The Limits of Protein Sequence Comparison?

William R. Pearson*

and

Michael L. Sierk

Department of Biochemistry and Molecular Genetics,
Jordan Hall, Box 800733
University of Virginia, Charlottesville, VA 22908, USA

January, 2005

Abstract

Modern sequence comparison algorithms are used routinely to identify homologous proteins, proteins that share a common ancestor. Homologous proteins always share similar structures, and often have similar functions. Since the advent of rapid sequence comparison programs and publicly available protein sequence databases more than 20 years ago, sequence comparison has become both more sensitive, largely because of profile-based methods, and more reliable, because of more accurate statistical estimates. Indeed, at low error rates, sequence comparison methods are beginning to approach structure comparison in their ability to recognize distant homologs. As sequence and structure databases become larger, and comparison methods become more powerful, reliable statistical estimates will become even more important for distinguishing similarities that are likely to have arisen because of homology from those that occur because of analogy (convergence). Unfortunately, statistical estimates for the most sensitive sequence and structure comparison methods often over estimate the statistical significance of the similarity, making it difficult to distinguish homology from analogy.

*FAX: (434) 924-5069; email: wrp@virginia.EDU

Introduction

It has been more than 20 years since the first rapid biological sequence comparison programs were developed [23, 46]. These programs, and their descendants [1, 2, 32], together with freely available DNA and protein sequence databases [3, 5], have revolutionized the practice of biochemistry, and molecular and evolutionary biology. Early sequence comparisons revealed extraordinary evolutionary relationships, e.g. the homology shared by v-sis and platelet-derived growth factor Since then, the inference of sequence homology from significant sequence similarity has become routine, and considerably more reliable.

Nonetheless, the inference of homology from similarity can be controversial. Perhaps this is expected, since such inferences often make assertions about molecules in organisms that lived billions of years in the past. Moreover, some of the links between similarity and homology include implicit assumptions about a fundamental biological process — the emergence of new protein structures and new protein families.

Discussions of homology are also confused because neither the relationship between sequence similarity and homology, nor the relationship between homology and common ancestry, are invertable.[1] Thus, sequences that share significant similarity are inferred to be homologous, but sequences lacking significant similarity need not be non-homologous. Likewise, proteins that are homologous share a common ancestor, but proteins that are not homologous (because they do not share significant similarity, so the inference of homology cannot be tested) may also share a common ancestor.

In this short review, we describe a logically consistent rationale for sequence and structure comparison, and some guidelines, for evaluating whether the inference of homology is likely to be justified. We show that current sequence comparison methods, and the DALI structure alignment program, provide accurate statistical estimates, which can be used to infer homology. Unfortunately, other methods, with high apparent sensitivity, are less useful in practice, because of unreliable statistical estimates. We suggest that future improvements in sequence, profile, and structure based homolog identification will involve a better understanding of random alignment scores.

[1] For the logical proposition, *if* $p \rightarrow q$, the inverse is *if* $\neg p \rightarrow \neg q$.

Homology and Statistical Significance - the argument from Parsimony

Whenever two protein sequences or protein structures are found that seem very similar, the similarity can be explained by one of two alternatives: (1) the two proteins are similar because they are homologous — they are both are descendents from a common ancestor; or (2) the proteins are not related, they are similar because some set of structural or functional constraints caused them to converge from independent origins to the observed similarity. Thus, in Fig. 1, panels A–C, three structures of trypsin-like serine proteases are shown. Sequence and structural similarity scores for those proteins are shown in Table I. From the structural perspective, all three trypsin like serine proteases are qualitatively very similar: they share the same symmetric beta-barrel structures; they have a single alpha-helix in a similar position, and the structures of the active sites are the same. Just as importantly, these trypsin-like serine protease structures look very different from the structures of other protein families. Trypsin-like serine proteases belong to the mainly-beta CATH [29] class of proteins, which includes 23 different mainly-beta barrel topologies distinct from the trypsin-like fold — ribbons, prisms, rolls, sandwiches, and propellors — and 813 different topologies altogether. Trypsin-like serine proteases have structures that are both similar to each other, and very different from other proteins.

The simplest explanation for the structural similarity of these proteins — the most *parsimonious* explanation — is that the trypsin-like serine protease structure arose once in evolution, and the proteins that share this structure do so because they diverged from that first trypsin-like serine protease. The alternative explanation — that the trypsin-line serine protease structure arose several times independently — requires the structure to be re-discovered, or re-invented, several times over evolutionary time — a less parsimonious explanation. This is the fundamental tension when homology, descent from a common ancestor, versus analogy, convergence from independent origins: Is the similarity sufficiently great that it seems unlikely that it could have occurred several times independently?

Subtilisin (Fig. 1E) demonstrates the alternative to homology. Subtilisin is also a serine protease, with exactly the same catalytic triad in the active site, but an overall three dimensional structure that is completely different from trypsin family members (CATH classifies subtilisin as an alpha-beta-alpha sandwich protein with a Rossman fold topology). Because there are thousands of other proteins more structurally similar to subtilisin than trypsin, parsimony makes no demand that the two structures share a common ancestor. Indeed, it would be more parsimonious to produce the subtilisin structure from some other protein family, one with more alpha-helices and no beta barrels. Subtilisin and trypsin are examples of convergent evolution to a common active site

from independent origins.

However, the case for independent origins can be much more subtle. For example, in the CATH classification of trypsin-like serine proteases, there are two other families of proteins that have a trypsin-like topologies, but differ from each other sufficiently to suggest that they arose independently One of the other families includes viral proteases (Fig. 1D), while the other includes ATP phosphorylases. While all three families have a similar beta-barrel topologies, the details of the strand geometry in the barrels suggest the three different families probably did not diverge from a common ancestor. Thus, the inference of homology — that two proteins share a common ancestor rather than arising independently — is based both on the degree of similarity that they share, *and* some sense of how unlikely that this similarity could have arisen independently. From this perspective, the inference of homology must be supported by some measure of statistical significance.

Sequence Similarity Statistics

The need to base the inference of homology on statistically significant similarity was recognized in the earliest days of protein sequence comparison, almost 40 years ago [10]. In those first comparisons, it was recognized that segments from unrelated proteins, or segments compared to random positions within the same protein, produced similarity scores that were indistinguishable from those produced from a normal distribution [10]. Today's most widely used sequence comparison algorithms [1,2,32,41] calculate local sequence alignment scores that are described by the extreme value distribution [18,25]. Once again, unrelated sequences have local alignment similarity scores that are very accurately described by mathematical models of random sequences [6,31]. This leads to a fundamental observation in pairwise sequence similarity searching:

> Sequence alignment scores for *unrelated* sequences are indistinguishable from scores from *random* sequences.
>
> Thus, if a similarity score is *not random*, then the sequences must be *not unrelated*
>
> Therefore, sequences that share *statistically significant similarity* are *homologous*.

We note that this syllogism does not make any statements about sequences that do not share statistically significant similarity; they may be related or unrelated. It simply states that because unrelated sequences have similarity scores that are indistiguishable from the scores of random sequences, statistically significant similarities come from homologous sequences.

This perspective on homology and statistical significance makes the implicit assumption that similarities readily appear by chance, so that finding a sequence with the highest similarity score out of 10^5 or 10^6 sequences is not surprising; indeed it occurs in every database search, whether a homolog is present or not. For protein sequences, this is a generally accepted assumption, in part because of the apparent lack of constraints on protein sequences [45], and supported by the observation that identical short sequences can adopt very different structures.

In structure comparison, however, there is less consensus that similar strucutures can arise independently. When surprising structural similarities are found, it it often suggested that these similarities may represent unrecognized ancient homologies [24, 36] or common functional roles. From this perspective, the re-discovery of a structural motif is extremely unusual, so that similar structural motifs may reflect either common ancestry, or convergence to a common (possibly structural) function. We believe that this perspective unnecessarily blurs the boundary between homology and analogy. In this review, we will rely on the argument from parsimony, so that homology is inferred only when there is more similarity (in sequence or structure) than is expected by chance.

Similarity, Significance and Alignments

Sequence and structural similarity scores are calculated from implicit alignments, so there is a strong relationship between the sensitivity of a sequence comparison method and the quality of the alignment it produces. However, from the statistical perspective outlined above, sensitivity — the ability assign statistically significant similarity scores to distant homologs — is distinct from alignment accuracy, for several reasons. First, methods that produce the most statistically significant scores for distant homologs balance two competing goals: producing good scores for homologs, and, at the same time, producing significantly worse scores for non-homologs. In contrast, alignment quality depends only on the behaviour of the method on homologs. Although, with improved statistical estimates and powerful multiple-sequence based methods like PSI-BLAST [2], it is now routine to identify homologs sharing considerably less than 25% amino-acid identity, it can still be difficult to produce accurate structural alignments for proteins that share less than 30 – 40% identity. [42, 43].

Recently, several authors have suggested that the accuracy of structural alignments, rather than search sensitivity, is a more useful measure of the effectiveness of sequence [9] and structure [20] comparison methods. This perspective highlights the difference between the inference of homology, and the associated alignments. For the molecular biologist or genome annotator, the

identification of a homologous sequence from a database search guarantees that the two proteins have similar structures, and often provides preliminary functional insights (homologous proteins often have similar functions, but many homologous proteins have different functions, and some proteins, e.g. Fig. 1, with similar functions are not homologous), even if the underlying alignment is wrong. But for the structural biologist interested in structure-function relationships or homology modeling, an accurate alignment, even between non-homologous proteins, can provide critical information.

Progress in sequence similarity searching

The development of Karlin-Altschul extreme value statistics [18] and their incorporation into the BLAST program [1] provided a firm statistical foundation for the inference of homology from local sequence similarity. Moreover, it provided the statistical foundation for additional improvements in search sensivity for other alignment programs as well [30, 31]. Moreover, new vectorized implementations of the Smith-Waterman algorithm [34, 47] have made it possible to do optimal protein similarity searches in a matter of minutes for comprehensive protein databases. Although additional improvements in pairwise sequence comparison statistics may be possible by treating low-complexity regions more accurately [48], it seems likely that the limits to searching with single sequences are near.

Today, the most powerful sequence-based comparison methods use sets of aligned sequences, either as profiles [12], Hidden Markov Models (HMM's) [8, 15, 21], or Position Specific Scoring Matrices [2, 13]. PSI-BLAST [2, 38] is an extremely sensitive comparison tool that has revealed similarites using sequences that previously were recognized only from structure comparison [4]. Like BLAST before it, PSI-BLAST seeks to provide accurate statistical estimates for the similarities it finds [38], though the estimates are less accurate than those provide by the SSEARCH implementation of Smith-Waterman [40]. Unfortunately, because of the iterative nature of PSI-BLAST, the inclusion of a non-homologous sequence in the position specific scoring matrix can be difficult to detect, with very misleading consequences.

Profile/HMM/PSSM methods are more sensitive than single-sequence comparison methods because they summarize the evolutionary history of a family and thus identify relatively invariant, and less-constrained, positions within the protein [28]. Recently, profile/HMM/PSSM based methods have been extended to provide profile-profile based comparisons [33, 35, 36, 50]. Like the profile-sequence based searching methods before them, they can provide tantalizing examples of unrecognized sequence similarities that may reflect structural similarity and homology [37]. Evaluation

of profile-profile comparison methods using Receiver Operator Characteristic (ROC) curves (see below) suggest that profile-profile methods perform about 20% better than PSI-BLAST [27, 44].

Nonetheless, profile-profile methods still fail to correctly identify similarities that can be identified through three-dimensional structure alignment (Fig. 3, 27) with programs like CE [39], DALI [14], Structal/LSQMAN [19, 22], and VAST [11]. The size of the gap, however, depends greatly on both the level of selectivity specified, and how overall performance is summarized — whether all errors in all families are grouped together, or considered separately. Moreover, there are large differences in the accuracy of the statistical estimates provided by the different approaches.

Evaluating Search Algorithms

If the inference of homology requires statistically significant sequence or structural similarity, then the best comparison methods must: (1) assign higher scores to homologous protein pairs than to non-homologs and (2) provide accurate statistical estimates, so that non-homologous proteins do not "appear" homologous based on an over estimate of statistical significance. The most common evaluations of sequence and structural comparison methods focus on the first criterion, the ability to rank related sequences above unrelated ones, frequently using "ROC" (Receiver Operator Characteristic) curves, which plot relationship between the number of false-positives and true positives (or false-negatives) [6, 16, 28, 40, 49, 50].

While ROC curves provide useful comparisons among different methods, identification of distant homologs poses some special problems: (1) if all pairwise alignments are plotted, protein families with many diverse members (e.g. globins, immunoglobulins, serine-proteases) contribute considerably more to the curve than families with fewer structures; (2) even when only one query is selected from each family, differences in family diversity can produce dramatically different ROC curves for different families [40]; (3) some ROC curves provide very little information about performance at low error rates, e.g. E()¡0.01, or one error per hundred queries; and (4) ROC curves provide comparative information when the correct answer is known; but they do not provide useful guidelines on how to select a score or statistical significance threshold that will produce the desired performance for novel protein families, or protein families lacking homologous three-dimensional structures.

To infer homology, or to identify pairs of sequences that are likely to have informative alignments, one needs an explicit statistical threshold that accurately predicts the performance of the method on novel protein families. For protein sequence comparison, the expectation or $E()$-value

calculated by SSEARCH [6, 31] provides a very accurate estimate of whether an alignment score is likely to occur by chance. PSI-BLAST [2, 38] provides less accurate estimates [40], but the false positive rate is quite low (Fig. 2). DALI [14] performs about as well as PSI-BLAST, and provides estimates that are considerably more accurate than other structure alignment methods (Fig. 2; [40]). In contrast, the estimates produced by COMPASS [36] and VAST [11] cannot be reliably used to identify homologs, because proteins with different topologies (which are very unlikely to share a common ancestor) can have similarities with expectation values many orders of magnitude lower than expected by chance.

The problem of family diversity, and the need for an accurate statistical threshold, is illustrated in Fig. 3, which summarizes the ability of different sequence, profile/PSSM, and structure comparison methods to identify homologs using the CATH [29] classification. Two criteria for identifying homologs are shown: reported statistical significance ($E() < 0.01$ and $E() < 1$), and empirical error rate (the first non-homolog or first non-topolog, using the CATH classification — the non-topolog criterion should avoid most miss-classified non-homologs). The median-bars in the middle of the boxes in Fig. 3 show the overall trends among the different sequence, profile, and structure-based search methods; the pairwise Smith-Waterman algorithm identifies only about 25% of the homologous proteins for the median performing family, while PSI-BLAST finds about 40% of homologs at the median, COMPASS finds around 60%, DALI almost 98% before the first non-topolog error, and VAST 50–60%, depending on the error criterion.

Although the overall trend is clear — structure comparison is better than profile-profile comparison, which is better than profile-sequence comparison, which is better than sequence-sequence comparison — the details of the trend are more complex. For example, the worst performing family with DALI identifies fewer homologs with any criterion than the worst performing family with SSEARCH or PSI-BLAST. Consistent with the results in Fig. 2, the median coverage of homologs for SSEARCH, PSI-BLAST, and DALI at $E() < 1$ is very close to the coverage at the first non-homolog, as expected for accurate statistical estimates (since the first non-homolog should have $E() \sim 1$. In contrast, both COMPASS and DALI appear to identify many more homologs at $E() < 1$ than at the first non-homolog, consistent with the observation that the $E() < 1$ statistical estimate greatly underestimates the number of false positives (and thus over estimating statistical signficance). While COMPASS is capable of finding many more homologs than SSEARCH or PSI-BLAST before scoring the first non-homolog, the statistical estimates it provides cannot be used to reliably set an error threshold.

Figs. 2 and 3 suggest that, while recent developments in profile-profile comparison appear capable of identifying distant relationships that cannot be detected by profile-sequence comparison

methods, to be reliable in practice, profile-profile methods will need much more accurate statistical estimates. More accurate statistics may reduce their apparent sensitivity; for COMPASS median coverage drops from about 85 to 60% at the first-non-homolog, which is lower than DALI, but considerably better than PSI-BLAST.

Although most structural alignment methods calculate unreliable statistical estimates, DALI estimates are comparable to those calculated by PSI-BLAST ([40], Fig. 2), a sequence-based method. The observation that one of the most sensitive structure comparison methods can also produce statistical estimates comparable in accuracy to the most reliable current sequence-based method, PSI-BLAST, supports the argument that the relationship between excess similarity and homology is not fundamentally different for sequences and structures. From this perspective, very "similar" structures can reoccur independently chance, just as similar sequences do, and arguments for homology, particularly for short domains in very different structural contexts, should be supported with accurate statistical estimates. Over the past 15 years, many of the most dramatic improvements in sequence similarity searching involved a better understanding of the statistical properties of unrelated sequences. It seems likely that future improvements in profile-profile searching, and structure comparison, will also involve a better understanding of the statistical behavior of unrelated sequences and structures.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

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

3. R. Apweiler, A. Bairoch, and C. H. Wu. Protein sequence databases. *Curr Opin Chem Biol*, 8:76–80, 2004.

4. L. Aravind and E. V. Koonin. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.*, 287:1023–40, 1999.

5. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucleic Acids Res*, 33 Database Issue:D34–D38, 2005.

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

7. R. F. Doolittle, M. W. Hunkapiller, L. E. Hood, S. G. Devare, K. C. Robbins, S. A. Aaronson, and H. N. Antoniades. Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, 221:275–277, 1983.

8. S. R. Eddy. Hidden markov models. *Curr. Opin. Struct. Biol.*, 6:361–365, 1996.

9. R. C. Edgar and K. Sjolander. A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, 20:1301–1308, 2004.

10. W. M. Fitch. An improved method of testing for evolutionary homology. *J Mol Biol*, 16:9–16, 1966.

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

12. M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, 84:4355–4358, 1987.

13. S. Henikoff and J. G. Henikoff. Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci*, 6:698–705, 1997.

14. L. Holm and C. Sander. Mapping the protein universe. *Science*, 273:595–603, 1996.

15. R. Hughey and A. Krogh. Hidden markov models for sequence analysis: extension and analysis of the basic method. *Comput Appl Biosci*, 12:95–107, 1996.

16. B. John and A. Sali. Detection of homologous proteins by an intermediate sequence search. *Protein Sci*, 13:54–62, 2004.

17. W. Kabsch and C. Sander. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA*, 81:1075–1078, 1984.

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

19. G. J. Kleywegt. Use of non-crystallographic symmetry in protein structure refinement. *Acta. Cryst.*, D52:842–857, 1996.

20. R. Kolodny, P. Koehl, and M. Levitt. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol.*, 346:1173–88, 2005.

21. A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology. applications to protein modeling. *J. Mol. Biol.*, 235:1501–1531, 1994.

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

23. D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

24. A. N. Lupas, C. P. Ponting, and R. B. Russell. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol*, 134:191–203, 2001.

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

26. S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

27. T. Ohlson, B. Wallner, and A. Elofsson. Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins*, 57:188–197, 2004.

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

29. F. Pearl, A. Todd, I. Sillitoe, M. Dibley, O. Redfern, T. Lewis, C. Bennett, R. Marsden, A. Grant, D. Lee, A. Akpor, M. Maibaum, A. Harrison, T. Dallman, G. Reeves, I. Diboun, S. Addou, S. Lise, C. Johnston, A. Sillero, J. Thornton, and C. Orengo. The cath domain structure database and related resources gene3d and dhs provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*, 33 Database Issue:D247–D251, 2005.

30. W. R. Pearson. Comparison of methods for searching protein sequence databases. *Prot. Sci.*, 4:1145–1160, 1995.

31. W. R. Pearson. Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, 276:71–84, 1998.

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

33. S. Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res*, 24:3836–3845, 1996.

34. T. Rognes and E. Seeberg. Six-fold speed-up of smith-waterman sequence database searches using parallel processing on common microprocessors. *Bioinformatics*, 16:699–706, 2000.

35. L. Rychlewski, L. Jaroszewski, W. Li, and A. Godzik. Comparison of sequence profiles. strategies for structural predictions using sequence information. *Protein Sci*, 9:232–241, 2000.

36. R. Sadreyev and N. Grishin. Compass: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, 326:317–336, 2003.

37. R. I. Sadreyev, D. Baker, and N. V. Grishin. Profile-profile comparisons by compass predict intricate homologies between protein families. *Protein Sci*, 12:2262–2272, 2003.

38. A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*, 29:2994–3005, 2001.

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

40. M. L. Sierk and W. R. Pearson. Sensitivity and selectivity in protein structure comparison. *Protein Sci*, 13:773–785, 2004.

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

42. A. Tramontano and V. Morea. Assessment of homology-based predictions in casp5. *Proteins*, 53 Suppl 6:352–368, 2003.

43. C. Venclovas. Comparative modeling in casp5: progress is evident, but alignment errors remain a significant hindrance. *Proteins*, 53 Suppl 6:380–388, 2003.

44. B. Wallner, H. Fang, T. Ohlson, J. Frey-Skott, and A. Elofsson. Using evolutionary information for the query and target improves fold recognition. *Proteins*, 54:342–350, 2004.

45. O. Weiss, M. A. Jimenez-Montano, and H. Herzel. Information content of protein sequences. *J Theor Biol*, 206:379–386, 2000.

46. W. J. Wilbur and D. J. Lipman. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA*, 80:726–730, 1983.

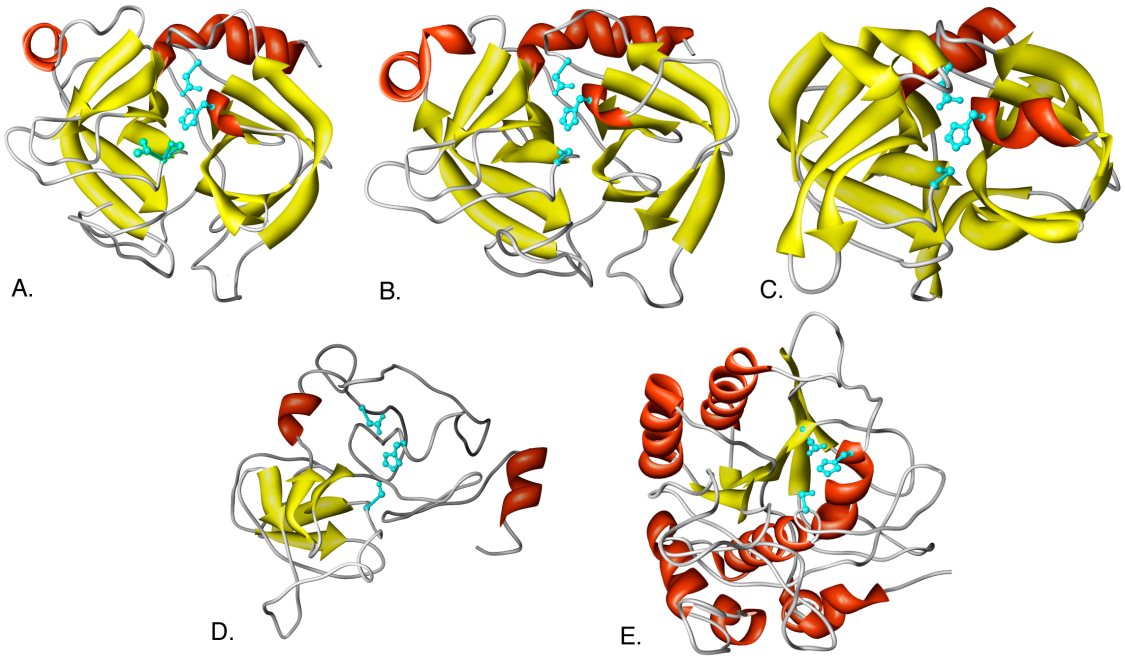
47. A. Wozniak. Using video-oriented instructions to speed up sequence comparison. *Comput Appl Biosci*, 13:145–150, 1997.

48. Yu Y.-K. and Altschul S. F. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, 21:(in press), 2005.

49. Y. Ye and A. Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19 Suppl 2:II246–II255, 2003.

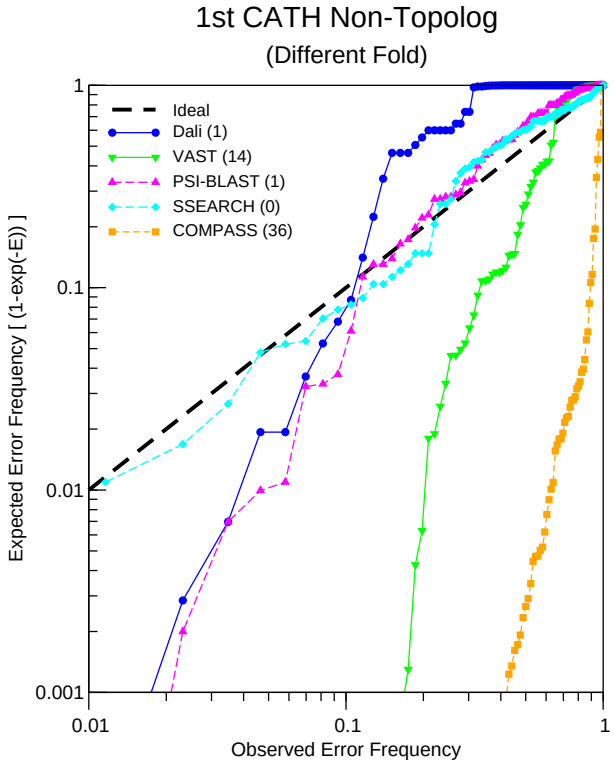
50. G. Yona and M. Levitt. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, 315:1257–1275, 2002.

Figure 1: Homologs, Analogs(?), and Convergent Evolution



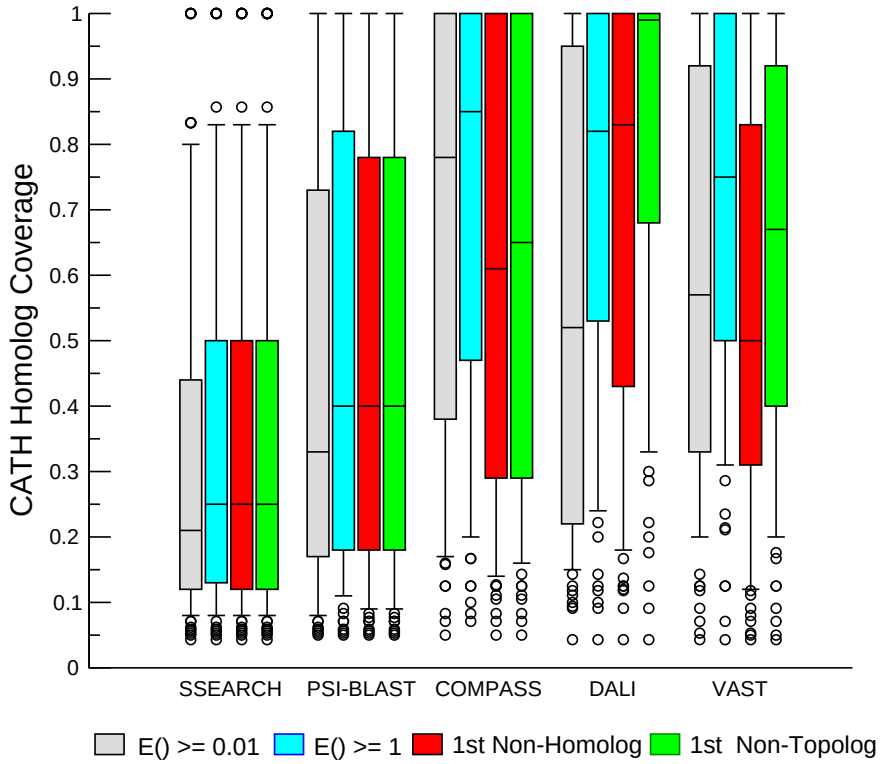
Three-dimensional structures of five serine proteases. (A) Bovine trypsin, PDB identifier 5PTP; (B) *S. griseus* trypsin (1SGT; (C) *S. griseus* protease A (1SGA); (D) viral serine protease (1BEP); (E) Subtilisin (1SBT). The CATH structure classification places 5PTP, SGT, SGA in the same homology category, while BEP has the same topology, but is classified as non-homologous to 5PTP. SCOP places 1BEF in the same superfamily. Subtilisin (1SBT) has a very different structure from the trypsin-like serine proteases and is clearly non-homologous. However, the active sites of subtilisin and trypsin are examples of convergent evolution.

Figure 2: Accuracy of statistical estimates



The expected Poisson probability of seeing the reported E()-value vs. the observed probability of seeing a domain with a different fold according to CATH (i.e. the domains have different CAT classifications) for SSEARCH, PSI-BLAST, COMPASS, Dali, and VAST. The E()-values for the highest-scoring false-positive (different topology) for each of 86 queries from different CATH homologous superfamilies are shown. The Z-scores reported by Dali were converted into E()-values assuming an extreme-value distribution (see 40 for details). The numbers in parentheses refer to the number of data points that have y-values less than 0.001.

Figure 3: Homologs found by different search methods



Boxplot of the CATH Homolog coverage achieved by 86 query domains from different CATH homologous superfamilies under different error criteria for SSEARCH [31], PSI-BLAST [2], COMPASS [36], DALI [14], and VAST [11]. The upper and lower edges of the boxes are at the 75th and 25th percentile, respectively, with the upper and lower whiskers at the 90th and 10th percentile, respectively. The middle line is the median amount of coverage, and the circles are the outliers. The fractions of CATH homologs identified at four thresholds are shown: the expectation value (E()-value) reported by the program greater than 0.01 (gray boxes); reported E()-value greater than 1 (blue); the first non-homolog according to CATH (red); the first non-topolog (different fold) according to CATH (green).