J. Mol. Biol. (1998) **276**, 71–84

JMB

Empirical Statistical Estimates for Sequence Similarity Searches

William R. Pearson

Department of Biochemistry
University of Virginia
Charlottesville, VA 22908
USA

The FASTA package of sequence comparison programs has been modified to provide accurate statistical estimates for local sequence similarity scores with gaps. These estimates are derived using the extreme value distribution from the mean and variance of the local similarity scores of unrelated sequences after the scores have been corrected for the expected effect of library sequence length. This approach allows accurate estimates to be calculated for both FASTA and Smith-Waterman similarity scores for protein/protein, DNA/DNA, and protein/translated-DNA comparisons. The accuracy of the statistical estimates is summarized for 54 protein families using FASTA and Smith-Waterman scores. Probability estimates calculated from the distribution of similarity scores are generally conservative, as are probabilities calculated using the Altschul-Gish λ, K, and H parameters. The performance of several alternative methods for correcting similarity scores for library-sequence length was evaluated using 54 protein superfamilies from the PIR39 database and 110 protein families from the Prosite/SwissProt rel. 34 database. Both regression-scaled and Altschul-Gish scaled scores perform significantly better than unscaled Smith-Waterman or FASTA similarity scores. When the Prosite/SwissProt test set is used, regression-scaled scores perform slightly better; when the PIR database is used, Altschul-Gish scaled scores perform best. Thus, length-corrected similarity scores improve the sensitivity of database searches. Statistical parameters that are derived from the distribution of similarity scores from the thousands of unrelated sequences typically encountered in a database search provide accurate estimates of statistical significance that can be used to infer sequence homology.

Keywords: sequence similarity; statistical estimates; FASTA; Smith-Waterman

Introduction

Sequence similarity searches today are the most effective method for exploiting the information in the rapidly growing DNA and protein sequence databases. One of the most dramatic improvements in similarity searching was the introduction of accurate statistical estimates for similarity searches for alignments without gaps in the BLAST sequence comparison package (Altschul *et al.*, 1990). Accurate statistical estimates make it possible to identify automatically sequences that are likely to be homologous (i.e. that share statistically significant similarity because of descent from a common ancestor). In general, if statistically significant similarity is found between two sequences and the similarity does not simply reflect a region with unusual amino acid composition, the sequences are likely to be homologous.

The BLAST package of sequence comparison programs (Altschul *et al.*, 1990, 1994) provides the most widely used similarity searching programs, in part because of its accurate statistical estimates. BLAST uses two parameters, K and λ, to estimate the statistical significance of a high scoring alignment using the formula (Karlin & Altschul, 1990; Altschul *et al.*, 1994; Altschul & Gish, 1996):

$$P(S > x) = 1 - \exp(-Kmne^{-\lambda x}) \qquad (1)$$

where x is the similarity score, and m and n are the lengths of the two sequences being compared. Unfortunately, the underlying statistical model used by BLAST for high scoring segment pairs is

0022–2836/98/060071–14 $25.00/0/mb971525

limited to alignment without gaps (Karlin & Altschul, 1990), although scores from several ungapped alignments can be evaluated as well (Karlin & Altschul, 1993). Because sequence alignments between distantly related proteins typically require gaps, and similarity searching with the Smith-Waterman algorithm and the FASTA program (with gaps) can perform better than BLAST on divergent protein families (Pearson, 1995), we sought a general strategy that would provide accurate statistical estimates for alignments with gaps that would work not only for Smith-Waterman scores but also for FASTA protein-protein, DNA-DNA comparisons, and for comparisons between protein sequences and translated DNA (FASTX, TFASTX, TFASTA).

Here, we evaluate several approaches for calculating the ''location'' (K) and ''scale'' (λ) parameters from the distribution of similarity scores from unrelated sequences that are calculated during a sequence database search. We show that statistical estimates for similarity scores that have been scaled to correct for the length-dependence of local similarity scores are very accurate, and that the empirical approach described here provides an internal calibration of the accuracy of the estimates. In addition, we show that length-corrected similarity scores are more effective than raw scores at identifying distantly related members of protein families. These estimation methods have been incorporated into versions 2.0 and 3.0 of the FASTA package of sequence comparison programs.

Results

Accurate statistical estimates

This paper describes a general method for determining the statistical significance of a local similarity score, based on the distribution of similarity scores obtained from a sequence database search. Current protein and DNA sequence databases contain many tens of thousands of sequences, almost all of which are unrelated to an individual query sequence (even the largest protein families comprise less than 5% of a comprehensive protein database like SwissProt or PIR). Thus, every database search provides tens of thousands of scores from unrelated, effectively random, protein and DNA sequences. For local similarity scores, these ''random'' sequence scores are expected to follow the extreme-value distribution (Mott, 1992; Altschul & Gish, 1996) with location and scale parameters† that reflect the lengths and compositions of the query and library sequences and the scoring matrix and gap penalties used. In this section, we show that the statistical significance values produced by several estimate procedures are accurate; thus, they can be used with confidence to infer homology from significant sequence similarity.

To calculate the statistical significance of a similarity score from the distribution of unrelated-sequence scores, the location and score parameters must be estimated. Previous workers have described estimation procedures based on searches with random sequences or pruned searches with ''real'' sequences (Mott, 1992), successive local sub-alignments (Waterman & Vingron, 1994), or by simulations with random sequences (Altschul & Gish, 1996). We prefer to estimate the parameters directly from the distribution of actual unrelated sequence similarity scores so that any local alignment procedure, including the heuristic methods used by FASTA, can be used. However, unrelated sequence similarity scores are often ''contaminated'' with high scores from unrelated sequences; these scores must be removed for accurate estimation. Several methods for estimating these parameters are outlined in Methods.

The distributions of Smith-Waterman, FASTA protein, FASTA DNA, and FASTX similarity scores are accurately described by the extreme-value distribution after parameter estimation from unrelated-sequence similarity scores (Figure 1). For SSEARCH, the statistical parameters are estimated from the Smith-Waterman local similarity scores. For the FASTA and FASTX programs, the location and scale parameters are estimated from the distribution of optimized similarity scores (*opt* scores in Table 1). FASTA and FASTX optimized scores result from a Smith-Waterman alignment in a band of 16 or 32 residues centered on the highest scoring initial region found in the FASTA/X scanning heuristic (Pearson, 1990).

We are most concerned with similarity scores in the upper tail of the distribution (Figure 1, insets), since the highest similarity scores are used to infer homology. There is excellent agreement between the observed and expected distribution of similarity scores not only for protein sequence comparison using the rigorous Smith-Waterman algorithm (Smith & Waterman, 1981; SSEARCH, Figure 1A), but also for similarity scores calculated by the heuristic FASTA procedure, for either protein-protein or DNA-DNA comparisons (Figure 1B and C). The FASTX program compares a DNA query sequence (such as an Expressed Sequence Tag cDNA sequence) to a protein sequence library by translating the DNA sequence in three frames and calculating the best alignment between the translated sequence and a protein sequence, allowing frameshifts (Zhang *et al*., 1997); the frequencies of actual FASTX similarity scores also agree closely with the expected frequencies of scores from the extreme-value distribution.

Because protein and DNA sequence identification rely on accurate statistics for the highest

† For the normal distribution, the location parameter is the mean μ and the scale parameter is the standard deviation σ. For the extreme-value distribution, $\exp(-e^{-(x-a)/b})$, the location parameter a and scale parameter b are related to μ and σ as: $\mu = a - b\Gamma'(1)$ and $\sigma^2 = b^2\pi^2/6$ (Evans *et al*., 1993).

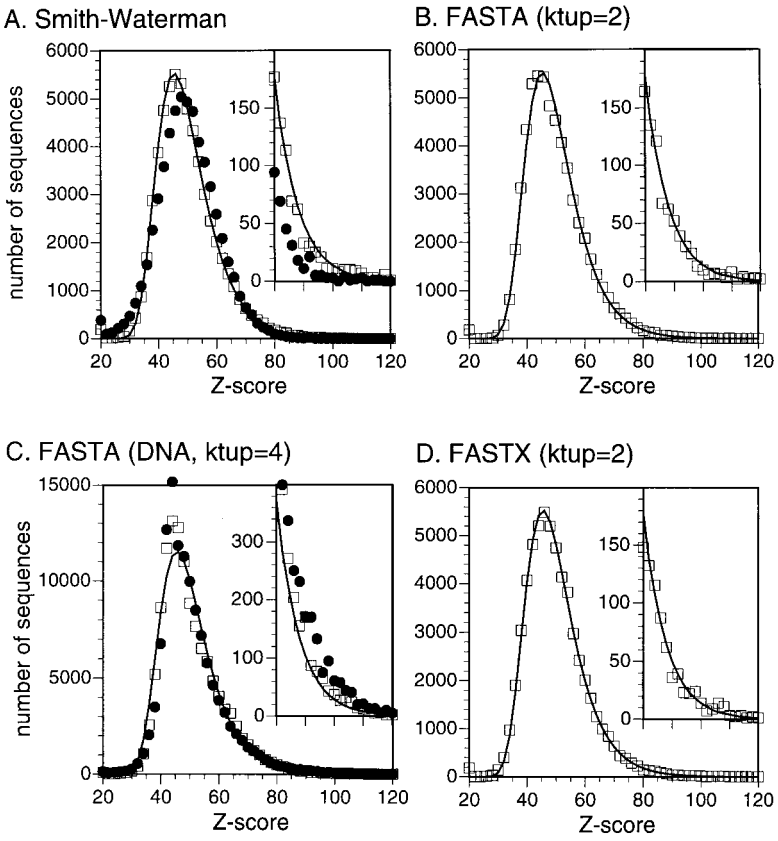

Figure 1. Distribution of sequence similarity *z*-scores. The number of sequences obtaining a similarity score (*z*-score), calculated using the regress1 method, in *x*-axis bins of two *z*-score units (A, B and C) or four units (C) are shown. Symbols show the observed number of sequences; the continuous line indicates the expected distribution of *z*-scores for an extreme value distribution. (*z*-scores are scaled to have a mean of 50 and a standard deviation of 10.) A, Smith-Waterman algorithm (SSEARCH). Comparison of `gtml_mouse` with SwissProt (rel.34) using the BLOSUM50 (Heniknoff & Henikoff, 1992) scoring matrix, −12 for the first residue in a gap and −2 for each additional residue (−12/−1,□) or −8/−1 (●). For the search with gap penalties of −12/2, the Kolmogorov-Smirnov (KS) statistic for the fit of the observed to expected distribution of scores was 0.0049 ($N = 29$); for the −8/−1 gap-penalty search KS = 0.062 ($N = 29$). B, Comparison of `gtml_mouse` with SwissProt using FASTA, *ktup* = 2, BLOSUM50 matrix, −12/−2 gap-penalties (□, KS = 0.017, $N = 29$). C, DNA sequence comparison of the `MUSGLUTA` cDNA sequence with the primate, rodent, and mammalian divisions of Genbank (rel. 102), *ktup* = 4, +5 for a match, −4 for a mismatch, −16/−4 gap-penalties (□, KS = 0.023, $N = 29$) or −12/−2 gap-penalties (●, KS = 0.050, $N = 29$). D, Comparison of the `MUSGLUTA` cDNA sequence with SwissProt using the FASTX program, which translates the DNA query sequence and calculates the best alignment between the three-frame translation and each SwissProt protein sequence, allowing frameshifts. The BLOSUM50 scoring matrix, gap-penalties of −15/−3, and a frameshift penalty of −30, was used (□, KS = 0.010, $N = 29$).

scoring sequences from a database, the expectation (*E*()-values) of the score of the highest scoring unrelated sequence, the number of times the score is expected by chance, is more important than the overall distribution of similarity scores. The *E*()-value for obtaining a similarity score $S > x$ in a sequence database search is $E(S > x) = P(S > x)N$, where $P(S > x)$ is the probability of obtaining a score $S > x$ in a single comparison (which can be estimated from the extreme-value distribution), and N is the number of tests that have been performed. For similarity searches against a protein sequence database, a similarity score is calculated (a test is performed) for each sequence in the database, so N is the number of entries in the database, Ideally, the highest scoring unrelated sequence should have an expectation value of ~1.0; an unrelated (random) sequence should have an expectation value of 0.02 about 2% of the time. The BLAST suite of programs also calculates an expectation, but typically reports the probability of obtaining the score in a database search: $P(S > x \mid N \text{ comparisons}) = 1 - e^{-E(S > x)}$. For *E*() or *P*()-values <0.05, the two values are approximately equal, but BLAST *P*()-values range from 0...1, while FASTA *E*()-values range from 0 to the number of sequences in the database.

Table 1 shows the highest scoring related and unrelated sequences found in each of the searches displayed in Figure 1. Because the glutathione transferase family used for these searches is large and diverse, it is relatively straightforward to identify the highest-scoring unrelated sequence by doing additional searches with each candidate unrelated sequence. For these examples, the expectation values for the highest scoring unrelated sequences ranged from 0.21 (FASTA DNA) to 2.4 (FASTX). A more comprehensive summary of expectation values for Smith-Waterman (SSEARCH) and FASTA scores is shown in Figures 2 and 3.

The excellent agreement between observed and expected distributions of similarity scores relies on the local character of the sequence alignment. Figure 1A and C also include examples where low gap penalties were used. When gap penalties are too low, alignments shift from local to global and the extreme value statistics no longer apply (Waterman *et al.*, 1987; Mott, 1992; Altschul & Gish, 1996). In contrast, very high gap penalties

Table 1. High-scoring related and unrelated sequences

Smith-Waterman (SSEARCH)					
The best scores are:		len	opt	*z*-sc	*E*(58,741)
GTM1_MOUSE	Glutathione S-transferase GT8.7 (mu)	217	1490	1929.5	10^{-100}
GTP_HUMAN	Glutathione S-transferase Pi (pi)	209	356	457.6	10^{-18}
GTA1_RAT	Glutathione S-transferase Ya (alpha)	221	238	303.8	10^{-10}
SC1_OCTDO	S-Crystallin 1 (OL1).	215	224	285.9	10^{-9}
GTS2_DROME	Glutathione S-transferase 2	247	164	206.7	10^{-4}
GTH3_ARATH	Glutathione S-transferase ERD13	215	142	179.5	0.002
GTT2_HUMAN	Glutathione S-transferase T2 (theta) 2	243	132	165.3	0.012
GTT4_MUSDO	Glutathione S-transferase 4	210	125	157.6	0.033
GTH4_MAIZE	Glutathione S-transferase IV	222	125	157.1	0.033
GTT1_MUSDO	Glutathione S-transferase 1	208	122	153.8	0.054
GTT2_RAT	Glutathione S-transferase YRS (theta)	243	123	153.6	0.056
GTH3_MAIZE	Glutathione S-transferase III	221	115	144.1	0.019
**YJY1_YEAST*	Hypothetical 30.5 kDa protein in SPC1-I	261	110	136.1	0.53
DCMA_METS1	Dichloromethane dehalogenase	266	103	127.9	1.3
GTT1_DROME	Glutathione S-transferase 1	209	100	126.3	1.6
GTH1_WHEAT	Glutathione S-transferase 1	229	98	121.7	3.3
LGUL_SOYBN	Lactoylglutathione lyase	219	97	120.9	3.7
**SLT_HAEIN*	Soluble lytic murein transglycosylase	593	103	119.2	4.6
**MOD5_YEAST*	tRNA isopentenyltransferase	427	100	118.4	5.1
**SPCB_HUMAN*	Spectrin beta chain	2137	108	113.4	9.7
FASTA, *ktup* = 2					
The best scores are:		len	opt	*z*-sc	*E*(58,742)
GTM1_MOUSE	Glutathione S-transferase GT8.7 (mu)	217	1490	1747.0	10^{-90}
GTP_HUMAN	Glutathione S-transferase Pi (pi)	209	356	426.2	10^{-16}
GTP_CRILO	Glutathione S-transferase Pi (pi)	209	352	421.6	10^{-16}
GTA1_RAT	Glutathione S-transferase Ya (alpha)	221	237	287.2	10^{-9}
GTA3_RAT	Glutathione S-transferase 8	222	179	219.6	10^{-5}
GTS2_DROME	Glutathione S-transferase 2	247	161	197.9	0.00019
GTA2_CHICK	Glutathione S-transferase	193	144	179.8	0.0019
SC1_OCTDO	S-Crystallin 1 (OL1).	215	132	165.1	0.013
GTT4_MUSDO	Glutathione S-transferase 4	210	125	157.1	0.036
GTH3_ARATH	Glutathione S-transferase ERD13	215	112	141.8	0.25
**YJY1_YEAST*	Hypothetical 30.5 kDa protein in SPC1-I	261	110	138.1	0.41
GTH4_MAIZE	Glutathione S-transferase IV	222	103	131.1	1.0
GTT1_DROME	Glutathione S-transferase 1	209	100	128.0	1.5
GTT2_RAT	Glutathione S-transferase YRS (theta)	243	97	123.9	2.2
GTT2_HUMAN	Glutathione S-transferase T2 (theta)	243	97	123.5	2.7
**SPCB_HUMAN*	Spectrin beta chain	2137	108	121.4	3.5
**YLB5_CAEEL*	Hypothetical 146.8 kDa protein C	1281	103	119.1	4.7
**DAPF_YERPE*	Diaminopimelate epimerase	198	90	116.7	6.3
**YHC9_YEAST*	Hypothetical 77.8 kDa protein	679	96	115.3	7.6
GT_ECOLI	Glutathione S-transferase	201	88	114.3	8.6
FASTX *ktup* = 2					
The best scores are:		len	opt	*z*-sc	*E*(58,760)
GTM1_MOUSE	Glutathione S-transferase GT8.7 (mu)	217	1490	1860.6	0
GTP_HUMAN	Glutathione S-transferase Pi (pi)	209	337	422.0	10^{-16}
GTP_CRILO	Glutathione S-transferase Pi (pi)	209	333	417.0	10^{-16}
GTA1_RAT	Glutathione S-transferase Ya (alpha)	221	208	260.2	10^{-7}
GTA3_RAT	Glutathione S-transferase 8	222	149	186.5	0.00082
GTA1_CAEEL	Probable glutathione S-transferase	207	125	157.5	0.030
SC1_OCTDO	S-Crystallin 1 (OL1).	215	120	150.5	0.083
GTH3_ARATH	Glutathione S-transferase ERD13	215	107	134.3	0.67
GTT1_MUSDO	Glutathione S-transferase 1 (theta)	208	101	127.0	1.7
SC3_OCTDO	S-Crystallin 3 (OL3).	215	99	124.3	2.4
**RPB1_CRIGR*	DNA-directed RNA polymerase II	467	103	124.2	2.4
**CA19_RAT*	Collagen alpha 1 (IX) chain	325	100	122.9	2.9
**TGFB_HUMAN*	Latent TGF-beta binding protein	1394	107	122.1	3.2
GTT2_HUMAN	Glutathione S-transferase (theta)	243	97	121.0	3.7
GTT2_RAT	Glutathione S-transferase YRS (theta)	243	97	121.0	3.7
GTH5_ARATH	Glutathione S-transferase PM239	218	96	120.5	3.9
**PRPS_MOUSE*	Proline-rich protein MP-3	296	96	118.5	5.1
**YJ9P_YEAST*	Hypothetical 118.4 kDa protein	1161	103	118.3	5.2
DNA FASTA *ktup* = 4					
The best scores are:		len	opt	*z*-sc	*E*(122,490)
MUSGLUTA	Mouse GST1-1 mRNA (mu)	1287	6435	5672.7	0
RATGSTY	Rat GST Yb (mu)	6294	372	307.9	10^{-10}
HSGSTM1B	*Homo sapiens* GSTM1b (mu)	2667	358	301.1	10^{-10}
MMGSTM3	Mouse GSTM3 gene (mu)	422	331	289.1	10^{-8}

Table 1—*Continued*

The best scores are:		len	opt	*z*-sc	*E*(122,490)
HSGSTMU3	Human GSTmu3 gene (mu)	1820	322	271.8	10^{-8}
HSGSTPI	Human mRNA for GST-pi mRNA (pi)	714	237	202.6	0.00025
RRGTS8	*R.rattus* mRNA for GST8 (alpha)	893	182	152.7	0.12
**BTRNAXOR*	*Bos taurus* xanthine oxidoreductase	4719	175	135.8	0.21
**HUMKAL2*	Human glandular kallikrein gene	6139	170	129.7	0.35
**HUMTROPI01*	Human troponin I TNNI1 gene	1475	170	132.4	0.54
RNGSTYC2F	Rat GST Yc2 (alpha)	1129	170	140.6	0.47
MUSGSTYC	Mouse GST Yc (alpha)	950	168	140.0	0.60
**MUSTHYGP*	Mouse Thy-1.2 glycoprotein	5572	163	124.1	0.78
**HS186D3R*	*Homo sapiens* CpG island DNA	334	163	142.3	1.3

High scoring sequences from the searches shown in Figure 1. Only selected related sequences are shown; all the highest-scoring unrelated sequences are shown. High scoring unrelated sequences are highlighted with an asterisk (*) and italics. The sequence length (len) and several similarity measures are shown. The opt column is the uncorrected Smith-Waterman score for SSEARCH searches, and the ''optimized'' FASTA score (see the text) for FASTA searches. The *z*-sc column reports the length-corrected *Z*-score for the alignment. The *E*(*N*) value reports the number of times the score should be obtained by chance for a search against a database of size *N*. For searches of the SwissProt database, $N \sim 58{,}750$; *N* varies slightly because of the different numbers of excluded sequences with the different searches.

simply produce fewer alignments with gaps and move the algorithm towards the BLAST HSP model, where the extreme value distribution was first shown to apply (Karlin & Altschul, 1990; Altschul & Gish, 1996). While high gap-penalties do not compromise the statistical model, they can reduce the effectiveness of the search Pearson, 1995).

The agreement between observed and expected numbers of sequences decreases when low gap penalties were used. In Figure 1A, there is an excess of sequences with scores from 50 to 70; in Figure 1C there are too many observed sequences with scores from 80 to 100. The effect on the expectation value for both distantly related and highest-scoring unrelated sequences can be dramatic. For the Smith-Waterman (SSEARCH) protein sequence comparison, reducing the gap penalty to $-8/-1$ raised the expectation value for the highest-scoring unrelated sequence from 0.4 (YJY1_YEAST, Table 1) to 81 (ABF2_YEAST, Table 2). Likewise, expectation values for related sequences increased several orders of magnitude, so that plant enzymes with expectation values from 0.002 to 0.19 using ''reasonable'' gap penalties increased to 6.8 to 16 when the gap penalties were reduced. For this protein sequence comparison, improper gap penalties increased the raw Smith-Waterman similarity score but decreased the significance of the match.

Low gap penalties can also produce erroneously low estimates of statistical significance. When gap penalties for the glutathione transferase cDNA database search are decreased from $-16/-4$ to $-12/-2$, the expectation value for the highest scoring unrelated sequence dropped from 0.21 to 0.025, thus indicating statistically significant similarity at the <0.05 criterion. As before, the expectation values for all the divergent related sequences increased substantially. With appropriate ''local'' gap penalties, DNA sequence comparison can detect significant similarities between the class-mu

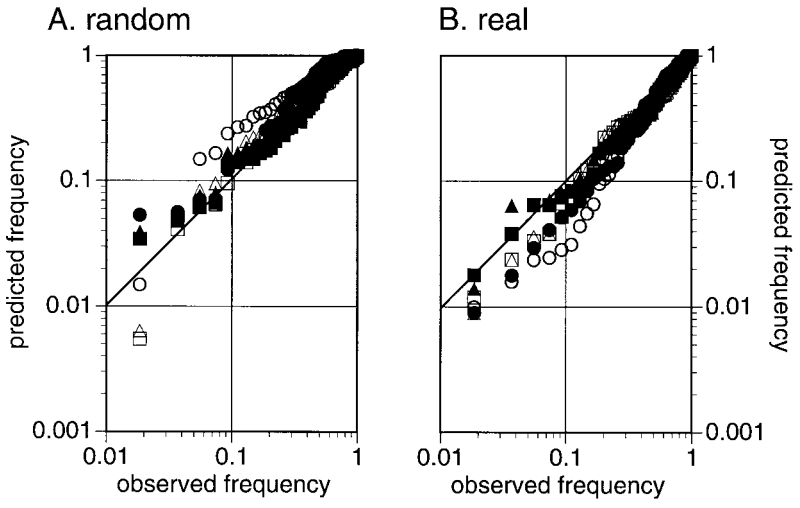

Figure 2. Distribution of *P*()-values, Smith-Waterman. Randomly shuffled (A) or unshuffled (B) sequences from 54 PIR39 superfamilies were used to search the PIR39b database. Expectation values for the highest scoring sequence (A) or the highest scoring unrelated sequence (B) were converted to probabilities using the Poisson formula: $P() = 1 - e^{-E}$. The probabilities (predicted frequency) were sorted from lowest to highest and plotted as a fraction of the total number of searches (54, observed frequency). Predicted frequencies (probabilities) for the highest unrelated sequence score expectation values calculated using regress1 ($-12/-2$, ■; $-14/-2$, □), regress2 ($-12/-2$, ▲; $-14/-2$ △), or Altschul-Gish (●, ○) parameters.

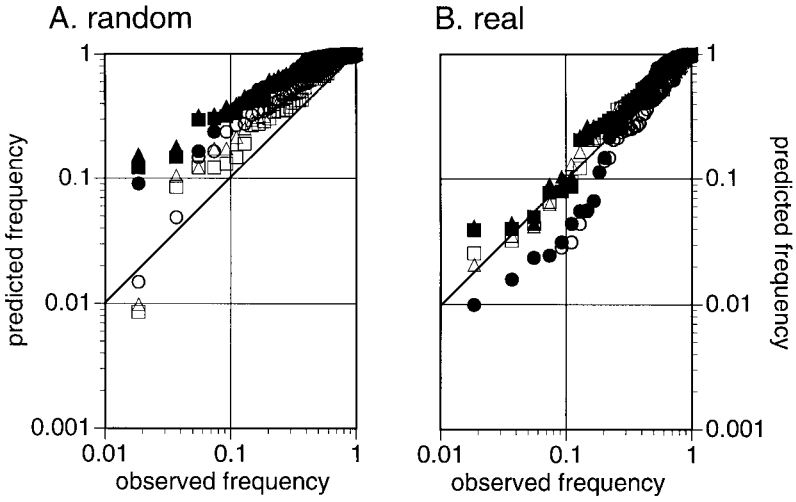

Figure 3. Distribution of *P*()-values. The procedure described for Figure 2 was performed with scores calculated with FASTA, *ktup* = 2 (filled symbols) and *ktup* = 1 (open symbols) with either regress1 (squares) or Altschul-Gish (circles) parameters.

query sequence and a class-pi mRNA (HSGSTPI, $E() < 0.00025$). With the lower gap-penalties, the HSGSTPI sequence has a non-significant value ($E() < 0.51$). thus, increasing the gap penalty to ensure that the highest-scoring unrelated sequence similarity score is ~1.0 can improve dramatically the detection of distantly related family members.

To confirm that the statistical estimates generated by SSEARCH and FASTA are accurate, we performed database searches with sequences from 54 protein families against an annotated protein sequence database (see Methods). Two groups of sequences were used, 54 random sequences derived by shuffling a sequence from each of the 54 families or the 54 original unshuffled sequences. For the random sequences, we determined the expectation of the highest-scoring library sequence (since the query sequence is random, this is the highest scoring unrelated sequence and the score should have an expectation value ~1). For the unshuffled sequences, we determined the expec-

Table 2. Low gap-penalties reduce sensitivity

SSEARCH, −8/ − 1					
The best scores are:		len	opt	z-sc	*E*(58,661)
GTM1_MOUSE	Glutathione S-transferase GT8.6 (mu)	217	1490	832.0	10^{-39}
GTP_HUMAN	Glutathione S-transferase Pi (pi)	209	378	212.8	10^{-5}
GTA1_RAT	Glutathione S-transferase Ya (alpha)	221	304	170.6	0.0063
SC1_OCTDO	S-Crystallin (OL1).	215	279	157.1	0.036
GTH3_ARATH	Glutathione S-transferase ERD13	215	190	115.2	6.8
GTH4_MAIZE	Glutathione S-transferase IV	222	205	115.3	7.6
DCMA_METS1	Dichloromethane dehalogenase	266	200	110.3	14
GTH3_MAIZE	Glutathione S-transferase III	221	194	109.3	16
GTT2_HUMAN	Glutathione S-transferase T2 (theta) 2	243	189	104.9	29
GTT1_MUSDO	Glutathione S-transferase 1	208	183	104.2	32
GTH1_WHEAT	Glutathione S-transferase 1	229	185	103.6	34
GTT2_RAT	Glutathione S-transferase YRS (theta)	243	177	98.2	68
**ABF2_YEAST*	ARS-binding factor 2 precursor.	183	166	96.9	81
LGUL_SOYBN	Lacoylglutathione lyase	219	171	96.6	84
DNA FASTA *ktup* = 4, −12/ − 2					
The best scores are:		len	opt	z-sc	*E*(123,689)
MUSGLUTA	Mouse GST1-1 mRNA (mu)	1287	6435	2514.3	10^{-133}
HUMGSTM3A	Human GSTM3 cDNA (mu)	1266	1608	641.2	10^{-28}
HSGSTMB1B	*Homo sapiens* GSTM1b gene (mu)	2667	462	191.8	0.00028
RATGSTY	Rat GSTYb (mu)	6294	386	157.0	0.01
**MMUI66249*	Mouse cut alternate splice (CASP)	1757	373	159.8	0.025
HSGSTMU3	Human GSTmu3 gene (mu)	1820	358	153.8	0.053
MMGSTM3	Mouse HSTM3 gene (mu)	422	354	161.3	0.087
**MMSPARCR*	Mouse mRNA for cysteine rich glycoprot.	2079	333	143.3	0.18
RATGSTPPS	Rat GST-P (pi) pseudogene	1213	323	142.7	0.33
HSGSTPI	Human class Pi GST (pi)	714	317	143.4	0.51
HSGSTMU2	Human GSTmu2 gene (mu)	1222	309	137.3	0.66

Selected high scoring related, and highest scoring unrelated sequences from the searches with low gap penalties from Figure 1A and C. High scoring unrelated sequences are highlighted with an asterisk (*) and italics.

tation value of the similarity score of the highest-scoring unrelated sequence.

In addition to examining the expectation values calculated by the regress1 regression strategy described in Methods, we examined expectation values based on the λ, *K*, and *H* parameters published by Altschul & Gish (1996) and values based on an alternative regression strategy (regress2, see Methods). The results for all 54 sequences were combined by plotting the cumulative fraction (observed frequency) of the 54 sequences *versus* the indicated probability (predicted frequency). The indicated probability (*P*()-value) was calculated from the expectation value (*E*) using the formula $P(E)\text{-}1 - e^{-E}$; this is the *P*()-value reported by the BLAST programs. With this transformation, a plot of the cumulative fraction of query sequences *versus* the *P*()-value of the highest scoring sequence should have a slope of 1.0, which is indicated as a diagonal line on the plot. Thus, expectation values with a *P*()-value < 0.02 would occur about 2% of the time, *P*() < 0.1 about 10% of the time, etc. (Figure 2). If the calculated expectation values (and their associated Poisson probabilities) are conservative, then the points will lie above and to the left of the diagonal. Points below the diagonal indicate that the expectation values are too low (and thus incorrectly imply statistical significance, Figures 2 and 3).

Figure 2 shows that the two fitting strategies (regress1 and regress2) and Altschul-Gish parameters produce accurate and conservative estimates for the expectation value for the highest scoring library sequence when randomly shuffled protein sequences are used. When unshuffled sequences are used and the scores of the highest scoring unrelated sequences are examined, the estimated regress1 and regress2 values are more conservative; with shuffled sequences, Altschul-Gish estimates are slightly more conservative. Thus, the regress1 and regress2 statistical estimates are at least as accurate as those calculated using the Altschul-Gish parameters for protein sequences when the Smith-Waterman algorithm is used.

The estimation strategies are quite flexible; they can be used for similarity scores calculated by FASTA for protein or DNA searches, and for translated-DNA protein sequence comparison. Figure 3 shows that the expectation values calculated for FASTA protein sequences, with either random (A) or unshuffled (B) sequences, are quite accurate, both when searches are performed with the faster $ktup = 2$ or the more sensitive $ktup = 1$ search strategy. As with the Smith-Waterman searches (Figure 2), expectation values from searches performed with real unshuffled sequences are somewhat lower than those obtained for random sequences, but the estimates for both random and real sequences rarely overestimate the significance of a similarity more than twofold.

Improved search performance

The best estimation strategy should not only provide accurate estimates, it should also improve similarity search performance by correcting for the expected increase in unrelated sequence similarity score with length. In an earlier paper (Pearson, 1995), we showed that scaling similarity scores by $\log(n_q)/\log(n_l)$, where n_l is the length of the library sequence and n_q is the length of the query sequence, significantly improved both Smith-Waterman and FASTA similarity searches, allowing more distantly related sequences to be identified. The extreme value distribution parameters calculated by regress1, regress2 fitting and the Altschul-Gish parameters can also be used to correct raw similarity scores for the expected effect of library sequence length†.

We compared the performance of similarity searches performed with five different length-scaling procedures (as well as unscaled scores) using 110 different PROSITE protein families identified as challenging by Henikoff & Henikoff (1993) (Figure 4). This set of protein families is twice as large as the earlier PIR39 based group; in addition, it includes more families containing modular protein domains. To evaluate the relative performance of a pair of length-scaling strategies, we examined both the equivalence number criterion used earlier (Pearson, 1995) and also a second measure of search quality, the number of unrelated sequences found plus related sequences missed at the *E*() < 0.02 significance level. The ''equivalence number'' is the number of related sequences that score at or below a similarity score (*z*-score) that balances the number of related sequences at or below the value and the number of unrelated sequences with scores above the value; i.e. the score where the number of false-positives equals the number of false-negatives. As before, we calculated the *z*-value for the difference in performance between the best length-scaling method and the alternatives; differences in performance are significant at the 0.05 level if $z > 2$.

When the equivalence number criterion is used; excellent performance is seen with the regress1 and regress2 scaling strategies and gap penalties of −14 for the first residue in a gap and −2 for each additional residue in the gap. Searches with regress1 scaling perform significantly better than comparison with unscaled-scores. With this database, the regress1 procedure also performs better than the older $\log(n_q)/\log(n_l)$-scaling (log()) that was the best performer in our earlier study (the difference is significant when the −14/−2 gap penalty is used). We attribute this difference to the presence of many more domain-shuffled sequences in the Prosite/SwissProt based test set and to improvements in the regression scaling strategy. (The method used in our earlier paper is shown as

† The $\log(n_q)/\log(n_l)$ scaling, though effective in practice, does not have any theoretical basis.

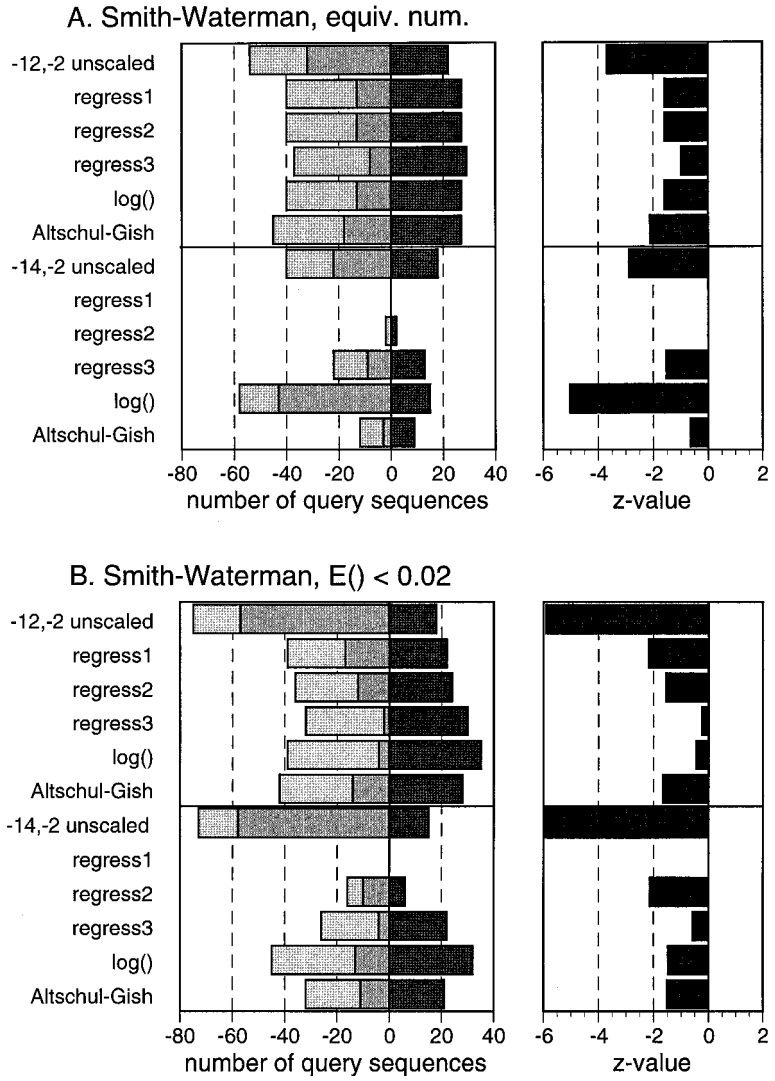

Figure 4. Search performance, Smith-Waterman. The relative performance of different length-scaling strategies with searches of 110 query sequences from the Prosite/SwissProt challenging families against the SwissProt rel. 34 database is shown. In each panel, the top half reports searches with a $-12/-2$ gap penalty; the bottom half $-14/-2$. All comparisons are against regress1, $-14/-2$. A, Performance using the equivalence number criterion. The number of sequences performing better or worse (left panel) and the z-value of the difference (right panel; Pearson, 1995) is shown. z-value differences of 2.0 or greater are statistically significant at the $P() < 0.05$ level. B, Performance evaluated using the $E() < 0.02$ criterion.

regress3.) The difference in performance for the three different regression scaling strategies are not significant. On this dataset, searches with similarity scores scaled using Altschul-Gish parameters (gap penalty $-14/-2$) are slightly less effective than regression scaling, but the difference is not statistically significant.

The equivalence number criterion balances the number of false-positive and false-negative results reported in a search, regardless of whether the ''related'' sequences have statistically significant similarity scores. For automated sequence classification, a more useful criterion is the number of related sequences missed (false negatives) plus the number of unrelated sequences ''found'' (false positives) at a specified confidence limit. e.g. $E() < 0.02$. Using this criterion, regress1 with gap penalties of $-14/-2$, or regress3 and Altschul-Gish with penalties of $-12/-2$ are the most effective (Figure 4B). As before, searching with unscaled scores is significantly less effective.

Figure 5 also compares searches with FASTA in its most sensitive mode ($ktup = 1$). When either the equivalence number or the $E() < 0.02$ criterion is used, the Smith-Waterman algorithm always performs significantly better than FASTA with $ktup = 1$ (and better than FASTA, $ktup = 2$, not shown). Not only does the Smith-Waterman algorithm rank distantly related sequences above high-scoring unrelated sequences, thus improving the equivalent number, Smith-Waterman also identifies additional distantly related sequences at the $E() < 0.02$ criterion.

We also examined the relative performance of the different scaling methods on the 54 protein families from the revised PIR39 dataset (Figure 6). On these protein families, the Altschul-Gish parameters provided the most effective searching when the equivalence number criterion was used; regress1 scaling did not perform as well, but the difference was not statistically significant. In contrast to our earlier study (Pearson, 1995), $\log(n_1)$ scaling is not significantly better than regress1 with $-14/-2$ gap penalties. This difference may reflect the more careful selection of the query sequences in the current study; we expect that log()-scaling performs better on this dataset

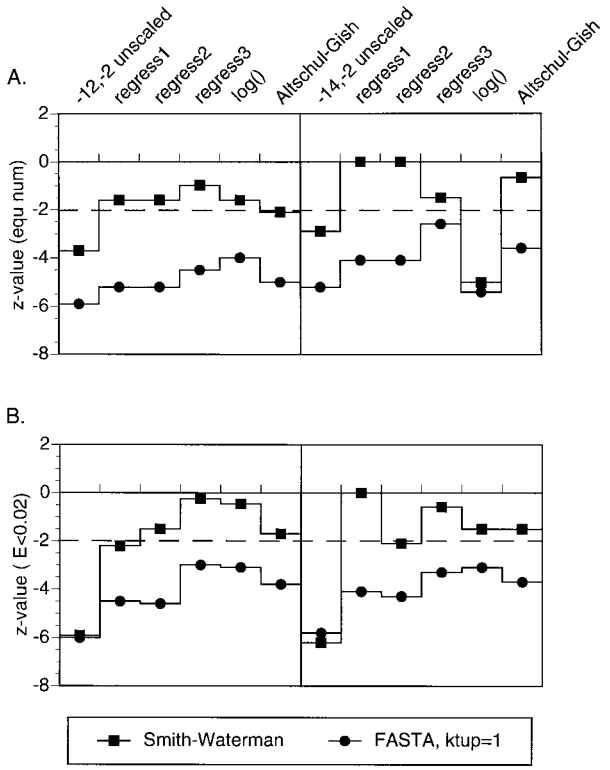

Figure 5. Search performance, Smith-Waterman and FASTA. The plot shows the difference in performance between reference Smith-Waterman searches using gap penalties of −14/−2 and regress1 statistical estimates. A, Equivalence number criterion; B, $E() < 0.02$ criterion, and the gap-penalties and normalization strategy named across the top of the panel. When reference search performs better, the *z*-values are negative and indicate the significance of the difference in performance. Searches were performed using Prosite/SwissProt families and SwissProt rel. 34. Searches on the left half of each panel used gap penalties of −12/−2; searches on the right half used −14/−2. Differences with *z*-values <−2.0 or >2.0 are statistically significant.

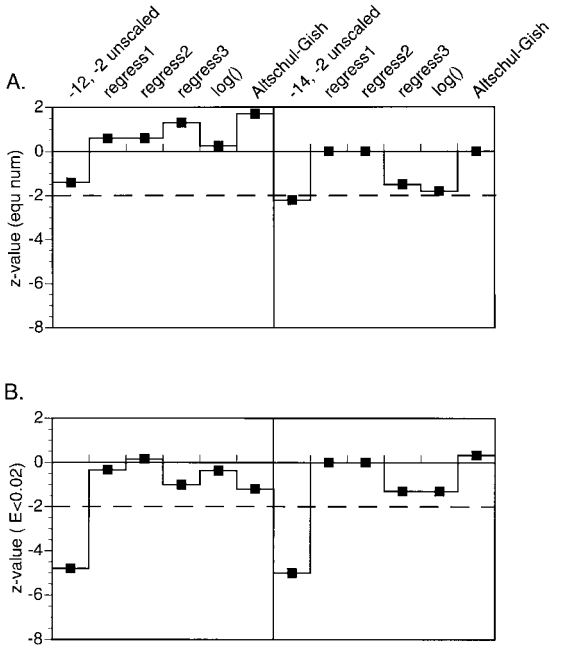

Figure 6. Search performance, PIR39. Search performance using two sequences from each of 54 PIR39b families is shown. Panels are labeled as for Figure 5 and all comparisons are with respect to reference Smith-Waterman searches (regress1, −14/−2). A, Comparison using the equivalence number. B, Comparison using the $E() < 0.02$ criterion.

because most of the query sequences share global similarity with their homologues in the PIR39 protein database.

Discussion

We have examined several strategies for correcting the length-dependence of local protein sequence similarity scores. The default method used by programs in versions 2.0 and 3.0 of the FASTA package (Pearson, 1996), regress1, produces accurate statistical estimates and significantly improves search performance over unscaled similarity scores. In addition, Altschul-Gish scaling and regress2 and regress3 scaling are available in current versions of the programs; however, Altschul-Gish scaling is available only for protein sequence comparisons.

The accurate and conservative statistical estimates shown in Figures 2 and 3 are due in part to our selection of the 54 query sequence families. Query sequences with highly biased amino acid compositions, such as histones, keratins, metallothioneins, produce ''significant'' ($E() \ll 0.01$) matches with unrelated library sequences because of sequence bias alone. This problem has been well documented (Wootton & Federhen, 1993; Altschul *et al.*, 1994) and tools are available to identify and remove highly biased regions in protein and DNA sequences (Wootton & Federhen, 1993).

Our results suggest that, once biased composition regions have been removed, the statistical estimates can be relied upon to reflect accurately the frequency of a similarity score occuring by chance. Investigators often wonder: what $P()$-value or $E()$-value should be used to infer homology? The answer to this question depends both on the number of searches that are being performed and the investigator's concern about inferring homology erroneously. If one search is performed each week on a newly sequenced protein, using the criterion $P() = E() < 0.02$ for inferring homology, then, on average, one sequence in 50, or one per year, will have a homology assigned incorrectly. These ''false positive'' errors can be reduced by using a more strict criterion for inferring homology (e.g. $P() = E() < 0.001$), but this will cause additional distant homologues to be missed. For example, when the $E() < 0.02$ criterion is used with the 110 Prosite/SwissProt query sequences, 4674 related

sequences are missed in the 110 searches; when the stricter $E() < 0.001$ criterion is used, an additional 732 sequences are missed.

Estimation of statistical significance of local similarity scores from the empirical distribution of similarity scores was first proposed by Collins *et al.* (1988). These authors recognized that the number of high-scoring sequences declined exponentially and suggested fitting a line to all but the highest scoring 3% of similarity scores. They did not correct for the effect of library sequence length on unrelated sequence similarity score; without length correction, search performance is significantly worse. In addition, they did not estimate a scale parameter, making it difficult to estimate accurately the statistical significance of very high scoring sequences.

Mott (1992) used maximum likelihood estimation of the parameters of the extreme value distribution to evaluate statistical significance. His approach is similar to the regress2 estimation evaluated here, except that maximum likelihood estimation was used instead of linear regression and a composition parameter (c) as well as library sequence length (n_l) was included in the likelihood estimation. Calculation of c may provide even more effective scaling of unrelated similarity scores with similar compositions, but it is computationally intensive, as c is the positive root of the equation $\Sigma_{u,v} p_u q_v e^{S(u,v)/c} = 1$ (Mott, 1992). Since c depends both on the amino acid composition of the query sequence p_u and each library sequence q_v, it must be recalculated for each sequence comparison. Length scaling clearly provides a significant improvement in search effectiveness; we plan to examine the additional benefit of composition-scaling in the future.

An alternative strategy for estimating λ and K was described by Waterman & Vingron (1994). Their approach estimates the statistical parameters by examining the distribution of sub-optimal alignment scores. The approach requires additional alignments to be calculated, and sequences with internal duplications can confuse the estimation procedure if the duplications are not recognized. Their approach is especially effective in examining the significance of a single pairwise alignment, since an arbitrary number of sub-optimal scores can be generated from a pair of sequences. The Waterman-Vingron estimate reflects the difference between the optimal alignment score for a sequence pair and alternative alignment scores for the same pair of sequences, and thus is similar to estimates produced by randomly shuffling one of the sequences and examining the distribution of scores. The database-based parameters described here estimate the significance of a similarity in the context of an entire protein sequence database search, rather than alternative alignments of two sequences. Expectation values calculated from database searches are often quite similar to those calculated by comparisons to shuffled sequences (Pearson, 1996).

A strength of our empirical approach is that, at least in theory, it can produce expectation values for any local similarity scoring function. For example, the regress1 statistical estimates are used in FASTA-SWAP and FASTA-PAT (Lafunga *et al.*, 1996), which are used to search consensus pattern databases. However, before relying on the empirical statistical estimates calculated by FASTA for unconventional scoring matrices, gap-penalties, or databases, an analysis similar to those shown in Figures 2 and 3 should be performed to ensure that the expectation values of high scoring and unrelated sequences are accurate.

Length scaling significantly improves similarity searching performance (Figures 5 and 6; Pearson, 1995) by correcting for the variance of unrelated sequence similarity scores are likely to improve search performance as well. Clearly, the effect of amino acid composition should be considered, as was done by Mott. In addition, it may be possible to correct for the effect of hydrophobic patches and of low complexity regions. Reducing the ''noise'' from high scoring, unrelated sequences may provide additional improvements in search performance.

Methods

Sequence libraries and similarity searching

Searches were performed on the annotated portion (PIR1) of the National Biomedical Research Foundation protein sequence database (Barker *et al.* (1990), release 39, 31 December 1993, 4,306,189 amino acid residues in 11,982 sequences), augmented as described by Pearson (1995). This older library, and the same set of query sequences, was used to provide consistency with the earlier work. The library has been annotated so that every sequence in the database has been assigned to a protein superfamily. The earlier experiments compared the performance of two comparison methods, a reference method and an experimental method, to reduce the effects of a superfamily misclassification. This report focuses on the statistics of high-scoring unrelated sequences, which required modifications to the original reference database. Searches were performed with the Smith-Waterman algorithm and when high-scoring, apparently unrelated sequences were found, additional searches were done to confirm that the sequences were in fact unrelated. In several cases (for example, serine proteases, protein kinases, immunoglobulins, and calcium binding proteins) it became clear that homologous members of the same protein family had been assigned different superfamily numbers. These sequences from the same superfamily were given the same superfamily number. This database is referred to as PIR39b, and is available from `ftp://ftp.virginia.edu/pub/fasta`.

Searches were performed with 54 of the 67 query sequences selected from the PIR39 database listed in Table I of Pearson (1995). Thirteen of the previous superfamilies were excluded either because they were homologous with other superfamilies (immunoglobulin kappa V-I, kappa C, class-I HLA) or because they contained regions with highly biased amino acid composition (HIV gag polyprotein, HPV L2 and E2 proteins, hepatitis core antigen, muc, keratin, protamine Y2, histone H1b, soybean protease inhibitor, and metallothionein). A database of random query sequences with the same length and amino acid composition as the 54 query PIR39 query sequences was constructed by uniformly shuffling each of the 54 sequences using the program `randseq`. The families with biased amino acid composition were identified when searches with the randomly shuffled sequences produced large numbers of sequences with high similarity scores.

A second set of query sequences was developed from the 254 challenging PROSITE (Bairoch, 1991) pattern families described by Henikoff & Henikoff (1993). Query sequence families were selected from the original PROSITE families if the family had 40 or more members in release 34 of SwissProt (Bairoch & Boechmann, 1991). A copy of SwissProt was modified to include PROSITE pattern numbers and additional sequences were labeled with the numbers if they shared significant similarity ($E() < 0.005$) with family members. The resulting set of Prosite/SwissProt-based query sequences contains 110 sequences. The PIR39 and Prosite/SwissProt34 query sets have 20 protein families in common. Both the revised PIR39 annotated database and the annotated SwissProt 34 database are available from `ftp://ftp.virginia.edu/pub/fasta`.

Similarity searches and scoring matrices

Searches with the FASTA (Pearson & Lipman, 1988; Pearson, 1990) and Smith-Waterman (Smith & Waterman, 1981; Pearson & Miller, 1992) algorithms were performed in parallel on a DEC alpha 2100 4/275 using a general platform for large-scale sequence comparison as described by Pearson (1995). Version 3.0t of the FASTA programs was used. Protein sequence comparisons were performed with the BLOSUM50 scoring matrix (Henikoff & Henikoff, 1992) using gap penalties of -12 or -14 for the first residue in a gap and -2 for each additional residue. DNA sequence comparisons used the scoring matrix used by BLASTN, which scores a match as $+5$ and a mismatch as -4. The standard DNA gap penalty was -16 for the first residue in a gap and -4 for additional residues. Sequence comparisons with FASTX, which compares a translated DNA sequence to a protein sequence library, used a gap penalty of $-15/-3$ with a frame-shift penalty of -30.

Statistical estimates for scaled similarity scores

Six methods, Altschul-Gish, log()-`scaled`, `scaled`, `unscaled`, regress1, regress2, and regress3, were used to calculate statistical estimates for similarity scores. Altschul-Gish estimates were calculated using the table of λ, K, and H parameters described by Altschul & Gish (1996). These parameters were than used in the following equation to calculate the probability of obtaining a score in a single sequence comparison:

$$P(S > x) = 1 - \exp(-Km'n'e^{-\lambda x}) \qquad (2)$$

where m, n, are the length of the query and library sequences, respectively, and $m' = m - \ln(mn)/H$ and $n' = n - \ln(mn)/H$. $P()$-values from equation (2) were converted expectation values for obtaining a score in a database search using equation:

$$E(S) = P(S)N \qquad (3)$$

where N is the number of sequences in the library database (the number of times a score was calculated).

We also evaluated four methods for estimating statistical significance based on the distribution of similarity scores that are calculated during a similarity search of a protein or DNA sequence database. The mean μ and standard deviation σ (which are related to Kmn and λ) can be estimated from the distribution of similarity scores produced in a search, but they will vary depending on the scoring matrix and gap penalties (Altschul & Gish, 1996). The estimation is straightforward if all the sequences are unrelated, as occurs when random sequences are compared to one another (Altschul & Gish, 1996), or are used to search a sequence database (Mott, 1992). However, the estimation is more difficult if a ''real'' protein sequence is used to calculate the similarity scores because some of the similarity scores may be very high when homologous proteins are present. If ''contaminating'' high similarity scores from homologous sequences are not removed from the estimation, the scale parameter is greatly overestimated, which reduces the statistical significance of related sequences.

The simplest estimation method (`unscaled`) calculates the mean μ and variance σ^2 for all the unscaled similarity scores and then calculated a z-value for each score using the formula: $z = (S - \mu)/\sigma$. Similarity scores from related sequences are then removed by excluding sequences with z-values >7.0 or <-3.0 and the process of estimation and exclusion is repeated as many as five times.

This z-value was converted to a probability using the extreme value distribution:

$$P(Z > z) = 1 - \exp(-e^{-z\pi/\sqrt{6} - \Gamma'(1)}) \qquad (4)$$

from which an expectation was calculated using equation (3).

Equation (4) can be used to convert any similarity z-value, including length normalized scores

calculated by the regress1, regress2, regress3, or log()-scaling methods described below, into the probability $P(z)$ of obtaining that score by chance in a single sequence comparison. Once $P(z)$ for one comparison is known, equation (3) can be used to estimate the number of times the score would be expected after the N (typically 10,000 to 250,000) sequence comparison performed in a database search.

We also examined the performance of a simple log-length correction (log()-scaled) described by Pearson (1995). Similarity scores were multiplied by the term $\ln(200)/\ln(n)$, where n was the length of the library sequence; the mean and variance were estimated as with the unscaled scores above. This method has no theoretical basis; it was included because is had performed well in earlier tests.

The regression-based methods for estimating the location and scale parameters use a pruning strategy to remove exceptionally high scores that are calculated when related sequences are compared. Because one does not know whether a score is exceptionally high until the length and scale parameters are estimated, if high scores are included in this estimation process, the scale parameter will be too large, with the result that scores from related sequences will not be excluded. Two pruning strategies were used: regress1 performs two regressions of the scores with respect to $\log(n_l)$ and excludes scores from library sequence length bins with high variance and then performs a final regression (Figure 7; regress3 uses the same pruning strategy); regress2 performs the same regression, but estimates the scale parameter from the variance of the center 90% of library sequence length bins ranked by residual variance, repeating this process up to five times (in practice, it is done twice, on average). Estimates for regress1 and regress3 regression-scaled scores are calculated by the following steps.

(1) The similarity scores are "binned" into a histogram as a function of $\ln n_l$, where n_l is the length of the library sequence: $10\ln(10) = 23$ bins were used for each tenfold change in length. The mean μ and standard error of the mean (SEM, $\sigma^2/\sqrt{N}$) for each bin are calculated.

(2) A line is fit using linear regression weighted by the SEM through the mean scores (μ) of the bins (Figure 7A and C):

$$\mu = \rho \ln n_l + \psi \tag{5}$$

(3) The z-value:

$$z = (S - \mu)/\sigma \tag{6}$$

of each similarity score is calculated and scores with z-values <-3.0 or >5.0 are removed from the bins (■, Figure 7A and C). For the 110 query sequences used to search the SwissProt database with the Smith-Waterman algorithm (gap

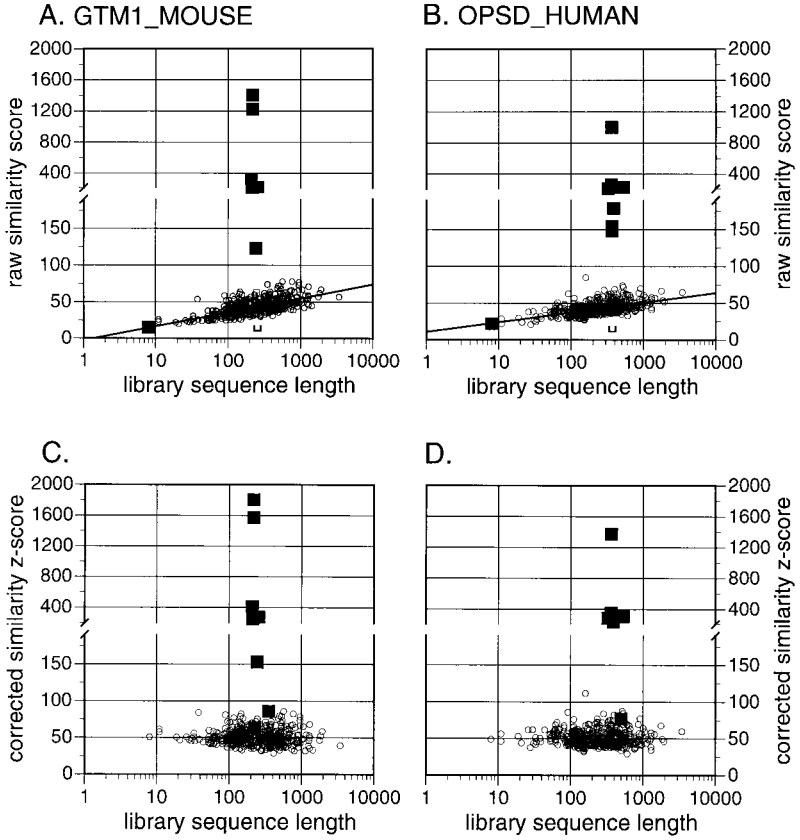

Figure 7. Length correction of similarity scores. The effect of the regress1 strategy on similarity scores for the `gtm1_mouse` (A and C) and `opsd_human` (B and D) query sequences. Searches were performed with a 500 sequence subset of the PIR39b protein sequence database; scores for every sequence in the database are shown. A and B, The length dependence of the raw Smith-Waterman similarity scores. Large filled squares indicate scores that were excluded because they had z-scores above the threshold used to prune "outliers" (step 3; seven scores for `gtm1_mouse`; seven for `opsd_human`). The bracket indicates the bin that was excluded because it had residual standard error higher than three times the average standard error (step 5; 55 scores in one bin at length 245 to 269 for `gtm1_mouse`; 59 scores in one bin at 365 to 402 for `opsd_human`). The final regression line used to calculate the expected mean score as a function of library sequence length is also shown. C and D, The length dependence and distribution of the length-corrected z-scores.

penalty −14/−2), this step removed 109(±10) (mean(±SEM)) of the 59,020 sequences per query.

(4) The linear regression coefficients (equation (5)) are recalculated.

(5) Bins with residual standard error three times the average residual standard error are removed from the calculations. For the 110 query sequences used to search SwissProt, this step removed 0.87(±0.06) bin with 1343(±131) sequences per query, an average. The bins removed in Figure 7A and C are indicated with a ⊔.

(6) The linear-regression formula (equation (5)) is calculated a third time, and the average residual variance $\hat{\sigma}^2$ is calculated.

(7) regress1 z-values are calculated using the formula:

$$z(S) = \frac{S - \rho \ln n_l - \psi}{\sqrt{\sigma^2}} \qquad (7)$$

regress3 (Pearson, 1995) calculates σ from a second regression against $\ln n_l$; while regress1 simply uses the average residual variance.

(8) Z-scores (Z) are derived from the z-values (z) using the formula: $Z(z) = 50 + 10z$. (As a result of this scaling, Z-scores have approximately the same magnitude as raw similarity scores calculated with the PAM250 or BLOSUM50 matrices.)

As can be seen in Figure 7B and D, the length-corrected Z-scores have an average 50, independent of length. As a result of the length correction, the variance of the similarity scores is typically reduced from 30 to 50%, which can improve the statistical significance of a related sequence score 5 to 20-fold or more, while the scores of long (>2000 amino acid residues) library sequences are often reduced from 0.5 to 1 standard deviation.

The regress2 strategy uses the following steps.

(1) As with regress1, the similarity scores are ''binned'' into a histogram as a function of $\ln n_l$, where n_l is the length of the library sequence. The mean μ and SEM of each bin are calculated.

(2) A line is fit using a weighted linear regression through the mean cores (μ) of the bins using equation (5).

(3) The average residual variance of the bins was calculated excluding 10% of the bins, 5% with the largest and 5% with the smallest variance.

(4) The z-value (equation (6)) of each similarity score is calculated and scores with z-values <−3.0 or >7.0 are removed form the bins and the mean and SEM of the bin is re-calculated.

(5) Steps (2) to (5) are repeated up to five times, or until fewer than five sequences have been removed. For the 110 query sequences used to search the SwissProt database, 97.4(±12.9) of 59,020 library sequences were excluded after 2.5(±0.07) iterations.

(6) The linear-regression formula (equation (5)) is calculated a final time, and the average residual variance $\hat{\sigma}^2$ of all the bins is calculated.

(7) Z-values are calculated using equation (7).

(8) Z-scores calculated from z-values as before.

Z-values for unscaled and log-length scaled scores are calculated from the mean and variance of the scores, after iteratively pruning scores with z-values >7.0 and <−3.0. (The mean and variance was recalculated up to five times after removing the high and low scoring sequences.) This iterative strategy is also used to recalculate the $\log(n_l)$ regression coefficients in the regression scaling instead of removing length ''bins'' with high variance.

Variations on the bin sizes and pruning limits for regress1, regress2, and regress3 strategies were examined; the limits chosen tend to exclude a small number of high or low scoring sequences in 0 to 2 length bins (regress1, regress3) or after only a few iterations (regress2). The bin size chosen yields a large number (>45 for library sequence lengths from 100 to 1000 amino acid residues) of well-populated bins from sequences that differ in length by about 10%. Because it excludes all the scores in a library sequence length bin, the regress1 strategy appears more aggressive. However, in the searches of SwissProt, scores from only 1 of the 62 length bins are excluded, on average, leaving more than 95% of the scores in the remaining 61 bins. The regress2 strategy pruned about 100 scores (<0.2% of the sample) per query sequence, which is similar to the average number of related sequences per query (99 ± 10). Of course, the most distantly related sequences are not excluded by pruning, and thus the 100 pruned scores include some scores from unrelated sequences.

Comparison of normalization methods

The ability of searches using different length normalization methods to identify distantly related sequences was evaluated as described (Pearson, 1995). Briefly, searches are done with a large number of query sequences (110 queries from the Prosite/SwissProt ''challenging'' families or 108 queries from the 54 PIR39b families) using two length normalization methods. For evaluations using the equivalence number (Pearson, 1995), the number of sequences missed with each method is compared, and a method receives a + or − depending on whether more or fewer sequences are missed. The distribution of pluses and minuses is compared to that expected from a binomial distribution or its normal approximation (the signtest). The distribution of pluses and minuses is represented by the z-value the normal approximation in Figure 4 (right panels) and Figures 5 and 6; in these Figures, z-values >2 are statistically significant in a two-tailed test at the $P() < 0.05$ level. Signtest z-values for the PIR queries were divided by $\sqrt{2}$ because two queries were used from each family (Pearson, 1995).

Search performance was also evaluated using the second criterion for ''finding'' a related sequence; the number of related sequences with $E() > 0.02$ plus the number of unrelated sequences

with $E() < 0.02$. When just the number of related sequences with $E() > 0.02$ was used, methods that produced very low $E()$-values for every sequence appeared to perform well. By adding the number of unrelated sequences with $E() < 0.02$ to the number of related sequences with $E() > 0.02$, methods with inaccurate $E()$-values do not have an advantage. Sums of the related sequences with $E() > 0.02$ and unrelated sequences with $E() < 0.02$ were compared between methods and evaluated using the signtest. These results are referred to as the $E() < 0.02$ criterion in Figures 4 to 6.

Acknowledgements

The author thanks Phil Green for very helpful discussions and code to calculate the regress3 regression-scaled scores. This work was supported by a grant from the National Library of Medicine (LM04969) with additional support from the Digital Equipment Corporation.

References

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

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

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

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **19**(suppl.), 2241–2245.

Bairoch, A. & Boechmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **19**(suppl.), 2247–2249.

Barker, W. C., George, D. G. & Hunt, L. T. (1990). Protein sequence database. In *Methods in Enzymology* (Doolittle, R. F., ed.), vol. 183, pp. 31–49, Academic Press, San Diego.

Collins, J. F., Coulson, A. F. W. & Lyall, A. (1988). The significance of protein sequence similarities. *Comput. Appl. Biosci.* **4**, 67–71.

Evans, M., Hastings, N. & Peacock, B. (1993). *Statistical Distributions,* John Wiley and Sons, New York.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitutions matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915–10919.

Henikoff, S. & Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins: Struct. Funct. Genet.* **17**, 49–61.

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

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA,* **90**, 5873–5877.

Ladunga, I., Wiese, B. A. & Smith, R. F. (1996). FASTA-SWAP and FASTA-PAT: pattern database searches using combinations of aligned amino acids, and a novel scoring theory. *J. Mol. Biol.* **259**, 840–854.

Mott, R. (1992). Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59–75.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. In *Methods in Enzymology* (Doolittle, R. F., ed.), vol. 183, pp. 63–98, Academic Press, San Diego.

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

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227–258.

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

Pearson, W. R. & Miller, W. (1992). Dynamic programming algorithms for biological sequence comparison. In *Methods in Enzymology* (Brand, L. & Johnson, M. L., eds), vol. 210, pp. 575–601, Academic Press, San Diego.

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

Waterman, M. S. & Vingron, M. (1994). Rapid and accurate estimates of the statistical significance for sequence database searches. *Proc. Natl Acad. Sci. USA,* **9192**, 4625–4628.

Waterman, M. S., Gordan, L. & Arratia, R. (1987). Phase transitions in sequence matches and nucleic acid structure. *Proc. Natl Acad. Sci. USA,* **84**, 1239–1243.

Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acids sequences and sequence database. *Comput. Chem.* **17**, 149–163.

Zhang, Z., Pearson, W. R. & Miller, W. (1997). Aligning a DNA sequence with a protein sequence. *J. Comput. Biol.* **4**, 339–349.

Edited by F. E. Cohen

(Received 1 May 1997; received in revised form 4 November 1997; accepted 5 November 1997)