Conservative, Non-Conservative and Average Pairwise Statistical Significance of Local Sequence Alignment

Ankit Agrawal and Xiaoqiu Huang
Department of Computer Science, Iowa State University,
226 Atanasoff Hall, Ames, IA 50011-1041, USA,
{ankitag,xqhuang}@iastate.edu

Abstract

Estimation of statistical significance of a pairwise alignment is an important problem in sequence comparison. Recently, it was shown that pairwise statistical significance does better in practice than database statistical significance in terms of retrieval accuracy of homologs. In this paper, we introduce the concept of conservative, non-conservative, and average pairwise statistical significance which can be easily derived from original pairwise statistical significance estimates and use more information specific to the sequence pair under consideration using multiple shuffle spaces. Experimental results for homology detection reveal that the proposed measures give at least comparable or significantly better retrieval accuracy than original pairwise statistical significance and database statistical significance reported by BLAST, PSI-BLAST, and SSEARCH. The use of the proposed measures is further shown to be extremely useful when using sequence-specific substitution matrices.

1. Introduction

1.1. Statistical Significance of Sequence Alignment Scores

Sequence alignment is an underlying application in the comparison of DNA and protein sequences [3]. There exist programs for sequence alignment that use popular algorithms [8] or their heuristic versions [3, 6]. The local sequence alignment programs typically report alignment scores for the alignments constructed, and related (homologous) sequences will have *higher* alignment scores. But whether a given score is *high* enough or not depends on the alignment score distribution, and hence estimating statistical significance of an alignment score is very useful. An alignment score is considered statistically significant if it has a low probability of occurring by chance. Since the alignment score distribution depends on various factors like alignment program, scoring scheme, sequence lengths, sequence compositions [4], it is possible to have two alignments of different sequence pairs with scores x and y with $x < y$, but x more significant than y. Therefore, compared to alignment score, the statistical significance of an alignment score is considered a better indicator of (potential) biological significance.

1.2. Database statistical significance versus pairwise statistical significance

Recently, a study of pairwise statistical significance and its comparison with database statistical significance was conducted [1]. In summary, the database statistical significance which is commonly reported by most database search programs is database-dependent, and hence the same pairwise alignment with same alignment score can be assessed different significance values in different database searches. Pairwise statistical significance, on the other hand is database-independent and specific to the sequence pair being aligned. In [1], various approaches to estimate pairwise statistical significance were compared to find that maximum likelihood fitting with censoring left of peak is the most accurate method for estimating pairwise statistical significance. Further, comparison with database statistical significance revealed that pairwise statistical significance performs comparable to and sometimes marginally better than database statistical significance using SSEARCH, and hence significantly better than BLAST and FASTA.

2. Conservative, Non-Conservative, and Average Pairwise Statistical Significance

In this paper, we introduce the concept of *conservative*, *non-conservative*, and *average* pairwise statistical significance, which can be derived using simple functions from original pairwise statistical significance estimates [1], and

978-0-7695-3452-7/08 $25.00 © 2008 IEEE
DOI 10.1109/BIBM.2008.19

give better results. Consider the pairwise statistical significance defined in [1] to be obtainable by the following function: $PairwiseStatSig(Seq1, Seq2, SC, N)$ where $Seq1$ is the first sequence, $Seq2$ is the second sequence, SC is the scoring scheme, and N is the number of shuffles. The function PairwiseStatSig, therefore generates a score distribution by aligning $Seq1$ with N shuffled versions of $Seq2$, fits the distribution to an extreme value distribution using censored maximum likelihood fitting to obtain statistical parameters K and λ, and returns the pairwise statistical significance estimate of the pairwise alignment score between $Seq1$ and $Seq2$ using the parameters K and λ. More details on pairwise statistical significance can be found in [1].

Using this function two times with different ordering of sequence inputs, we can define conservative, non-conservative, and average pairwise statistical significance. Let

$$S1 = PairwiseStatSig(Seq1, Seq2, SC, N)$$
$$S2 = PairwiseStatSig(Seq2, Seq1, SC, N)$$

Then,
Conservative Pairwise Statistical Significance

$$= max\{S1, S2\}$$

Non-Conservative Pairwise Statistical Significance

$$= min\{S1, S2\}$$

Average Pairwise Statistical Significance

$$= avg\{S1, S2\}$$

Using the $PairwiseStatSig$ function two times in this way makes sure that both sequences are shuffled separately to generate two different distributions to get two different pairwise statistical significance estimates for the same sequence pair ($S1$ and $S2$), and the final reported pairwise statistical significance estimate is a simple function of these two individual estimates. Conservative pairwise statistical significance is termed as 'conservative' because it reports the maximum of $S1$ and $S2$, which means that two sequences would be declared as related only if both $S1$ and $S2$ are low enough. Similarly, non-conservative pairwise statistical significance is termed as 'non-conservative' because it reports the minimum of $S1$ and $S2$, which means that even if one of $S1$ or $S2$ is low enough, $Seq1$ and $Seq2$ would be declared related. The definition of average pairwise statistical significance follows naturally as the average of $S1$ and $S2$.

This very simple modification of using the $PairwiseStatSig$ function two times with different shuffle spaces is expected to capture more information specific to the sequence pair being aligned, and hence give better performance in terms of retrieval accuracy. Intuitively, this approach is expected to be most effective when the individual estimates $S1$ and $S2$ are sufficiently different, since if they are almost equal, all the three proposed estimate measures would be roughly the same. Note that this approach facilitates the use of sequence-specific/position-specific substitution matrices, and further enhances its benefits. Since during the calculation of $S1$, only $Seq2$ is shuffled, the sequence-specific substitution matrix for $Seq1$ can be used for generating the empirical score distribution, even if it is position-specific. Similarly, during the calculation of $S2$, only $Seq1$ is shuffled, and the sequence-specific substitution matrix for $Seq2$ can be used for generating the score distribution. Since $S1$ and $S2$ are expected to be most different when using sequence-specific substitution matrices for alignment, this approach is expected to be very useful when sequence-specific substitution matrices are available for both the sequences being aligned.

3. Experiments and Results

To evaluate the performance of the proposed significance measures, we used the experiment setup used earlier in [7], and subsequently in [1]. A non-redundant subset of the CATH 2.3 database (Class, Architecture, Topology, and Hierarchy, [5]) available at ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/prot_sci_04/ was selected in [7] to evaluate seven structure comparison programs and two sequence comparison programs. This benchmark dataset consists of 2771 domain sequences and includes 86 query sequences.

Following [7], Error per Query (EPQ) versus Coverage plots were used to visualize the results. To create these plots, the list of pairwise comparisons was sorted based on decreasing statistical significance (increasing P-values). Traversing the sorted list from top to bottom, the count of true homologs detected is increased by one if the two sequences of the pair are homologs, else the error count is increased by one. At any given point in the list, EPQ is the total number of errors incurred so far, divided by the number of queries; and coverage is the fraction of total homolog pairs so far detected. Te ideal curve would go from 0% to 100% coverage, without incurring any errors, which would correspond to a straight line on the x-axis. Therefore, a curve more to the right is better. The EPQ vs. Coverage curves for the proposed significance measures using four BLOSUM substitution matrices are presented in Fig. 1. For comparison purposes, the corresponding curves using original pairwise statistical significance is also presented in the same figures. The curves are quite close to each other, which means that the individual estimates $S1$ and $S2$ are very close to each other as expected, because of using general substitution matrices (same scoring scheme SC). Still, in all the four sub-figures of Fig. 1, the curve for original

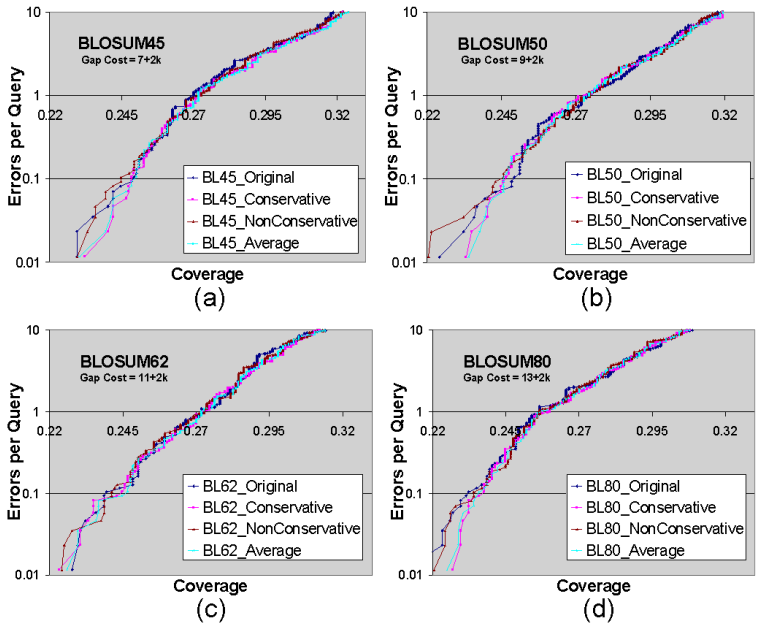

Figure 1. EPQ vs. Coverage plot for original, conservative, non-conservative, and average pairwise statistical significance using four substitution matrices. (a) BLOSUM45; (b) BLOSUM50; (c) BLOSUM62; (d) BLOSUM80. Although the curves are very close to each other, in all the four figures, the curve for original pairwise statistical significance is towards the left for most error levels.

pairwise statistical significance is towards the left at most error levels. To further demonstrate the impact of using the proposed measures, we also present an example of using sequence-specific substitution matrices. Fig. 2 shows the EPQ vs. Coverage plot for different kinds of pairwise statistical significance using sequence-specific substitution matrices. The details of deriving sequence-specific substitution matrices can be found in [2]. Fig. 2 clearly reveals the significant improvement in retrieval accuracy using the proposed significance measures. Notably, non-conservative pairwise statistical significance outperforms all other measures. Therefore, it suggests that using the proposed significance measures gives performance at least comparable, and many times significantly better than original pairwise statistical significance. 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) [7], to compare the performance of the proposed measures with database statistical significance, we examined the performance of the methods with individual queries, following the work in [7]. The coverage of each of the 86 queries at the 1st, 3rd, 10th, 30th, and 100th error was recorded, and the median coverage for each error level across the 86 queries was plotted to obtain EPQ vs. Coverage curves for the method to be evaluated.

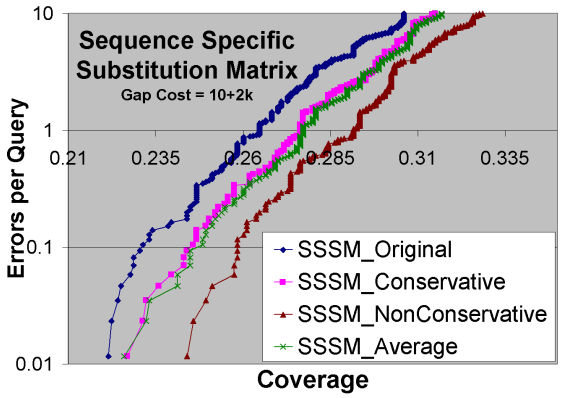

Figure 2. EPQ vs. Coverage plot for different kinds of pairwise statistical significance using sequence specific substitution matrices. Non-conservative pairwise statistical significance outperforms other variants of pairwise statistical significance. All the three variants proposed in this paper are better than original pairwise statistical significance.

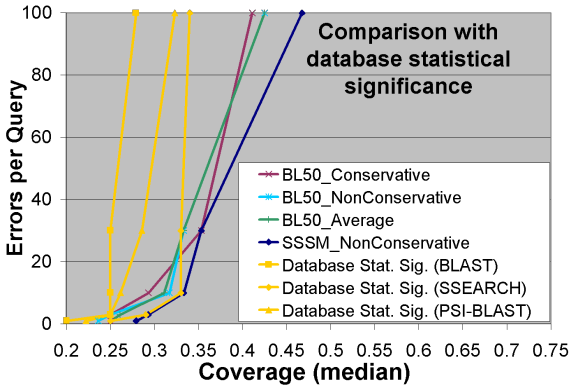

Figure 3. Comparison of proposed significance measures with database statistical significance using BLAST, PSI-BLAST and SSEARCH. The proposed meausres are significantly better than BLAST and PSI-BLAST. With general substitution matrices, the performance of the proposed measures is significantly better than SSEARCH only for higher error levels. Using sequence-specific substitution matrices with non-conservative pairwise statistical significance is better than SSEARCH at all error levels.

Fig. 3 shows the median coverage level at the 1st, 3rd, 10th, 30th, and 100th false positive for homologs (i.e. 43 of the queries have worse coverage, and 43 have better coverage). The curve for SSEARCH in Fig. 3 is derived from the figure 2A in [7]. All other curves were obtained by experimentation. The curves suggest that using the proposed variants gives significantly better results than database statistical significance using BLAST and PSI-BLAST at all error levels, and better than SSEARCH only at higher error levels. Further, non-conservative pairwise statistical significance using sequence-specific substitution matrices is significantly better than all three. According to experiments reported in [7], it is possible to improve PSI-BLAST results by using position-specific scoring matrices (PSSMs) derived against the BLAST non-redundant protein database rather than against the (smaller) benchmark database.

4. Conclusion and Future Work

This paper extends the work on pairwise statistical significance by introducing the concept of conservative, non-conservative, and average pairwise statistical significance, and compares them with database statistical significance for the knowledge discovery application of homology detection. Results indicate that deriving more sequence-pair-specific information by using the proposed measures is slightly better than original pairwise statistical significance and also better than database statistical significance using BLAST, PSI-BLAST and SSEARCH, but the accuracy of PSI-BLAST can be further improved using more information from larger universal databases.

Since PSI-BLAST results can be improved by using better quality PSSMs derived from larger universal protein databases, we believe that the performance of pairwise statistical significance can also be improved using position-specific substitution matrices, which is a significant part of our future work. Another important contribution can be to speed up the estimation process, since the variants proposed in this work take about twice the time compared to original pairwise statistical significance.

Acknowledgments

The authors would like to thank Dr. Sean Eddy for making the HMMER routines of censored maximum likelihood fitting available online, Dr. William R. Pearson for making the benchmark protein comparison database available online, and Dr. Volker Brendel for helpful discussions and providing links to the data.

References

[1] A. Agrawal, V. Brendel, and X. Huang. Pairwise Statistical Significance Versus Database Statistical Significance for Local Alignment of Protein Sequences. In *Bioinformatics Research and Applications*, volume 4983 of *LNCS(LNBI)*, pages 50–61. Springer Berlin/Heidelberg, 2008.

[2] A. Agrawal and X. Huang. Using sequence-specific substitution matrices for estimating pairwise statistical significance of local sequence alignment. 2008. in preparation.

[3] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[4] R. Mott. Alignment: Statistical Significance. *Encyclopedia of Life Sciences*, 2005. available at http://mrw.interscience.wiley.com/emrw/9780470015902/els/article/a0005264/current/abstract.

[5] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH - A Hierarchic Classification of Protein Domain Structures. *Structure*, 28(1):1093–1108, 1997.

[6] W. R. Pearson. Flexible Sequence Similarity Searching with the FASTA3 Program Package. *Methods in Molecular Biology*, 132:185–219, 2000.

[7] M. L. Sierk and W. R. Pearson. Sensitivity and Selectivity in Protein Structure Comparison. *Protein Science*, 13(3):773–785, 2004.

[8] T. F. Smith and M. S. Waterman. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.