



Increased coverage obtained by combination of methods for protein sequence database searching

Caleb Webber¹ and Geoffrey J. Barton^{1, 2,*}

¹EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England, UK and ²School of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, Scotland, UK

Received on April 12, 2002; revised on September 23, 2002; February 4, 2003; accepted on February 6, 2003

ABSTRACT

Motivation: Sequence alignment methods that compare two sequences (pairwise methods) are important tools for the detection of biological sequence relationships. In genome annotation, multiple methods are often run and agreement between methods taken as confirmation. In this paper, we assess the advantages of combining search methods by comparing seven pairwise alignment methods, including three local dynamic programming algorithms (PRSS, SSEARCH and SCANPS), two global dynamic programming algorithms (GSRCH and AMPS) and two heuristic approximations (BLAST and FASTA), individually and by pairwise intersection and union of their result lists at equal p -value cut-offs.

Results: When applied singly, the dynamic programming methods SCANPS and SSEARCH gave significantly better coverage ($p = 0.01$) compared to AMPS, GSRCH, PRSS, BLAST and FASTA.

Results ranked by BLAST p -values gave significantly better coverage compared to ranking by BLAST e -values.

Of 56 combinations of eight methods considered, 19 gave significant increases in coverage at low error compared to the parent methods at an equal p -value cutoff.

The union of results by BLAST (p -value) and FASTA at an equal p -value cutoff gave significantly better coverage than either method individually.

The best overall performance was obtained from the intersection of the results from SSEARCH and the GSRCH62 global alignment method. At an error level of five false positives, this combination found 444 true positives, a significant 12.4% increase over SSEARCH applied alone.

Contact: geoff@compbio.dundee.ac.uk

INTRODUCTION

The amount of publicly available DNA sequence data continues to double every 10 months (Stoesser *et al.*, 2002). Once probable open reading frames have been identified in a newly sequenced genome, the task is to assign a putative function to the coding regions on the basis of similarity to previously annotated proteins (e.g. Bork and Koonin, 1998).

Similarities may be found by pairwise sequence searches where a single sequence is scanned against each sequence in a database by dynamic programming or heuristic methods (for reviews, see Barton, 1996; Durbin *et al.*, 1998). More effective methods for identifying sequence similarity exploit multiple alignment profiles (Gribskov *et al.*, 1987; Barton and Sternberg, 1990) or hidden markov models (Sonnhammer *et al.*, 1997), but these rely on pairwise searches as the first step in finding sequences from which the initial profile is constructed (Altschul *et al.*, 1997; Sonnhammer *et al.*, 1997; Rychlewski *et al.*, 2000).

Although studies have been performed into the number of true homologues versus false identified by sequence searching methods (Pearson, 1991, 1995, 1998; Brenner *et al.*, 1998), less consideration has been given to whether such methods are identifying the same homologues at a given level of error. Evaluation is complicated since methods vary, not only in terms of the algorithm and its implementation (Barton, 1996; Durbin *et al.*, 1998), but also in the techniques applied to assess significance (Pagni and Jongeneel, 2001). For example, while BLAST (Altschul *et al.*, 1990, 1997) employs preset parameters to estimate significance from Extreme Value (EV) distributions, FASTA (Pearson and Lipman, 1988) and SSEARCH (Pearson, 1991, 1995) estimate these parameters from EV distributions fitted to scores for sequences thought to be unrelated to the query sequence in each database search.

*To whom correspondence should be addressed.

Differences in performance may also result from the use of alternative substitution matrices and gap penalties. The use of multiple substitution matrices for sequence searching has been investigated previously (Altschul, 1991; Henikoff and Henikoff, 1993). Altschul (Altschul, 1991) suggested that employing multiple matrices may find more homologues over a wider evolutionary distance and supported this with evidence from four sequence searches. However, Henikoff and Henikoff (1993) tested Altschul's suggestion on a subset of PIR 9.0 (Barker *et al.*, 2001) with BLAST 1.2.9 (Altschul *et al.*, 1990), and found no improvement over the best single matrix.

Variations in any chosen search method may lead to different homologues detected at the same error or *p*-value. Exploiting these differences could result in increased sensitivity for sequence searching. This is of particular importance in genome annotation systems that employ multiple pairwise alignment methods to assign similarity (Andrade *et al.*, 1999; Gaasterland and Sensen, 1996).

In this paper, the effect of combining multiple pairwise sequence search methods is evaluated. Score-ordered lists for eight pairwise alignment methods were obtained and these lists examined by intersection and union to investigate the effect on coverage of true positives compared to the parent methods at equivalent levels of error.

METHODS

Benchmarking

The success of methods for sequence similarity searching was tested by a SCOP-based (Murzin *et al.*, 1995) benchmark. This benchmark is derived from earlier benchmarks by Brenner *et al.* (1998) and makes use of the relationships defined between protein domains within the SCOP database. The SCOP database provides a detailed hierarchical description of the structural and evolutionary relationships between proteins whose 3D structure is known. The lowest level of the SCOP hierarchy is the 'family' where proteins are grouped together that share clear sequence, structural, and functional similarity. Above this is the 'superfamily', formed from families whose structural and functional features suggest a probable evolutionary relationship. Superfamilies are then classified into 'folds' if they share the same major secondary structures in the same arrangement and with the same topological connections.

The benchmark data set comprises 1113 protein domain sequences, taken from the PDB40D-B data set, version 1.37 (Brenner *et al.*, 1998), representing 479 SCOP superfamilies across 343 SCOP folds. Within this benchmark there are 2528 true positives, defined as a pair of protein domain sequences belonging to the same SCOP superfamily, and 616 293 true negatives, defined as a pair of sequences belonging to different SCOP folds.

The benchmark is performed by searching the benchmark data set with each of the protein domain sequences in turn. The score for each of the 618 821 benchmark pairs is collected and ranked from best to worst. This ordered list is then parsed with each pair scored as either a true positive or true negative (i.e. false positive.) For clarity in this work, all benchmarking results are simply reported as the number of true positives identified against the number of false positives reported at a given cut-off, rather than conversion to other measures such as percentage coverage and percentage error. However, given a total of 2528 true-positives within this benchmark, each 25.28 true positives correctly identified by a method at a given cut-off is equal to additional 1% coverage of the total.

Search methods

Seven sequence similarity search methods were considered, BLAST 2.0 (Altschul *et al.*, 1997), FASTA 3.0 (Pearson and Lipman, 1988), SSEARCH (Pearson, 1996), PRSS (Pearson and Lipman, 1988), AMPS *p*-values (Barton and Sternberg, 1987; Webber and Barton, 2001), SCANPS (Barton, 1993), and GSRCH (unpublished). In order to aid the empirical statistical estimates of FASTA, SSEARCH, and SCANPS, the benchmark data set was embedded within the NRDB90 database (Park *et al.*, 2000).

BLAST (Altschul *et al.*, 1990, 1997) and FASTA (Pearson and Lipman, 1988; Pearson, 1991, 1998) are both heuristic alignment methods. Both BLAST and FASTA make an initial calculation to estimate whether aligning the query sequence against a particular database sequence is likely to give an insignificant score and if so, further alignment is not carried out. A consequence of this is that the methods do not report an alignment score for every pairwise sequence comparison.

SCANPS implements the Smith–Waterman algorithm (Smith and Waterman, 1981), while PRSS (Pearson and Lipman, 1988) and SSEARCH (Pearson, 1996) implement the algorithm as modified by Green (1993).

AMPS and GSRCH are both global alignment methods. AMPS implements the Needleman and Wunsch (1970) dynamic programming algorithm while GSRCH implements the Sellers algorithm (Sellers *et al.*, 1974; Waterman *et al.*, 1976). Both methods are employed here with length-independent gap penalties that do not penalize overhanging sequence at the ends of the alignment and 100 randomizations were performed to calculate the *Z*-score. However, while AMPS randomizes both sequences to derive a *Z*-score, GSRCH only randomizes the query sequence. Global alignment *Z*-scores were subsequently converted to *p*-values (Webber and Barton, 2001). Both methods can be obtained from the authors.

BLAST, AMPS, and GSRCH calculate the significance of an alignment score on the basis of pre-calculated prob-

ability distributions. While BLAST bases estimates on extreme value statistics parameterized from distributions of random sequence scores (Altschul and Gish, 1996; Altschul *et al.*, 1997), AMPS and GSRCH base their estimates on distributions of 3D-structurally unrelated protein sequence alignment scores (Webber and Barton, 2001). FASTA, SCANPS, and SSEARCH estimate the significance of a given score by comparison to an extreme value distribution (EVD) fitted to scores obtained from the database search which are deemed unrelated to the query. PRSS estimates score significance by fitting an EVD to a distribution of alignment scores obtained by repeatedly shuffling and re-aligning the original sequences (Pearson, 1998).

p -Values were used throughout in order to equate the scoring schemes reported by each of the methods. Since BLAST, FASTA, and SSEARCH only report e -values, the p -values were back-calculated. For FASTA and SSEARCH, the p -value was calculated as the e -value divided by the number of sequences within the database deemed unrelated, as reported in the program output. For BLAST, the p -value was approximately calculated as the e -value divided by N/n , where N is the edge-corrected cumulative length of the database sequences, as reported in the program output, and n is the length of the database sequence (Altschul and Gish, 1996; Altschul *et al.*, 2001).

Combining search methods

From the benchmark, p -value ordered lists of sequence pairs were obtained for each method. Sets from each method were formed by taking all those pairs scoring below a given p -value threshold. The pairwise union and intersection of these sets was found and the number of false and true positive pairs recorded for each. This process was repeated for each of 1400 values of p between $p = 10^{-7}$ and $p = 5 \times 10^{-3}$ and plots drawn of true positives against true negatives for the pairwise intersection and union of the methods at equal p -value.

For example, Figure 1 shows the sets formed by BLAST and FASTA at a p -value cut-off of 3.5×10^{-5} . At this p -value threshold, BLAST hits 386 true positives and 10 false positives, and FASTA hits 355 true positives and nine false positives. Taking the intersection of these sets gives 331 true and four false positives, while the union yields 410 true and 15 false positives. The new sets formed by intersection and union at equal p -value thresholds can then be compared to the coverage yielded by the parent methods, in this example BLAST and FASTA, at the corresponding new error. For example, at 15 false positives, BLAST hits 392 true positives and FASTA hits 382 true positives, compared to 410 true positives for the union of BLAST and FASTA results at equal p -value. At an error of four false positives, BLAST hits 371 true positives and FASTA hits 337 true positives, compared to

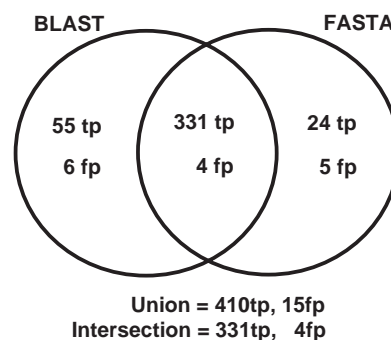


Fig. 1. Set interactions between BLAST and FASTA hits found below a p -value cut-off of 3.5×10^{-5} . The numbers of true and false positives are given as **tp** and **fp**, respectively.

the 331 true positives for the intersection of BLAST and FASTA results at equal p -value.

Given two equally efficient search methods, taking the union of the output of these methods will give increased coverage of true positives if the methods find dissimilar true positives but similar false positives, above a given threshold. Likewise, taking the intersection of their output will increase coverage if the methods find similar true positives but dissimilar false positives, above a given threshold.

Statistical methods

The significance of differences found between methods, or combinations of methods, can be determined by McNemar's test (Bland, 1987). McNemar's test is similar to a Sign-Test, and is applicable where little is known about the underlying distributions of the samples tested except that the samples are paired, as is the case here. To determine whether the difference in the number of correctly-identified homologous relationships (true positives) between two methods is significant, the number of true positives found by method A but not by method B (**A not B**), and the number found by method B but not method A (**B not A**) is recorded. The χ^2 value is then calculated as

$$\frac{|\mathbf{A \ not \ B} - \mathbf{B \ not \ A}|^2}{\mathbf{A \ not \ B} + \mathbf{B \ not \ A}} \quad (1)$$

For expected values less than 20 (i.e. small samples) a continuity correction factor may be applied, such that the χ^2 value is calculated as

$$\frac{(|\mathbf{A \ not \ B} - \mathbf{B \ not \ A}| - 1)^2}{\mathbf{A \ not \ B} + \mathbf{B \ not \ A}} \quad (2)$$

The resulting χ^2 value can then be compared at the desired level of significance using standard tables at one degree of freedom.

Table 1. Number of true positives found at an error level of 5, 20, and 50 false positives

Method	Error level (false positives)		
	5	20	50
AMPS	356	410	440
BLAST <i>p</i> -value	374	406	431
BLAST <i>e</i> -value	374	387	414
FASTA	346	389	405
GSRCH50	394	413	439
GSRCH62	395	421	440
PRSS	382	422	444
SCANPS	401	432	473
SSEARCH	388	430	471

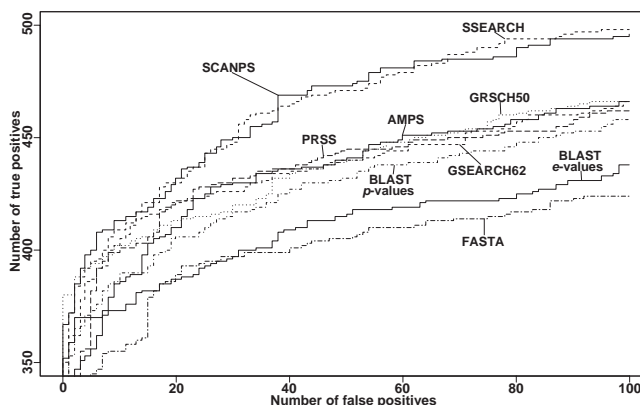
RESULTS AND DISCUSSION

Exploiting multiple similarity search programs

Each of the seven search methods were run with default parameters. GSRCH was run with an additional BLOSUM50 substitution matrix (GSRCH50) due to the exceptional performance of this combination at very low error. However, the default BLOSUM62 matrix (GSRCH62) performs better over a wider range of error. Figure 2 shows false positives plotted against true positives hit at a given threshold for each of the eight methods. In order to emphasize the differences between the methods, data below 350 true positives are not shown. The absolute numbers of true positives found at an error level of 5, 20, and 50 false positives are shown in Table 1. Table 1 illustrates that the highest scoring method at five false positives is SCANPS which finds 401 true positives. This gives a McNemar χ^2 test statistic (Bland, 1987) of 7.36 (significant at $p = 0.01$) or higher against all other methods except GSRCH50 and GSRCH62.

Figure 2 shows that from 20 false positives the methods begin to segregate into three bands. At 20 false positives the difference between the best-performing band, consisting of SCANPS and SSEARCH, and the middle band, consisting of PRSS, GSRCH62, GSRCH50, AMPS, and BLAST, is not significant. However, at an error of 50 false positives the lowest χ^2 test statistic between either SCANPS or SSEARCH, and any other method is 8.21, which is significant at a threshold of $p = 0.01$. FASTA is the poorest performing method over the entire plot.

Figure 2 also plots true against false positives for both BLAST (Altschul *et al.*, 1990, 1997) *p*-values and *e*-values. Figure 2 shows that the use of BLAST *p*-values gives a significantly higher coverage of true positives than BLAST *e*-values in this single-domain benchmark, with a χ^2 test statistic of 19.0 at a 20 false positive cut-off, which is significant at the $p = 0.001$ threshold. If BLAST benchmark pair scores are ranked by *p*-value

**Fig. 2.** False positives (*x*-axis) plotted against true positives (*y*-axis) for eight pairwise alignment search methods, run with default parameters.

rather than *e*-value, there is no statistically-significant difference between BLAST and PRSS, which implements the Smith–Waterman local alignment method (Smith and Waterman, 1981). As a consequence we use BLAST *p*-values throughout this work.

The coverage of true positives found by all 56 pairwise combinations of the eight sequence similarity search programs by union and intersection at equal *p*-value cut-offs was calculated. The results were compared to the performance of the parent methods on this benchmark at low levels of error (0–100 false positives.)

Of the 56 pairwise combinations of methods considered, 19 show significant increases in true positives found over both parent methods at an error of five false positives, detailed in Table 2. The selectivity of these combinations at this error, defined as true positives/(true + false positives), is 98.7–98.9%. 11/19 combinations are set intersections and eight are set unions. 17/19 combinations are formed by combining a local and global alignment method, of which eight combinations are formed from both the intersection and union of four methods, GSRCH50/SCANPS, GSRCH50/SSEARCH, GSRCH62/SCANPS, and GSRCH62/SSEARCH. This shows that while SSEARCH and SCANPS hit similar false positives and dissimilar true positives compared to GSRCH50 and GSRCH62 towards the top of their score-ranked pair lists, further down these pairs of methods hit dissimilar false positives and similar true positives.

All of the combinations shown in Table 2 continue to yield more true positives than either parent method up to an error 100 false positives, except the union of BLAST/SSEARCH, GSRCH50/SSEARCH, and GSRCH62/SSEARCH, which return fewer true positives than SSEARCH from an error of 26, 30, and 31 false positives, respectively. Seventeen of the combinations

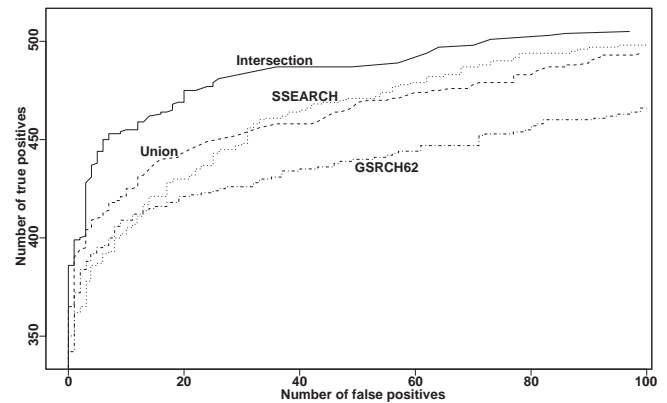
Table 2. Pairwise set interactions formed from the output of eight sequence search methods

Method A	Method B	Set Operation	True positives hit	%increase over parent methods	<i>p</i> -value cut-off
GSRCH62	SSEARCH	Intersection	444	12.4/14.4	3.5e-05
AMPS	SCANPS	Intersection	435	22.2/8.5	1.3e-04
AMPS	SSEARCH	Intersection	426	19.7/9.8	1.7e-04
GSRCH62	SCANPS	Intersection	426	7.8/6.2	1.3e-04
GSRCH50	SCANPS	Union	424	7.6/5.7	3.5e-06
GSRCH50	SCANPS	Intersection	423	7.4/5.5	9.3e-05
GSRCH62	SCANPS	Union	423	7.1/5.5	3.5e-06
GSRCH50	SSEARCH	Intersection	423	7.4/9.0	1.74e-04
BLAST	GSRCH50	Union	421	12.6/6.9	6.2e-06
BLAST	GSRCH62	Union	421	12.6/6.6	5.8e-06
GSRCH50	SSEARCH	Union	413	4.8/6.4	4.2e-06
AMPS	PRSS	Intersection	412	15.7/7.9	1.3e-04
GSRCH62	SSEARCH	Union	410	3.8/5.7	4.2e-06
FASTA	GSRCH50	Union	407	17.6/3.3	5.9e-06
GRSCH50	PRSS	Union	407	3.3/6.5	4.7e-06
AMPS	BLAST	Intersection	403	13.2/7.8	2.5e-04
BLAST	SSEARCH	Union	403	7.8/3.9	1.1e-05
BLAST	FASTA	Union	388	3.7/12.1	1.3e-05
AMPS	FASTA	Union	368	3.4/6.4	1.0e-06

These methods yield a significant increase (*p*-value of >0.01) in the coverage of true positives over both parent methods at an error of five false positives. The percentage increase (%increase) is given over Method A/Method B. Hits returning a *p*-value lower than the *p*-value cut-off were used to form the sets. The table is sorted by the number of true positives found.

shown in Table 2 find more true positives than the best single method at this error, SCANPS, which finds 401. This result suggests that for profile-based search methods where the initial profile is constructed from hits found by pairwise alignment methods (Altschul *et al.*, 1997; Sonnhammer *et al.*, 1997; Rychlewski *et al.*, 2000), combining the results of two pairwise methods in the initial database search may yield more true homologues and so improve the subsequent profile.

At an error of five false positives the intersection of GSRCH62 and SSEARCH yields the most true positives of any combination, finding 444. Figure 3 plots false against true positives found by combining GSRCH62 and SSEARCH and shows that the intersection of these methods yields significant gains over both parent methods up to an error of 41 false positives. Above this error the intersection continues to find more true positives than either of the parent methods, although this increase is not significantly greater than SSEARCH alone. Figure 3 also shows that up to an error of 25 false positives, the union returns significantly more true positives than either parent method, and continues to return an increase until an error of 31 false positives. However, above 31 false positives the union performs worse than SSEARCH alone.

**Fig. 3.** False positives (*x*-axis) plotted against true positives (*y*-axis) for SSEARCH and GSRCH62, and the union and intersection of these score-ranked pair lists at equal *p*-value.

Two of the 19 significant combinations are formed by union of BLAST and FASTA, and BLAST and SSEARCH. This shows that at low error these methods are finding dissimilar true positives but similar false positives at low *p*-values.

While the combination of BLAST and SSEARCH is used in some applications (Enright and Ouzounis, 2000), BLAST and FASTA are of particular interest due to their frequent use in genome annotation (Andrade *et al.*, 1999; Gaasterland and Sensen, 1996). Figure 4 illustrates the results for the combination of BLAST and FASTA at equal *p*-value. Figure 4 shows that while up to four false positives, the union of BLAST and FASTA performs as well as BLAST alone, above this error the union of these methods performs better than either method. This increase is statistically significant at a *p*-value of 0.01 or greater, except for the region from 19 to 29 false positives. Figure 4 also shows that the intersection of BLAST and FASTA at equal *p*-value consistently yields a lower coverage of true positives than BLAST alone. These results show that on this benchmark it is better to combine hits found by either BLAST or FASTA rather than only to take those hits that are found by both methods.

Is it better to take the intersection or union of the eight methods? Intersection at equal *p*-value yields on average 2.8 (SD of 24.1), 4.22 (SD of 20.3), and -3.1 (SD of 23.3) more true positives than union at equal *p*-value, at an error of 5, 20 and 50 false positives, respectively. This suggests that in general, taking the intersection of these methods yields higher gains in coverage of true positives at lower error, while taking the union of these sets yields on average slightly higher gains at higher error. However, the high variance shows that the choice largely depends on the particular methods being combined.

Many of the methods examined employ different

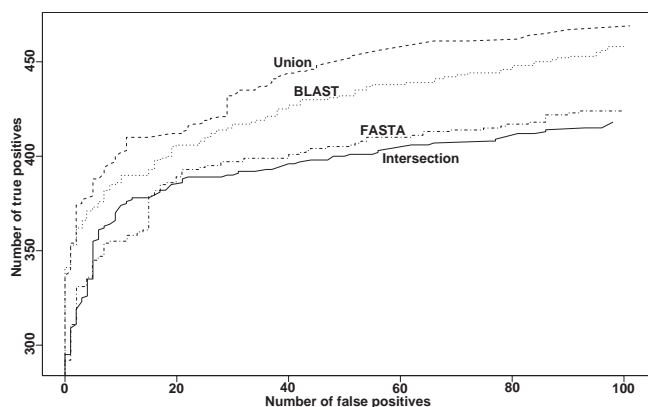


Fig. 4. False positives (x -axis) plotted against true positives (y -axis) for BLAST and FASTA, and the union and intersection of these score-ranked pair lists at equal p -value.

substitution matrices and gap penalties as default. To examine whether the differences in true and false homologues detected at a given level of error are due solely to variation in the matrix and gap penalty, BLAST, FASTA, and SCANPS were run against the benchmark, each with varying matrix and gap penalties. For each method, the resulting score-ordered lists were examined by set operations as above. While sporadic increases in the detection of homology were found between pairwise combinations of matrix and gap penalty, no consistent significant increases were detected (results not shown.) This broadly confirms the earlier work of Henikoff and Henikoff (1993). While variation in matrix and gap penalty may contribute to variation in homology detection, the algorithm and calculation of alignment statistical significance are the primary factors responsible for the significant differences found between the methods examined here. While analysis of exactly which components of each search program give rise to these differences is non-trivial, it may be useful research. Furthermore, given the myriad of available sequence search methods and the limited set considered in this work, analysis of a broader range of search methods may also prove worthwhile.

The benchmark employed in this work allows direct evaluation of sequence search methods against a well understood and reliable protein classification. However, the benchmark considers SCOP domains rather than the complete protein sequence, so is not directly testing the ability of methods to locate domains within larger proteins, either with a multidomain or single domain query. Ideally, one would employ a benchmark that could also test these classes of problem, but unfortunately, the design of such a benchmark is non-trivial. For example, the distribution and order of domains would influence results. One may seek to represent combinations of

domains previously observed in nature but this may optimize the benchmark towards better characterized organisms and may be skewed by the availability of the protein structures needed to define homology. The alternative of randomly combining domains may have little biological relevance where domains may only be observed with a limited set of partners. Construction of structurally validated, multidomain benchmarks of this type will be the subject of future studies.

ACKNOWLEDGEMENTS

This work was supported by a European Molecular Biological Laboratory studentship to CW.

REFERENCES

- Altschul, S. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **199**, 555–565.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Altschul, S.F., Bundschuh, R., Olsen, R. and Hwa, T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.
- Altschul, S.F. and Gish, W. (1996) Local alignment statistics. *Meth. Enzymol.*, **266**, 460–480.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. *et al.* (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.
- Barker, W., Garavelli, J., Hou, Z., Huang, H., Ledley, R., McGarvey, P., Mewes, H., Orcutt, B., Pfeiffer, F., Tsugita, A. *et al.* (2001) Protein information resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**, 29–32.
- Barton, G.J. (1993) An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Comput. Appl. Biosci.*, **9**, 729–734.
- Barton, G.J. (1996) Protein sequence alignment and database scanning. In Sternberg, M.J.E. (ed.), *Protein structure prediction: a practical approach*. IRL Press at Oxford University Press.
- Barton, G.J. and Sternberg, M.J.E. (1987) Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng.*, **1**, 89–94.
- Barton, G.J. and Sternberg, M.J.E. (1990) Flexible protein sequence patterns—a sensitive method to detect weak structural similarities. *J. Mol. Biol.*, **212**, 389–402.
- Bland, M. (1987) *An introduction to medical statistics*. Oxford University Press, Oxford.
- Bork, P. and Koonin, E.V. (1998) Predicting functions from protein sequences—where are the bottlenecks? *Nat. Genet.*, **18**, 313–332.

- Brenner,S.E., Chothia,C. and Hubbard,T. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
- Durbin,R., Eddy,S., Krogh,A. and Mitchinson,G. (1998) *Biological Sequence Analysis*. Cambridge University Press.
- Enright,A.J. and Ouzounis,C.A. (2000) GenAge. *Bioinformatics*, **16**, 451–457.
- Gaasterland,T. and Sensen,C. (1996) MAGPIE: automated genome interpretation. *Trends Genet.*, **12**, 76–78.
- Green,P. (1993) <http://www.genome.washington.edu/uwgc/analysistools/swat.htm>.
- Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355.
- Henikoff,S. and Henikoff,J.G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins*, **17**, 49–61.
- Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
- Pagni,M. and Jongeneel,C.V. (2001) Making sense of score statistics for sequence alignments. *Briefings in Bioinformatics*, **2**, 51–67.
- Park,J., Holm,L., Heger,A. and Chothia,C. (2000) RSDB: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.
- Pearson,W. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.
- Pearson,W. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.*, **4**, 1145–1160.
- Pearson,W. (1996) Effective protein sequence comparison. *Meth. Enzymol.*, **266**, 227–258.
- Pearson,W. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.
- Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
- Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
- Sellers,P.H. (1974) On the theory and computation of evolutionary distances. *J. Appl. Math.*, **26**, 787–793.
- Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Sonnhammer,E., Eddy,S. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
- Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. *et al.* (2002) The EMBL nucleotide sequence database. *Nucleic Acid Res.*, **30**, 21–26.
- Waterman,M.S., Smith,T.F. and Beyer,W.A. (1976) Some biological sequence metrics. *Adv. Math.*, **20**, 367–387.
- Webber,C. and Barton,G.J. (2001) Estimation of *P*-values for global alignments of protein sequences. *Bioinformatics*, **17**, 1158–1167.