Vol. 17 no. 12 2001
Pages 1158–1167

Estimation of P-values for global alignments of protein sequences

Caleb Webber and Geoffrey J. Barton*

EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on March 21, 2001; revised on May 14, 2001; accepted on May 24, 2001

ABSTRACT

Motivation: The global alignment of protein sequence pairs is often used in the classification and analysis of full-length sequences. The calculation of a Z-score for the comparison gives a length and composition corrected measure of the similarity between the sequences. However, the Z-score alone, does not indicate the likely biological significance of the similarity. In this paper, all pairs of domains from 250 sequences belonging to different SCOP folds were aligned and Z-scores calculated. The distribution of Z-scores was fitted with a peak distribution from which the probability of obtaining a given Z-score from the global alignment of two protein sequences of unrelated fold was calculated. A similar analysis was applied to subsequence pairs found by the Smith–Waterman algorithm. These analyses allow the probability that two protein sequences share the same fold to be estimated by global sequence alignment.

Results: The relationship between Z-score and probability varied little over the matrix/gap penalty combinations examined. However, an average shift of $+4.7$ was observed for Z-scores derived from global alignment of locally-aligned subsequences compared to global alignment of the full-length sequences. This shift was shown to be the result of pre-selection by local alignment, rather than any structural similarity in the subsequences. The search ability of both methods was benchmarked against the SCOP superfamily classification and showed that global alignment Z-scores generated from the entire sequence are as effective as SSEARCH at low error rates and more effective at higher error rates. However, global alignment Z-scores generated from the best locally-aligned subsequence were significantly less effective than SSEARCH. The method of estimating statistical significance described here was shown to give similar values to SSEARCH and BLAST, providing confidence in the significance estimation.

Availability: Software to apply the statistics to global alignments is available from http://barton.ebi.ac.uk.

Contact: geoff@ebi.ac.uk

INTRODUCTION

Genome sequencing projects have stimulated a massive growth in the amount of publicly available DNA sequence data, with the total number of bases doubling every ten months (Baker *et al.*, 2000). Given a new sequence, the challenge is to identify the location of coding regions and assign functions to the protein products. While the experimental determination of all gene functions may take decades, identification of similarity to previously well characterized proteins provides a valuable first-step in function assignment. The most reliable techniques for identifying sequence similarity exploit multiple alignment profiles (Gribskov *et al.*, 1987; Barton and Sternberg, 1990) or Hidden Markov Models (Sonnhammer *et al.*, 1997) derived from well-constructed and annotated alignment data collections (Apweiler *et al.*, 2000). Despite this, only 63% of known proteins in SWISS-PROT and TREMBL are matched to the Pfam database by this approach (Bateman *et al.*, 2000), so the alignment of sequence pairs remains important. Pair-wise alignments are also an essential step in the clustering of sequences and for multiple alignment by hierarchical methods (Barton, 1998).

Commonly used alignment algorithms produce a raw score which is a function of the chosen similarity matrix and gap penalty. However, since the raw scores depend upon the length of the original sequences, a high raw score may be due only to aligning long sequences rather than a measure of the quality of the alignment. For local alignments, there are two widely-used methods for evaluating significance. Karlin and Altschul (1990) apply Extreme Value (EV) statistics to describe the probability distribution of scores from locally aligning random sequences without gaps. Given the EV distribution and the lengths of the sequences aligned, a raw score is expressed as the likelihood of achieving such a score, or higher, by aligning random sequences of the same lengths. While EV statistics have only been proven for alignments without gaps, they have also been applied to

*To whom correspondence should be addressed. Present address: School of Life Sciences, University of Dundee, Dow St., Dundee, DD1 5EH.

gapped alignments with some success (Altschul *et al.*, 1997). An alternative approach has been directly to model scores obtained from each database search (Collins *et al.*, 1988; Pearson, 1998). Scores in the search deemed to be from alignments to unrelated sequences are used to derive statistical estimates for the likelihood of obtaining any score. An advantage of this method is that it can correct for compositional bias in the database that might otherwise lead to spurious high significance being assigned to an alignment.

Brenner *et al.* (1998) tested the ability of the statistical methods of Karlin and Altschul (1990); Pearson (1998), and of the raw alignment score and percentage identity to identify homologues in the SCOP database (Murzin *et al.*, 1995) in which the relationships of the proteins are known from their three-dimensional structures and functions. Brenner *et al.* (1998) concluded that on their benchmark, the statistical scoring schemes were able to identify more homologues with fewer errors than either the raw score or percentage identity. They also concluded that the method of Pearson (1998) provides more accurate statistical estimates than that of Karlin and Altschul (1990).

While there has been much research on the statistics of local alignments (Mott, 2000; Altschul *et al.*, 2001), the statistics underlying global alignments are still unknown (Waterman, 1995; Durbin *et al.*, 1998). The significance of a global alignment of two sequences is often given as a Z-score (Dayhoff *et al.*, 1978; Barton and Sternberg, 1987a). To obtain a Z-score, the score for aligning the two sequences is found. Each sequence is then shuffled to generate new sequences which are then aligned and scored. This process of shuffling and aligning is repeated a predetermined number of times, often 100. The Z-score can then be calculated from the equation

$$\frac{s-\mu}{\sigma}, \tag{1}$$

where s is the score found from aligning the original sequences, and μ and σ are the mean and standard deviation of the raw score distribution of shuffled and aligned sequences. Shuffling the original sequences guarantees that the length and the composition is maintained in the shuffled sequences. Randomizations require additional alignments to be calculated so there is a cost in computation time approximately equal to number of randomizations performed.

The Z-score gives a measure of significance against a background of randomly generated sequences with the same composition and length as the original sequences, but, like other statistical estimates, says nothing about the biological significance of the alignment.

In this paper, the distribution of Z-scores observed for the global alignment of protein sequence pairs known to be of unrelated three-dimensional structure is used to estimate the significance of pair-wise Z-scores.

METHODS

Dataset

The SCOP database (Murzin *et al.*, 1995) provides a detailed and comprehensive hierarchical description of the structural and evolutionary relationships between all proteins whose three-dimensional structure is known. The lowest level of the SCOP hierarchy is the ‘family’ where proteins are grouped together that share clear sequence, structural, and functional similarity. Above this is the ‘superfamily’, formed from families whose structural and functional features suggest a probable evolutionary relationship. Superfamilies are then classified into ‘folds’ if they share the same major secondary structures in the same arrangement and with the same topological connections.

The dataset of structurally-unrelated protein sequences applied in this study was derived from the PDB40D-B dataset, version 1.37 (Brenner *et al.*, 1998). Brenner *et al.* assembled PDB40D-B, a dataset of 1434 domain sequences, by taking Protein Data Bank (Bernstein *et al.*, 1977) domain sequences via the SCOP database (Murzin *et al.*, 1995) and then removing the lowest quality sequence from any pair of sequences showing greater than 40% sequence identity to any other.

For this work, the PDB40D-B dataset was further reduced. Domains were only included whose structure had been experimentally determined by X-ray crystallography, and single segment sequences (i.e. continuous domains) were selected. For each remaining SCOP fold, one sequence was chosen at random, leaving 303 sequences. In order to allow the dataset to be used directly in other studies, low resolution structures (>2.5 Å) were also removed, leaving 250 protein domain sequences. This dataset is available from the authors.

Figure 1 shows a histogram of the distribution of chain lengths in the dataset. The chain lengths range from 26 to 613 residues, with an average of 184 residues and standard deviation of 115 residues.

Global alignments

All global alignments were carried out by the Multalign program from the AMPS package (Barton and Sternberg, 1987b; Barton, 1990). This implements the Needleman–Wunsch (Needleman and Wunsch, 1970) dynamic programming algorithm with length-independent gap penalties that do not penalize overhanging sequence at the ends of the alignment.

The number of randomizations needed to safely calculate a Z-score was investigated and it was determined that performing as few as 75 randomizations would be adequate (data not shown). This broadly confirms the work of

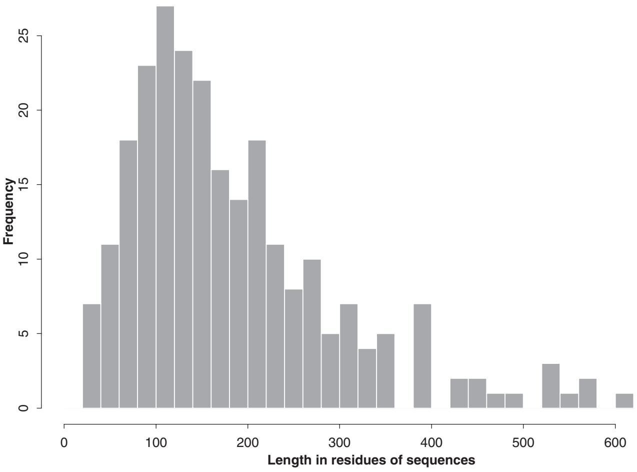

Fig. 1. Histogram showing the distribution of lengths of sequences within the test set of structurally-unrelated protein domain sequences.

Feng *et al.* (1985). In this study, 100 randomizations were performed throughout, unless stated otherwise.

Generation of Z-scores from local alignments

Z-scores were obtained from local alignments by first finding the highest scoring local alignment between two sequences with SCANPS (Barton, 1993), a program which implements the Smith–Waterman algorithm (Smith and Waterman, 1981). The local alignment was then excised, the two subsequences were aligned globally, and a Z-score calculated by AMPS (Barton and Sternberg, 1987b; Barton, 1990) with a BLOSUM62 matrix and length-independent gap penalty of 10.

Matrix/gap-penalty combinations

Eight matrices were investigated to generate global alignments. The PAM30, PAM120, PAM180, and PAM250 matrices (Dayhoff *et al.*, 1978) were chosen to represent a broad range of evolutionary distances as recommended by Altschul (1991). The BLOSUM50, BLOSUM62, and BLOSUM75 matrices (Henikoff and Henikoff, 1992) were also chosen since these were the matrices that gave the highest quality alignments in the alignment benchmarking study by Raghava *et al.* (2000a). In addition, the Gonnet matrix (Gonnet *et al.*, 1992) was chosen as the equivalent of a modern PAM matrix. From the rankings found by Raghava *et al.*, the gap penalty that gave the best quality alignments with each matrix was selected. Additionally, two further gap penalties were chosen either side of the highest ranking penalty.

The matrix/gap penalty combinations to find local alignments were those tested previously for their effectiveness in local similarity searches by Pearson (1998) and Brenner *et al.* (1998).

Benchmarking

The effectiveness of global alignments for sequence similarity searching was tested by a benchmark developed by Raghava *et al.* (2000b). This benchmark makes use of the relationships defined between protein domains within the SCOP database. The benchmark dataset comprises 1091 protein domain sequences, taken from the PDB40D-B dataset, version 1.37 (Brenner *et al.*, 1998), representing 469 SCOP superfamilies across 338 SCOP folds. This dataset is available from the authors. Within this benchmark there are 2528 true positives, defined as a pair of protein domain sequences belonging to the same SCOP superfamily, and 589 172 true negatives, defined as a pair of sequences belonging to different SCOP folds. To reduce the number of comparisons, and thus the computational overhead, the set of true negatives was reduced to 6945 by only accepting unrelated pairs of sequences with a BLAST (Altschul *et al.*, 1997) e-value less than 30. Accordingly, the true negative pairs selected are those with the greatest sequence similarity and so presents a tougher than normal test for sequence similarity search methods.

The benchmark is performed by searching the benchmark dataset with each of the protein domain sequences in turn. The score for each of the 9473 benchmark pairs is collected and ranked from best to worst. This ordered list is then parsed with each pair scored as either a true positive or true negative (i.e. false positive).

Statistical analysis

All data were processed and analyzed in the S-plus package from Mathsoft (StatSci, 1988). Curve-fitting was carried out by the Levenberg–Marquardt method implemented within the TableCurve-2D software from SPSS (SPSS Inc., 1989).

RESULTS AND DISCUSSION

Derivation of probability estimates from global alignment Z-scores

In order to derive empirically the probability of obtaining a given Z-score from a global alignment of two structurally-unrelated protein sequences it is necessary to generate, and statistically model the distributions of such scores. The probabilities of obtaining a Z-score under the conditions of the model can then be found by integration.

Z-scores were calculated for each of the 31 125 pairs of structurally-unrelated protein domain sequence, with 24 different matrix/gap-penalty combinations (see Section **Methods**). Figure 2 illustrates a Z-score distribution calculated with standard parameters and is shown fitted with a Gaussian curve. The mean Z-score of the distribution shown in Figure 2 is 0.184, the highest and lowest Z-scores are 7.09 and -3.23 respectively, and the distribu-

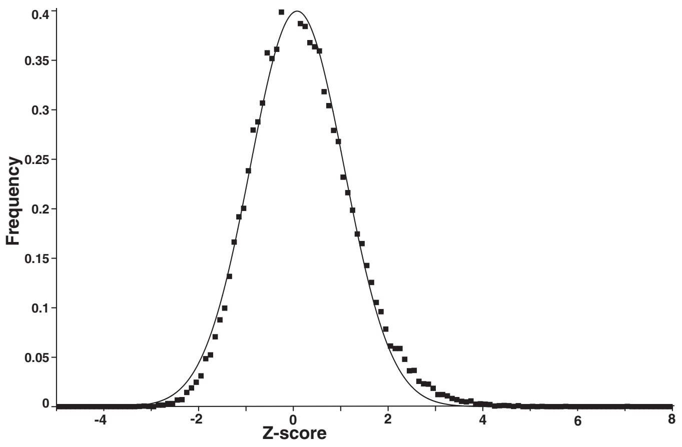

Fig. 2. Distribution of Z-scores calculated from global pairwise alignments of a set of structurally-unrelated protein domain sequences with a BLOSUM62 matrix and a length-independent gap penalty of 10. The distribution has been fitted with a Gaussian curve, indicated by a solid line.

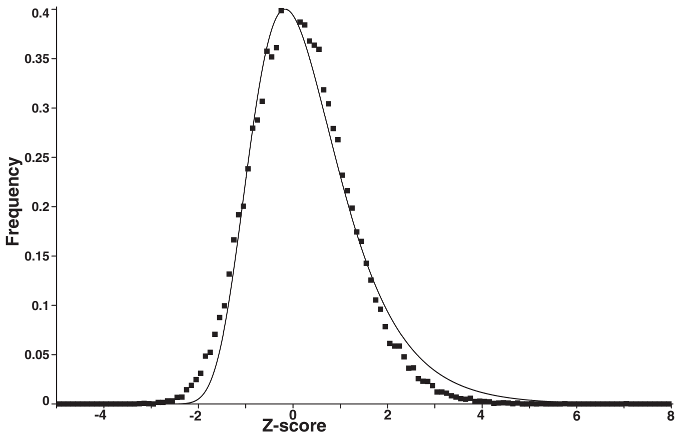

Fig. 3. Distribution of Z-scores calculated from global alignments found with a BLOSUM62 matrix and a length-independent gap penalty of 10 between a set of structurally-unrelated protein domain sequences. The distribution has been fitted with an extreme value curve.

tion has a positive skew of 0.22. The maxima of all 24 Z-score distributions calculated with different matrix/gap-penalty combinations occurred close to zero at an average Z-score of 0.019. The mean global Z-score over all calculated distributions was 0.13 with a standard deviation of 1.1. The close proximity to zero of the mean Z-scores of all the global methods tested here means that the shuffled alignments score as highly as the original structurally-unrelated global alignments.

It can be seen from Figure 2 that the Gaussian curve fitted to the Z-score distribution describes the data well over the majority of the Z-score frequencies. However, at the tail of the distribution (Z-score > 2) Figure 2 illustrates that the Gaussian curve deviates from the experimental data. Differentiating between Z-scores likely to be from related or unrelated sequences is most difficult towards the tail of the distribution and so it is in this region that the fit of the curve should be most accurate. Accordingly, the Gaussian curve shown fitted in Figure 2 is unsuitable for derivation of probabilistic estimates and other functions must be considered.

The Extreme Value Distribution (EVD) is popular in applications to the statistics of length-corrected local alignment raw scores (Karlin and Altschul, 1990; Pearson, 1998). Figure 3 shows an EVD, given as

$$\int(x) = 1/\beta \mathrm{e}^{-\frac{x-\alpha}{\beta}} \exp\left[-\exp\left(-\frac{x-\alpha}{\beta}\right)\right], \qquad (2)$$

where $\alpha(-\inf < \alpha < +\inf)$ is the location parameter and $\beta(\beta > 0)$ is the scale parameter, fitted to a Z-score distribution calculated with standard parameters. Figure 3 demonstrates that the EVD is skewed relative to the observed Z-score distribution.

A range of alternative distributions were tried against the data. The best-fitting distributions, as ranked by least-squares, over all 24 matrix/gap-penalty combinations were the log-Normal, Chi-squared, and gamma distributions. The gamma distribution, given as

$$\int(x) = \frac{1}{\Gamma(\alpha)\lambda^{\alpha}}(x+\beta)^{\alpha-1}\mathrm{e}^{-(x+\beta)/\lambda} \qquad (3)$$

where $0 \leqslant (x+\beta) < \inf, \alpha > 0$ is the shape parameter, and $\lambda > 0$ is the scale parameter, was chosen as the model distribution since it fitted with a significantly higher F-statistic (33 991.788) than any other distribution. Additionally, it is straightforward to integrate numerically. However, choice of the gamma distribution is purely based on the quality of fit. There is no proof that this distribution represents the true nature of the statistics underlying the distribution of global alignment Z-scores.

Figure 4 shows the gamma distribution fitted to a Z-score distribution calculated with standard parameters. The probability values (P-values) derived from this fit have been added to the x-axis of Figure 4. From Figure 4, the area lying to the right of a Z-score of 0 is 0.545, which is the probability of obtaining a Z-score of at least 0 from a global alignment between two structurally-unrelated proteins using a BLOSUM62 matrix and a length-independent gap penalty of 10. From Figure 4 it can be seen that the probabilities derived fall rapidly, with the probability of obtaining a Z-score at least as high as 5 calculated to be 5.1×10^{-5}. This supports

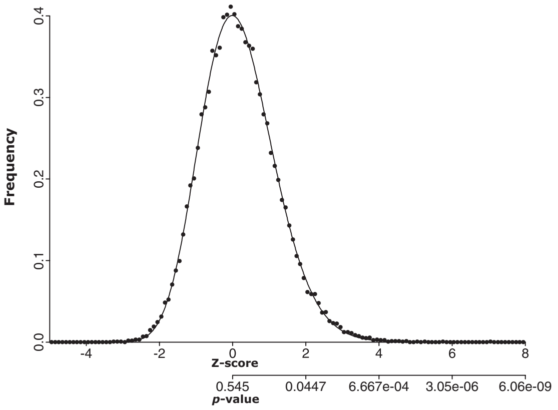

Fig. 4. Distribution of Z-scores calculated from global alignments found with a BLOSUM62 matrix and a length-independent gap penalty of 10 between a set of structurally-unrelated protein domain sequences. The distribution has been fitted with a gamma curve from which probabilities have been derived and added to the x-axis.

early empirical observations that a Z-score greater than 5 indicates strong structural similarity between the globally-aligned sequences (Barton and Sternberg, 1987a).

Table 1 summarizes the probabilities derived for 10 of the most commonly used matrix/gap-penalty combinations examined for global Z-scores 0–11, inclusive. The remaining 14 matrix/gap-penalty combinations examined can be found in Table A1 of the Appendix. Table 1 shows that for the 10 Z-score distributions, at any given Z-score all the probabilities are very similar, with a small standard deviation. This similarity and the observation that all the global Z-score distributions are centred close to 0 may be due to there being little common information to align in these sequences or that the significance of any information is lost when aligning the entire sequence.

Deriving probabilistic estimates from Z-scores calculated from local alignment subsequences

A database search with a local alignment method, such as Smith–Waterman (Smith and Waterman, 1981), will produce alignments between subsequences. To see what Z-score would be expected for the sequence fragments found in these subsequence alignments when aligned by a global algorithm, Z-scores were calculated for the highest-scoring local alignments found between each of the 31 125 structurally-unrelated pairs of sequences (see Section **Methods**). The average length of the highest scoring local alignments found varied considerably with the matrix and gap penalty. For example, changing the extension penalty with a BLOSUM50 matrix and a gap opening penalty of 12, from 2 to 1 changed the mean alignment length of 31.38 residues with an SD of 23.95 to a mean length of 43.39 with an SD of 36.18 respectively.

The mean Z-score of all the matrix/gap-penalty combinations was 4.70 with a standard deviation of 0.20, showing a positive shift in Z-score for these distributions compared to those obtained from global alignment of the full-length sequences.

As with the method used for global alignment of the full sequences, probabilities of obtaining a given Z-score between two structurally-unrelated protein sequences were derived by fitting the distributions of Z-scores calculated from local alignments with a well-fitting distribution.

Figure 5 illustrates the distribution for a BLOSUM50 matrix and an affine gap penalty of 12 and 2 fitted with a gamma distribution from which probabilities have been derived and added to the x-axis. Table 1 shows the probabilities derived from all 4 matrix/gap penalty combinations for Z-scores of 0–11. The probabilities derived from the Z-score distributions calculated from local alignments for three of the matrix/gap penalty combinations (BLOSUM45 matrix with an affine gap penalty of 12 and 1, and the BLOSUM50 matrix gap penalties of 12 and 2, and 12 and 1) are similar. As the Z-score increases, differences begin to appear but the probabilities of these matrix/gap-penalty combinations stay within the same order of magnitude as each other. From Table 1 it can be seen that the difference is greater between the Z-score distribution calculated with a BLOSUM62 matrix and an affine gap penalty of 11 and 1, and those calculated with the other three matrix/gap penalty combinations tested. However, it is not clear why this is observed. Figures 5 and 6 illustrate the most striking difference between global and local Z-scores. For example, the probability of obtaining a Z-score of 6 or greater calculated from a global alignment of best locally-aligned subsequences between two structurally-unrelated protein domain sequences is 0.139 compared to a probability 3.05×10^{-6} for a global alignment of the full-length sequences.

The significantly positive Z-scores for local alignments show that the local alignment found scores higher than the alignments calculated from shuffling the original subsequences. To investigate whether these alignments indicate structural or biological significance, each sequence in the set of 250 protein sequences was shuffled. This set of shuffled sequences was then used to generate a distribution of local alignment Z-scores, as above. The distribution obtained from the shuffled sequences had a mean Z-score of 4.81 with an SD of 1.06, which was very similar to the mean of 4.79 and SD of 1.12 found for the distribution obtained from the biological sequences. A Welch-modified two-sample t-test (Hogg and Craig, 1970) between the shuffled and non-shuffled distributions gave a P-value of 0.0385 strongly supporting the hypothesis that there is no difference between these two distributions. This suggests that the shift in Z-scores seen between the full-length sequences and locally-aligned subsequences is simply a

Table 1. Probability of a structurally-unrelated protein sequence match for a given Z-score, matrix, and gap penalty. Global alignment method indicates global alignment of entire sequences, while local alignment method indicates global alignment (BLOSUM62 matrix and a gap penalty of 10) of best locally-aligned subsequences (parameters given in table)

Alignment		Gap	Gamma PDF fit parameters			Z-score											
method	Matrix	penalty	α	β	λ	0	1	2	3	4	5	6	7	8	9	10	11
Global	BLO50	12	27.00	5.07	0.194	0.547	0.201	0.0454	0.00674	0.000712	5.66e − 05	3.57e − 06	1.84e − 07	8.05e − 09	3.05e − 10	1.02e − 11	2.98e − 13
Global	BLO62	10	29.43	5.31	0.186	0.545	0.2	0.0447	0.00652	0.000668	5.1e − 05	3.05e − 06	1.49e − 07	6.06e − 09	2.12e − 10	6.51e − 12	1.73e − 13
Global	BLO62	12	25.54	4.96	0.200	0.535	0.195	0.0441	0.00664	0.000719	5.93e − 05	3.9e − 06	2.13e − 07	9.87e − 09	3.99e − 10	1.43e − 11	4.62e − 13
Global	BLO75	10	31.05	5.43	0.181	0.549	0.202	0.0449	0.00645	0.000643	4.74e − 05	2.72e − 06	1.25e − 07	4.83e − 09	1.58e − 10	4.52e − 12	1.14e − 13
Global	GONNET	100	29.44	5.20	0.184	0.562	0.208	0.0462	0.0066	0.000656	4.84e − 05	2.77e − 06	1.29e − 07	5.01e − 09	1.67e − 10	4.84e − 12	1.18e − 13
Global	PAM30	15	30.60	5.43	0.183	0.548	0.203	0.0457	0.0067	0.000687	5.24e − 05	3.11e − 06	1.5e − 07	6.02e − 09	2.07e − 10	6.22e − 12	1.72e − 13
Global	PAM120	10	29.48	5.33	0.187	0.551	0.206	0.047	0.00702	0.000737	5.79e − 05	3.57e − 06	1.79e − 07	7.51e − 09	2.71e − 10	8.58e − 12	2.48e − 13
Global	PAM180	10	52.56	7.03	0.138	0.567	0.213	0.0468	0.00641	0.000587	3.82e − 05	1.86e − 06	7.09e − 08	2.18e − 09	5.56e − 11	1.2e − 12	1.43e − 14
Global	PAM250	8	45.51	6.55	0.149	0.568	0.212	0.0454	0.00591	0.0005	2.93e − 05	1.25e − 06	4.06e − 08	1.04e − 09	2.18e − 11	3.85e − 13	8.77e − 15
Global	PAM250	10	35.73	5.84	0.169	0.555	0.206	0.045	0.0062	0.000587	3.97e − 05	2.04e − 06	8.28e − 08	2.74e − 09	7.62e − 11	1.81e − 12	3.84e − 14
Mean						0.5527	0.2046	0.04552	0.0065	0.00065	4.8e − 05	2.8e − 06	1.2e − 07	5.3e − 09	1.9e − 10	5.9e − 12	3.0e − 13
SD						0.01	0.005	0.0009	0.0003	7.0e − 05	9.7e − 06	8.4e − 07	6.5e − 08	2.8e − 09	1.2e − 10	4.3e − 12	2.3e − 13
Local	BLO45	12–1	59.50	3.56	0.137	1	1	0.996	0.937	0.694	0.327	0.0894	0.0142	0.00138	8.56e − 05	3.57e − 06	1.04e − 07
Local	BLO50	12–1	91.88	5.70	0.112	1	1	0.996	0.938	0.705	0.345	0.101	0.0174	0.00187	0.00013	6.1e − 06	2.04e − 07
Local	BLO50	12–2	95.87	5.55	0.108	1	1	0.998	0.965	0.783	0.428	0.139	0.0261	0.00297	0.000213	1.01e − 05	3.34e − 07
Local	BLO62	11–1	129.21	7.49	0.0967	1	0.998	0.971	0.818	0.489	0.18	0.0393	0.00516	0.000422	2.24e − 05	8.04e − 07	2.02e − 08
Mean						1	1	0.99	0.91	0.67	0.32	0.09	0.016	0.0017	0.00011	5.1e − 06	1.7e − 07
SD						0	0.001	0.013	0.066	0.13	0.10	0.041	0.0086	0.0011	8.0e − 05	3.9e − 06	1.4e − 07

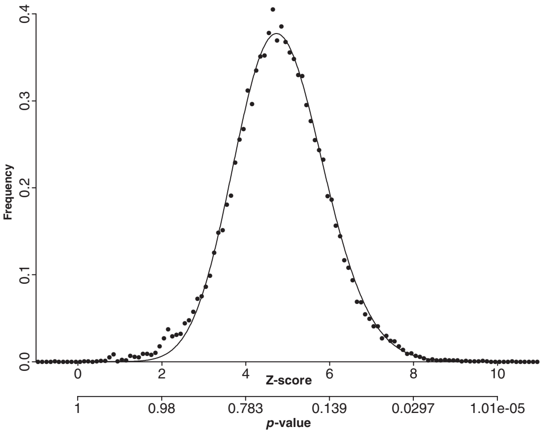

Fig. 5. Distribution of Z-scores calculated from local alignments found with a BLOSUM50 matrix with an affine gap penalty of 12 and 2 between a set of structurally-unrelated protein domain sequences. The distribution has been fitted with a gamma curve from which probabilities have been derived and added to the *x*-axis.

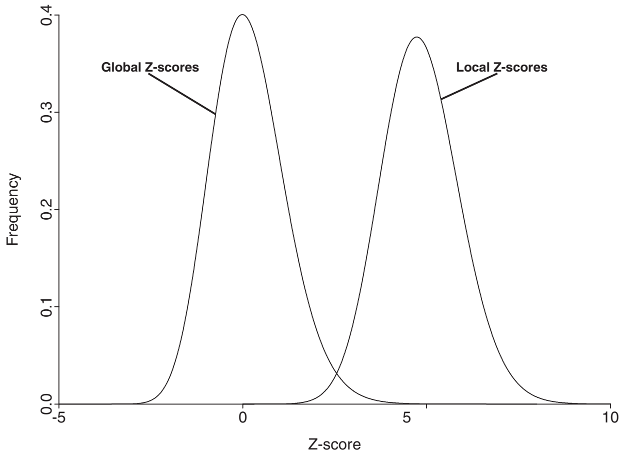

Fig. 6. Plot showing the gamma distributions fitted in Figure 4 and Figure 5 overlaid on the same axis.

result of the local alignment method and that local alignments found between structurally-unrelated proteins at the fold level do not imply that these regions share any local structural similarity. This finding supports studies carried out by Sternberg and Islam (1990) and Cuff and Barton (unpublished data) who found no obvious correlation between regions of sequences aligning well locally and those regions coding for residues involved in α-helix or β-strand secondary structures.

Benchmarking

Figure 7 compares the performance of SSEARCH (default parameters) (Pearson, 1995) with the global alignment Z-scores, and Z-scores calculated from the best locally-aligned subsequence, by plotting the number of false positives found (x-axis) against the number of true positives (y-axis) for each method at a given threshold. To aid the empirical statistical scoring employed by SSEARCH, the benchmark dataset was embedded within the NRDB90 database (Park *et al.*, 2000). SSEARCH has been shown by Brenner *et al.* to be an effective sequence similarity detection method (Brenner *et al.*, 1998). From Figure 7 it can be seen that there is little difference between the discriminatory abilities of global alignment Z-scores and SSEARCH on the benchmark set of single domain protein sequences at less than 80 false positives. For example, from Figure 7 at a threshold of 20 false positives, SSEARCH finds 443 true positives while global alignment Z-scores find 445 true positives. The difference between these methods at this threshold gives a χ^2-test statistic of 0.04 (Bland, 1987), which is not statistically significant at one degree of freedom. Figure 7 shows that above 80 false positives the number of true positives found by global alignment Z-scores increases at a higher rate compared to SSEARCH. At a threshold of 300 false positives, SSEARCH finds 549 true positives while global alignment Z-scores find 597 true positives. At this threshold a χ^2-test statistic of 14.40 shows the difference is statistically significant. However, it is unlikely that this level of error is acceptable within most search applications. This analysis shows that Z-scores calculated by global alignment are as good at detecting sequence homology as SSEARCH at low error rates and better at higher error rates, within this benchmark. However, Figure 7 shows that Z-scores calculated by global alignment of the best locally-aligned subsequence offer poor discriminatory performance compared to the other methods, finding only 301 true positives at a cut-off of 20 false positives. This is supported by a χ^2-test statistic of 134.65 between this method and SSEARCH at this threshold, which is significant at the 0.1% level for one degree of freedom. It should be noted that the matrix/gap penalties chosen for the global alignment methods were selected on the basis of their performance in an alignment benchmarking study (Raghava *et al.*, 2000b), and have not been optimized for sequence similarity searches.

It is not possible to assess the accuracy of the statistical estimates for the significance of global alignment Z-scores and Z-scores calculated by global alignment of the best locally-aligned subsequence within this benchmark. This is because the reduced number of true negatives (6945 pairs non-randomly selected from a possible 589 172 pairs) would skew the results. However, because of the potential inaccuracy due to extrapolation from the observed Z-score distribution on which the statistical model was developed, it is important to verify that the model is able to provide reasonable probability

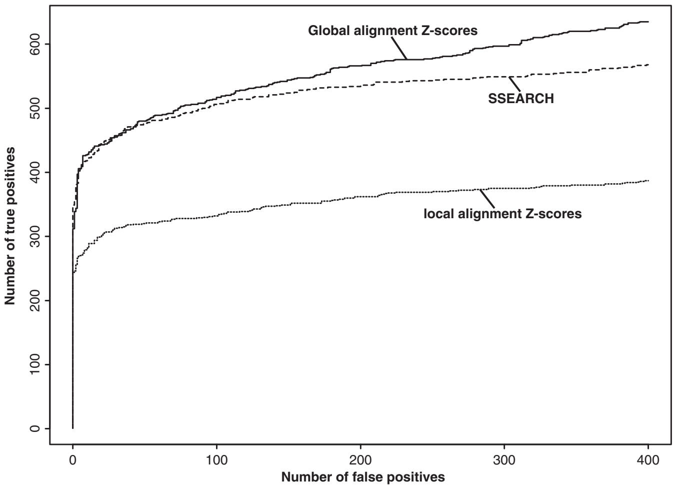

Fig. 7. Plot showing number of true positives found against the number of false positives at a given threshold for global alignment Z-scores, global alignment Z-scores for the best locally-aligned subsequence, and SSEARCH (Pearson, 1995), a Smith–Waterman implementation.

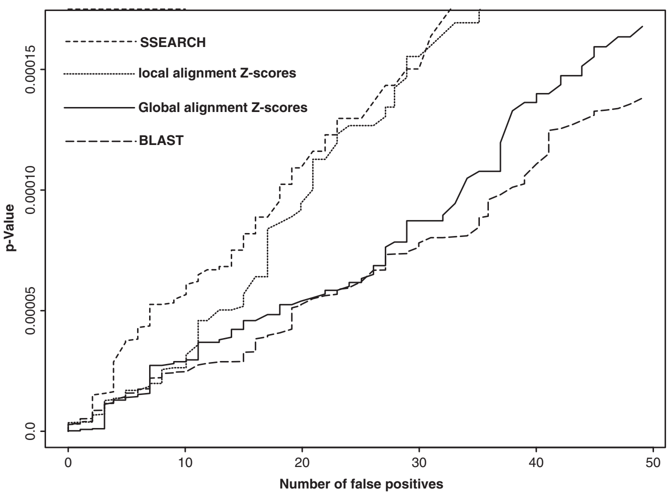

Fig. 8. Plot showing the P-value reported against the number of false positives found, determined from the benchmark, for global alignment Z-scores, global alignment Z-scores for the best locally-aligned subsequence, and SSEARCH (Pearson, 1995), and BLAST (Altschul *et al.*, 1997).

estimates at low error rates. The statistical estimates of SSEARCH and BLAST (Altschul *et al.*, 1997) have been benchmarked previously by Brenner *et al.* (1998). Brenner *et al.* showed that SSEARCH provides a good approximation, although slight underestimation, of the ideal statistical score, while BLAST is more inaccurate and overestimates the statistical significance. By using the statistical predictions of these methods as upper and lower bounds in a plot of P-value against the number of false positives, as shown in Figure 8, it is possible to estimate the accuracy of the statistical significance predicted within this paper at low error rates. From Figure 8 it can be seen that the plots for both the global alignment P-values and P-values calculated for global alignment of the best locally-aligned subsequence lie between those of SSEARCH and BLAST, in the area shown by the work of Brenner *et al.* that the ideal statistical score should be found. This lends confidence to the statistical estimates described here.

DISCUSSION

In this paper we have developed a method for estimating the statistical significance of global alignment Z-scores based on the observed distribution of scores between structurally-unrelated protein sequences. While the meaning of a Z-score is not yet fully understood, the conversion to a P-value derived from known structural relationships provides a readily understood figure—the probability of obtaining a given Z-score or higher by chance alignment of two structurally-unrelated sequences. This allows a more rational decision to be made in any method that employs Z-scores to provide a cut-off.

Benchmarking has shown global alignment Z-scores to be as effective in the detection of structural similarity as other popular local alignment statistics. This may in part be due to the length and composition correction implicit in the calculation. These probabilities could be used to provide a stand-alone method for updating the clustering of large protein sequence databases where addition of new sequences would require a complete recalculation of the statistical significance of all scores if on-the-fly statistics, such as in SSEARCH, were employed. Additionally, global alignment with P-values could be used in partnership with other methods of estimating significance in a confirmatory role.

In order to facilitate the application of this work in other research and methods, a C program has been written which takes a Z-score, the matrix, and the gap penalties as arguments, and returns a probability. This program is available from http://barton.ebi.ac.uk.

ACKNOWLEDGEMENTS

We would like to thank Steven Searle, Uwe Dengler, and James Cuff for their discussions on this work, in particular we thank James Cuff for his help with preparation of the dataset. We would also like to thank Lorenz Wernisch for his statistical advice. This work was supported by a European Molecular Biological Laboratory studentship to C.W.

REFERENCES

Altschul,S. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **1991**, 555–565.

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

Altschul,S., Bundschuh,R., Olsen,R. and Hwa,T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.

Apweiler,R., Attwood,T., Bairoch,A., Bateman,A., Birney,E., Bucher,P., Codani,J.-J., Corpet,F., Croning,M., Durbin,R., Etzold,T., Fleischmann,W., Gouzy,J., Hermjakob,H., Jonassen,I., Kahn,D., Kanapin,A., Schneider,R., Servant,F. and Zdobnov,E. (2000) InterPro—an integrated documentation resource for protein families, domains and functional sites. *CCP11 Newsletter*, **10**, 1–1.

Baker,W., van den Broek,A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G. and Tuli,M. (2000) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **28**, 19–23.

Barton,G.J. (1990) Protein multiple sequence alignment and flexible pattern matching. *Meth. Enzymol.*, **183**, 403–428.

Barton,G.J. (1993) An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Comput. Appl. Biosci.*, **9**, 729–734.

Barton,G.J. (1998) Protein sequence alignment techniques. *Acta Crystallogr.*, **54**, 1139–1146.

Barton,G.J. and Sternberg,M.J.E. (1987a) Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng.*, **1**, 89–94.

Barton,G.J. and Sternberg,M.J.E. (1987b) A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, **198**, 327–337.

Barton,G.J. and Sternberg,M.J.E. (1990) Flexible protein sequence patterns—a sensitive method to detect weak structural similarities. *J. Mol. Biol.*, **212**, 389–402.

Bateman,A., Birney,E., Durbin,R., Eddy,S., Howe,K. and Sonnhammer,E. (2000) The Pfam protein families database. *NAR*, **28**, 263–266.

Bernstein,F., Koetzle,T., Williams,G., Meyer Jr.,E., Brice,M., Rodgers,J., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Bland,M. (1987) *An Introduction to Medical Statistics*. Oxford University Press, Oxford.

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

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

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A Model of evolutionary change in proteins. Matrices for detecting distant relationships. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Washington, DC, pp. 345–358.

Durbin,R., Eddy,S., Krogh,A. and Mitchinson,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Feng,D.F., Johnson,M.S. and Doolittle,R.F. (1985) Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.*, **21**, 112–125.

Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1444.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10 915–10 919.

Hogg,R.V. and Craig,A.T. (1970) *Introduction to Mathematical Statistics*, 3rd edn, Macmillan, New York.

SPSS Inc (1989) TableCurve2D. SPSS Inc, Chicago, http://www.spss.com.

Karlin,S. and 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.

Mott,R. (2000) Accurate formula for P-values of gapped local sequence and profile alignments. *J. Mol. Biol.*, **300**, 649–659.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

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

Park,J., Holm,L., Heger,A. and Chothia,C. (2000) RSDB: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.

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

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

Raghava,G., Clamp,M. and Barton,G. (2000a) Measurement and significance of protein structural similarity, in preparation.

Raghava,G., Searle,S. and Barton,G. (2000b) A strategy for the evaluation of protein sequence alignments: datasets and quality measures, submitted.

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

Sonnhammer,E., Eddy,S. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

StatSci (1988) Splus v3.4. Stat Sci is a division of MathSoft, Inc., Seattle.

Sternberg,J. and Islam,S. (1990) Local protein similarity does not imply a structural relationship. *Protein Eng.*, **4**, 125–131.

Waterman,M. (1995) *Introduction to Computational Biology*. Chapman and Hall, London.

APPENDIX

Table A.1. Probability of a structurally-unrelated protein sequence match for a given Z-score, matrix, and gap penalty. Global alignment method indicates global alignment of entire sequences, while local alignment method indicates global alignment of best locally-aligned subsequences

Alignment		Gap	Gamma PDF fit parameters			Z-score											
method	Matrix	penalty	α	β	λ	0	1	2	3	4	5	6	7	8	9	10	11
Global	BLO50	10	33.10	5.59	0.175	0.555	0.206	0.0456	0.006 5	0.000 64	4.63e − 05	2.59e − 06	1.16e − 07	4.33e − 09	1.37e − 10	3.76e − 12	8.77e − 14
Global	BLO50	14	25.78	4.96	0.199	0.538	0.196	0.0441	0.006 59	0.000 704	5.72e − 05	3.71e − 06	1.98e − 07	9.02e − 09	3.57e − 10	1.25e − 11	3.98e − 13
Global	BLO75	8	33.59	5.62	0.173	0.559	0.208	0.0462	0.006 58	0.000 645	4.63e − 05	2.56e − 06	1.14e − 07	4.18e − 09	1.31e − 10	3.53e − 12	9.26e − 14
Global	BLO75	12	25.74	4.98	0.120	0.541	0.2	0.0455	0.006 93	0.000 756	6.29e − 05	4.17e − 06	2.29e − 07	1.07e − 08	4.33e − 10	1.56e − 11	5.01e − 13
Global	GONNET	80	38.16	5.891	0.161	0.575	0.214	0.0465	0.006 28	0.000 569	3.69e − 05	1.8e − 06	6.92e − 08	2.16e − 09	5.65e − 11	1.27e − 12	2.25e − 14
Global	GONNET	120	29.12	5.20	0.185	0.552	0.202	0.0445	0.006 34	0.000 633	4.7e − 05	2.73e − 06	1.29e − 07	5.08e − 09	1.72e − 10	5.11e − 12	1.4e − 13
Global	PAM30	12	32.72	0.177	5.58	0.556	0.208	0.0466	0.006 75	0.000 678	5.02e − 05	2.88e − 06	1.33e − 07	5.1e − 09	1.67e − 10	4.73e − 12	1.03e − 13
Global	PAM30	18	29.00	0.189	5.31	0.54	0.199	0.0447	0.006 62	0.000 692	5.43e − 05	3.35e − 06	1.69e − 07	7.14e − 09	2.6e − 10	8.34e − 12	2.4e − 13
Global	PAM120	8	33.17	5.63	0.176	0.562	0.212	0.0479	0.006 95	0.000 692	5.04e − 05	2.82e − 06	1.26e − 07	4.65e − 09	1.45e − 10	3.9e − 12	7.16e − 14
Global	PAM120	12	25.86	4.98	0.199	0.542	0.2	0.0454	0.006 88	0.000 747	6.16e − 05	4.05e − 06	2.2e − 07	1.02e − 08	4.09e − 10	1.46e − 11	4.65e − 13
Global	PAM180	8	52.54	7.03	0.139	0.579	0.219	0.0474	0.006 13	0.000 509	2.88e − 05	1.17e − 06	3.59e − 08	8.56e − 10	1.65e − 11	2.58e − 13	5.33e − 15
Global	PAM180	12	35.06	5.79	0.171	0.557	0.208	0.0464	0.006 64	0.000 653	4.7e − 05	2.59e − 06	1.15e − 07	4.17e − 09	1.29e − 10	3.43e − 12	6.87e − 14
Global	PAM250	12	30.61	5.41	0.182	0.544	0.200	0.044	0.006 4	0.000 643	4.81e − 05	2.80e − 06	1.32e − 07	5.22e − 09	1.76e − 10	5.17e − 12	1.24e − 13
Mean						0.554	0.206	0.0458	0.006 58	0.000 659	4.9e − 05	2.86e − 06	1.37e − 07	5.60e − 09	1.92e − 10	6.32e − 12	1.78e − 13
SD						0.01	0.007	0.001	0.000 3	7e − 05	9.3e − 06	8.3e − 07	5.5e − 08	2.9e − 09	1.3e − 10	4.9e − 12	1.69e − 13