structural alignment programs are biased. Dali, Structal, and VAST are biased towards errors in the α/β Class, whereas CE is biased towards errors in the all-α Class. The interesting exception is Matras, which has errors evenly distributed among the three classes, with only a slight bias towards the β class.

Equivalence scores

An alternative way to evaluate the performance of a method is to calculate the equivalence number (Pearson 1995) for each query. The equivalence number is the point at which the number of false positives is equal to the number of false negatives. One records the number of false positives at this point, not the reported score, as different methods have different scoring schemes. Thus, when a method performs perfectly, finding all of the family members before finding any nonfamily members, it will receive an equivalence score of 0 for that family.

We calculated the equivalence score for each of the 86 queries for each method, and then tabulated which method had a better (lower) equivalence score. Tables 2A and 2B show the results of a binomial distribution sign test performed between different methods over the 86 families. The sign-test results (Table 2A) show that Dali performs significantly better than each of the other approaches and generally support (Table 2B) the same pattern shown in the EPQ versus Coverage plots (Dali > Matras > Structal > CE > VAST > PSI-BLAST > SSEARCH > SGM). The notable exception is SGM, which performs worse than SSEARCH on the Errors versus Coverage plots, but the same in the equivalence test. This indicates that the overall performance of the methods is not greatly affected by extremely poor or excellent performance on a small number of protein families. Table 2B presents a pairwise comparison for all the approaches; we note that our analysis is consistent with earlier comparisons of PSI-BLAST and SSEARCH (Park et al. 1998) which show that PSI-BLAST is significantly more sensitive.

Alternate query set

We also checked the extent to which the performance is affected by the specific set of queries used in these experiments. To this end, we randomly selected five different sets of 100 queries from our library and analyzed them with the Structal method as implemented in the LSQMAN program (Kleywegt 1996). The EPQ versus Coverage plots are shown in Figure 3, which shows that the original data set represents the best performance, as the curves from the randomly selected sets generally lie to the left of the original set. In detecting Homologs, there is considerable variation, with the different curves crossing each other many times. At an EPQ of 0.1, coverage ranges from 0.149 to 0.229, and at an EPQ of 1, coverage ranges from 0.269 to 0.424. Thus, the variation in the performance of the structure comparison programs is similar to the variation produced by selecting different queries. Topolog detection efficiency also varies with the query set, but only below an EPQ of 1 (Fig. 3B). We found a similar variation in the curves generated by SSEARCH and PSI-BLAST with the alternative query sets (data not shown).

For each query set, the coverage of either Homologs or Topologs detected by each of the three methods (Structal/

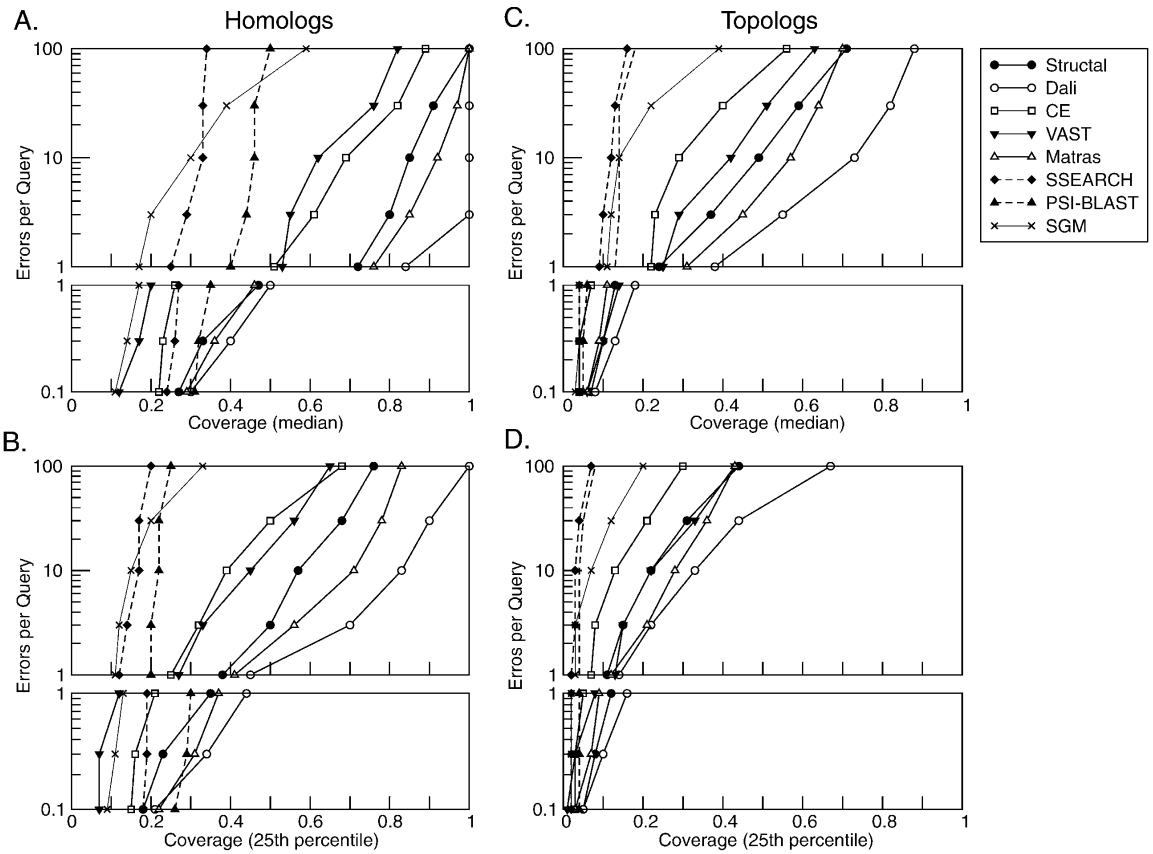

Figure 2. Errors per Query vs. Coverage plots for individual families. (*A*) The median level of coverage generated by the 86 queries is shown at a given number of errors (false positives) for CATH Homologs. (*B*) The same as *A*, except that the level of coverage is shown at the 25th percentile (with the families ranked by percent coverage). (*C*,*D*) The same as *A* and *B*, respectively, with CATH Topologs used as the set of true positives. The portions of the plot with EPQ <1 were made by grouping the families into groups of 10 by the length of the query (see Materials and Methods).

LSQMAN, PSI-BLAST, and SSEARCH) was noted at 0.1 and 1 EPQ (data not shown). In each case, the relative performance was the same as in the original query set (i.e., at 0.1 EPQ in detecting Homologs, PSI-BLAST always had the highest coverage, followed by SSEARCH, followed by Structal/LSQMAN), with the exception of the coverage at 1 EPQ in detecting Homologs, where either PSI-BLAST or both PSI-BLAST and SSEARCH had higher coverage than Structal/LSQMAN in some of the query sets.

We conclude from this that, although different sets of queries may result in moderate differences in absolute coverage levels, the relative performance of the different methods was essentially unchanged, and there were no drastic changes in the shapes of the curves or overall performance generated by the three methods tested with alternative query sets. Thus, although some details might change with different sets of queries or different target libraries (which we did not test), the overall performances of the methods presented here are generally indicative of their true performance.

Table 2A. *Equivalence scores comparing Dali and seven of the eight other methods tested*

Dali vs.:	+	−	Tie	Binomial P()
Matras	26	12	48	0.017
Structal	44	6	36	1.6e-8
CE	59	4	23	6.9e-14
VAST	78	4	4	0
PSI-BLAST	64	4	18	2.9e-15
SSEARCH	73	4	9	0
SGM	76	3	7	0

Plus (+), Minus (−), and Tie are the number of times that Dali had a better (lower), worse (higher), or the same equivalence number as the method being compared when searching for CATH Homologs. P() is the probability, based upon a binomial distribution, that Dali and the compared method perform equivalently.

RMSD values

The most widely used statistic in protein structure comparisons is the root-mean-square distance (RMSD). It has been

Table 2B. *Equivalence score performance for all methods*

						Homologs		
	Dali	Matras	Structal	CE	VAST	PSI-BLAST	SGM	SSEARCH
Dali	—	26/12	44/6	59/4	78/4	64/4	76/3	73/4
		0.017	1.6e-8	6.9e-14	0	2.9e-15	0	0
Matras	26/8	—	31/17	52/8	70/7	62/6	72/6	72/6
	0.002		0.03	2.6e-9	1.8e-14	4.1e-13	8.9e-16	8.9e-16
Structal	36/6	28/14	—	46/10	68/11	60/4	69/3	71/5
	1.4e-6	0.022		6.2e/7	1.8e-11	3.7e-14	0	2.2e-16
CE	52/3	46/8	47/21	—	47/25	41/19	58/12	52/13
	7.7e-13	6.9e-8	0.001		0.006	0.003	1.1e-8	5.8e-7
VAST	56/6	49/10	53/28	43/37	—	42/31	59/17	54/20
	1.5e-11	1.4e-7	0.004	0.29		0.12	7.0e-7	4.8e-5
PSI-BLAST	66/2	64/6	73/2	60/14	61/17	—	43/25	39/14
	0	1.2e-13	0	3.1e-8	2.8e-7		0.019	4.0e-4
SGM	73/1	67/5	72/5	59/14	67/14	28/46	—	29/28
	0	3.2e-15	1.1e-16	5.1e-8	9.4e-10	0.024		0.5
SSEARCH	74/3	75/6	76/4	64/10	67/14	41/19	48/21	—
	0	1/1e-16	0	4.5e-11	9.4e-10	0.003	7.8e-4	
		Topologs						

The upper numbers are the plus (+) and minus (−) values (see Table 2A). The lower numbers are the probability, based upon the binomial distribution, that the two methods perform equivalently. The upper half of the matrix is for the Homolog true positive set, the lower half (shaded), for the Topolog set. The better-performing method is on the left when searching for Homologs, and on the top when searching for topologs (e.g., when searching for Homologs, Dali has a better equivalence number than Structal in 44 families, whereas Structal has a better equivalence number in six families. When searching for Topologs, Dali does better in 36 families, whereas Structal does better in six). The exception is PSI-BLAST vs. SGM, where PSI-BLAST does better with Homologs, but SGM does better with Topologs.

known for some time that the RMSD value by itself is often a misleading indicator of the significance of a structural alignment, because one can achieve a very low RMSD by aligning a small number of residues (Levitt and Gerstein 1998; Carugo and Pongor 2001). Consequently, one must also take into account the length of the alignment. Here, we divide the RMSD by the number of aligned residues (N_{align}) and compare the performance of this normalized RMSD value to that of the scores reported by the five structure alignment programs (Fig. 4). Dali and CE Z-scores do slightly better than RMSD/N_{align}, though the improvement is small, whereas Matras Z-scores perform significantly better than RMSD/N_{align}. The performance of Structal E()-values is indistinguishable from the RMSD scores, which is perhaps not surprising, as it is a method based on minimizing the distance between Cα atoms, and its structural similarity score is essentially the inverse of this distance, plus penalties for gaps. Surprisingly, the VAST E()-values actually perform worse than RMSD scores in detecting CATH Homologs, but significantly better than RMSD scores in detecting Topologs. A comparison of the equivalence points for the statistical score (E()-value or Z-score) and the RMSD/N_{align} for each method largely corroborates these findings (data not shown); the only difference is that the VAST E()-values and RMSD/N_{align} are indistinguishable in the equivalence point test.

Statistical reliability

Sequence comparison programs are widely used to infer homology, or descent from a common ancestor. In general, if a sequence in a comprehensive sequence database shares sequence similarity that is expected $<10^{-6}$–10^{-3} times in a database search by chance, the most parsimonious explanation is that the two sequences diverged from a common ancestor. For tertiary structure comparison, both the definition of excess similarity and its implications are less well understood. Often, there is some question of whether the high similarity reflects common ancestry, or convergence to a similar structure from independent origins (analogy).

The reliability of statistical estimates can be evaluated by examining the estimates given in pairwise comparisons to unrelated sequences. The sequence comparison programs, SSEARCH and PSI-BLAST, report expectation values (E()-values) that are quite accurate (Altschul et al. 1997; Brenner et al. 1998; Pearson 1998). Some of the structure comparison programs also report probabilistic scores for their hits. Structal and VAST report probabilities, which can be converted to an E()-value by multiplying by the size of the database. Dali, Matras, and CE report only Z-scores, which can be converted to E()-values by assuming a distribution of similarity scores. For Dali and Matras, we assumed that the Z-scores were derived from an underlying extreme-value

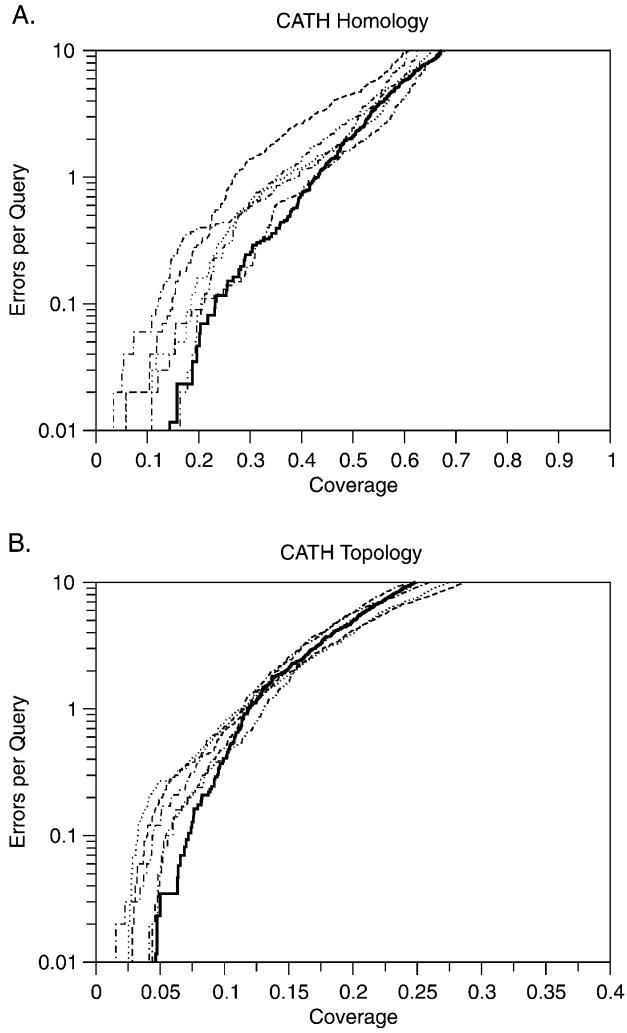

Figure 3. Errors per Query vs. Coverage plots for five independent query sets using the Structal method/LSQMAN program. (*A*) CATH Homologs and (*B*) CATH Topologs as the set of true positives. The data for the original set of queries is shown in bold.

distribution (EVD); Z-scores can then be converted to P() values using the formula

$$p(x > V) = 1 - \exp(-\exp(-x)) \quad (1)$$

where $p(x>V)$ is the probability of finding a score greater than V, and V is

$$V = \frac{\pi}{\sqrt{6}} \cdot Z - \gamma \quad (2)$$

where Z is the Dali Z-score and γ is the Euler constant. For CE, the authors derive their Z-scores from a normal distribution (Shindyalov and Bourne 1998), which we used as well. Pairwise-alignment probabilities were converted to E()-values for a database search by multiplying by the size of the database (2771). We then collected the expectation values of the best-scoring non-Homolog or non-Topolog for each of the 86 queries, calculated the Poisson probability of seeing the given E()-value,

$$p_{Poisson} = 1 - \exp(-E) \quad (3)$$

and plotted the expected versus observed (1/86, 2/86, 3/86,. . .) distribution of E()-values (Fig. 5). As has been shown previously (Brenner et al. 1998), SSEARCH provides very accurate statistical estimates; SSEARCH E()-values follow the ideal line almost perfectly. PSI-BLAST also is close to the ideal line, with a small number of alignments with E()-values less than 0.01 (the expected lowest E()-value within a set of 86 queries, as $1/86 \cong 0.011$). These are all short alignments between domains that are not structurally similar, and are either true non-homologs or may reflect errors in CATH domain boundaries. After these errors, the PSI-BLAST expectation values follow the ideal line closely.

In contrast, the distribution of structural expectation values deviates considerably from the ideal line. Structal, Dali, CE, VAST, and Matras all report E()-values that are below the diagonal line, indicating that they tend to overestimate the significance of a match. As seen in Figure 1, the simplest explanation for these errors is misclassification of non-Homologs. However, the pattern of E()-values for Topologs is similar to that for Homologs, with the exception that Dali improves significantly. Dali has only three E()-values less than 0.01, and, although not following the ideal line as well as SSEARCH or PSI-BLAST, comes closest to it of any of the structural comparison methods when searching for Topologs. Thus, when using a more generous definition of non-homology—non-Topology—most of the structural comparison methods appear to overestimate statistical significance.

Discussion

We analyzed the ability of various structural comparison programs to detect Homologs and Topologs as defined by the CATH database using 86 non-homologous protein families. We chose to use 86 queries, rather than do the more common all-versus-all comparison, to minimize the chance of bias by large families. Large families can also bias the results because small families contribute relatively less to the overall coverage. To see whether this was affecting our results, we created Errors versus Coverage plots in which the contribution of each family to the overall coverage was normalized for the size of the family, and we found no substantial changes in relative performance (data not shown).

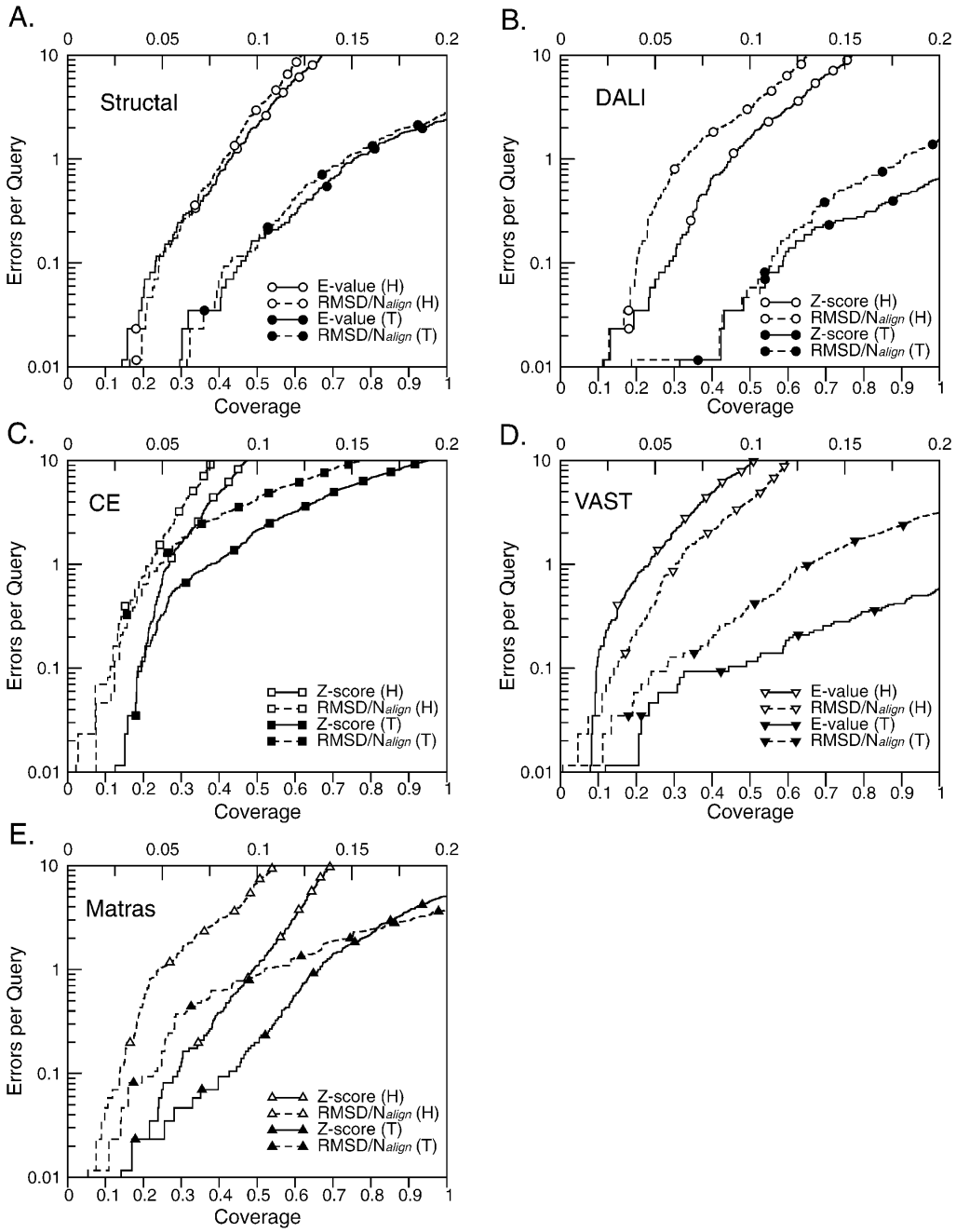

Figure 4. Errors per Query vs. Coverage plots comparing statistical (E()-value or Z-score) scores vs. RMSD/N$_{align}$ for Structal, Dali, CE, VAST, and Matras. (*A*) Structal/LSQMAN, (*B*) Dali, (*C*) CE, (*D*) VAST, and (*E*) Matras. RMSD/N$_{align}$ is shown by dashed lines; E()-value (Structal/VAST) or Z-scores (Dali/CE/Matras), by solid lines. Homolog true positive set, open symbols; Topolog true positive set, closed symbols. The coverage for Homologs is shown on the *lower* x-axis; that for Topologs is shown on the *upper* x-axis.

As pointed out by Brenner et al. (1998), there is a chicken-and-egg problem when evaluating structure comparison methods—one needs a "correct" homology classification against which to measure comparison programs, yet homology classifications are inferred from sequence and structure comparison. Our gold-standard subset of the CATH database is constructed with a structural alignment program, SSAP (Orengo et al. 1997), and, as stated in the introduction, there is no optimal alignment between two protein structures. Thus, some of the homology assignments may be incorrect; some proteins labeled "nonhomologous" may share a common ancestor, and other proteins may be labeled "homologous" because of similarity that arose by convergence. To address some of this uncertainty, we ex-

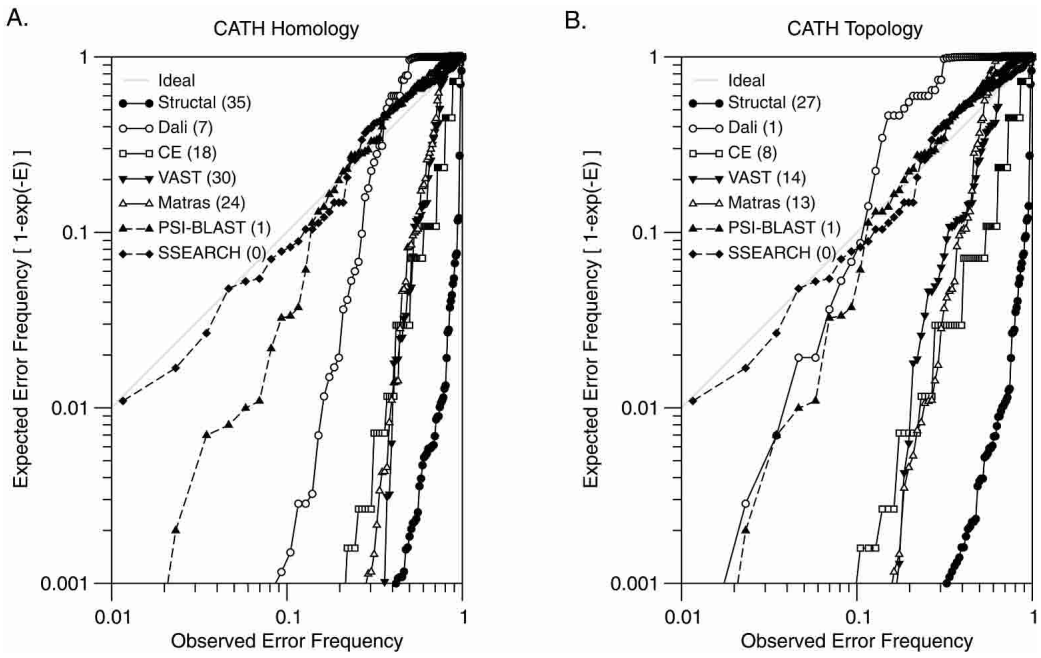

Figure 5. The expected Poisson probability of seeing the reported E()-value vs. the observed probability when searching for (*A*) CATH Homologs and (*B*) CATH Topologs for LSQMAN/Structal, Dali, CE, VAST, Matras, SSEARCH, and PSI-BLAST. The E()-values for the highest-scoring false positive for each query are shown. Lines and symbols are as in Fig. 1, except that the Z-scores for Dali, CE, and Matras (open symbols) were converted into E()-values (see text for details). The numbers in parentheses refer to the number of data points that have *y*-values less than 0.001.

amined four different error models (Fig. 1): (1) CATH Homologs are true positives (TP) and non-Homologs are false positives (FP; Fig. 1A), (2) Homologs (TP), non-Topologs (FP; Fig. 1B), (3) Topologs (TP), non-Topologs (FP; Fig. 1C), and (4) Topologs that are non-Homologs (TP), non-Topologs (FP; Fig. 1D). Using non-Topologs as false positives is more conservative, because it is unlikely that non-Topologs are homologous. With some exceptions, the relative performance of the methods tested was the same regardless of the error model, and we conclude that these results accurately reflect the general characteristics of the methods.

Evaluation of sequence and structural comparison methods addresses a fundamental biological question: "How can we reliably distinguish homology (divergence from a common ancestor) from analogy (convergence)?" This question differs from the more common focus on search coverage—identifying the largest number of homologs—as well as from the common goal of detecting structural similarity for the purpose of studying the protein folding problem (in which case one is less concerned about common ancestry). It is well recognized that sequences that share significant structural similarity—and are thus inferred to be homologous—may not share significant sequence similarity; homologous proteins often lack significant *sequence* similarity. Fortunately, however, there are no examples of proteins that share statistically significant sequence similarity and fold into different structures (a report by Sternberg and Islam [1990], to the contrary, uses an inaccurate measure of statistical significance; none of the examples in that paper share statistically significant sequence similarity). Hence sequence comparison is not concerned with analogy; nonhomologous sequences have E()-values less than 10^{-2} about 1 time in 100 (Fig. 5A; Pearson 1998).

Structures may also diverge such that they do not share significant similarity, but, unlike sequences, it is presumed that they can attain similar structures via convergent evolution. The question is, how similar? Can we determine a baseline of statistical significance above which we do not expect to find any nonhomologous domains? The results presented here suggest that we can, but with lower sensitivity than is generally acknowledged.

Previous studies (Russell et al. 1997; Matsuo and Bryant 1999) have shown that it is difficult to distinguish between homologous and analogous protein structures using automatic pairwise structure comparison, and our present results bear this out. The best-performing structural alignment program, Dali, can detect ~75% (840 out of 1120) of the Homolog pairs, but requires 860 errors, or 10 errors per query, to do so. However, 84% of those 860 errors are hits to domains with the same Topology as the query, with 65% of those domains being classified as TIM barrels or Rossmann folds. (Similar numbers obtain for Structal and VAST, but CE and Matras have much lower numbers of errors to Topologs).

A more accurate estimate of the true number of non-homologs found at 75% coverage can be inferred from Figures 1 and 5. Figure 5B indicates that Dali generally does not give statistically significant (i.e., e()-value <0.1) scores to domains that are non-Topologs, and Figure 1D shows that non-Homologous Topologs are generally not detected by Dali with statistical significance: The coverage is only ~3% at 0.1 EPQ and ~15% at 1 EPQ. Thus ~85% of non-Homologous Topolog pairs (probable analogs) are as different as non-Topolog pairs (clear analogs). This implies that about 15% of the non-Homologous Topologs might be misclassified; these are responsible for the differences between Figures 1A and 1B. Likewise, 165 of the 860 errors have Dali expectations <0.1, and 395 have expectations <1.0, suggesting that to identify 75% of homologs, 500 to 700 non-homologs will also be found.

Thus, to identify 75% of the homologs, Dali must accept two false positive errors for every three true positive Homologs. In contrast, the best sequence comparison method, PSI-BLAST, can only identify about half as many homologs, but achieves this sensitivity with only 10 false positives; PSI-BLAST is almost 30 times more selective. Indeed, at 0.1 EPQ, PSI-BLAST finds two-thirds of the Homologs found by Dali, the best-performing structure comparison method.

These conclusions must be qualified by considering Figure 2A, which shows the performance of each family independently. With Dali, half of the queries achieve greater than 80% coverage before hitting the first false positive, Structal and Matras achieve greater than 70%, and CE and VAST greater than 50%. Hence, the performance in Figure 1 is dominated by the worst-performing families. Nonetheless, at low error rates (EPQ <0.1), sequence comparison still performs as well as structure comparison. Thus, conclusions based on the data in Figure 1 (and other EPQ versus Coverage results) are conservative; homologs from the "average" protein family are more easily identified. Unfortunately, there is no simple way to predict in advance which families will perform poorly; family performance varies with different comparison methods, and does not depend upon simple criteria such as domain length or family size.

The difficulty in distinguishing homologs from analogs is reflected in the generally poor estimates of statistical significance available for structure comparison. Structural alignment programs often give statistical estimates that overestimate the significance of a hit, at least when searching CATH for Homologs, by several orders of magnitude. When searching for Topologs, converting Dali Z-scores to probabilities (equations 1 and 2) gives estimates that are reasonably close to the observed distribution of errors, but the other methods still dramatically overestimate the significance of structural similarities. (This remains true when counting errors only when the two domains are not in the same CATH Architecture; data not shown.) Although misclassification of Homologs by CATH may account for some of the poor statistical accuracy in Figure 5A, the poor performance on Topologs (Fig. 5B) probably reflects genuine error.

An expectation value estimates the probability of seeing a given level of structural similarity by chance, but it is unclear how to theoretically estimate the distribution of scores one expects to see by chance. Practically, however, one can establish some general upper and lower bounds. A common model for the baseline level of structural similarity seems to be that seen between domains in different classes (Levitt and Gerstein 1998). This is perhaps reasonable if one is interested in looking at structural similarity from a protein folding perspective, but we would argue that this baseline level is too low if one is interested in detecting homology. Given the physical constraints on protein folding, one would expect a certain level of structural similarity between two unrelated domains in the same class, and thus the distribution of scores between unrelated domains should include scores between domains in the same class, and possibly with the same architecture. Figure 5B suggests that, for Dali at least, more structural similarity than is seen for non-Topologs would reliably indicate homology, rather than analogy.

Thus, although structures diverge more slowly than sequences, making sequence comparison less sensitive than structure comparison, one must take care to account for structural analogy. At conservative statistical significance thresholds (E() <0.01), PSI-BLAST finds 32.0% of CATH Homologs and Dali detects 43.6%, an improvement of about 33%. At much higher false-positive rates, structure alignment can "identify" many more homologs; at E() <1.0, DALI finds 65.3% of CATH Homologs, whereas PSI-BLAST finds 36.3%, but the similar structures may include many convergent analogs. If statistically accurate significance is consistently required when inferring homology, the number of independent ancient homologs will be substantially larger than when homologs and analogs are grouped together.

Materials and methods

Selection of queries/library

A nonredundant subset of the CATH 2.3 (available at http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html) protein domain database was created using the CD-HI program (Li et al. 2001), such that none of the domains had more than 40% sequence identity with one another. First, all NMR structures, theoretical models, and X-ray structures with a resolution poorer than 3.0 Å were removed, as well as domains shorter than 20 residues, leaving a total of 25,506 of 27,396 domains. This list of sequences was then made nonredundant at 40% sequence identity by the CD-HI program, leaving 2870 domains. Domains not in classes 1–4 (all-α,

all-β, mixed α/β, and few secondary structures, respectively) were removed, leaving 2771 domains. These 2771 domains represent 1099 Homologous superfamilies and 623 Topologies. We then selected the longest domain from each of 86 CATH Homology families that had more than five members in their respective CATH Sequence family (35% sequence identity) as a query. (The queries were left in the target library.) These 86 queries represented 57 different CATH Topologies. Pairwise comparisons were carried out, with each query being compared to each member of the library. We refer to the 86 queries as our query set, to the 2771 library domains as our target library, and to pairwise comparisons as alignments.

Programs used

There are numerous programs available for comparing two protein structures. Time and space considerations limited our analysis to a subset of these. We evaluated programs that (1) are widely used, (2) report statistical scores, and (3) provided a variety of algorithms and scoring schemes. We used the standalone version of the Dali program (Holm and Sander 1996) called DaliLite (Holm and Park 2000), obtained from the Web site ftp://ftp.ebi.ac.uk/pub/contrib/holm/dl/, with default parameters. We used the Linux version of the Combinatorial Extension (CE) program (Shindyalov and Bourne 1998), obtainable at http://cl.sdsc.edu/ce.html, also with default parameters. We used the Structal method as implemented in the LSQMAN program (Kleywegt 1996) from the Uppsala Software Factory: http://xray.bmc.uu.se/usf/. Specifically, we used the Fast Force and Improve commands to get an initial alignment, then the DP command to implement the dynamic-programming method of Levitt and Gerstein (1998). We then used the Global command to calculate the statistics based on the Gerstein and Levitt structural similarity score (Levitt and Gerstein 1998). For VAST (Gibrat et al. 1996), we set up a local version of the program (available at ftp://ftp.ncbi.nih.gov/pub/pkb/). After printing out the text files in Splus that list the domains and secondary structure elements (SSEs) as calculated by the PKB system, we modified the domain definitions to fit the CATH domains. The statistics reported by VAST depend upon a specific definition of SSEs and their frequency in the database (T. Madej, pers. comm.), so in some cases the CATH segment or domain boundaries had to be trimmed or lengthened slightly so that they would not interrupt a VAST SSE definition. Also, for 67 of the domains in the library, VAST was unable to assign SSEs, meaning that the library for VAST consisted of 2704 domains. However, only one of the omitted domains is a true positive for any of the queries, and none of them appear in the first 100 errors of any of the other methods. Thus these omissions do not affect the results reported here. For Matras, we used the Linux version of the program provided by the authors (Kawabata and Nishikawa 2000) with default parameters. For SGM (Rogen and Fain 2003), we used an executable provided by the authors that calculates the Gauss integrals of input domains, then calculated the Euclidean distance between each of the query and target domains to derive the SGM, as described in (Rogen and Fain 2003). We also evaluated the PRIDE method (Carugo and Pongor 2002), which calculates a structural distance between proteins that is based upon internal contacts. This method did not perform as well as the SGM method, so we do not show the data from PRIDE.

For the sequence alignment programs, we used locally installed versions of the SSEARCH program (Smith and Waterman 1981; Pearson 1991; part of the FASTA package, available at ftp://ftp.virginia.edu/pub/fasta) and the PSI-BLAST program (Altschul et al. 1997; available at http://www.ncbi.nlm.nih.gov/blast/). SSEARCH was run using the BLOSUM50 matrix with gap open and extension penalties of −10 and −2, respectively. PSI-BLAST was run for a maximum of five cycles against an 95% identity nonredundant version of the BLAST nonredundant (nr) database, with a cutoff expectation value of 10^{-3} for inclusion in the next round of searching, to derive a position-specific scoring matrix (pssm) for each query. The BLOSUM62 scoring matrix was used with gap open/extension penalties of −11/−1. The query was then run once with the same scoring matrix and gap penalties against the target CATH library using the previously defined pssm.

Errors per Query versus Coverage plots

To create an Errors per Query (EPQ) versus Coverage plot, the list of pairwise comparisons is sorted on the score of interest (e.g., the Dali Z-score). Then the list is examined, from best score to worst. The coverage is increased if the two members of the pair are homologs, and the error is increased if they are not. The ideal curve would go from zero to 100% coverage before finding any errors. EPQ is the total number of errors at a given point, divided by the number of queries. Coverage is the total number of hits at a given point, divided by the total number of possible homolog pairs (Brenner et al. 1998).

For the individual query Errors versus Coverage plots (Fig. 2), the coverage for each query was recorded at the 1st, 3rd, 10th, 30th, and 100th false positive, and either the median of the coverage values or the coverage at the 25th percentile was plotted for each method. For the lower portion of the plots, the queries were sorted according to length and binned into groups of 10 (the group with the longest lengths had only six members). For each grouping, the coverage was recorded at the 1st, 3rd, and 10th false positive, with either the median or second worst coverage plotted.

Acknowledgments

We thank Steve Bryant and Tom Madej for help with VAST; Gerard Kleywegt for help with LSQMAN; and the other program authors for making their programs available. We also thank two anonymous reviewers for helpful comments on the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Altschul, S.F. and Gish, W. 1996. Local alignment statistics. *Methods Enzymol.* **266:** 460–480.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Brenner, S.E. and Levitt, M. 2000. Expectations from structural genomics. *Protein Sci.* **9:** 197–200.

Brenner, S.E., Chothia, C., and Hubbard, T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95:** 6073–6078.

Carugo, O. and Pongor, S. 2001. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* **10:** 1470–1473.

———. 2002. Protein fold similarity estimated by a probabilistic approach based on C(α)-C(α) distance comparison. *J. Mol. Biol.* **315:** 887–898.

Eidhammer, I., Jonassen, I., and Taylor, W.R. 2000. Structure comparison and structure patterns. *J. Comp. Biol.* **7:** 685–716.

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

Gribskov, M. and Robinson, N.L. 1996. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.* **20:** 25–33.

Holm, L. and Park, J. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics* **16:** 566–567.

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

———. 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26:** 316–319.

Karlin, S. and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.* **87:** 2264–2268.

Kawabata, T. and Nishikawa, K. 2000. Protein structure comparison using the markov transition model of evolution. *Proteins* **41:** 108–122.

Kendrew, J.C., Dickerson, R.E., Strandberg, G., Hart, R.G., Davies, D.R., Phillips, D.C., and Shore, V.C. 1960. Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Ångstroms resolution. *Nature* **185:** 422–427.

Kleywegt, G.J. 1996. Use of noncrystallographic symmetry in protein structure refinement. *Acta Crystallogr. D* **52:** 842–857.

Koehl, P. 2001. Protein structure similarities. *Curr. Opin. Struct. Biol.* **11:** 348–353.

Levitt, M. and Gerstein, M. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci.* **95:** 5913–5920.

Li, W., Jaroszewski, L., and Godzik, A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17:** 282–283.

Madej, T., Gibrat, J.F., and Bryant, S.H. 1995. Threading a database of protein cores. *Proteins* **23:** 356–369.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29:** 291–325.

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

Mott, R. 1992. Maximum-likelihood-estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull Math. Biol.* **54:** 59–75.

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. 2001. Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins* **Suppl. 5:** 2–7.

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

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

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284:** 1201–1210.

Pearson, W.R. 1991. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11:** 635–650.

———. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci.* **4:** 1145–1160.

———. 1998. Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276:** 71–84.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Perutz, M.F., Bolton, W., Diamond, R., Muirhead, H., and Watson, H. 1960. Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5 Ångstroms resolution, obtained by x-ray analysis. *Nature* **185:** 416–422.

Rogen, P. and Fain, B. 2003. Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci.* **100:** 119–124.

Russell, R.B., Saqi, M.A., Sayle, R.A., Bates, P.A., and Sternberg, M.J. 1997. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269:** 423–439.

Shindyalov, I.N. and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11:** 739–747.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Sternberg, M.J. and Islam, S.A. 1990. Local protein sequence similarity does not imply a structural relationship. *Protein Eng.* **4:** 125–131.

Vitkup, D., Melamud, E., Moult, J., and Sander, C. 2001. Completeness in structural genomics. *Nat. Struct. Biol.* **8:** 559–566.

Wu, C.H., Zhao, S., Chen, H.L., Lo, C.J., and McLarty, J. 1996. Motif identification neural design for rapid and sensitive protein family search. *Comput. Appl. Biosci.* **12:** 109–118.