



ELSEVIER

# The limits of protein sequence comparison?

William R Pearson and Michael L Sierk

Modern sequence alignment 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. Over the past 20 years, sequence comparison has become both more sensitive, largely because of profile-based methods, and more reliable, because of more accurate statistical estimates. 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 due to homology from those that are due to analogy (convergence). The newest sequence alignment methods are more sensitive than older methods, but more accurate statistical estimates are needed for their full power to be realized.

## Addresses

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

Corresponding author: Pearson, William R (wrp@virginia.edu)

**Current Opinion in Structural Biology** 2005, 15:1–7

This review comes from a themed issue on Sequences and topology

Edited by Steven E Brenner and Anna Tramontano

0959-440X/\$ – see front matter

© 2005 Published by Elsevier Ltd.

DOI 10.1016/j.sbi.2005.05.005

## Introduction

It has been more than 20 years since the first rapid biological sequence comparison programs were developed [1,2]. These programs, and their descendants [3,4], together with freely available DNA and protein sequence databases [5,6], have revolutionized the practice of biochemistry, and molecular and evolutionary biology. Early sequence comparisons revealed extraordinary evolutionary relationships (e.g. the homology shared by  $\nu$ -sis and platelet-derived growth factor) [7]. Since then, the inference of homology from significant sequence similarity has become routine and considerably more reliable.

Nonetheless, the inference of homology from similarity can be controversial. Perhaps this should be expected, as such inferences often make assertions about molecules in organisms that lived billions of years in the past. More-

over, some of the links between similarity and homology include implicit assumptions about a fundamental biological process — the emergence of new (non-homologous) protein structures and new protein families.

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 SSEARCH, PSI-BLAST and the DALI structure alignment program provide accurate statistical estimates that can be used to infer homology. Some of the newer methods, although apparently more sensitive, are currently less useful in practice due to their 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.

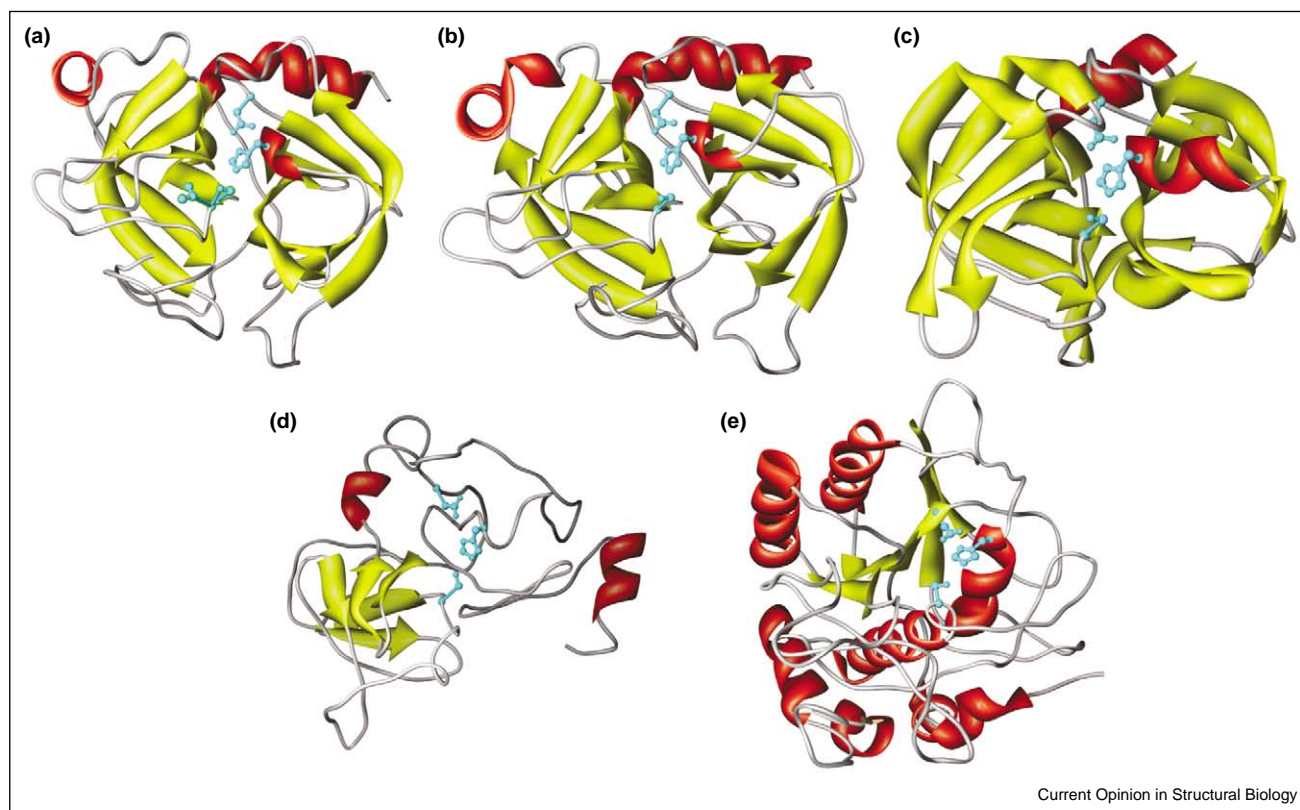
## Homology and statistical significance – the argument from parsimony

Whenever two protein sequences or protein structures seem very similar, the similarity can be explained by one of two alternatives: the two proteins are similar because they are homologous — both are descendants from a common ancestor; or 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 Figure 1a–c, the three trypsin-like serine proteases appear to be very similar to each other; this similarity is supported by statistical estimates (E(2775)) from DALI [8] and weakly supported by VAST [9] (Table 1). Structurally, all three trypsin-like serine proteases share the same symmetric  $\beta$ -barrel structure; they have two  $\alpha$  helices in similar positions and the structures of the active sites are the same. Just as importantly, these serine protease structures look very different from the structures of other protein families. Trypsin-like serine proteases belong to the mainly- $\beta$  CATH [10] class of proteins, which includes 23 different mainly  $\beta$ -barrel topologies distinct from the trypsin-like fold — ribbons, prisms, rolls, sandwiches and propellers — and 813 different topologies altogether. Trypsin-like serine proteases have structures that are both similar to each other and different from other proteins.

The simplest — most parsimonious — explanation for this structural similarity 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 alterna-

## 2 Sequences and topology

Figure 1



Current Opinion in Structural Biology

Homologs, analogs(?) and convergent evolution. Three-dimensional structures of five serine proteases: **(a)** bovine trypsin (PDB code 5PTP), **(b)** *Streptomyces griseus* trypsin (PDB code 1SGT), **(c)** *S. griseus* protease A (PDB code 2SGA), **(d)** viral serine protease (PDB code 1BEF) and **(e)** subtilisin (PDB code 1SBT). The CATH structure classification places 5PTP, 1SGT and 2SGA in the same homology category, whereas 1BEF has the same topology, but is classified as non-homologous to 5PTP. SCOP places 1BEF in the same superfamily as 5PTP. Subtilisin (1SBT) has a very different structure to the trypsin-like serine proteases and is clearly non-homologous. However, the active sites of subtilisin and trypsin are examples of convergent evolution.

tive explanation — that the trypsin-like 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. Thus, there is a fundamental tension between homology, descent from a common ancestor, and analogy, convergence from independent origins: is the observed similarity sufficiently great that it seems unlikely that it could have occurred several times independently?

Subtilisin is also a serine protease, with exactly the same catalytic triad in its active site; however, it demonstrates the alternative to homology — convergent evolution. Its overall three-dimensional structure is completely different from that of trypsin (CATH classifies subtilisin as an  $\alpha$ - $\beta$ - $\alpha$  sandwich protein with a Rossmann fold topology). With thousands of other proteins more structurally similar to subtilisin than trypsin, it would be more parsimonious to produce the subtilisin structure from some other protein family. Subtilisin and trypsin are examples of

convergent evolution to a common active site from independent origins.

However, the case for independent origins is often more subtle. In the CATH classification of trypsin-like serine proteases, there are two additional families of proteins with a trypsin-like topology; however, these differ from trypsin sufficiently to suggest that they arose independently. One comprises viral proteases (Figure 1d), the other ATP phosphorylases. Whereas all three families have similar  $\beta$ -barrel topologies, the details of the strand geometry of the barrels suggest that the three different families probably did not diverge from a common ancestor. SCOP [11] places the viral proteases in the same homologous superfamily as trypsin, perhaps due to their similar functions. Thus, the inference of homology — that two proteins share a common ancestor rather than arising independently — is based on both the degree of similarity that they share and some sense of how unlikely it is that this similarity could have arisen independently. From this perspective, the inference of

Table 1

Similarity for homologs and non-homologs<sup>a</sup>.

	5PTP versus:	1SGT	2SGA	1BEF	1SBT
DALI	Z	32.7	13.7	8.8	<2
	E(2775)	10 <sup>-14</sup>	10 <sup>-4</sup>	0.02	>100
	N <sub>align</sub> (%id)	209 (34)	147 (19)	131 (10)	NA
VAST	RMSD (Å)	1.4	2.8	2.9	NA
	E(2775)	10 <sup>-21</sup>	0.017 <sup>b</sup>	1.94	NA
	N <sub>align</sub> (%id)	208 (34)	130 (22)	122 (14)	NA
COMPASS	RMSD (Å)	1.5	2.3	2.8	NA
	E(10 000)	10 <sup>-114</sup>	10 <sup>-13</sup>	0.056	13
PSI-BLAST	E(2775)	10 <sup>-48</sup>	2.5	>10	>10
	N <sub>align</sub>	231	40	NA	NA
SSEARCH	E(10 000)	10 <sup>-19</sup>	2.6	>10	>10
	N <sub>align</sub> (%id)	223 (36)	181 (25)	68 (33)	159 (25)

<sup>a</sup> N<sub>align</sub> is the number of aligned residues, %id is the percent sequence identity and RMSD is the root mean square distance. E(N) is the expectation value, the number of times a score is expected by chance in a search of a database of size N. 5PTP, 1SGT and 2SGA are trypsin-like serine protease homologs; 1BEF is a viral protease that CATH says is not homologous to the trypsin-like proteases, but SCOP says is homologous to them. 1SBT is subtilisin, which has the same catalytic triad as the trypsin-like serine proteases, but a completely different global domain structure.

<sup>b</sup> Based on alignment with 5SGA, which is 100% sequence identical to 2SGA. NA: not available — no alignment was calculated.

homology should 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, when it was realized 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 [12]. Today's most widely used sequence comparison algorithms [3,4,13,14] calculate local sequence alignment scores that are described by the extreme value distribution [15,16]. Once again, unrelated sequences have local alignment similarity scores that are very accurately described by mathematical models of random sequences [17]. This leads to a fundamental observation on pairwise sequence similarity searching:

- Sequence alignment scores for *unrelated* sequences are indistinguishable from scores for *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.

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 indistinguishable from the scores of random sequences, statistically significant similarities come from homologous sequences. Implicit in this view is the assumption that similarities between protein sequences readily appear by chance.

For structure comparison, however, there is less consensus that similar structures can arise independently. When surprising structural similarities are found, it is often suggested that these similarities may represent unrecognized ancient homologies [18,19<sup>\*</sup>] or common functional roles. From this perspective, the re-discovery of a structural motif is extremely unusual, such that similar structural motifs may reflect either common ancestry or convergence to a common (possibly structural) function. We believe that this perspective blurs the distinction between homology and analogy, by invoking the argument from parsimony for objects that do not share statistically significant similarity. In this review, we invoke parsimony (and infer homology) only when sequences or structures share more similarity than is expected by chance.

### Similarity, significance and alignments

Recently, several authors have suggested that the accuracy of structure alignments, rather than search sensitivity, is a more useful measure of the effectiveness of sequence [20,21] and structure [22] comparison methods. Sequence similarity scores are calculated from implicit alignments; however, sensitivity — the ability to assign statistically significant similarity scores to distant homologs — is distinct from alignment accuracy. 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. By contrast, alignment quality depends on the behavior of the method on homologs only. Although it is now routine to identify homologs that share considerably less than 25% amino acid identity, it can still be difficult to produce accurate structure alignments for proteins that share less than 30–40% identity [23,24].

This focus on alignment quality underscores the difference between the inference of homology and the associated alignment. 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, 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, is more important than significant similarity.

### Progress in sequence similarity searching

The development of Karlin–Altschul extreme value statistics [15] and their incorporation into the BLAST

## 4 Sequences and topology

program [3] 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 sensitivity for other alignment programs as well [17,25]. In addition, new vectorized implementations of the Smith–Waterman algorithm [26,27] have made it possible to carry out optimal protein similarity searches with sequences or profiles in a matter of minutes for comprehensive protein databases. Although additional improvements in pairwise sequence comparison statistics may be possible by treating compositionally biased regions more accurately [28], it seems likely that the limits of searching with single sequences are near.

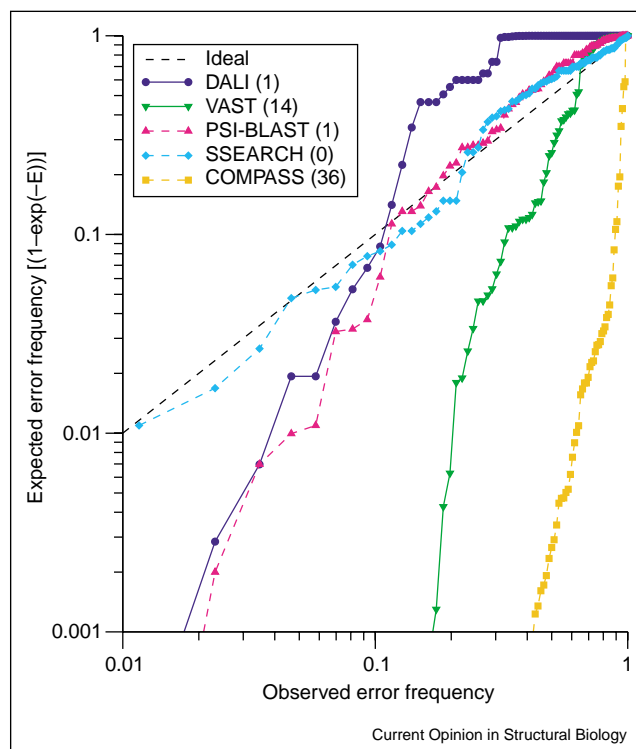
Today, the most powerful sequence-based comparison methods use sets of aligned sequences, either as profiles [29], hidden Markov models (HMMs) [30–32] or position-specific scoring matrices (PSSMs) [4,33]. There are also several profile databases, including PFAM [34], SUPER-FAMILY [35] and IMPALA [36], which can be searched using these methods.

PSI-BLAST [4,37] is an extremely sensitive comparison tool that has revealed homologies between sequences that previously were recognized only from structure comparison [38]. PSI-BLAST seeks to provide accurate statistical estimates for the similarities it finds [37], although it occasionally gives good scores to unrelated sequences (Figure 2). This sometimes produces misleading results, as the iterative nature of PSI-BLAST makes the inclusion of a non-homologous sequence in the PSSM difficult to detect.

Profile/HMM/PSSM methods are more sensitive than single-sequence comparison methods because they summarize the evolutionary history of a family, identifying more and less conserved positions within the protein [39]. Recently, profile/HMM/PSSM-based methods have been extended to provide profile-profile based comparisons [19\*,40–43]. 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 [44]. Evaluation of profile-profile comparison methods using receiver operator characteristic (ROC) curves (see below) suggests that profile-profile methods can identify about 20–30% more homologs than PSI-BLAST [19\*,41,42,45].

Nonetheless, profile-profile methods still fail to correctly identify similarities that can be identified through three-dimensional structure alignment (Figure 3) with programs such as CE [46], DALI [8], Strucal [47] and VAST [9]. The magnitude of the difference in performance, however, depends greatly on both the level of selectivity specified and how overall performance is summarized. There are also large differences in the accuracy of the statistical estimates provided by the different approaches.

Figure 2



Accuracy of statistical estimates. The expected Poisson probability of seeing the reported  $E_0$ -value versus the observed probability of seeing a domain with a different fold according to CATH (i.e. the domains have different CATH topology classifications) for SSEARCH, PSI-BLAST, COMPASS, DALI and VAST. The  $E_0$ -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  $E_0$ -values assuming an extreme value distribution (see [51\*\*] for details). The numbers in parentheses show the number of non-homologs with reported  $E_0 < 0.001$ .

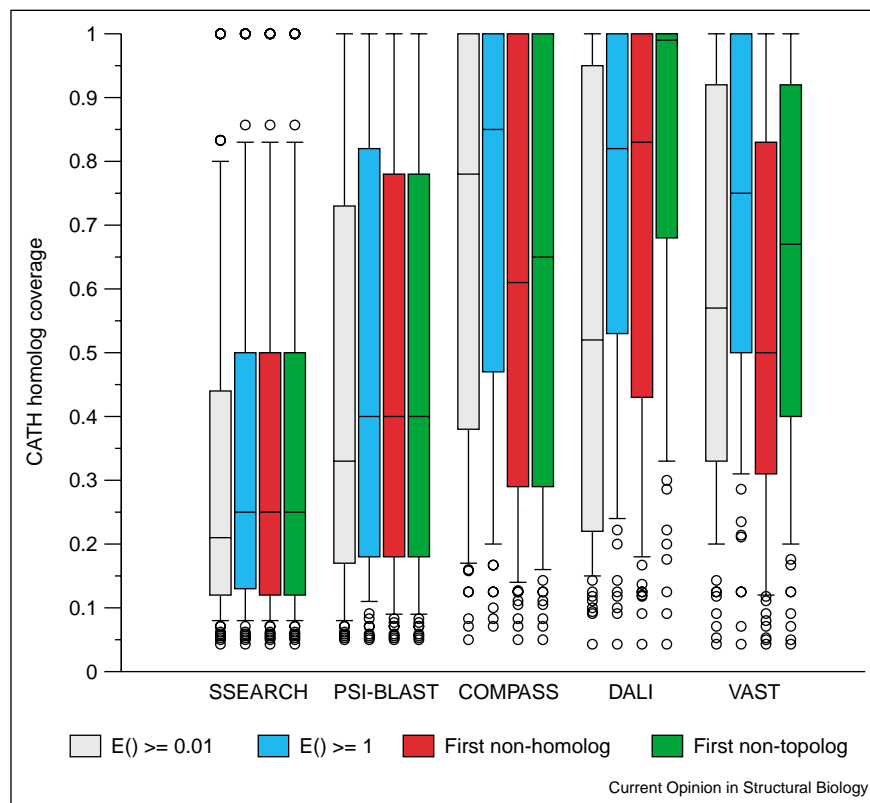
### Evaluating search algorithms

If the inference of homology requires statistically significant sequence or structural similarity, then the best comparison methods must: assign higher scores to homologous protein pairs than to non-homologs; and provide accurate statistical estimates, so that non-homologous proteins do not ‘appear’ homologous as a result of an overestimate of statistical significance. Many evaluations of sequence and structural comparison methods focus on the first criterion, the ability to rank related sequences above unrelated ones, frequently using ROC curves, which plot the relationship between the number of false-positives and true positives (or false-negatives) [19\*,21,37,39,42,48–50,51\*\*,52\*,53].

Although ROC curves provide useful comparisons of different methods, the identification of distant homologs poses some special problems. Firstly, if all pairwise alignments from the database are plotted, protein families with



Figure 3



Homologs found by different search methods. Box plot of the CATH homolog coverage achieved by 86 query domains from different CATH homologous superfamilies under different error criteria for SSEARCH [54], PSI-BLAST [4], COMPASS [19\*], DALI [8] and VAST [9]. 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. 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: reported  $E() > 0.01$  (gray boxes);  $E() > 1$  (blue); the first non-homolog according to CATH (red); the first non-topolog (different fold) according to CATH (green).

many diverse members (e.g. globins, immunoglobulins, serine proteases) contribute considerably more to the curve than families with fewer known structures. Secondly, even when only one query is selected from each family, differences in family diversity can produce dramatically different ROC curves for different families [49,51\*\*]. Thirdly, some ROC curves provide very little information about performance at the low error rates typically used in sequence comparison, for example,  $E() < 0.01$  or one error per one hundred queries (one error per 10 000 true positives, assuming 100 homologs per query). Finally, 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 provides a very accurate estimate of whether an alignment score is likely to occur by chance [17,48,49,51\*\*]. PSI-BLAST [4] provides less accurate estimates [51\*\*], but the false-positive rate is quite low (Figure 2). DALI [8] performs about as well as PSI-BLAST and provides estimates that are considerably more accurate than other structure alignment methods (Figure 2) [51\*\*]. By contrast, the estimates produced by COMPASS [19\*] and VAST [9] 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 consequences of family diversity and the need for accurate statistical thresholds are illustrated in Figure 3. 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 misclassified non-homologs). The median bars in the middle of the boxes in Figure 3 show the overall trends for the different search methods: the pairwise Smith–Waterman algorithm identifies only about 25% of homologous proteins for the median performing family, whereas 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. As expected for accurate statistical estimates, 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 the first non-homolog should have  $E() \sim 1$ , Figure 2). By contrast, both COMPASS and VAST appear to identify many more homologs at  $E() < 1$  than at the first non-homolog threshold, consistent with the observation that their statistical calculations greatly underestimate the number of false-positives (and thus overestimate statistical significance). Whereas COMPASS, which uses only sequence information, 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 set a reliable error threshold.

Figures 2 and 3 suggest that recent profile-profile comparison methods can identify distant relationships that cannot be detected by profile-sequence comparison methods. To be reliable in practice, however, these methods 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 worse than DALI, but considerably better than PSI-BLAST.

## Conclusions

Using the argument from parsimony, proteins can be inferred to be homologous — to share a common ancestor — when they share statistically significant similarity or more similarity than is expected by chance. Although many structure alignment methods calculate unreliable statistical estimates [51\*\*], DALI estimates are comparable to those calculated by PSI-BLAST (Figure 2). The observation that one of the most sensitive structure comparison methods can also produce statistical estimates comparable in accuracy to a reliable 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, similar structures can occur independently by chance, just as similar sequences do, and arguments for homology, particularly for short domains in 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 in structure comparison will also involve a better understanding of the statistical behavior of unrelated structures.

## Acknowledgements

The authors thank Bob Kretsinger for reviewing the manuscript. The National Institutes of Health (LM04969) funded this work.

## References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
  - of outstanding interest
1. Wilbur WJ, Lipman DJ: **Rapid similarity searches of nucleic acid and protein data banks.** *Proc Natl Acad Sci USA* 1983, **80**:726-730.
  2. Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches.** *Science* 1985, **227**:1435-1441.
  3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
  4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
  5. Apweiler R, Bairoch A, Wu CH: **Protein sequence databases.** *Curr Opin Chem Biol* 2004, **8**:76-80.
  6. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2005, **33**:D34-D38.
  7. Doolittle RF, Hunkapiller MW, Hood LE, Devare SG, Robbins KC, Aaronson SA, Antoniades HN: **Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor.** *Science* 1983, **221**:275-277.
  8. Holm L, Sander C: **Mapping the protein universe.** *Science* 1996, **273**:595-602.
  9. Gibrat JF, Madej T, Bryant SH: **Surprising similarities in structure comparison.** *Curr Opin Struct Biol* 1996, **6**:377-385.
  10. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH—a hierarchic classification of protein domain structures.** *Structure* 1997, **5**:1093-1108.
  11. Hubbard TJ, Murzin AG, Brenner SE, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res* 1997, **25**:236-239.
  12. Fitch WM: **An improved method of testing for evolutionary homology.** *J Mol Biol* 1966, **16**:9-16.
  13. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85**:2444-2448.
  14. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195-197.
  15. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87**:2264-2268.

16. Mott R: **Maximum-likelihood-estimation of the statistical distribution of Smith-Waterman local sequence similarity scores.** *Bull Math Biol* 1992, **54**:59-75.
17. Pearson WR: **Empirical statistical estimates for sequence similarity searches.** *J Mol Biol* 1998, **276**:71-84.
18. Lupas AN, Ponting CP, Russell RB: **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* 2001, **134**:191-203.
19. Sadreyev R, Grishin NV: **COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance.** *J Mol Biol* 2003, **326**:317-336.  
 A new profile-profile alignment program that extends the PSI-BLAST sequence-profile alignment algorithm is presented. The authors demonstrate its ability to detect remote homologies not detected by PSI-BLAST.
20. Edgar RC, Sjolander K: **A comparison of scoring functions for protein sequence profile alignment.** *Bioinformatics* 2004, **20**:1301-1308.
21. Ohlson T, Wallner B, Elofsson A: **Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods.** *Proteins* 2004, **57**:188-197.
22. Kolodny R, Koehl P, Levitt M: **Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures.** *J Mol Biol* 2005, **346**:1173-1188.
23. Tramontano A, Morea V: **Assessment of homology-based predictions in CASP5.** *Proteins* 2003, **53**(suppl 6):352-368.
24. Venclovas C: **Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance.** *Proteins* 2003, **53**(suppl 6):380-388.
25. Pearson WR: **Comparison of methods for searching protein sequence databases.** *Protein Sci* 1995, **4**:1145-1160.
26. Wozniak A: **Using video-oriented instructions to speed up sequence comparison.** *Comput Appl Biosci* 1997, **13**:145-150.
27. Rognes T, Seeberg E: **Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors.** *Bioinformatics* 2000, **16**:699-706.
28. Yu YK, Wootton JC, Altschul SF: **The compositional adjustment of amino acid substitution matrices.** *Proc Natl Acad Sci USA* 2003, **100**:15688-15693.
29. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins.** *Proc Natl Acad Sci USA* 1987, **84**:4355-4358.
30. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235**:1501-1531.
31. Hughey R, Krogh A: **Hidden Markov models for sequence analysis: extension and analysis of the basic method.** *Comput Appl Biosci* 1996, **12**:95-107.
32. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
33. Henikoff S, Henikoff JG: **Embedding strategies for effective use of information from multiple sequence alignments.** *Protein Sci* 1997, **6**:698-705.
34. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al.*: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32**:D138-D141.
35. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J: **The SUPERFAMILY database in 2004: additions and improvements.** *Nucleic Acids Res* 2004, **32**:D235-D239.
36. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF: **IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices.** *Bioinformatics* 1999, **15**:1000-1011.
37. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.** *Nucleic Acids Res* 2001, **29**:2994-3005.
38. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.** *J Mol Biol* 1999, **287**:1023-1040.
39. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.** *J Mol Biol* 1998, **284**:1201-1210.
40. Pietrokovski S: **Searching databases of conserved sequence regions by aligning protein multiple-alignments.** *Nucleic Acids Res* 1996, **24**:3836-3845.
41. Rychlewski L, Jaroszewski L, Li W, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information.** *Protein Sci* 2000, **9**:232-241.
42. Yona G, Levitt M: **Within the twilight zone: a sensitive profile-profile comparison tool based on information theory.** *J Mol Biol* 2002, **315**:1257-1275.
43. Edgar RC, Sjolander K: **COACH: profile-profile alignment of protein families using hidden Markov models.** *Bioinformatics* 2004, **20**:1309-1318.
44. Sadreyev RI, Baker D, Grishin NV: **Profile-profile comparisons by COMPASS predict intricate homologies between protein families.** *Protein Sci* 2003, **12**:2262-2272.
45. Wallner B, Fang H, Ohlson T, Frey-Skott J, Elofsson A: **Using evolutionary information for the query and target improves fold recognition.** *Proteins* 2004, **54**:342-350.
46. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Eng* 1998, **11**:739-747.
47. Levitt M, Gerstein M: **A unified statistical framework for sequence comparison and structure comparison.** *Proc Natl Acad Sci USA* 1998, **95**:5913-5920.
48. Brenner SE, Chothia C, Hubbard TJ: **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.** *Proc Natl Acad Sci USA* 1998, **95**:6073-6078.
49. Green RE, Brenner SE: **Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison.** *Proc IEEE* 2002, **90**:1834-1847.
50. John B, Sali A: **Detection of homologous proteins by an intermediate sequence search.** *Protein Sci* 2004, **13**:54-62.
51. Sierk ML, Pearson WR: **Sensitivity and selectivity in protein structure comparison.** *Protein Sci* 2004, **13**:773-785.  
 The first attempt to assess the accuracy of the statistical estimates provided by various structure alignment programs, in terms of detecting homologous domains in the CATH database. The authors demonstrate that many structure alignment programs overestimate the significance of structural similarities when searching the database for potential homologs.
52. Wang G, Dunbrack RL Jr: **Scoring profile-to-profile sequence alignments.** *Protein Sci* 2004, **13**:1612-1626.  
 A systematic assessment of different statistical models for comparing sequence profiles. The authors demonstrate that there is little difference between scoring systems once gap scores are optimized.
53. Madera M, Gough J: **A comparison of profile hidden Markov model procedures for remote homology detection.** *Nucleic Acids Res* 2002, **30**:4321-4328.
54. Pearson WR: **Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.** *Genomics* 1991, **11**:635-650.