Pairwise Statistical Significance Versus Database Statistical Significance for Local Alignment of Protein Sequences

Ankit Agrawal[1], Volker Brendel[2], and Xiaoqiu Huang[1]

[1] Department of Computer Science, Iowa State University,
226 Atanasoff Hall, Ames, IA 50011-1041, USA
{ankitag,xqhuang}@iastate.edu
[2] Department of Genetics, Development, and Cell Biology and Department of Statistics, Iowa State University,
2112 Molecular Biology Building, Ames, IA, 50011-3260, USA
vbrendel@iastate.edu

Abstract. An important aspect of pairwise sequence comparison is assessing the statistical significance of the alignment. Most of the currently popular alignment programs report the statistical significance of an alignment in context of a database search. This database statistical significance is dependent on the database, and hence, the same alignment of a pair of sequences may be assessed different statistical significance values in different databases. In this paper, we explore the use of pairwise statistical significance, which is independent of any database, and can be useful in cases where we only have a pair of sequences and we want to comment on the relatedness of the sequences, independent of any database. We compared different methods and determined that censored maximum likelihood fitting the score distribution right of the peak is the most accurate method for estimating pairwise statistical significance. We evaluated this method in an experiment with a subset of CATH2.3, which had been previoulsy used by other authors as a benchmark data set for protein comparison. Comparison of results with database statistical significance reported by popular programs like SSEARCH and PSI-BLAST indicate that the results of pairwise statistical significance are comparable, indeed sometimes significantly better than those of database statistical significance (with SSEARCH). However, PSI-BLAST performs best, presumably due to its use of query-specific substitution matrices.

Keywords: Database statistical significance, Homologs, Pairwise local alignment, Pairwise statistical significance.

1 Introduction

Sequence alignment is extremely useful in the analysis of DNA and protein sequences [1,2,3]. Sequence alignment forms the basic step of making various high level inferences about the DNA and protein sequences - like homology, finding protein function, protein structure, deciphering evolutionary relationships, etc.

I. Măndoiu, R. Sunderraman, and A. Zelikovsky (Eds.): ISBRA 2008, LNBI 4983, pp. 50–61, 2008.
© Springer-Verlag Berlin Heidelberg 2008

There are many programs that use some well known algorithms [4,5] or their heuristic version [3,6,7]. Recently, some enhancements in alignment program features have also become available [8,9] using difference blocks and multiple scoring matrices. Quality of a pairwise sequence alignment is gauged by the statistical significance rather than the alignment score alone, i.e., if an alignment score has a low probability of occurring by chance, the alignment is considered statistically significant.

For ungapped alignments, rigorous statistical theory for the alignment score distribution is available [10], and it was shown that the statistical parameters K and λ can be calculated analytically for a pair of sequences with given amino acid composition and scoring scheme. However, no perfect theory currently exists for gapped alignment score distribution, and for score distributions from alignment programs using additional features like difference blocks [8], and which use multiple parameter sets [9]. The problem of accurately determining the statistical significance of gapped sequence alignment has attracted a lot of attention in the recent years [11,12,13,14,15]. There exist a couple of good starting points for statistically describing gapped alignment score distributions [16,17], but a complete mathematical description of the optimal score distribution remains far from reach [17]. Some excellent reviews on statistical significance in sequence comparison are available in the literature [18,19,20].

Pairwise protein local sequence alignment programs give the optimal or suboptimal alignment of a given sequence pair. In the case of database searches, the second sequence is the complete database consisting of many sequences. Many approaches exist currently to estimate the statistical significance of a database hit (match of the query sequence with part of the database). For the database searches, the statistical significance of a pairwise alignment score is reported in terms of E-value, which is the expected number of hits in the database with a score equal or higher arising by chance, or the P-value, which is the probability of getting at least one score equal or higher arising by chance. These E-values and P-values are corresponding to the database, and although these can be converted to the pairwise E-values and P-values [15], they cannot estimate the true statistical significance of the specific pairwise alignment under consideration, since the database E-values and P-values depend on the average sequence features like length, amino acid composition, and not the features of sequence pair under consideration.

In particular, BLAST2.0 [3] reports the statistical significance as the likelihood that a similarity as good or better would be obtained by two random sequences with average amino-acid composition and lengths similar to the sequences that produced the score. However, if either of the two sequences has amino acid composition significantly different from the average, the statistical significance may be an over or underestimate. Similarly, the statistical estimates provided by the FASTA package [6,21] report the expectation that a sequence would obtain a similarity score against an unrelated sequence drawn at random from the sequence database that was searched, which again is dependent on the average sequence composition of the entire database and not on the specific sequence pair.

Accurate estimates of the statistical significance of pairwise alignments can be very useful to comment on the relatedness of a pair of sequences aligned by an alignment program independent of any database. And thus, pairwise statistical significance can also be used to compare different alignment programs independently. In addition to the standard local alignment programs [4,5], some recent programs have been developed [8,9] that take into account other desirable biological features in addition to gaps, like difference blocks, and the use of multiple parameter sets (substitution matrices, gap penalties). These features of the alignment programs enhance the sequence alignment of real sequences by suiting to different conservation rates at different spatial locations of the sequences. As pointed out earlier, rigorous statistical theory for alignment score distribution is available only for ungapped alignment, and not even for its simplest extension, i.e., alignment with gaps. Accurate statistics of the alignment score distribution from newer and more sophisticated alignment programs therefore is not expected to be straightforward. For comparing the performance of newer alignment programs, accurate estimates of pairwise statistical significance are needed.

The statistical significance of a pairwise alignment depends upon various factors: sequence alignment method, scoring scheme, sequence length, and sequence composition [19]. The straightforward way to estimate statistical significance of scores from an alignment program for which the statistical theory is unavailable is to generate a distribution of alignment scores using the program with randomly shuffled versions of the pair of sequences and compare the obtained score with the generated score distribution, either directly or by fitting an extreme value distribution (EVD) curve to the generated distribution to calculate the statistical significance of the obtained score (as described in the next section).

The PRSS program in the FASTA package [6,7,21] calculates the statistical significance of an alignment by aligning them, shuffling the second sequence up to 1000 times, and estimating the statistical significance from the distribution of shuffled alignment scores. It uses maximum likelihood to fit an EVD to the shuffled score distribution. A similar approach is also used in HMMER [22]. It also uses maximum likelihood fitting [23] and also allows for censoring of data left of a given cutoff, for fitting only the right tail of the histogram. A heuristic approximation of the gapped local alignment score distribution is also available [11], and based on these statistics, accurate formulae for statistical parameters K and λ for gapped alignments are derived and implemented in a program called ARIADNE [12]. These methods can provide an accurate estimation of statistical significance for gapped alignments, but currently do not incorporate the additional features of sequence alignment, like using difference blocks and multiple parameter sets [8,9].

The contribution of this paper is two-fold: First, we compare various existing methods to estimate pairwise statistical significance and determine the most accurate method for estimating it. We found that maximum likelihood fitting of score distribution censored left of peak (fitting right of peak) is the most accurate method. Secondly, we used this method in the experiments reported in [24] on a subset of the CATH2.3 database to compare the retrieval accuracy for pairwise statistical significance and database statistical significance. [24] had earlier created this database to evaluate seven protein structure comparison methods and the two sequence comparison programs SSEARCH and PSI-BLAST. Comparison of the results with those reported in [24] show that pairwise statistical significance gives comparable and at times better accuracy than the SSEARCH program, but less than PSI-BLAST.

2 The Extreme Value Distribution for Ungapped and Gapped Alignments

Just as the distribution of the sum of a large number of independent identically distributed (i.i.d) random variables tends to a normal distribution (central limit theorem), the distribution of the maximum of a large number of i.i.d. random variables tends to an extreme value distribution (EVD). This is an important fact, because it allows us to fit an EVD to the score distribution from any local alignment program, and use it for estimating statistical significance of scores from that program. The distribution of Smith-Waterman local alignment score between random, unrelated sequences is approximately a Gumbel-type EVD [10]. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP (High-scoring Segment Pairs which correspond to the ungapped local alignment) scores are characterized by two parameters, K and λ. The probability that the optimal local alignment score S exceeds x is given by the P-value:

$$\Pr(S > x) \sim 1 - e^{-E},$$

where E is the E-value and is given by

$$E = Kmne^{-\lambda x}.$$

For E-values less than 0.01, both E-value and P-values are very close to each other. The above formulae are valid for ungapped alignments [10], and the parameters K and λ can be computed analytically from the substitution scores and sequence compositions. An important point here is that this scheme allows for the use of only one substitution matrix. For the gapped alignment, no perfect statistical theory has yet been developed, although there exist some good starting points for the problem as mentioned before [16,17]. Recently, researchers have also looked closely at the low probability tail distribution, and the work in [25] applied a rare-event sampling technique and suggested a Gaussian correction to the Gumbel distribution to better describe the rare event tail, resulting in a considerable change in the reported significance values. However, for most practical purposes, the original Gumbel distribution has been widely used to describe gapped alignment score distribution [26,21,12,27,9].

From an empirically generated score distribution, we can directly observe the E-value E for a particular score x, by counting the number of times a score x or higher was attained. Since this number would be different for different number of random shuffles N (or number of sequences in the database in case of database search), a normalized E-value is defined as

$$E_{normalized} = \frac{E}{N} \ .$$

3 Tools and Programs Used

We worked with the alignment programs SIM [28], which is an ordinary alignment program (similar to SSEARCH), GAP3 [8], which allows dynamically finding similarity blocks and difference blocks, and GAP4 [9], which can also use multiple parameter sets (scoring matrices, gap penalties, difference block penalties) to generate a single pairwise alignment. For estimating the statistical parameters K and λ, we used several programs. First is PRSS from the FASTA package [6,7,21], which takes two protein sequences and one set of parameters (scoring matrix, gap penalty), generates the optimal alignment, and estimates the K and λ parameters by aligning up to 1000 shuffled versions of the second sequence, and fitting an EVD using maximum likelihood. In addition to uniform shuffling, it also allows for windowed shuffling. We also used ARIADNE [12], that uses an approximate formula to estimate gapped K and λ from ungapped K and λ. Both these methods are currently applicable only for alignment methods using one parameter set. We also used the linear regression fitting program used in [9] to estimate K and λ from an empirical distribution of alignment scores. Finally, we also used the maximum likelihood method [23] and corresponding routines in the HMMER package [22] to fit an EVD to the empirical distribution. We compared all these methods on the basis of accuracy in estimating K and λ values for a pair of sequences.

4 Experiments and Results

4.1 Accurate Estimation of K and for λ a Specific Sequence Pair

For each sequence pair, we need to find accurate estimates of the statistical parameters K and λ. Here, we are not too much concerned with the time taken for estimating K and λ since we are interested in determining the method which gives the most accurate estimates of the parameters. Therefore, we can afford to spend more time for accurate estimates.

To decide on the method for estimating statistical parameters for a sequence pair, we used the following approach: a pair of remotely homologous protein sequences was selected using PSI-BLAST by giving a G protein-coupled receptor sequence (GENE ID: 55507 GPRC5D) as query and running two iterations of PSI-BLAST. The second sequence was selected from the new results after second iterations that were not present in the results of the first iteration. The sequence was a novel protein similar to vertebrate pheromone receptor protein [Danio rerio] (emb|CAM56437.1|). We used this pair of real protein sequences to generate eleven large scale simulations of alignment score distributions using different alignment programs and scoring schemes described in Section 3. Each of the eleven simulations involved aligning one million pairs of randomly shuffled versions of the sequence pair (with different seeds for the random number generator). Because we are mostly interested in the tail distribution of scores, we looked at the distribution of scores for which the normalized E-value was less than 0.01. We got eleven empirically derived random distributions, and although theoretically they should have been same, there was slight variation within the eleven distributions (because of random sampling). Here we combined the eleven distributions by taking the mean of the E-values for each score from each of the eleven distributions. This is equivalent to doing one big simulation with eleven million shuffles. We assume that the resulting mean distribution is the most accurate representation of the actual distribution and subsequently used this distribution to validate the predicted E-values from different methods of estimating K and λ. Fig. 1 shows the mean score distribution (complementary distribution function in terms of statistics) based on the simulations, which is same as the normalized E-value, for three alignment schemes. The solid line curve shows the mean of the normalized E-values from the eleven different simulations. The vertical bars for each alignment score indicates the variation in normalized E-values observed within the eleven different simulations.

For evaluating various methods of estimating statistical parameters, the K and λ estimates from different programs for the same sequence pair were examined. For the PRSS program, both uniform and windowed shuffling was used with two values of window size: 10 and 20. The ARIADNE program was also used to estimate gapped K and λ. Since we are interested in accurate fitting of the tail distribution, for the curve fitting methods like maximum likelihood (ML) and linear regression (LR), we used the censored distribution for fitting. Here type-I censoring is defined as the one in which we fit only the data right of the peak of the histogram [23], and type-II censoring is defined as one where the cutoff is set to the score that corresponds to a normalized E-value of 0.01. We also show results for uncensored fitting with ML method, applied to the eleven empirical distributions (with a million shuffles each) to make a realistic comparison of other fitting schemes with the methodology used in PRSS, which also uses maximum likelihood method, but only up to 1000 shuffles. Since we generated eleven independent score distributions, we used them individually to estimate eleven pairs of K and λ using both ML and LR, so that we can perform the best case, worst case and average case prediction analysis for fitting methods. The estimated K and λ values from each program are used to predict the E-values for different alignment scores using the EVD formula, and the resulting distribution is compared with the mean empirical distribution generated from eleven independent simulations as described above.

Table 1 shows the comparison of the sum of squares of differences (SSD) between predicted normalized E-values and actual normalized E-values for different methods and alignment schemes. Since we had eleven estimates of K and λ for the ML and LR methods, we report the minimum, maximum and average SSD. PRSS and ARIADNE report one set of parameters, and thus there is

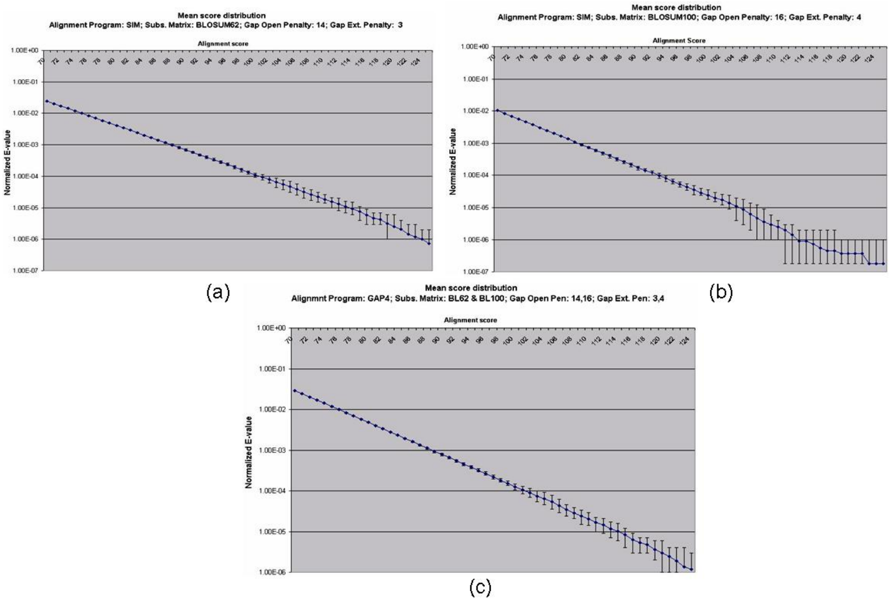

Fig. 1. Distribution of alignment scores generated (a) using SIM program and BLOSUM62 matrix, (b) using SIM program and BLOSUM100 matrix and (c) using GAP4 program and BLOSUM62 and BLOSUM100 matrices. The solid line curve represents the mean of the eleven distributions generated, and the vertical bars represent the variation within the eleven distributions.

Table 1. Comparison of the Sum of Squares of Differences (SSD) between predicted normalized E-values and actual normalized E-values for different methods and alignment schemes

Program: SIM		Matrix: BLOSUM62			GapOpenPen.: 14, GapExtPen.: 3				
Statistic	Ariadne	PRSS			Maximum Likelihood			LinRegr	Minimum
		Uniform	-w 10	-w 20	Full	Censor-I	Censor-II	Censor-II	
Min(SSD)					**8.05E-09**	9.11E-09	2.67E-08	8.58E-08	**8.05E-09**
Max(SSD)	5.6×	3.46×	4.22×	7.5×	6.03E-07	**2.75E-07**	2.15E-06	5.20E-06	**2.75E-07**
Avg(SSD)	E-04	E-05	E-02	E-03	3.02E-07	**7.91E-08**	6.08E-07	1.48E-06	**7.91E-08**

Program: SIM		Matrix: BLOSUM100			GapOpenPen.: 16, GapExtPen.: 4				
Statistic	Ariadne	PRSS			Maximum Likelihood			LinRegr	Minimum
		Uniform	-w 10	-w 20	Full	Censor-I	Censor-II	Censor-II	
Min(SSD)					1.88E-09	1.76E-09	**8.16E-10**	8.27E-09	**8.16E-10**
Max(SSD)	1.02×	4.58×	8.3×	4.38×	3.90E-08	**2.50E-08**	1.62E-07	4.20E-07	**2.50E-08**
Avg(SSD)	E-05	E-05	E-04	E-04	**8.51E-09**	9.18E-09	4.54E-08	1.13E-07	**8.51E-09**

Program: GAP4		Matrix: BL62,BL100			GapOpen:14,16 GapExt:3,4				
Statistic	Ariadne	PRSS			Maximum Likelihood			LinRegr	Minimum
		Uniform	-w 10	-w 20	Full	Censor-I	Censor-II	Censor-II	
Min(SSD)					2.20E-07	2.05E-08	**1.35E-08**	9.34E-08	**1.35E-08**
Max(SSD)	NA	NA	NA	NA	1.62E-06	**6.86E-07**	2.97E-06	9.77E-06	**6.86E-07**
Avg(SSD)					9.88E-07	**2.42E-07**	6.49E-07	2.83E-06	**2.42E-07**

only one SSD corresponding to these methods. Further, for alignment method GAP4 which can use multiple parameter sets, there is no entry corresponding to ARIADNE and PRSS, as these methods do not currently support the use of multiple parameter sets. The last column gives the minimum SSD obtained, and its second and third entries correspond to the minimum worst case and minimum average case error in prediction. We can see that the minimum SSD is obtained for the ML method in all cases. Specifically, ML fitting with type-I censoring gives the minimum Max(SSD), (i.e. minimum worst case error) for all the three cases. Therefore, we conclude that ML fitting with type-I censoring gives the most accurate estimates of statistical parameters K and λ.

4.2 Using Pairwise Statistical Significance to Infer Homology

To evaluate our method, we used a non-redundant subset of the CATH 2.3 database (Class, Architecture, Topology, and Hierarchy, [29]) provided by [24] and available at ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/prot_sci_04/. As described in [24], this dataset consists of 2771 domain sequences and includes 86 selected test query sequences, each representing at least five members of their respective CATH sequence family (35% sequence identity) in the data set. We used this database and query set for experimenting with pairwise statistical significance. For each of the 86×2771 comparisons, we used the maximum likelihood method with type-1 censoring with 2000 shuffles to fit the score distribution from the GAP3 program with a very high difference block penalty (to not use that feature), which essentially reduces it to an ordinary alignment program like SIM. Alignments were obtained using the BLOSUM50 substitution matrix (in 1/3 bit units as used by SSEARCH) with gap open penalty as 10, and gap extension penalty as 2. The same combination of parameters was used in [24] to report the results obtained with the SSEARCH program. The parameters K and λ resulting from the ML fitting were then used to find the pairwise statistical significance of the pairwise comparison, and the P-value was recorded. Following [24], Error per Query (EPQ) versus Coverage plots were used to present the results. To create these plots, the list of pairwise comparisons were sorted based on statistical significance, and subsequently, the lists were examined, from best score to worst. Going down the list, the coverage count is increased by one if the two members of the pair are homologs, and the error count is increased by one if they are not. At a given point in the list, EPQ is the total number of errors incurred so far, divided by the number of queries. Coverage at that point is the fraction of homolog pairs detected at this significance level.

For each of the 86 queries, 2771 comparisons were done, and EPQ vs. Coverage curves were plotted. Since the EPQ vs. Coverage curves on the complete dataset can be distorted due to poor performance by one or two queries (if those queries produce many errors at low coverage levels), reference [24] examined the performance of the methods with individual queries. Fig. 2(a) shows the level of coverage generated by the median query (43 queries performed better, 43 worse) at the 1st, 3rd, 10th, 30th, and 100th false positive for homologs. Fig. 2(b) shows the same results for 25th percentile of coverage (i.e. 21 of the queries have worse

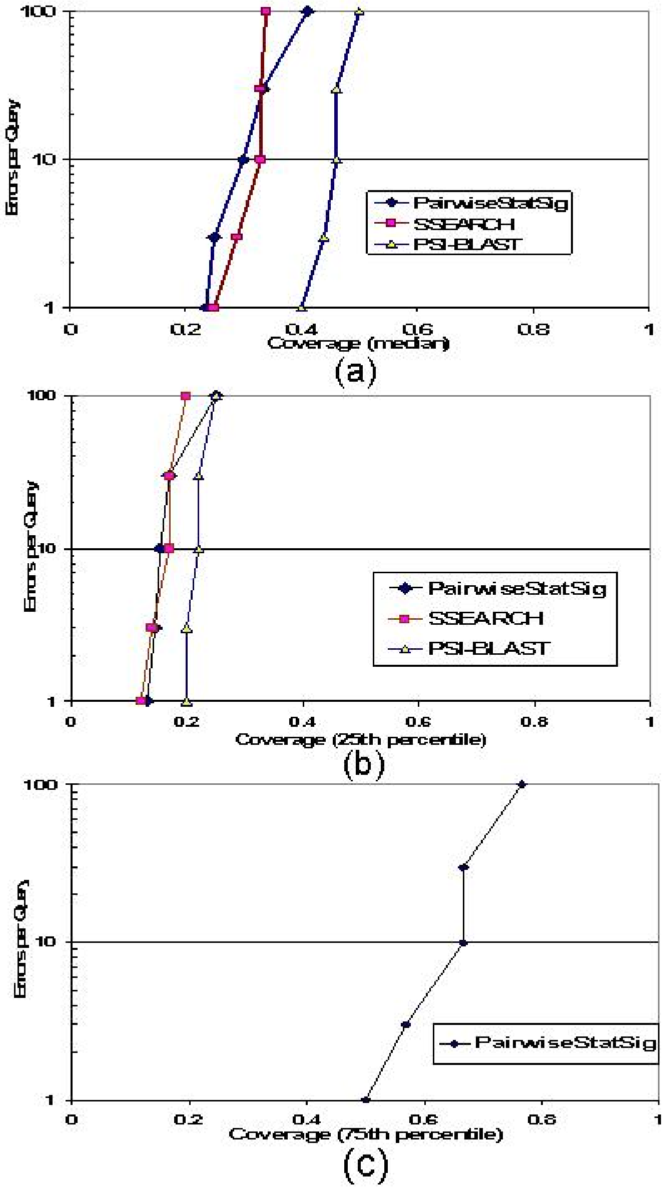

Fig. 2. Errors per Query vs. Coverage plots for individual queries. (a) The median level of coverage for 86 queries; (b) 1st quartile (25th percentile) coverage; (c) 3rd quartile (75th percentile) coverage. Along with the curve for pairwise statistical significance, the curves for SSEARCH and PSI-BLAST in (a) and (b) are derived from figures 2A and 2B in [24]. The corresponding results for (c) were not available in [24].

coverage, and 65 have better coverage). And fig. 2(c) shows the same results for 75th percentile of coverage (i.e. 65 of the queries have worse coverage, and 21 have better coverage). Along with the curve for pairwise statistical significance, the curves for SSEARCH and PSI-BLAST in fig. 2(a) and (b) are derived from the figures 2A and 2B in [24]. The results corresponding to Fig. 2(c) were not available in [24], and hence, only the results of pairwise statistical significance are reported. This figure shows that pairwise statistical significance performs comparable to and sometimes significantly better than database statistical significance (with SSEARCH program), particularly at higher error rates. However, the results using PSI-BLAST are clearly the best.

Since the SSEARCH program used the same substitution matrix as we used for our experiments (BLOSUM50) [24], the results indicate that pairwise statistical significance works better in practice than database statistical significance. However, even better results with PSI-BLAST using database statistical significance indicates that sequence specific substitution matrices should be used for the pairwise comparisons, and to fairly compare pairwise statistical significance with the database statistical significance reported by PSI-BLAST, more experiments need to be performed with pairwise statistical significance using sequence specific substitution matrices.

The time required to estimate pairwise statistical significance for a given pair of sequences is certainly expected to depend on the length of the two sequences. Therefore, to get an idea of the average time needed to estimate pairwise statistical significance using the proposed method, we used the following approach. We took a real sequence from the CATH2.3 database of length 135 (1que01) and estimated its pairwise statistical significance with more than a thousand other real sequences. It took 2574.151 seconds for finding 1013 pairwise statistical significance estimates on an Intel processor 2.8GHz, which means on an average 2.54 seconds per comparison. Certainly, this is much faster than a database search, if we are only interested in a specific (or a few) pairwise comparison(s), but will take a huge amount of time if applied for all pairwise comparisons in a large database search.

The program PairwiseStatSig is available for free academic use at
www.cs.iastate.edu/∼ankitag/PairwiseStatSig.html

5 Conclusion and Future Work

This paper explores the use of pairwise statistical significance, and compares it with database statistical significance for the application of homology detection. Large scale experimentation was done to determine the most accurate method for determining pairwise statistical significance. Further, preliminary experimentation for homology detection with a benchmark database (a subset of CATH2.3 database) shows that the pairwise statistical significance performs better than database statistical significance (using SSEARCH program), but still the accuracy of retrieval results is best for PSI-BLAST.

We believe that PSI-BLAST gives best results because of the use of sequence specific substitution matrices, although it also uses database statistical significance to estimate the E-value. Using pairwise statistical significance is shown to be better than database E-value (used in SSEARCH), and thus, we believe that the results of pairwise statistical significance can be further improved by using sequence specific substitution matrices, which is the significant part of our future work. Also, more experimentation with other standard databases such as SCOP can be done to compare the performance. Another major contribution can be to estimate the pairwise statistical significance accurately in less time, as the method used in this paper was to use maximum likelihood to fit a score distribution generated by simulation, which is not time-efficient. Faster methods for determining pairwise statistical significance are thus required. We have made some progress in this direction [30]. Another aspect of future work is to experiment with other sample space for shuffling of protein sequences for generating score distribution, which may provide better significance estimates.

References

1. Pearson, W.R., Lipman, D.J.: Improved Tools for Biological Sequence Comparison. Proceedings of the National Academy of Sciences, USA 85(8), 2444–2448 (1988)
2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic Local Alignment Search Tool. Journal of Molecular Biology 215(3), 403–410 (1990)
3. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Research 25(17), 3389–3402 (1997)
4. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology 147(1), 195–197 (1981)
5. Sellers, P.H.: Pattern Recognition in Genetic Sequences by Mismatch Density.. Bulletin of Mathematical Biology 46(4), 501–514 (1984)
6. Pearson, W.R.: Effective Protein Sequence Comparison. Methods in Enzymology 266, 227–259 (1996)
7. Pearson, W.R.: Flexible Sequence Similarity Searching with the FASTA3 Program Package.. Methods in Molecular Biology 132, 185–219 (2000)
8. Huang, X., Chao, K.-M.: A Generalized Global Alignment Algorithm. Bioinformatics 19(2), 228–233 (2003)
9. Huang, X., Brutlag, D.L.: Dynamic Use of Multiple Parameter Sets in Sequence Alignment. Nucleic Acids Research 35(2), 678–686 (2007)
10. Karlin, S., Altschul, S.F.: Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. Proceedings of the National Academy of Sciences, USA 87(6), 2264–2268 (1990)
11. Mott, R., Tribe, R.: Approximate Statistics of Gapped Alignments. Journal of Computational Biology 6(1), 91–112 (1999)
12. Mott, R.: Accurate Formula for P-values of Gapped Local Sequence and Profile Alignments. Journal of Molecular Biology 300, 649–659 (2000)
13. Altschul, S.F., Bundschuh, R., Olsen, R., Hwa, T.: The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Research 29(2), 351–361 (2001)

14. Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F.: Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-based Statistics and Other Refinements. Nucleic Acids Research 29(14), 2994–3005 (2001)
15. Yu, Y.K., Gertz, E.M., Agarwala, R., Schäffer, A.A., Altschul, S.F.: Retrieval Accuracy, Statistical Significance and Compositional Similarity in Protein Sequence Database Searches. Nucleic Acids Research 34(20), 5966–5973 (2006)
16. Kschischo, M., Lässig, M., Yu, Y.: Toward an Accurate Statistics of Gapped Alignments. Bulletin of Mathematical Biology 67, 169–191 (2004)
17. Grossmann, S., Yakir, B.: Large Deviations for Global Maxima of Independent Superadditive Processes with Negative Drift and an Application to Optimal Sequence Alignments. Bernoulli 10(5), 829–845 (2004)
18. Pearson, W.R., Wood, T.C.: Statistical Significance in Biological Sequence Comparison. In: Balding, D.J., Bishop, M., Cannings, C. (eds.) Handbook of Statistical Genetics, pp. 39–66. Wiley, Chichester, UK (2001)
19. Mott, R.: Alignment: Statistical Significance. Encyclopedia of Life Sciences (2005), http://mrw.interscience.wiley.com/emrw/9780470015902/els/article/a0005264/current/abstract
20. Mitrophanov, A.Y., Borodovsky, M.: Statistical Significance in Biological Sequence Analysis. Briefings in Bioinformatics 7(1), 2–24 (2006)
21. Pearson, W.R.: Empirical Statistical Estimates for Sequence Similarity Searches. Journal of Molecular Biology 276, 71–84 (1998)
22. Eddy, S.R.: Multiple Alignment Using Hidden Markov Models. In: Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T., Wodak, S. (eds.) Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, pp. 114–120. AAAI Press, Menlo Park (1995)
23. Eddy, S.R.: Maximum Likelihood Fitting of Extreme Value Distributions (1997), unpublished manuscript, citeseer.ist.psu.edu/370503.html
24. Sierk, M.L., Pearson, W.R.: Sensitivity and Selectivity in Protein Structure Comparison. Protein Science 13(3), 773–785 (2004)
25. Wolfsheimer, S., Burghardt, B., Hartmann, A.K.: Local Sequence Alignments Statistics: Deviations from Gumbel Statistics in the Rare-event Tail. Algorithms for Molecular Biology 2(9) (2007), http://www.almob.org/content/2/1/9
26. Altschul, S.F., Gish, W.: Local Alignment Statistics. Methods in Enzymology 266, 460–480 (1996)
27. Olsen, R., Bundschuh, R., Hwa, T.: Rapid Assessment of Extremal Statistics for Gapped Local Alignment. In: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pp. 211–222. AAAI Press, Menlo Park (1999)
28. Huang, X., Miller, W.: A Time-efficient Linear-space Local Similarity Algorithm. Advances in Applied Mathematics 12(3), 337–357 (1991)
29. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH - A Hierarchic Classification of Protein Domain Structures. Structure 28(1), 1093–1108 (1997)
30. Agrawal, A., Ghosh, A., Huang, X.: Estimating Pairwise Statistical Significance of Protein Local Alignments Using a Clustering-Classification Approach Based on Amino Acid Composition. In: Măndoiu, I., Sunderraman, R., Zelikovsky, A. (eds.) ISBRA 2008. LNCS(LNBI), vol. 4983, pp. 62–73. Springer, Heidelberg (2008)