BIOINFORMATICS

Vol. 14 no. 1 1998
Pages 40–47

Comparative accuracy of methods for protein sequence similarity search

Pankaj Agarwal[1] and David J. States

Institute for Biomedical Computing, Washington University, Box 8036, 700 South Euclid Avenue, St Louis, MO 63110, USA

Received on November 11, 1996; revised on May 28, 1997; accepted on July 18, 1997

Abstract

Motivation: *Searching a protein sequence database for homologs is a powerful tool for discovering the structure and function of a sequence. Two new methods for searching sequence databases have recently been described: Probabilistic Smith–Waterman (PSW), which is based on Hidden Markov models for a single sequence using a standard scoring matrix, and a new version of BLAST (WU-BLAST2), which uses Sum statistics for gapped alignments.*

Results: *This paper compares and contrasts the effectiveness of these methods with three older methods (Smith–Waterman: SSEARCH, FASTA and BLASTP). The analysis indicates that the new methods are useful, and often offer improved accuracy. These tools are compared using a curated (by Bill Pearson) version of the annotated portion of PIR 39. Three different statistical criteria are utilized: equivalence number, minimum errors and the receiver operating characteristic. For complete-length protein query sequences from large families, PSW's accuracy is superior to that of the other methods, but its accuracy is poor when used with partial-length query sequences. False negatives are twice as common as false positives irrespective of the search methods if a family-specific threshold score that minimizes the total number of errors (i.e. the most favorable threshold score possible) is used. Thus, sensitivity, not selectivity, is the major problem. Among the analyzed methods using default parameters, the best accuracy was obtained from SSEARCH and PSW for complete-length proteins, and the two BLAST programs, plus SSEARCH, for partial-length proteins.*

Availability: *The data and search tools are available from their original authors.*

Contact: *agarwal@mh.us.sbphrd.com, states@ibc.wustl.edu*

Introduction

Searching a molecular sequence database for homologous sequences is a powerful and widely used tool for determining the possible structure and function of a new sequence. There are a variety of algorithms and tools available to conduct these searches. The traditional algorithms are guaranteed to find the 'optimal' alignment and are based mostly on dynamic programming (Needleman and Wunsch, 1970; Smith and Waterman, 1981). The optimality is guaranteed under a specific scoring criterion that includes the scoring matrix and gap penalties. This optimal alignment is quite often not the true biological alignment. Several researchers have noted the importance of considering suboptimal alignments in assessing the significance and value of the optimal alignment (Argos *et al.*, 1991; Saqi and Sternberg, 1991; Zuker, 1991; Agarwal and States, 1996). FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990; Altschul and Gish, 1996) are heuristic search tools that find the 'optimal' alignment with high probability and are computationally less expensive. Hidden Markov model (HMM)-based search methods (Krogh *et al.*, 1994; Eddy *et al.*, 1995; Eddy, 1996) improve both the sensitivity and selectivity of sequence database searches. They use position-dependent scores to characterize and build a model for an entire family of sequences. HMMs work best with large sequence families.

Bucher and Hoffman (1996) have proposed an algorithm combining aspects of HMMs and dynamic programming to build models for single sequences using position-independent scoring and affine gap penalties. The scores are drawn from a traditional scoring matrix, such as BLOSUM62. Instead of computing in the log-odd score space, they compute the probabilities for the alignment. In the innermost loop of the dynamic programming, rather than choose a step corresponding to the most likely alignment, they add the probabilities from all the possible local alignments. This is similar to both Bishop and Thompson's (1986) method and the forward algorithm for HMMs. Thus, the number at each cell in the matrix is proportional to the sum of the probabilities of all the alignments ending at that point.

[1] *Present address: SmithKline Beecham Pharmaceuticals R&D, UW2230, 709 Swedeland Road, PO Box 1539, King of Prussia, PA 19406-0939, USA*

The sum of these numbers in all the cells of this matrix is an estimate of the relatedness of the two sequences. This number is divided by the null estimate, which is the number calculated in a similar fashion for two equal-length random sequences. This technique relies less on the score for an 'optimal' alignment, and includes the probabilities from the suboptimal alignments in evaluating the relatedness of two sequences.

A new developmental version of BLAST (WU-BLAST2 available from http://blast.wustl.edu/) has been released that incorporates Sum statistics for gapped alignments (Altschul and Gish, 1996).

Pearson (1995) has made available a curated database and query sequences that may be used as a 'standard' for sequence comparison tools. This is the annotated portion of the National Biomedical Research Foundation protein sequence database (PIR1, Release 39, December 31, 1993) (Barker *et al.*, 1990) augmented with 237 additional sequences for a total of 12 219 sequences. Viral polyproteins (because of their mosaic nature) and 23 other families (for which all the members were so closely related that none of the programs have difficulty discriminating) were not utilized as query sequences by Pearson. Sequences from 67 of the remaining largest families were used as queries. Two query sequences were selected from each family. In addition to these complete query sequences, two partial sequences from each family augmented on either side with random sequence were also used. Details are available from Pearson (1995).

System and methods

Metrics for comparing methods

Pearson (1995) has introduced the equivalence number as a criterion for measuring the search accuracy of a tool. The equivalence score is the threshold score at which the number of false positives equals the number of false negatives. The equivalence number (EN) is the number of false positives (equal to the number of false negatives) using the equivalence score as the classification threshold. Thus, the equivalence number balances the sensitivity (false negatives) and selectivity (false positives) of an algorithm.

The sequence similarity search problem is a special case of the sequence classification problem with the object being to classify sequences from the database as homologous or non-homologous to the query sequence. An alternative to the equivalence number is the error rate (i.e. the sum of the number of false positives and false negatives), which is often used in Bayesian classification (Duda and Hart, 1973). The minimum number of errors (MER) is upper bounded by twice the equivalence number, thus $\mathrm{MER} \leq 2 \times \mathrm{EN}$. For minimum errors, the threshold score must lie between the score for a true positive and the score for a true negative, and that true negative must follow that true positive in the rank order of the tool. For example, if there are 100 homologs and 1000 non-homologs, and the search algorithm ranks them in the order 80 homologs, 1 non-homolog, 20 homologs and 999 non-homologs, then the minimum number of errors is 1, with the threshold score between the last homolog and the first of the 999 non-homologs. The equivalence number is also 1, but the threshold score is set before the last homolog. The MER thus provides more biologically plausible thresholds. The MER also makes it possible to compare the sensitivity and selectivity of two algorithms/tools. At the MER threshold, the number of both the false positives and false negatives may be measured separately, indicating which of them has the larger contribution to the error.

Gribskov and Robinson (1996) have proposed the receiver operating characteristic (ROC) as a quantitative measure of the usefulness of a sequence classification. This measure, like the previous two, incorporates sensitivity and selectivity, but it depends significantly on the exact order of the positive hits and negative hits. ROC is a real number between 0 and 1, which is evaluated as the area under the parametric curve of the fraction positive (i.e. the number of homologs observed divided by the true number of homologs) plotted as a function of the fraction negative (i.e. the number of non-homologs observed divided by the total number of non-homologs). The data points in this plot correspond to the rank order from the tool; e.g. the plot may be parameterized on the ordering of sequences according the BLAST p-value. A perfect tool will list all the positives before the negatives; thus, the ROC plot will be a vertical line at $x = 0$ followed by a horizontal line at $y = 1$. The corresponding ROC value is 1. A poor tool will mix positives and negatives, resulting in an ROC value close to 0. Gribskov and Robinson propose using ROC_{50} which scans down the list until 50 negatives are observed. This is useful because the number of non-homologs is 100–10 000 times more than the number of homologs, and the ROC value is dominated by the non-homologs and is always very close to 1. The ROC must then be computed to a large number of significant digits to compare the methods. For the remainder of this paper, we use ROC to imply ROC_{50}.

Search tools

We evaluated five tools according to the above three characteristics.

(i) SSEARCH Release 3.0t74 (Pearson and Lipman, 1988). This is an implementation of the local alignment dynamic programming technique proposed by Smith and Waterman (1981). It includes code optimizations from Phil Green. The default scoring matrix is BLOSUM 50, and gap penalties are –12 for initiation and –2 for continuation. The sequences were ranked according to the '*z*-score'.

(ii) FASTA Release 3.0t74 (Pearson and Lipman, 1988). The default scoring matrix is BLOSUM50, and the gap penalties are (–12, –2). The sequences were ranked according to the 'z-score'.

(iii) Probabilistic Smith–Waterman (PSW) (Bucher and Hoffman, 1996). The default scoring matrix is BLOSUM 45, and gap penalties are (–8, –4).

(iv) BLASTP Release 1.4.9MP (Altschul *et al.*, 1990). The default scoring matrix is BLOSUM 62, and gaps are not permitted [i.e. gap penalty (–∞, –∞)]. The sequences were ranked according to the probability estimate from the Sum statistics.

(v) WU-BLASTP2 Release 2.0a1MP (Altschul *et al.*, 1990; Altschul and Gish, 1996). The default scoring matrix is BLOSUM 62, and gap penalties are (–9, –2). The sequences were ranked according to the probability estimate from the Sum statistics of gapped alignments.

The tools were compared with default scoring matrices and gap penalty options. This assumes that each tool has been configured with the scoring matrix and gap penalties that provide it with maximum accuracy. The performance of various tools as functions of the scoring parameters have been compared earlier (Altschul, 1993; Altschul *et al.*, 1994; Pearson, 1995; Gribskov and Robinson, 1996). It could be argued that the observed differences between the tools are primarily the function of their choices of scoring systems. To neutralize the effect of the scoring system, the tools were also compared with the same scoring system [BLOSUM 62 and gap penalty (–9, –2)] for a single set of query sequences (e0-b62).

Results and discussion

Table 2 (spanning two pages) contains the names for the families, the number of family members in the examined database, the length of the query sequence in set e0, and the corresponding minimum number of errors for the various methods. For a number of families, all the methods had perfect discrimination (zero error rate), but, in general, there is considerable variation in the error rates, depending on the family. Each tool outperforms every other tool for at least some protein families. It is quite likely that some of the tools are actually better than other tools. The observed variation is probably because of sampling errors due to the small sizes for some of the protein families. PSW performs better than any other method in four of the seven families that have >100 members. This suggests that PSW actually performs better than any other method when complete protein query sequences are provided; the smaller families for which it is observed not to perform as well may be chance observations, especially considering the small sample. PSW is incorporating information from a large number of suboptimal alignments in evaluating the similarity relationship. Related protein sequences often have conserved regions (corresponding to regions with secondary and tertiary structure) interspersed with non-conserved regions (corresponding to turns). There are numerous equally valid ways of aligning the non-conserved regions, leading to an exponentially large number of high-scoring suboptimal alignments.

Table 1. A summary of the statistics for the various search methods and sequence query sets

Method		PSW		SW		FASTA		BLASTP		BLASTP2	
Query set	Stat	+	+ =	+	+ =	+	+ =	+	+ =	+	+ =
e0	MER	**10**	41	4	**47**	1	34	2	32	3	42
e0	ROC	**10**	34	4	**35**	7	25	4	26	7	32
e0	EN	**11**	39	1	**46**	2	34	2	29	3	43
e1	MER	**8**	41	3	**43**	1	35	6	35	2	42
e1	ROC	**13**	**31**	4	**31**	6	27	6	26	5	**31**
e1	EN	**6**	44	2	**48**	2	38	4	36	2	45
e0-b62	MER	**12**	33	2	**43**	2	32	2	32	3	42
e0-b62	ROC	**10**	29	7	**33**	5	26	4	26	5	**33**
e0-b62	EN	**11**	33	1	**44**	3	32	2	30	3	**44**
l0	MER	3	23	2	**37**	3	36	**11**	34	3	35
l0	ROC	8	11	8	**21**	6	18	**12**	17	**12**	18
l0	EN	3	25	6	33	3	34	**8**	32	3	**39**
l1	MER	5	24	4	32	5	31	**13**	32	4	**35**
l1	RO	10	15	8	21	7	21	**13**	21	6	**23**
l1	EN	**8**	23	2	37	5	34	**8**	33	2	**40**

The first row in this table is a summary of the data from Table 2. The number in the '+' column is the number of protein families (maximum 67) for which the method had better discrimination (i.e. ability to rank family members ahead of non-family members) than any of the other methods. The '+ =' column indicates the number of protein families (maximum 67) for which the method had better or equal discrimination in pairwise comparison with all other methods. The best numbers for each statistic in a query set are in bold. e0 and e1 are the complete length query sequences, one from each family; e0-b62 is the same set of query sequences as e0, but all the methods were evaluated under uniform scoring conditions, namely BLOSUM 62 matrix and gap penalties (–9,–2); l0 and l1 are the partial-length query sequence sets. Explanations of the statistics are provided in System and methods: Metrics for comparing methods.

Table 2. The minimum number of errors (MER) for the various methods and families using a single set of query sequences (e0) and the default scoring parameters. The Best Method column is left blank if all the methods performed equally well

Description/superfamily	Length	Size	PS	SW	FA	BP	B2	Best method
Hemoglobin α/β	141	505	14	20	34	36	31	PS
Ig kappa chain V-I region	108	280	11	20	49	67	22	PS
G-prot. coupled receptors	348	165	29	26	47	34	27	SW
Cytochrome C	105	142	18	24	25	29	24	PS
Snake neurotoxin	74	109	0	0	0	5	0	PS/SW/FA/B2
Calcium binding EF-hand	159	106-	15	11	12	12	11	SW/B2
Glutathione transferase	222	106	3	4	6	9	5	PS
Protein kinase, cAMP-dependent	351	97	74	70	72	77	75	SW
Ferredoxin	54	93	58	57	68	60	58	SW
Ribulose-bisphosphate carboxylase	139	77	2	2	2	2	2	
Ig kappa chain C region	106	74	18	18	24	26	18	PS/SW/B2
Hemagglutinin	567	73	0	0	0	1	0	PS/SW/FA/B2
Histocompatibility antigen	338	71	0	0	2	2	3	PS/SW
Insulin	110	69	3	3	3	3	3	
α-Crystallin chain A	173	67	2	4	4	8	3	PS
Phospholipase A2	148	58	1	0	1	2	0	SW/B2
Glyceraldehyde-3-P DH	335	46	0	0	1	0	0	PS/SW/BP/B2
Transforming prot. (N-ras)	189	45	1	1	1	1	1	
Serine protease	246	45	20	22	16	22	19	FA
Glucagon precursor	180	44	14	10	11	7	6	B2
H^+-transporting ATP synthase α chain precursor	553	43	1	1	1	3	2	PS/SW/FA
Hemagglutinin-neuraminidase	576	42	0	0	0	0	0	
Ribonuclease	124	40	0	0	0	0	0	
Interferon α-I-6	189	39	0	0	0	0	0	
Glutamate-ammonia ligase	373	39	0	0	0	1	0	PS/SW/FA/B2
Azurin	129	38	19	17	23	27	22	SW
Fusion protein—Sendai virus	565	36	0	0	0	0	0	
Cytochrome P450	497	35	0	0	3	2	2	PS/SW
Outer capsid protein VP8	280	34	0	0	0	0	0	
gag polyprotein	512	33	3	3	5	3	3	PS/SW/BP/B2
Keratin	471	32	5	5	8	4	5	BP
Nucleoprotein-influenza A	498	31	0	0				
Acidic ribosomal protein P2	115	29	3	6	8	8	7	PS
E6 protein papillomavirus	158	29	0	0	1	0	0	PS/SW/BP/B2
Lysozyme	130	28	0	0	0	0	0	
N-Cadherin	906	27	0	0	0	0	0	
Exo-α-sialidase	454	27	0	0	0	0	0	
L2 protein papillomavirus	507	27	0	0	0	0	0	
Scorpion neurotoxin	64	26	5	5	5	6	5	PS/SW/FA/B2
E7 protein papillomavirus	98	26	0	0	0	1	1	PS/SW/FA
H^+-transporting ATP synthase lipid-binding	75	26	1	1	1	1	0	B2

Table 2. *Continued.*

Description/superfamily	Length	Size	PS	SW	FA	BP	B2	Best method
L-Lactate dehydrogenase	333	26	0	0	2	3	0	PS/SW/B2
E2 protein papillomavirus	322	26	0	0	0	0	0	
Core antigen—hepatitis B	183	25	0	0	0	0	0	
Antithrombin-III	464	25	0	0	1	0	0	PS/SW/BP/B2
Thymidine kinase	376	25	1	1	1	0	1	BP
Phycocyanin	162	25	0	0	0	0	0	
Protamine Y2	34	24	2	1	1	3	1	SW/FA/B2
Transforming prot. (myc)	439	24	0	0	0	0	0	
Matrix protein	348	24	0	0	6	0	0	PS/SW/BP/B2
H^+-transporting ATP synthase P6	226	23	1	1	8	4	1	PS/SW/B2
Alcohol dehydrogenase A	375	23	0	0	0	0	0	
Glycoprotein B	857	23	0	0	0	0	0	
Ionotropic acetylcholine receptor	457	23	0	0	0	0	0	
Non-structural protein NS2	121	22	0	1	1	2	2	PS
Annexin I	346	22	4	4	4	4	4	
Histone H1b	218	22	3	2	3	2	2	SW/BP/B2
Metallothionein	61	21	6	6	6	6	3	B2
β-Crystallin chain Bp	204	21	0	0	0	0	0	
Proteinase inhibitor	71	21	2	3	3	3	3	PS
Hepatic lectin H1	291	20	2	1	1	7	1	SW/FA/B2
E2 glycoprotein precursor	1447	20	0	0	0	0	0	
α-2u-Globulin precursor	181	20	6	9	10	10	9	PS
Pepsin	388	20	0	0	0	0	0	
DNA-directed DNA polymerase	1462	20	5	1	1	1	2	SW/FA/BP
Prolactin	227	20	0	0	0	0	0	
Vitamin B12 trans. btuD	249	20	3	7	9	6	10	PS
Total		3544	355	367	490	510	394	
False positives			109	126	110	159	102	
False negatives			246	241	380	351	292	

'Length' is the length of the query sequence for that family in e0 and 'Size' is the size of the family in the Pearson database. PS, Probabilistic Smith–Waterman; SW, Smith–Waterman; FA, FASTA; BP, BLASTP 1.4; B2, WU-BLASTP2. The false-positives and false-negatives in the last two rows are the break-up of the total errors across all the families. Sequence descriptions are from Pearson (1995).

The information from Table 2 is summarized in the first row of Table 1. The method with the minimum MER for 'hemoglobin α/β' (see the first row in Table 2) is PS. Thus, PS has the superior statistic (lower MER) for this family. The number of solo occurrences of PS in the Best method column in Table 2 add up to 10, which is the number in the top left (e0—MER,+) cell of Table 1. The number 41 in the adjacent column (+ =) is a count of the number of families for which the method had as low an MER as any other method. This is equal to the number of solo and/or joint occurrences of the tool in the Best method column in Table 2. Table 1, in addition to this summary line for e0 and MER, contains a summary of the results for the five data sets and using the three statistical criteria.

Smith–Waterman, not surprisingly, is at least as accurate as any other method for the most number of families over the five query sets (examine the '+ =' columns in Table 1). The new version of BLAST, WU-BLAST2 from Warren Gish, which includes Sum statistics of gapped alignments, is also accurate in this measure. Neither BLASTP 1.4 nor FASTA performed as accurately. Some of this may be related to the dependencies between the methods, and consequent failure to outperform a very similar method. Table 3 includes the pairwise comparisons for the various methods to help assess this possibility.

Table 3. Pairwise comparisons of the five methods. The number in the cell in row *i*, column *j* is the number of families (from 67) in which method *i* did better than method *j* according to the statistic. For example, in the upper left-most subtable, PS had lower error rates compared to FA in 24 families, while FA had lower error rates in seven families

	Minimum no. of errors e0—MER						False positives e0—FP						False negatives e0—FN				
	PS	SW	FA	BP	B2		PS	SW	FA	BP	B2		PS	SW	FA	BP	B2
PS	–	11	24	29	16	PS	–	10	11	15	7	PS	–	8	23	25	15
SW	11	–	26	27	15	SW	5	–	7	13	5	SW	11	–	23	26	17
FA	7	1	–	20	8	FA	7	7	–	15	3	FA	9	4	–	9	9
BP	6	4	14	–	5	BP	7	9	8	–	4	BP	10	6	17	–	9
B2	11	5	26	27	–	B2	7	11	9	17	–	B2	11	5	22	21	–
	e1—MER						e1—FP						e1—FN				
	PS	SW	FA	BP	B2		PS	SW	FA	BP	B2		PS	SW	FA	BP	B2
PS	–	11	20	23	12	PS	–	7	5	10	9	PS	–	12	22	21	12
SW	11	–	23	22	13	SW	8	–	5	9	9	SW	13	–	24	19	12
FA	11	4	–	18	6	FA	8	6	–	11	10	FA	8	4	–	16	4
BP	12	11	17	–	8	BP	7	4	5	–	10	BP	12	12	20	–	11
B2	12	8	20	19	–	B2	7	6	5	10	–	B2	13	10	20	18	–
	e0-b62—MER						e0-b62—FP						e0-b62—FN				
	PS	SW	FA	BP	B2		PS	SW	FA	BP	B2		PS	SW	FA	BP	B2
PS	–	16	21	23	16	PS	–	13	10	15	8	PS	–	13	22	20	13
SW	20	–	27	27	9	SW	4	–	7	15	5	SW	21	–	27	22	13
FA	11	6	–	21	7	FA	6	9	–	15	5	FA	13	3	–	17	5
BP	13	6	17	–	5	BP	5	8	–	4	BP	18	9	20	–	9	
B2	19	10	27	27	–	B2	6	9	8	17	–	B2	21	6	26	21	–
	l0—MER						l0—FP						l0—FN				
	PS	SW	FA	BP	B2		PS	SW	FA	BP	B2		PS	SW	FA	BP	B2
PS	–	14	11	11	10	PS	–	11	12	11	14	PS	–	9	12	9	8
SW	35	–	13	17	17	SW	13	–	9	11	12	SW	34	–	15	15	12
FA	35	12	–	15	13	FA	14	9	–	8	12	FA	32	11	–	16	12
BP	31	21	22	–	20	BP	14	10	11	–	12	BP	28	20	19	–	17
B2	35	13	18	16	–	B2	14	10	8	7	–	B2	33	14	19	16	–
	l1—MER						l1—FP						l1—FN				
	PS	SW	FA	BP	B2		PS	SW	FA	BP	B2		PS	SW	FA	BP	B2
PS	–	12	13	12	8	PS	–	15	12	13	14	PS	–	8	13	10	8
SW	27	–	18	16	11	SW	5	–	5	10	10	SW	28	–	19	17	9
FA	27	12	–	18	10	FA	10	15	–	10	12	FA	25	5	–	16	7
BP	29	18	18	–	16	BP	8	11	7	–	12	BP	30	14	18	–	15
B2	27	14	20	17	–	B2	10	12	8	8	–	B2	27	12	20	18	

For complete-length query sequences, according to all three statistics (MER, ROC and EN), PSW (Bucher and Hoffman, 1996) has higher accuracy for more families than any other method (examine the '+' columns in Table 1). When a partial-length protein sequence packed between random sequences was used as a query (l0 and l1), PSW lost its advantage, and both WU-BLAST2 and BLASTP 1.4 became competitive in the '+ =' criteria. SSEARCH continued to perform well, but not as well as it did for complete-length query sequences. BLASTP 1.4 did better in the + criterion than the other programs. This may indicate that the results from the other programs are more related to each other, perhaps because they all use gapped alignments. BLASTP 1.4 may be the most orthogonal to the other programs for partial-length protein queries.

It is important to utilize more than one tool when confronted with either ambiguous similarity results or a complete lack of significant similarities. PSW may have an advantage for complete-length proteins, but it is expensive computationally; WU-BLAST2 offers comparable performance at a lower cost. The three statistical measures mostly agree on the method that performs better for each class of query sequences (such as e0, e1, etc.). Thus, the choice of the statistical measure appears to have little influence on this part of the analysis.

In Table 2, the error rates are split up into false positives and false negatives. The total number of false positives (i.e. the number of non-homologous sequences classified as homologous) over all the families for query set e0 is very similar across all the tools, except for BLASTP 1.4 which has a high false-positive rate. The major variation observed is in the false-negative rates, which are the best for SW and PSW for full-length queries. Moreover, between two-thirds and three-quarters of the errors are due to false negatives, irrespective of the search tool. Ostensibly, these are family members (distant to the query sequence) that none of the search tools can classify as similar. It would be useful to quantify the errors that are common to all the search tools.

The results from the pairwise tests of the various methods in Table 3 indicate that PSW and SSEARCH perform similarly on query sets e0 and e1, but SSEARCH does a little better on the e0-b62 query set. Thus, if family sizes are ignored, there is no clear distinction between the performances of PSW and SSEARCH. With partial-length queries, PSW is comparable to the other methods only if false positives are considered; if either false negatives or total errors are considered, PSW's performance is the worst. SSEARCH and WU-BLAST2, on the other hand, perform consistently well in all categories (the numbers in the SW and B2 columns are low, and they are high in the SW and B2 rows). The numbers in the five subtables in the middle column are all remarkably similar, indicating that there is little or no discrimination for false positives among the various methods.

Both FASTA and SSEARCH performed much better with *z*-scores (the default option in Version 3) then with *opt* scores (analysis not shown). Thus, higher accuracy is achieved by using a good implementation (with *z*-scores) of Smith–Waterman, such as SSEARCH, rather than a simplistic in-house implementation. The default word size (*ktup* setting of two) for FASTA was used for this entire analysis. *ktup* = 1 does provide higher accuracy.

It would be useful to examine closely both the false positives and false negatives for each of the tools, and evaluate the decrease in the error rate caused by using combinations of these tools. In other words, are all the tools missing the same sequences, or how much can be gained by using multiple tools? Statistical tests to evaluate which tool is best given the error rates and/or ROC values would be valuable. Pearson (1995) has used the sign test to estimate the significance for equivalence numbers, but the sign test ignores the magnitude of the difference between families, the size of the families, and the data from all the families for which the tool performed equally well. The results in Table 1 are consistently different in the '+' and '+ =' categories. Even though the sign test is used for pairwise comparisons, while the data in this table are for all-against-all comparisons, the sign test would still overlook the effects of the columns. The results in Tables 1 and 3 also, unfortunately, overlook the sizes of the families.

In summary, assuming that these tools are used with their default options, for complete-length proteins, we would recommend using SSEARCH and PSW. We would substitute PSW with WU-BLAST2 if computational time was a consideration. For partial-length proteins, we would recommend the two BLAST programs and possibly SSEARCH.

Acknowledgements

We would like to thank William Pearson for providing the curated database, Tom Blackwell, Marcus B.Feldman and Michael Zuker for useful discussions, Sean Eddy and Warren Gish for comments on a draft, the anonymous reviewers for many excellent suggestions, and Laureen C.Treacy for proofreading. This work was supported in part by the Department of Energy (DOE) grant DE-FG02-94ER61910.

Note added in proof

The new version of BLAST [Altschul *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **25**, 3389–3402] was not included in this analysis, as it was released after this paper had been accepted.

References

Agarwal,P. and States,D. (1996) A Bayesian evolutionary distance for parametrically aligned sequences. *J. Comput. Biol.*, **3**, 1–17.

Altschul,S. (1993) A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.*, **36**, 290–300.

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

Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S., Boguski,M., Gish,W. and Wootton,J. (1994) Issues in searching molecular sequence databases. *Nature Genet.*, **6**, 119–129.

Argos,P., Vingron,M. and Vogt,G. (1991) Protein sequence comparisons: methods and significance. *Protein Eng.*, **4**, 375–383.

Barker,W., George,D. and Hunt,L. (1990) Protein sequence database. *Methods Enzymol.*, **183**, 31–49.

Bishop,M. and Thompson,E. (1986) Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.*, **190**, 159–165.

Bucher,P. and Hoffman,K. (1996) A sequence similarity algorithm based on a probabilistic interpretation of an alignment scoring system. In States,D., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R. (eds), *ISMB-4.* AAAI Press, Menlo Park, CA.

Duda,R. and Hart,P. (1973) *Pattern Classification and Scene Analysis*. John Wiley and Sons.

Eddy,S. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.

Eddy,S., Mitchison,G. and Durbin,R. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.*, **2**, 9–23.

Gribskov,M. and Robinson,N. (1996) The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–34.

Krogh,A., Brown,M., Mian,I., Sjölander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Needleman,S. and Wunsch,C. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, **48**, 444–453.

Pearson,W. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.*, **4**, 1145–1160.

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

Saqi,M. and Sternberg,M. (1991) A simple method to generate non-trivial alternate alignments of protein sequences. *J. Mol. Biol.*, **219**, 727–732.

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

Zuker,M. (1991) Suboptimal sequence alignment in molecular biology: Alignment with error analysis. *J. Mol. Biol.*, **221**, 403–420.