Assessing sequence comparison methods with reliable structurally-identified distant evolutionary relationships

Steven E. Brenner[*†‡], Cyrus Chothia[*], and Tim J. P. Hubbard[§]

[*] MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

and

[§] Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, UK

[†] Current Address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126, USA

[‡] To whom reprints requests should be addressed. Email: brenner@hyper.stanford.edu

Author responsible for page proofs: Steven E. Brenner, Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5400, USA. Email: brenner@hyper.stanford.edu. Phone: +1 650-725-0754. Fax: +1 650-723-8464.

Classification: we request listing under both
Biological Sciences: Genetics and Biological Sciences: Biochemistry

23 Pages (including figure legends); 6 Figures; 1 Table

Abbreviations:

EPQ - Errors Per Query
CVE - Coverage Versus Error

Abstract

The pairwise sequence comparison procedures SSEARCH, FASTA, BLAST, and WU-BLAST2 have been assessed using a database of sequences whose relationships are known reliably from their structures and functions. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup=1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are greater than 30%. For more distantly related proteins, they do much less well; only half of the relationships between proteins with 20% to 30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Introduction

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that different procedures' overall and relative capabilities are largely unknown. It is difficult to verify algorithms on sample data, because this requires large data-sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that while previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment has also been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST(1) have changed, and WU-BLAST2(2)—which produces gapped alignments—has become available. The latest version of FASTA(3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports have also left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. This means that the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties which have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the scop database(4), which is derived from structural and functional characteristics(5). This provides a uniquely reliable set of homologs which are known independently of sequence comparison. Second, we employ an assessment method which jointly measures both sensitivity and specificity. This allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches, and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison

Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson(6, 7), who compared the three most commonly used programs. Of these, the Smith-Waterman algorithm(8) implemented in SSEARCH(3) is the oldest and slowest, but the most rigorous. Modern

heuristics have provided BLAST(1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA(3), which may be run in two modes offering either greater speed (ktup=2) or greater effectiveness (ktup=1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database(9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith-Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff & Henikoff(11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a pre-determined score, but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the Swiss-Prot database(12) and used Prosite(13) to define homologous families. Their results showed that the Blosum62 matrix(14) performed markedly better than the extrapolated PAM-series matrices(15) which had previously been popular.

A crucial aspect of any assessment is the data which are used to test the programs' ability to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and Prosite are principally created by using the same sequence comparison methods which are being evaluated. This creates a "chicken and egg" problem, and means for example, that new methods

would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies(16).

To surmount these sorts of difficulties, Sander and Schneider(17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, while shorter alignments require higher identity. (Other studies have also employed structures(18-20), but these focused on a small number of model proteins and were principally oriented towards evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution(21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics(22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins(24, 25), the mathematical tractability of statistical scores “is a crucial feature of the BLAST algorithm”(1). The validity of this scoring procedure has been tested analytically and empirically (see (2) and references in (24)). However, all large empirical tests used random sequences which may lack the subtle structure found within biological sequences(26, 27) and obviously do not contain any real homologs. Thus, while many researchers have suggested that statistical scores be used to

rank matches(24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A database for testing homology detection

Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not(29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the scop: structural classification of proteins database(4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The scop database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From scop, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB)(30) and created two databases. One (PDB90D-B) has domains which were all less than 90% identical to any other, while (PDB40D-B) had those less than 40% identical. The databases were created by first sorting all protein domains in scop by their quality and making a list. The

highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1323 domains which have 9,044 ordered pairs of distant relationships, or roughly 0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program(27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of scop may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the PDB's heavy over-representation of a small number of families(31, 32), while PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. While the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment data and procedure

Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we

compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST(1) version 1.4.9MP and WU-BLAST2(2) version 2.0a13MP, and the FASTA package version 3.0t76(3) provided FASTA and the SSEARCH implementation of Smith-Waterman(8). For SSEARCH and FASTA, we used Blosum45 with gap penalties -12/-1 (7, 16). The default parameters and matrix (Blosum62) were used for BLAST and WU-BLAST2.

The "coverage versus error" (CVE) plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice this is impossible to achieve, so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally-determined homologs which have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of non-homologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage versus error (CVE) plots, were devised to

understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of ROC plots(33, 34), but better represent the high degrees of accuracy required in sequence comparison and the huge background of non-homologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. This is an aspect which has been largely ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes

All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Figure 1.

Sequence Identity. Though it has long been established that percentage identity is a poor measure(35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold(17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Figure 2 shows one of the many pairs of proteins with very different structures which have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Figure 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Since one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percent identity detect just a fraction of the distant homologs found by statistical scoring. If one considers only the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith-Waterman raw scores perform better than percentage identity (Figure 1), but ln-scaling(7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores, for a 20% change in cutoff score could yield a ten-fold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds are also affected by matrix and gap parameters.

Statistical scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores), but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Figure 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus an E-value of 0.01 indicates that roughly one pair of non-homologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and this validates the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST should also be directly interpretable, but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database.

Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall detection of homologs and comparison of algorithms

The results in Figure 5a and Table I show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup=1 is nearly as effective as SSEARCH. FASTA ktup=2 and WU-BLAST2 are intermediate in ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup=1. WU-BLAST2 is slightly faster than FASTA ktup=2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally-known homologs (Figure 5b). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup=1, SSEARCH, and WU-BLAST2 programs are not significant when compared with variation in database composition and scoring reliability.

Figure 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize more than 90% of the homologous

pairs with 30-40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with less than 50 residues. Of sequences having 25-30% identity, 75% are identified by SSEARCH E-values. However while the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20-25% identity are detected and only 10% of those with 15-20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP(37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

Conclusion

The general consensus amongst experts (see (7, 24, 25, 27) and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity-masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup=1 perform best, though BLAST and FASTA

ktup=2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.[*]

Acknowledgments

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge) Benefactors' Scholarship and by the American Friends of Cambridge University.

[*]* Additional and updated information about this work, including larger and supplementary figures, may be found at http://sss.stanford.edu/sss/

References

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403-410.

2. Altschul, S. F. & Gish, W. (1996) *Meth. Enzymol.* **266,** 460-480.

3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444-2448.

4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536-540.

5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Meth. Enzymol.* **266,** 635-643.

6. Pearson, W. R. (1991) *Genomics* **11,** 635-650.

7. Pearson, W. R. (1995) *Protein Sci.* **4,** 1145-1160.

8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195-197.

9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Meth. Enzymol.* **266,** 41-59.

10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249,** 816-831.

11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17,** 49-61.

12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24,** 21-25.

13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24,** 189-196.

14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915-10919.

15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, suppl. 3, pp. 345-352.

16. Brenner, S. E. (1996) *Molecular Propinquity: Evolutionary and Structural Relationships of Proteins* (Ph.D. Thesis. University of Cambridge, Cambridge, England).

17. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56-68.

18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233,** 716-738.

19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1,** 89-94.

20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1,** 77-78.

21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14,** 971-993.

22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264-2268.

23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5873-5877.

24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nature Genet.* **6,** 119-129.

25. Pearson, W. R. (1996) *Meth. Enzymol.* **266,** 227-258.

26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12,** 215-226.

27. Wootton, J. C. & Federhen, S. (1996) *Meth. Enzymol.* **266,** 554-571.

28. Waterman, M. S. & Vingron, M. (1994) *Statistical Science* **9,** 367-381.

29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669-678.

30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Commission of the International Union of Crystallography, Cambridge), pp. 107-132.

31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Cur. Opin. Struct. Biol.* **7,** 369-376.

32. Orengo, C., Michie, A., Jones S, Jones DT, Swindells MB & Thornton, J. (1997) *Structure* **5,** 1093-1108.

33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39,** 561-577.

34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20,** 25-33.

35. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9-16.

36. Chung, S. Y. & Subbiah, S. (1996) *Structure* **4,** 1123-1127.

37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389-3402.

38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131,** 417-433.

39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32,** 9906-9916.

Figure Legends

1. **Coverage versus error (CVE) plots of different scoring schemes for SSEARCH Smith-Waterman.** All proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes, and assessed. The graphs show the coverage and errors per query for statistical scores, raw scores, and three measures using percentage identity.

In the CVE plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. PDB40D-B, contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of errors made per query (EPQ). Since there are 1323 queries made in the PDB40D-B all-versus-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores which correspond to the levels of EPQ and coverage are shown in Figure 4 and Table I.

The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower-right corner of the graph which corresponds to identifying many evolutionary relationships without selecting unrelated proteins.

Three measures of percentage identity are plotted. Percent identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the

aligned region as a percentage of the average length of the query and target proteins. The HSSP equation(17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith-Waterman raw scores and E-values were taken directly from the sequence comparison program.

a: Analysis of PDB40D-B database. **b:** Analysis of PDB90D-B database.

2. **Unrelated proteins with high percentage identity**. Hemoglobin β-chain (PDB code 1hds chain b, (38), left) and cellulase E2 (PDB code 1tml, (39), right) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by rasmol.

3. **Length and percentage identity of alignments of unrelated proteins in PDB90D-B**: Each pair of non-homologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

4. **Reliability of statistical scores in PDB90D-B**: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, while P-values are shown for BLAST and WU-BLAST2. If the scoring

were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ, but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration depending upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

5. **Coverage versus error plots of different sequence comparison methods**: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). **a**: PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup=1 and WU-BLAST2 are almost as good. **b**: PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, though at higher levels of error it becomes slightly worse than FASTA ktup=1 and SSEARCH.

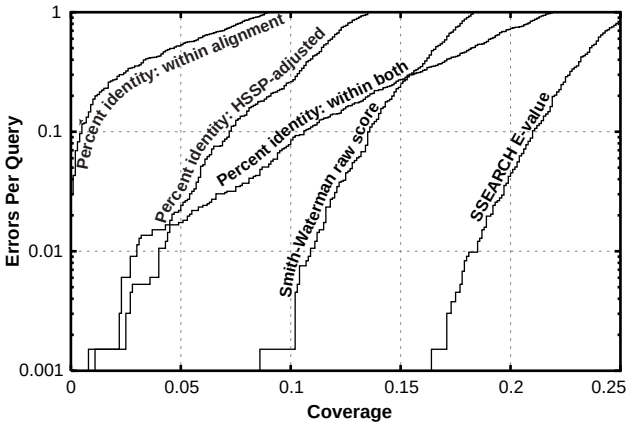
6. **Distribution and detection of homologs in PDB40D-B**. Bars show the distribution of homologous pairs in PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with less than 40% identity, and as shown on this graph, most structurally-identified homologs in the database have diverged extremely far in sequence and have less than 20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. The lower bars show that

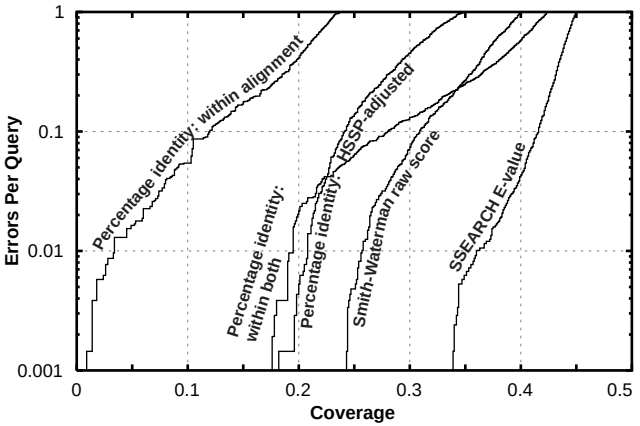
SSEARCH can identify most relationships which have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally-identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

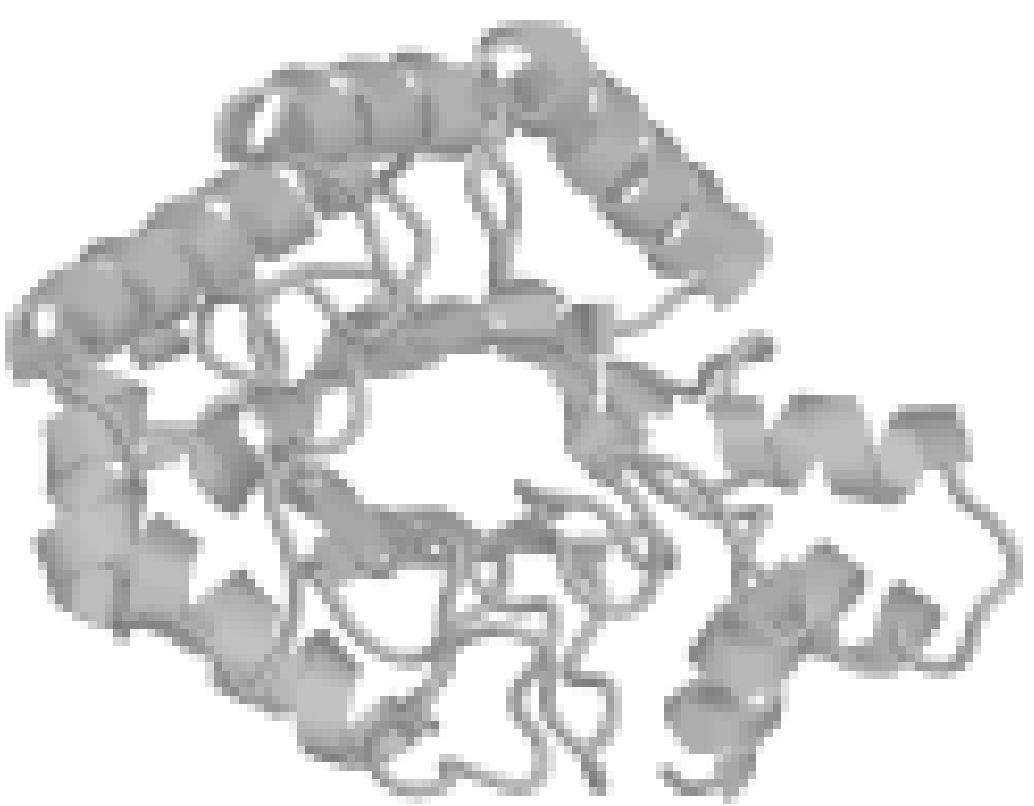
Table I: Summary of Sequence Comparison Methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP+9.8)	4.0
SSEARCH Smith-Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup=1 E-values	3.9	0.03	17.9
FASTA ktup=2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins

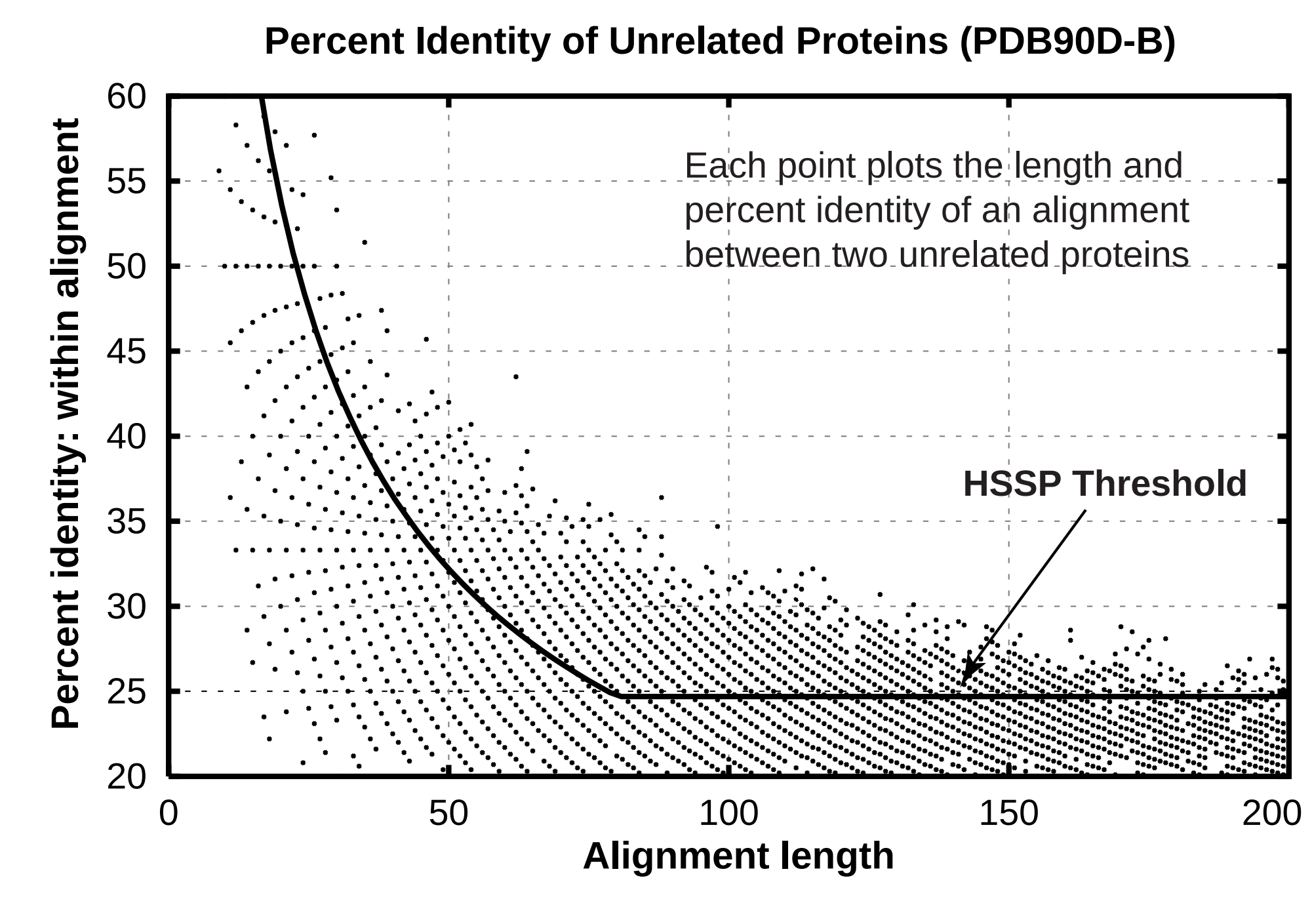

Smith-Waterman Scoring Schemes (PDB40D-B)
Errors Per Query
Coverage
1
0.1
0.01
0.001
0
0.05
0.1
0.15
0.2
0.25
Percent identity: within alignment
Percent identity: HSSP-adjusted
Percent identity: within both
Smith-Waterman raw score
SSEARCH E-value

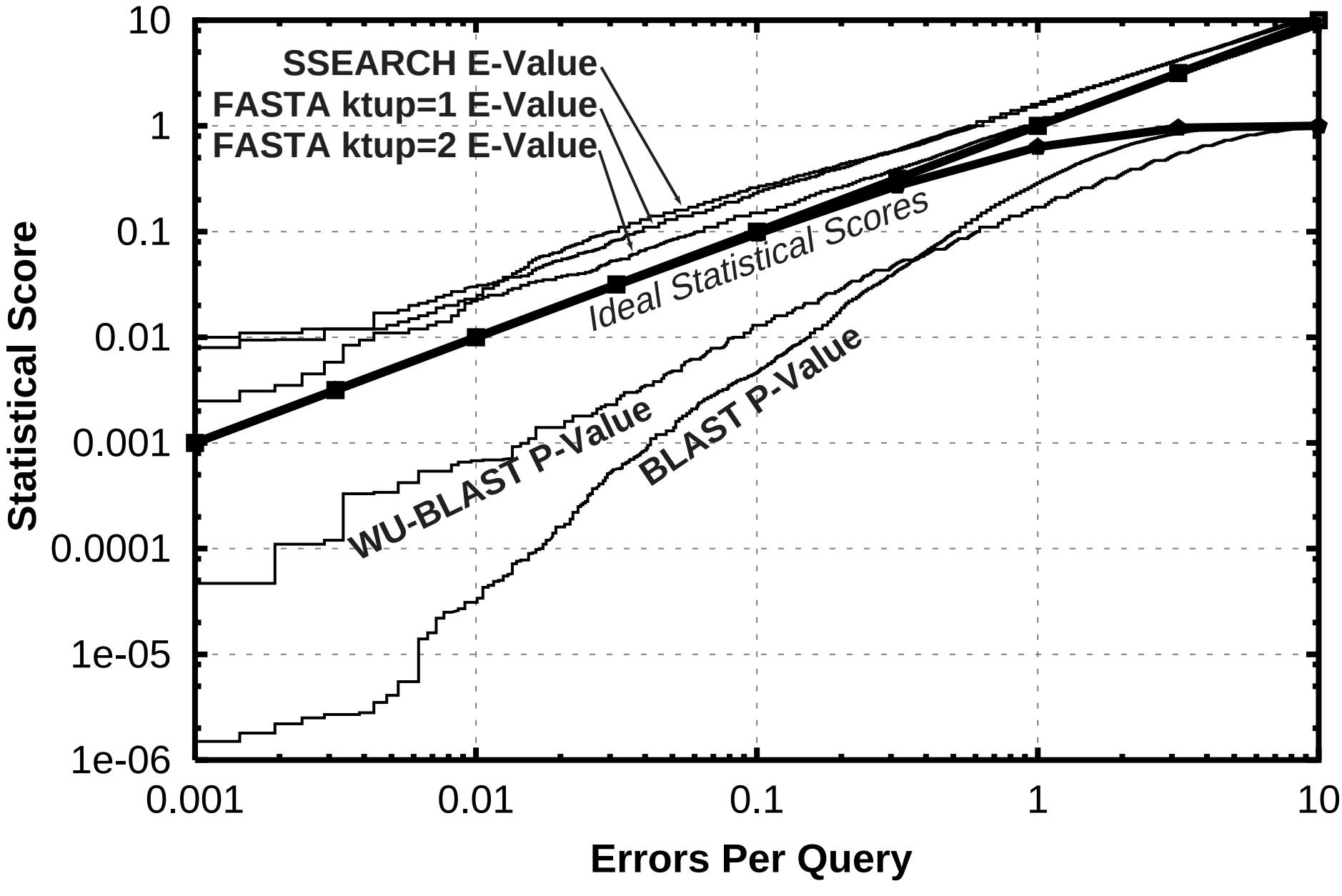

Smith-Waterman Scoring Schemes (PDB90D-B)
Errors Per Query
Coverage
1
0.1
0.01
0.001
0
0.1
0.2
0.3
0.4
0.5
Percentage identity: within alignment
Percentage identity: within both
Percentage identity: HSSP-adjusted
Smith-Waterman raw score
SSEARCH E-value

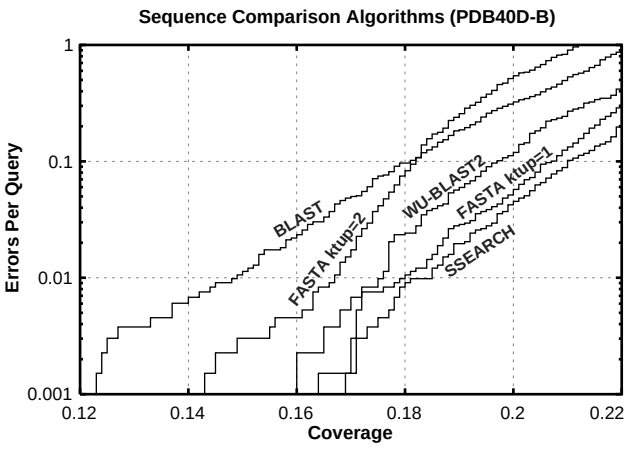


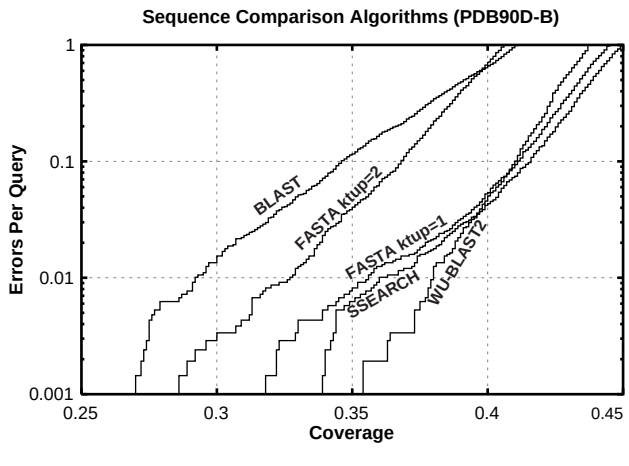
Hemoglobin β-chain (1hdsb) **Cellulase E2 (1tml_)**

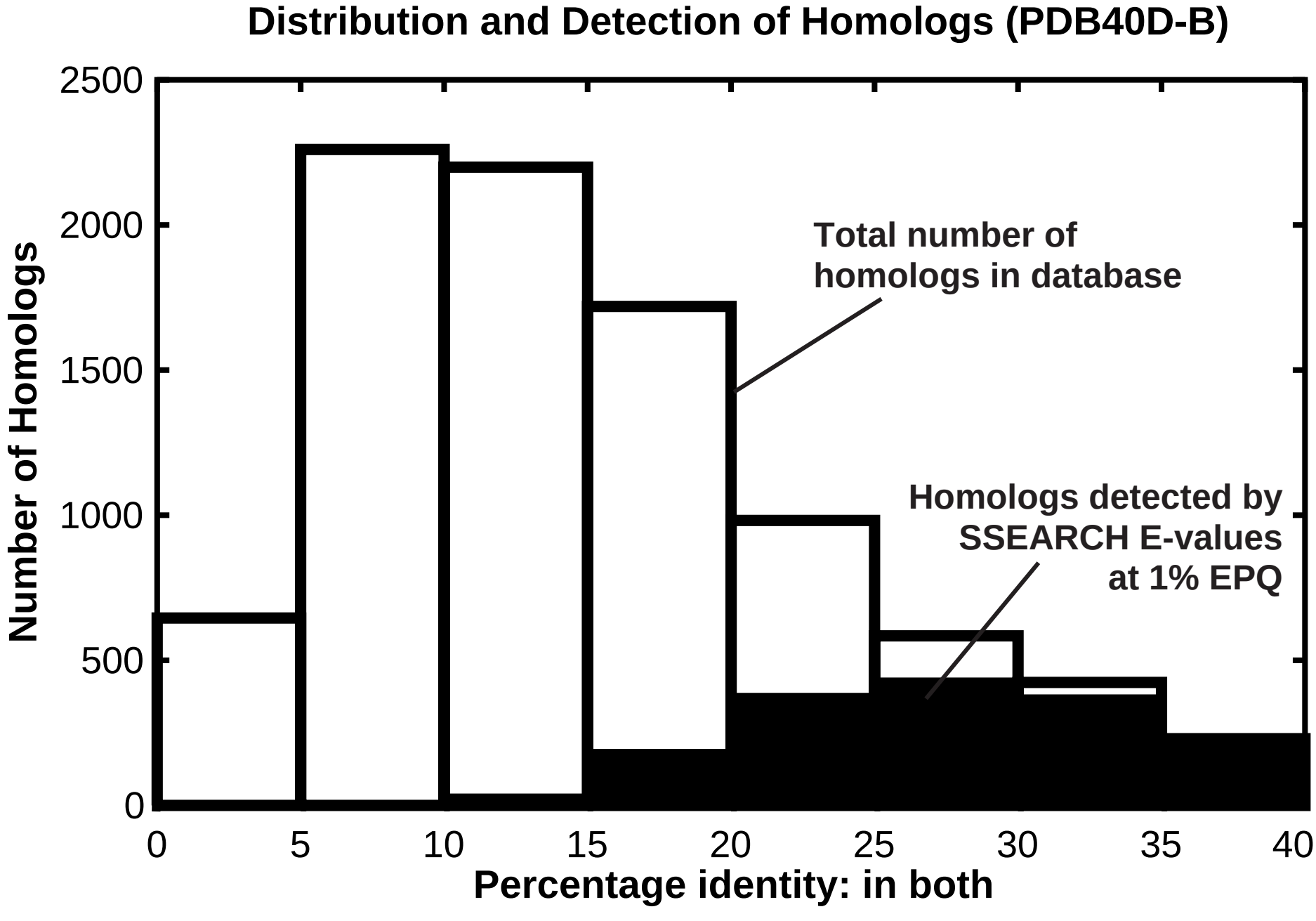
```
1hdsb GKVDVDVVGAQALGR--LLVVYPWTQRFFQHFGNLSSAGAVMNNPKVKAHGKRVLDAFTQGLKH
      :.::. . .::: :.  .:::: :    :: :::::   :. .:. .  .: :. :::.
1tml_ GQVDALMSAAQAAGKIPILVVYNAPGR---DCGNHSSGGA----PSHSAY-RSWIDEFAAGLKN
```


Percent Identity of Unrelated Proteins (PDB90D-B)
Percent identity: within alignment
60
55
50
45
40
35
30
25
20
0
50
100
150
200
Alignment length
Each point plots the length and
percent identity of an alignment
between two unrelated proteins
HSSP Threshold


Reliability of Statistical Scores (PDB90D-B)
SSEARCH E-Value
FASTA ktup=1 E-Value
FASTA ktup=2 E-Value
Ideal Statistical Scores
WU-BLAST P-Value
BLAST P-Value
Statistical Score
Errors Per Query
10
1
0.1
0.01
0.001
0.0001
1e-05
1e-06
0.001
0.01
0.1
1
10


Sequence Comparison Algorithms (PDB40D-B)
Errors Per Query
Coverage
1
0.1
0.01
0.001
0.12
0.14
0.16
0.18
0.2
0.22
BLAST
FASTA ktup=2
WU-BLAST2
FASTA ktup=1
SSEARCH


Sequence Comparison Algorithms (PDB90D-B)
Errors Per Query
Coverage
1
0.1
0.01
0.001
0.25
0.3
0.35
0.4
0.45
BLAST
FASTA ktup=2
FASTA ktup=1
SSEARCH
WU-BLAST2


Distribution and Detection of Homologs (PDB40D-B)
Number of Homologs
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30
35
40
Percentage identity: in both
Total number of
homologs in database
Homologs detected by
SSEARCH E-values
at 1% EPQ