Sequence-specific sequence comparison using pairwise statistical significance

by

Ankit Agrawal

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Science

Program of Study Committee:
Xiaoqiu Huang, Major Professor
Volker Brendel
David Fernández-Baca
Dimitris Margaritis
Arka Ghosh

Iowa State University

Ames, Iowa

2009

DEDICATION

To my teacher

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my major professor and advisor Dr. Xiaoqiu Huang for his constant support and guidance throughout the course of this study. I really appreciate his gentleness and considerate nature, and at the same time being highly professional at work. He always gave me the necessary time whenever I needed it, and has been always supportive to give the necessary freedom for pursuing my ideas, along with actively giving his own constructive inputs and suggestions. My experience with him here during this study has been extremely helpful and enriching for me. He really considers deeply about the welfare of his student. I am also grateful to Dr. Volker Brendel. His course Bioinformatics-II was extremely helpful for me while I was trying to enter the research arena, and I continued to receive his valuable inputs since then through numerous personal meetings for which he has always extended himself. He provided the links to the experimental data which was used in the experiments reported in this thesis. Having also worked with him during preparation of a couple of manuscripts, I have a learnt a lot from his research writing skills.

I sincerely thank all my committee members for their support and time. They always spared their valuable time whenever I needed their help. Special thanks are due to Dr. Arka Ghosh, whose course on statistics proved extremely helpful for this work, and he has been always willing to extend himself to help me even after the course. I would also like to acknowledge the computational facilities available here at ISU that really helped me. I would like to thank all my teachers, friends, and colleagues here whose well-wishings and friendships made my whole stay here very pleasant.

I am most grateful to Dr. P.V. Krishnan, whose personal example and association has been the inspiring and motivating factor for me in all aspects of life. From him I have not

only learnt how to be a successful researcher, but more importantly the need of good character as a responsibility that comes with education. I am also very grateful to Dr. Ankush Mittal, my advisor in my undergraduate days at IIT Roorkee, whose example was instrumental in inspiring me to pursue higher studies.

I am very grateful to my parents and brother for supporting me to pursue higher studies, without which it would not have been possible to do this work. I am very grateful for the love and affection of my friends here, especially Dr. Siddhartha, Dr. Kasthurirangan, Amit, Abhisek, Sparsh, Sidharath, Venkat, Chetan, Dr. Siva, Dr. Lakshminarasimhan, Dr. Tanay, Dr. Balasubramaniam, Mahantesh, Rakesh, Sandeep, Ganesh, Sudhindra, Tu-Liang, whose loving support has always been unwavering and without material expectations. I cannot forget the help and more than brotherly support I have received from Dr. Siddhartha and Dr. Kasthurirangan when I had just come to the US, which made me experience a home away from home.

I am grateful to God for everything and hope to be able to use all these gifts wisely for the purpose they are given.

ABSTRACT

Sequence comparison is one of the most fundamental computational problems in bioinformatics for which many approaches have been and are still being developed. In particular, pairwise sequence alignment forms the crux of both DNA and protein sequence comparison techniques, which in turn forms the basis of many other applications in bioinformatics. Pairwise sequence alignment methods align two sequences using a substitution matrix consisting of pairwise scores of aligning different residues with each other (like BLOSUM62), and give an alignment score for the given sequence-pair. The biologists routinely use such pairwise alignment programs to identify similar, or more specifically, related sequences (having common ancestor). It is widely accepted that the relatedness of two sequences is better judged by statistical significance of the alignment score rather than by the alignment score alone. This research addresses the problem of accurately estimating statistical significance of pairwise alignment for the purpose of identifying related sequences, by making the sequence comparison process more sequence-specific.

The major contributions of this research work are as follows. Firstly, using sequence-specific strategies for pairwise sequence alignment in conjunction with sequence-specific strategies for statistical significance estimation, wherein accurate methods for pairwise statistical significance estimation using standard, sequence-specific, and position-specific substitution matrices are developed. Secondly, using pairwise statistical significance to improve the performance of the most popular database search program PSI-BLAST. Thirdly, design and implementation of heuristics to speed-up pairwise statistical significance estimation by an factor of more than 200. The implementation of all the methods developed in this work is freely available online.

With the all-pervasive application of sequence alignment methods in bioinformatics using

the ever-increasing sequence data, this work is expected to offer useful contributions to the research community.

1. INTRODUCTION

Motivation for Present Research

Pairwise sequence alignment is an extremely important and common application in the analysis of DNA and protein sequences [52, 12, 65, 13, 50, 48, 40, 24]. It forms the basic step of many other bioinformatics applications like multiple sequence alignment, database search, finding protein function, protein structure, phylogenetic analysis, etc. for making various high level inferences about the DNA and protein sequences. Biological sequence data is also far more abundant as compared to other kinds of biological data, for example, protein structure data or microarray data.

In all applications making use of pairwise sequence alignment, the pairwise alignment step is primarily used to identify related sequences, i.e., sequences evolved from a common ancestor. A typical pairwise alignment program aligns two sequences and constructs an alignment with maximum similarity score. Although related sequences will have high similarity scores, the threshold alignment score T below which the two sequences can be considered unrelated depends on the probability distribution of alignment scores between random, unrelated sequences [42]. Therefore, the biological significance of a pairwise sequence alignment is gauged by the statistical significance rather than the alignment score alone. This means that if an alignment score has a low probability of occurring by chance, the alignment is considered statistically significant, and hence biologically significant. Of course, it is important to note here that although statistical significance may be a good preliminary indicator of biological significance which may be helpful in identifying potential homologs, statistical significance does not necessarily imply biological significance [9, 48, 42, 40].

The alignment score distribution depends on various factors like alignment program, scoring

scheme, sequence lengths, sequence compositions [42]. Fig. 1.1 shows two alignment score distributions (probability density functions) X and Y. Consider scores x and y in the score distributions X and Y respectively. Clearly $x < y$, but x is more statistically significant than y, since x lies more in the right tail of the distribution and is less probable to have occurred by chance. The shaded region represents the probability that a score equal or higher could have been obtained by chance. Therefore, instead of simply using the alignment score as the metric for homology, it is very useful to estimate the statistical significance of an alignment score to comment on the relatedness of the two sequences being aligned.

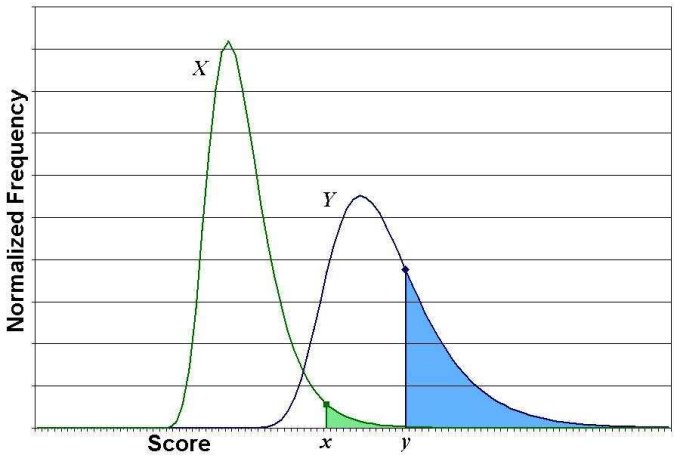

Figure 1.1 Two alignment score distributions depicting the advantage of statistical significance over alignment scores. $x < y$, but x is more statistically significant than y. A more statistically significant score is less likely to have occurred by chance, and hence is considered potentially biologically significant as well. The shaded region represents the probability that a score equal or higher could have been obtained by chance.

For ungapped alignments, rigorous statistical theory for the alignment score distribution is available [34]. However, no precise statistical theory currently exists for the simplest extension of ungapped alignment, which is alignment with gaps, although there exist a couple of good starting points for statistically describing gapped alignment score distributions for simple scoring schemes [36, 25]. But a complete mathematical description of the optimal score distribution remains far from reach [25]. Accurate statistics of the alignment score distribu-

tion from newer and more sophisticated alignment programs using difference blocks [32] and multiple parameter sets [31] therefore is not expected to be straightforward.

With the all-pervasive application of sequence alignment methods in bioinformatics using the ever-increasing sequence data, and the development of more sophisticated alignment methods whose statistics are not expected to be straightforward, we think that new approaches for accurate estimation of statistical significance of pairwise alignment would be highly useful for the bioinformatics community. This motivates the present research.

Nature of the Problem

The evolution of DNA and proteins in living organisms is influenced by a number of random factors, and observed patterns in DNA or protein sequences may be due to such factors, rather than from the selective pressure maintaining a certain function [40]. This means that not all sequence pairs which resemble each other will be homologs. Therefore, usually for the purpose of homology detection, the biologist is interested in finding if there are any biologically relevant targets sharing important structural and functional characteristics, which are common to the two sequences. Thus, pairwise local alignment [63] is more commonly used for this purpose which finds highly similar regions between two sequences, rather than global alignment [45] which are optimized along the whole length of the two sequences. Local pairwise sequence alignment methods align two sequences using a substitution matrix consisting of pairwise scores of aligning different residues with each other (like BLOSUM62), and give an alignment score for the given sequence-pair. The question here arises how likely it is that a high local similarity score is obtained by chance, which gives importance to the concept of statistical significance.

Statistical significance of a pairwise alignment score is commonly assessed by its P-value, which denotes the probability that an alignment with this score or higher occurs by chance alone. The notion of a P-value stems from the general statistical methodology of hypothesis testing [37]. Here, the test statistic is the alignment score, the null hypothesis is that the aligned sequences are unrelated, and the alternate hypothesis is that the sequences are related

(homologous). Therefore, the P-value is the probability of seeing an alignment score as extreme as the observed value, assuming that the null hypothesis is true. Thus, a close to zero P-value for an alignment score suggests that it could not have arisen by chance, and the sequences are, with good probability related. In Fig. 1.1, the shaded region represents the P-value. It is easy to see that estimation of the P-value for an alignment score requires the knowledge of the distribution of the local alignment scores. Since, the related sequences will (generally) have high scores, the right tail behavior of the score distribution is most crucial.

Score distribution for ungapped local alignment is known to follow a Gumbel-type EVD [34], as shown in Fig. 1.1 with analytically calculable parameters. For the gapped alignment, no perfect statistical theory has yet been developed, although there is ample empirical evidence that the gapped alignment score distribution also closely follows Gumbel-type EVD [65, 11, 50, 41, 46, 44, 31].

A good pairwise alignment based sequence comparison strategy should therefore, have the following characteristics:

1. *Sequence-specificity*: Since the distribution of alignment scores and hence the statistical significance depends on various factors like alignment program, scoring scheme, sequence lengths, sequence compositions [42], a good statistical significance estimation strategy should take all these factors for the specific sequence-pair being aligned into account.

2. *Statistical significance accuracy*: The approach should be able to estimate the P-values for high scores in the tail region of the distribution accurately.

3. *Retrieval accuracy*: Most importantly, the approach should be able to perform well for the central application of sequence comparison - identifying related sequences. Thus, it should assign lower P-values to pairs of related sequences than to pairs of unrelated sequences, which is commonly measured by retrieval accuracy (also known as coverage).

4. *Speed*: The estimation process should be fast enough to be usable in practice.

Recent Research Relevant to the Problem

It is a well-known result in statistics that 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). Similarly, the distribution of the maximum of a large number of i.i.d. random variables tends to an extreme value distribution (EVD) [26, 35]. This fact is theoretically well-founded for the ungapped local sequence alignment, where the distribution of Smith-Waterman local alignment score between random, unrelated sequences is known to follow a Gumbel-type EVD [34], as shown in Fig. 1.1. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP (High-scoring Segment Pairs which correspond to 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:

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

where E is the E-value and is given by

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

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 [34], and the parameters K and λ can be computed analytically from the substitution scores and sequence compositions.

For the gapped alignment, no perfect statistical theory has yet been developed, although there is ample empirical evidence that the gapped alignment score distribution also closely follows Gumbel-type EVD [65, 11, 50, 41, 46, 44, 40, 31]. Therefore, the frequently used approach has been to fit the score distribution to an extreme value distribution to get the parameters K and λ. In general, the approximations thus obtained are quite accurate [40]. The currently available theoretical results [60, 61] also support the assumption that gapped alignment score distribution follows Gumbel-type EVD. As mentioned before, there exist a couple of good starting points for statistically describing gapped alignment score distributions for simple scoring schemes [36, 25]. But currently, no rigorous statistical theory is available for the general case of local gapped pairwise alignments [40].

Some excellent reviews on statistical significance in sequence comparison are available in the literature [48, 53, 42, 40]. Here, we discuss some of the recent key developments in the field related to the problem. The HMMER program [21] uses maximum likelihood fitting [22] of extreme value distribution and also allows for censoring of data left of a given cutoff, for fitting only the right tail of the histogram. In addition to direct distribution fitting methods, a popular method is the island method [46, 10], which derives multiple island scores from a single optimal alignment, which are subsequently used to estimate the statistical parameters. Although efficient than numerical simulation, these methods are still known to be rather time-consuming [18] and hence the parameters K and λ have to be pre-computed for some specific scoring schemes. The latest versions of the widely popular database search program BLAST [57] uses the island method, along with a rescaling technique where the substitution matrix is scaled by an appropriate factor to take into account the variation in sequence composition, so that the relative entropy of the scaled matrix is close to that of the matrix originally used to numerically derive K and λ for gap penalties being used. Subsequently the corresponding originally derived K and λ are used for statistical significance estimation. An importance sampling based method for estimating λ was used in [18] and was shown to be at least five times faster than traditional methods. The method was further theoretically justified in [25]. [14] developed a maximum likelihood for the simultaneous estimation of K, λ, and H, where H is the relative entropy of the scoring system. The work in [43, 41] uses heuristic approximations used to derive approximate formulae for gapped K and λ from ungapped K and λ, taking into account the sequence length and composition for arbitrary gap penalties and substitution matrices. Recently, researchers have also looked closely at the low probability tail distribution, and the work in [66] applied a rare-event sampling technique earlier used in [27] 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 [65, 11, 50, 41, 46, 44, 31].

Most of the above mention approaches are designed to estimate statistical significance in

context of a database search. Database search programs typically use heuristics to obtain a sub-optimal local alignment in less time. Popular database search programs are BLAST [13], FASTA [49, 51, 50], SSEARCH (using full implementation of Smith-Waterman algorithm [63]), and PSI-BLAST [13, 57]. The database statistical significance so obtained for a pairwise comparison is dependent on the database, and will be different for different database, and even for the same database at different times, since the size of the database keeps on changing. In particular, BLAST2.0 [13] 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 [49, 50] 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.

In contrast to database statistical significance, the approach of fitting an extreme value distribution to a empirically generated score distribution from random shuffles of a specific pair has also been used to obtain pairwise statistical significance. The PRSS program in the FASTA package [49, 51, 50] 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. In addition to maximum likelihood fitting, linear regression has also been used [31] to fit score distributions to estimate statistical parameters. Some of the methods described earlier reporting database statistical significance can also be tweaked to estimate pairwise statistical significance by giving the second sequence as the database. For example, there exist the BL2Seq program [64], which uses the BLAST engine to align two sequences. Another example is ARIADNE program [41] which uses a formula for gapped K and λ.

Research Problem Statement

The research problem statement is *"to use sequence-specific strategies for pairwise sequence alignment and statistical significance estimation to accurately and quickly estimate the statistical significance of pairwise local sequence alignment for the purpose of identifying related sequences by using computational, statistical, and heuristic methods"*. In accordance with the characteristics of a good pairwise sequence alignment based sequence comparison strategy outlined before, this work pursues the goal in terms of the following intermediate goals:

1. Comparing existing approaches for estimating pairwise statistical significance in terms of *statistical significance accuracy*.
2. Comparing pairwise statistical significance with database statistical significance in terms of *retrieval accuracy*.
3. Using *sequence-specific* strategies for pairwise sequence alignment.
   (a) Multiple parameter sets.
   (b) Sequence-specific substitution matrices.
   (c) Position-specific substitution matrices.
4. Using *sequence-specific* strategies for statistical significance estimation
   (a) Using pairwise statistical significance instead of database statistical significance.
   (b) Using multiple shuffle spaces for generating empirical distribution.
5. Refining the results of a *fast* database search program like PSI-BLAST using pairwise statistical significance.
6. *Speeding up* pairwise statistical significance
   (a) Derived distribution points heuristic.
   (b) Database search heuristic.

Thesis organization

The thesis is organized as follows:

1. Chapter 1: Introduction

   It gives a general introduction to the thesis, presenting the motivation for present research, nature of the problem, recent relevant research, research problem statement and solution strategy, and thesis organization.

2. Chapter 2: Pairwise statistical significance and empirical determination of effective gap opening penalties for protein local sequence alignment

   This chapter compares different methods for estimating pairwise statistical significance in terms of statistical significance accuracy, and compares the best method (censored maximum-likelihood fitting) with database statistical significance in terms of retrieval accuracy. It also presents an application of pairwise statistical significance to empirically determine the most effective gap opening penalties for protein local sequence alignment.

3. Chapter 3: Pairwise statistical significance of local sequence alignment using multiple parameter sets and empirical justification of parameter set change penalty

   This chapter uses the censored maximum-likelihood fitting method in conjunction with dynamic use of multiple parameter sets (substitution matrix, gap opening penalty, gap extension penalty) to estimate pairwise statistical significance. Further, it also provides empirical justification of parameter set change penalty, which is an alignment parameter used for alignment with multiple parameter sets.

4. Chapter 4: Conservative, non-conservative and average pairwise statistical significance of local sequence alignment

   This chapter presents and demonstrates the usefulness of novel sequence-specific strategies for pairwise statistical significance estimation of a pairwise alignment of sequence pair by using shuffle spaces specific to both the sequences.

5. Chapter 5: Pairwise statistical significance of local sequence alignment using sequence-

specific and position-specific substitution matrices

This chapter furthers the approach of using sequence-specific pairwise sequence alignment strategies for pairwise statistical significance estimation by using sequence-specific and position-specific substitution matrices. One possible approach to construct sequence-specific substitution matrices is shown to improve retrieval accuracy results as compared to using standard substitution matrices. Using position-specific substitution matrices for pairwise statistical significance estimation further improves the results significantly.

6. Chapter 6: PSIBLAST_PairwiseStatSig: Reordering PSI-BLAST hits using pairwise statistical significance

   This chapter uses pairwise statistical significance to reorder the hits obtained by a BLAST/PSI-BLAST database search, thereby giving more accurate estimates of statistical significance after quickly filtering potentially unrelated sequences using a fast database search program.

7. Chapter 7: Fast pairwise statistical significance estimation using derived distribution points and database search heuristics

   This chapter proposes and implements two heuristics to speedup pairwise statistical significance estimation. The heuristics are designed taking advantage of the nature of pairwise statistical significance estimation, and are shown to give a speedup of more that 200 without significant loss in retrieval accuracy.

8. Chapter 8: Conclusions

   This final chapter presents the conclusions and some future ideas.

2. PAIRWISE STATISTICAL SIGNIFICANCE AND EMPIRICAL DETERMINATION OF EFFECTIVE GAP OPENING PENALTIES FOR PROTEIN LOCAL SEQUENCE ALIGNMENT

A paper published in International Journal of Computational Biology and Drug Design

Ankit Agrawal, Volker P. Brendel and Xiaoqiu Huang

Abstract

We evaluate various methods to estimate pairwise statistical significance of a pairwise local sequence alignment in terms of statistical significance accuracy and compare it with popular database search programs in terms of retrieval accuracy on a benchmark database. Results indicate that using pairwise statistical significance using standard substitution matrices is significantly better than database statistical significance reported by BLAST and PSI-BLAST, and that it is comparable and at times significantly better than SSEARCH. An application of pairwise statistical significance to empirically determine effective gap opening penalties for protein local sequence alignment using the widely used BLOSUM matrices is also presented.

Introduction

Sequence alignment is a very important and common application in the analysis of DNA and protein sequences [52, 12, 13]. The primary application of sequence alignment is homology detection, i.e., identifying sequences evolved from a common ancestor. Homology detection further forms the basis of many other bioinformatics applications for making various high level

A conference version of this paper appeared in the Proceedings of 4th Intl. Symposium on Bioinformatics Research and Applications, 2008 [1]

inferences about the sequences, like finding protein function, protein structure, deciphering evolutionary relationships, drug design, etc. There exist several programs for sequence alignment that use well known algorithms [63, 58] or their heuristic versions [13, 49, 51]. Recently, some enhancements in alignment program features have also become available [32, 31] using difference blocks and multiple scoring matrices, in an attempt to capture more biological features in the alignment algorithm.

Why Statistical Significance?

Usually, the sequence alignment programs report alignment scores for the alignments constructed, and related (homologous) sequences will have higher alignment scores. But the threshold alignment score T below which the two sequences can be considered unrelated depends on the probability distribution of alignment scores between random, unrelated sequences [42]. Therefore, the biological significance of a pairwise sequence alignment is gauged by the statistical significance rather than the alignment score alone. This means that if an alignment score has a low probability of occurring by chance, the alignment is considered statistically significant, and hence biologically significant. The alignment score distribution depends on various factors like alignment program, scoring scheme, sequence lengths, sequence compositions [42]. It is thus possible to have two alignment scores x and y with $x < y$, but x more statistically significant than y. Therefore, instead of simply using the alignment score as the metric for homology, it is very useful to estimate the statistical significance of an alignment score to comment on the relatedness of the two sequences being aligned. Of course, it is important to note here that although statistical significance may be a good preliminary indicator of biological significance which may be helpful in identifying potential homologs, statistical significance does not necessarily imply biological significance [9, 42].

For ungapped alignments, rigorous statistical theory for the alignment score distribution is available [34]. However, no precise statistical theory currently exists for the gapped alignment score distribution and for score distributions from alignment programs using additional features like difference blocks [32] or multiple parameter sets [31]. The problem of accurately

determining the statistical significance of gapped sequence alignment has attracted a lot of attention in the recent years [65, 11, 50, 43, 41, 18, 10, 57, 59, 54, 68]. There exist a couple of good starting points for statistically describing gapped alignment score distributions for simple scoring schemes [36, 25], but a complete mathematical description of the optimal score distribution remains far from reach [25]. Some excellent reviews on statistical significance in sequence comparison are available in the literature [48, 53, 42, 40].

Database statistical significance

The commonly available pairwise protein local sequence alignment programs give the optimal or suboptimal alignment of two given sequences. Database searches are a special case of pairwise local sequence alignment, where one sequence is the query sequence, and the second sequence is a database consisting of many component sequences. Many approaches exist currently to estimate the statistical significance of a database hit (match of the query sequence with a sub-sequence of the database). For database searches, the statistical significance of a pairwise alignment score is reported in terms of the E-value, which is the expected number of hits in the database with a score equal to or higher than 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 for a database hit are corresponding to the database, and generally not for the specific pairwise alignments.

BLAST2.0 [13] 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 [49, 50] 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.

There have been a lot of improvements to the BLAST programs in the last few years, primary of which is the use of composition-based statistics and substitution matrices [57, 67, 68]. These methods have resulted in an increase in accuracy of database searches by rescaling the substitution matrices for individual alignments [57] and combining different measures of similarity by deriving pairwise statistical significance values from database statistical significance values [68]. These methods have intelligently avoided time-consuming simulations of individual sequence pairs by using pre-computed statistical parameters and substitution matrix rescaling techniques, which may not always be able to estimate the true pairwise statistical significance as it is derived from database statistical significance without generating the score distribution for individual sequence pairs, but generally give good results in context of a database search.

Pairwise statistical significance

Pairwise statistical significance is the statistical significance of the specific pairwise alignment under consideration and is independent of any database. 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 combination of alignment parameters - like the alignment program itself, substitution matrices, gap costs. In addition to the standard local alignment programs [63, 58], some recent programs have been developed [32, 31] that take into account other desirable biological features in addition to gaps - like difference blocks or 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 [42]. 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).

There exist quite a few approaches to estimate pairwise statistical significance of a pairwise alignment. The PRSS program in the FASTA package [49, 51, 50] 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 [21]. It also uses maximum likelihood fitting [22] and also allows for censoring of data left of a given cutoff, for fitting only the right tail of the histogram. In addition to maximum likelihood fitting, linear regression has also been used [31] to fit score distributions to estimate statistical parameters, and hence statistical significance. A heuristic approximation of the gapped local alignment score distribution is also available [43], 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 [41]. 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 [32, 31].

Contributions

There contribution of this paper is three-fold. First, we compare various existing methods for statistical significance accuracy and determine the most accurate method to be maximum likelihood fitting of score distribution censored left of peak (fitting right of peak). Secondly,

we compared this method with database statistical significance reported by common database search programs like BLAST, PSI-BLAST, and SSEARCH in terms of retrieval accuracy (of homologs) using a benchmark database earlier used in [62]. Comparison of pairwise statistical significance results with database statistical significance show that pairwise statistical significance gives significantly better retrieval accuracy compared to BLAST and PSI-BLAST, and comparable and at times significantly better accuracy than SSEARCH as well. BLAST and PSI-BLAST are heuristic based methods for database search whereas SSEARCH uses a full implementation of Smith-Waterman algorithm [63], and hence takes much more time. Comparable and at times significantly better retrieval accuracy than SSEARCH makes it feasible to get statistical significance estimates at least as good as database statistical significance without doing a time-consuming database search. Thirdly, we use pairwise statistical significance to empirically determine the effective gap opening penalties under an affine gap penalty model for pairwise protein comparison with the most commonly used BLOSUM substitution matrices [28] - BLOSUM45, BLOSUM50, BLOSUM62 and BLOSUM80, using the same benchmark database. A similar empirical study for database searches using SSEARCH with PAM matrices [20] is available in the literature [55]. The first two contributions of this paper with preliminary results were earlier presented in the conference version of this paper [1].

The remainder of the paper is organized as follows: In Section 2, an introduction to the extreme value distribution in the context of estimating statistical significance for gapped and ungapped alignments is presented. Section 3 describes the existing tools and programs used in this work, followed by the experiments and results in Section 4, which contains the main contributions of this paper. Finally, the conclusion and future work is presented in Section 5.

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) [35]. This is an important and useful fact, because in principle 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 [34]. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP (High-scoring Segment Pairs which correspond to 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:

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

where E is the E-value and is given by

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

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 [34], and the parameters K and λ can be computed analytically from the substitution scores and sequence compositions. 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 [36, 25]. Recently, researchers have also looked closely at the low probability tail distribution, and the work in [66] applied a rare-event sampling technique earlier used in [27] 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 [65, 11, 50, 41, 46, 44, 31].

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

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

It is clear that in theory, this normalized E-value is same as the P-value (for large N).

Tools and Programs Used

We worked with the alignment programs SIM [33], which is an ordinary alignment program (similar to SSEARCH), GAP3 [32], which allows dynamically finding similarity blocks and difference blocks, and GAP4 [31], 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 [49, 51, 50], 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 [41], 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 described in [31] to estimate K and λ from an empirical distribution of alignment scores. Finally, we also used the maximum likelihood method [22] and corresponding routines in the HMMER package [21] to fit an EVD to the empirical distribution.

Experiments and Results

Accurate estimation of K and λ for 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.

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 the second iteration 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. 2.1 shows the mean score distribution (complementary cumulative 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. Because 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 [22], and type-II censoring is defined as one where the cutoff is set to the score that corre-

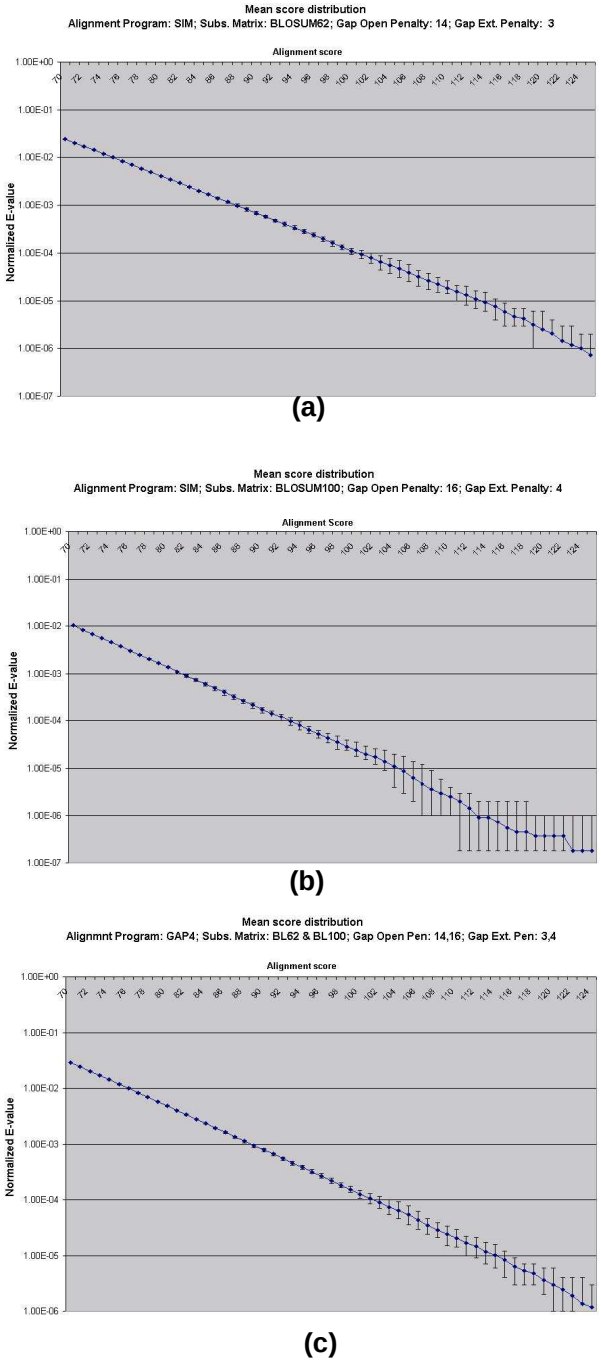

Figure 2.1 Distribution of alignment scores generated (a) using SIM program and BLOSUM62 matrix, (b) using SIM program and BLOSUM100 matrix and (c) using GAP4 program and BLOSUM62 and BLOSUM100 matrices. The solid line curve represents the mean of the eleven distribu1tions generated, and the vertical bars represent the variation within the eleven distributions.

Table 2.1 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. Maximum likelihood fitting with type-I censoring (censoring left of peak) gives the minimum SSD in most comparisons.

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**

sponds 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 2.1 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. Because we had eleven estimates of K and λ for the ML and LR methods, we report the minimum, maximum and average SSD for these methods. PRSS and ARIADNE report

one set of parameters, and thus there is 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.

Pairwise statistical significance versus database statistical significance for homology detection

More important than statistical significance accuracy of alignment scores is their retrieval accuracy for homology detection since it is the primary application of sequence alignment. To evaluate pairwise statistical significance and compare it with database statistical significance, we used a typical homology detection experiment setup as follows. We used a non-redundant subset of the CATH 2.3 database (Class, Architecture, Topology, and Hierarchy, [47]) provided by [62] and available at ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/prot_sci_04/, which was earlier selected in [62] to evaluate seven structure comparison programs and two sequence comparison programs. As described in [62], 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. This domain set is considered as a valid benchmark for testing protein comparison algorithms [56].

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 1000 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 implementing the Smith-Waterman algorithm.

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 [62] to report the results using the SSEARCH program. The parameters K and λ resulting from each censored ML fitting were then used to find the pairwise statistical significance of the corresponding pairwise comparison, and the P-value was recorded. Following [62], Errors per Query (EPQ) versus Coverage plots were used to present the results. To create these plots, the list of pairwise comparisons was sorted based on statistical significance, and subsequently, the lists were examined, from best score to worst. Going down the list, the count of true homologs detected 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, errors per query (EPQ) is the total number of errors incurred so far, divided by the number of queries; and coverage is the fraction of homolog pairs so far detected. The ideal situation would be to go from 0% to 100% coverage, without incurring any errors, which would correspond to the curve being a straight line on the x-axis. Therefore, the more the curve is towards the right, the better it is.

We compare the performance of pairwise statistical significance with database statistical significance reported by popular database search programs like BLAST, PSI-BLAST, and SSEARCH in terms of retrieval accuracy. BLAST and FASTA are heuristic based database search approaches and SSEARCH is a rigorous database search program using full implementation of Smith-Waterman algorithm [63]. PSI-BLAST is an iterative approach to BLAST, where position-specific scoring matrices (PSSMs) are constructed and refined over multiple iterations. The performance of SSEARCH is significantly better than BLAST and FASTA at the cost of search time. PSI-BLAST results depend heavily on the quality of the PSSMs but usually are significantly better, because of its use of position-specific scoring matrices constructed and refined over multiple iterations of BLAST. We performed experiments with PSI-BLAST both using the benchmark database used in our experiments and using good-quality pre-trained PSSMs constructed against non-redundant protein database.

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) [62], we examine the performance of the methods with individual queries, following the work in [62]. The coverage of each of the 86 queries at the 1st, 3rd, 10th, 30th, and 100th error was recorded, and percentile analysis was done for each error level across the 86 queries, which is presented in Fig. 2.2. Fig. 2.2(a) shows the 25th percentile coverage level at the 1st, 3rd, 10th, 30th, and 100th false positive for homologs (i.e. 21 of the queries have worse coverage, and 65 have better coverage). Fig. 2.2(b) shows the same results for 50th percentile of coverage, i.e. the median coverage (43 queries performed better, 43 worse), and Fig. 2.2(c) shows the same results for 75th percentile of coverage (i.e. 65 of the queries have worse coverage, and 21 have better coverage). The curves for SSEARCH in Fig. 2.2(a) and Fig. 2.2(b) are derived from the figures 2A and 2B in [62]. The results for SSEARCH corresponding to Fig. 2.2(c) were not available in [62].

Because the experiments with BLAST, SSEARCH, and PairwiseStatSig were conducted using a standard substitution matrix, the results indicate that pairwise statistical significance performs significantly better than database statistical significance with heuristic based database search approaches (like BLAST and FASTA), and at least comparable to database statistical significance with rigorous database search approach like SSEARCH. This implies that statistical significance estimates at least as good as database statistical significance can be obtained by pairwise statistical significance without having to do a time-consuming database search. This can be very useful to estimate accurate pairwise statistical significance of two (or a few) sequences, which is a common scenario in many pairwise alignment based applications like phylogenetic tree construction, progressive multiple sequence alignment.

The results further indicate that on the benchmark database used in our experiments, pairwise statistical significance also gives better results than PSI-BLAST, which uses position-specific scoring matrices, even though the experiments with pairwise statistical significance were conducted using standard substitution matrices. Since PSI-BLAST results heavily depend on the quality of PSSMs, we also conducted experiments with PSI-BLAST using pre-trained PSSMs against the non-redundant protein database provided along with the BLAST suite of

programs. The similar PSI-BLAST results with pre-trained PSSMs presented in the conference version of this paper [1] were taken from [62], but here we present the results obtained by repeating the experiments with the new version (2.2.17) of the PSI-BLAST program and the non-redundant database. As it is clear from the Errors per Query vs. Coverage plots, using PSI-BLAST with pre-trained PSSMs gives significantly better results than SSEARCH and pairwise statistical significance with standard substitution matrices, which is not surprising since pre-trained PSSMs use much more information than just a pair of sequences.

It is important to note here that the PairwiseStatSig program is not a database search program but a pairwise statistical significance estimation program, and its comparison with database search programs like BLAST, PSI-BLAST, and SSEARCH as presented in this paper is of their statistical significance estimation strategies.

Using pairwise statistical significance to evaluate alignment parameter combinations

Similar experiments as reported in the previous subsection can be used to evaluate and compare different parameter combinations for sequence alignment - like alignment program, substitution matrix, gap penalties. And therefore, it can be used to empirically determine the optimal value of a specific parameter, given other parameters. We conducted a series of experiments on the same benchmark database (a subset of CATH 2.3 database) with different alignment parameters to determine the effective gap opening penalties for the commonly used BLOSUM matrices - BLOSUM45, BLOSUM50, BLOSUM62 and BLOSUM80 (in 1/3 bit units). Clearly, both the extreme cases of very low gap penalty and very high gap penalty (corresponding to gapless alignment) should give poor coverage. Therefore, if we plot coverage vs. gap opening penalty curves for any specific substitution matrix, it should be able to give us the range of best gap opening penalties, on either side of which the coverage decreases. A related study was done earlier [55] to determine effective gap penalties for database search application with SSEARCH using PAM matrices. Here we attempt to do a similar analysis for the application of pairwise sequence alignment using four of the commonly used BLOSUM

matrices.

Fig. 2.3(a), 2.3(b), 2.4(a) and 2.4(b) show the coverage vs. gap open penalty curves at different errors per query for the four substitution matrices BLOSUM45, BLOSUM50, BLOSUM62 and BLOSUM100. All the pairwise comparisons in the experiments were conducted using the PairwiseStatSig program with 1000 random shuffles. Apart from the variable alignment parameters (substitution matrix, gap open penalties), the other important parameter, the gap extension penalty, was set to 2, which is the default in FASTA and SSEARCH programs. Although the number of shuffles used is not very large, due to which the curves are quite noisy, they still clearly show a concave downward behavior as expected. The coverage is poor with both the extremes of gap opening penalties - too low and too high. The curves suggest that at least on the benchmark database that we used, for each substitution matrix there exists a small range of gap opening penalties $[g_l, g_h]$ within which the coverage is nearly constant, below which the coverage is very poor, and above which the coverage gradually decreases to the minimum coverage at infinite gap penalty, corresponding to the gapless alignment. We report the g_l and g_h values for each of the four substitution matrices under consideration. Let the best gap opening penalty for a given matrix be g_m. Based on the above graphs we also choose a $g_m \in [g_l, g_h]$ for each matrix which seems to give the best aggregate coverage and should be used for pairwise alignment with the corresponding matrix. It is important to emphasize here that the values for g_m reported in this paper have been empirically determined by performing experiments with the subset of CATH 2.3 database, and may not be the best values universally. But since this database is a valid benchmark for protein comparison, we believe that the reported values should hold good for many pairwise sequence alignment applications.

Table 3.1 lists the range of effective gap opening penalties determined for pairwise protein sequence alignment by the above experiments. Because PAM and BLOSUM matrices can be related based on their relative entropy [28], we also list the effective gap opening penalties for the corresponding PAM matrices as obtained from [55]. [55] gave a formula for effective gap penalties for PAM matrices for distances 20 to 200, which is not valid for distances greater than 200, and hence the first entry for PAM250 is not available. Although the effective gap opening

penalties for pairwise alignments determined in this work are close to those reported in [55] for database search application, for some substitution matrices the values are quite different. In addition to different applications (pairwise sequence alignment in this work and database search in [55]), the difference in the results may also be because of different databases used for the experiments.

Table 2.2 Effective gap opening penalties for commonly used BLOSUM matrices determined on a benchmark database.

BLOSUM matrix	$[g_l, g_h]$	$g_m \in [g_l, g_h]$	**Equivalent PAM matrix**	g_m **from Reese et.al,'02**
BLOSUM45	[6,12]	7	PAM250	NA
BLOSUM50	[6,12]	9	PAM200	5
BLOSUM62	[8,14]	11	PAM160	9
BLOSUM80	[10,17]	13	PAM120	13

Running Time Analysis

The time required to estimate pairwise statistical significance for a given pair of sequences certainly depends on the number of random shuffles generated for constructing alignment score distribution and the length of the two sequences. We used the same value (1000) for the number of shuffles for this experiment as was used for the homology detection and effective gap opening penalty determination experiments. 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 six real sequences from the CATH 2.3 database of varying length - from 59 to 512, and estimated the pairwise statistical significance of each of them with other real sequences from the database for one hour. Fig. 2.5 shows the average time per comparison for each of the six sequences. As expected, the relation between length of sequence and running time is linear. All computations were done on an Intel processor 2.8GHz. It can be seen that PairwiseStatSig program can estimate pairwise statistical significance for two sequences in a few seconds. 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.

As mentioned earlier, PairwiseStatSig is not a database search program like SSEARCH

or PSI-BLAST, but is useful in quickly estimating pairwise statistical significance for two (or a few) sequences. Thus, it can be used in conjunction with a fast database search program like BLAST, where the top hits from BLAST can be further ranked according to the pairwise statistical significance estimates from PairwiseStatSig program.

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. The results show that pairwise statistical significance performs significantly better than database statistical significance using BLAST and PSI-BLAST, and comparable and at times significantly better than SSEARCH as well, but still the accuracy of retrieval results is better for PSI-BLAST when used with pre-trained PSSMs. Further, the program PairwiseStatSig was used in multiple homology detection experiments with several different alignment schemes, on a benchmark database (a subset of CATH 2.3 database) to determine the effective gap opening penalties for the commonly used substitution matrices BLOSUM45, BLOSUM50, BLOSUM62 and BLOSUM80.

Regarding the comparison of pairwise statistical significance and database statistical significance, significantly better performance than heuristic-based database search approaches like BLAST and PSI-BLAST, and comparable performance to rigorous database search approach (SSEARCH) indicates the clear advantage of pairwise statistical significance over database statistical significance. Using pairwise statistical significance with standard substitution matrices is shown to be better than database statistical significance using both standard substitution matrices (BLAST) and query-specific substitution matrices (PSI-BLAST), and thus, we believe that the results of pairwise statistical significance can be further improved by using sequence specific substitution matrices, which is a significant part of our future work. Another important 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 would be very useful. Another aspect of future work is to experiment with other sample spaces for shuffling of protein sequences for generating score distribution, which may provide better significance estimates.

As mentioned in the paper, the PairwiseStatSig program can be used to test and validate different combinations of alignment parameters - alignment program, substitution matrix, gap penalties, and/or any other parameters that the alignment program may use. In this work, we have only experimented with varying the gap opening penalties for common substitution matrices. This can be further extended to determine other optimal parameters for pairwise alignment. Further, the PairwiseStatSig program can be used for pairwise sequence alignment based applications like multiple sequence alignment and phylogenetic tree construction. It can also be used in conjunction with fast database search programs like BLAST to refine the results.

Acknowledgments.

The authors would like to thank Dr. Sean Eddy for making the maximum likelihood fitting routines available online, and Dr. W. R. Pearson for creating and making available online the benchmark dataset for protein comparison used in this work.

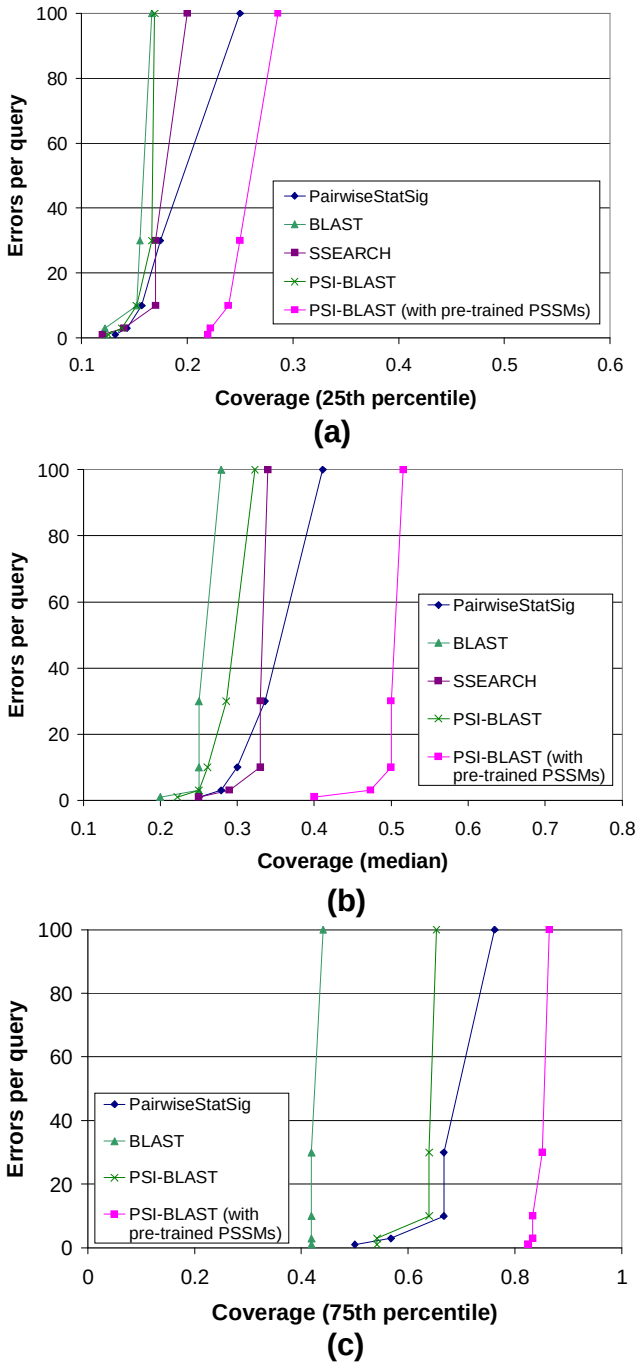

Figure 2.2 Errors per Query vs. Coverage plots at individual query level. (a) The 25th percentile coverage level for 86 queries; (b) 50th percentile (median) coverage level; (c) 75th percentile coverage level. PairwiseStatSig performs significantly better than BLAST and PSI-BLAST, and comparable and at times significantly better than SSEARCH as well, but poorer than PSI-BLAST using pre-trained PSSMs.

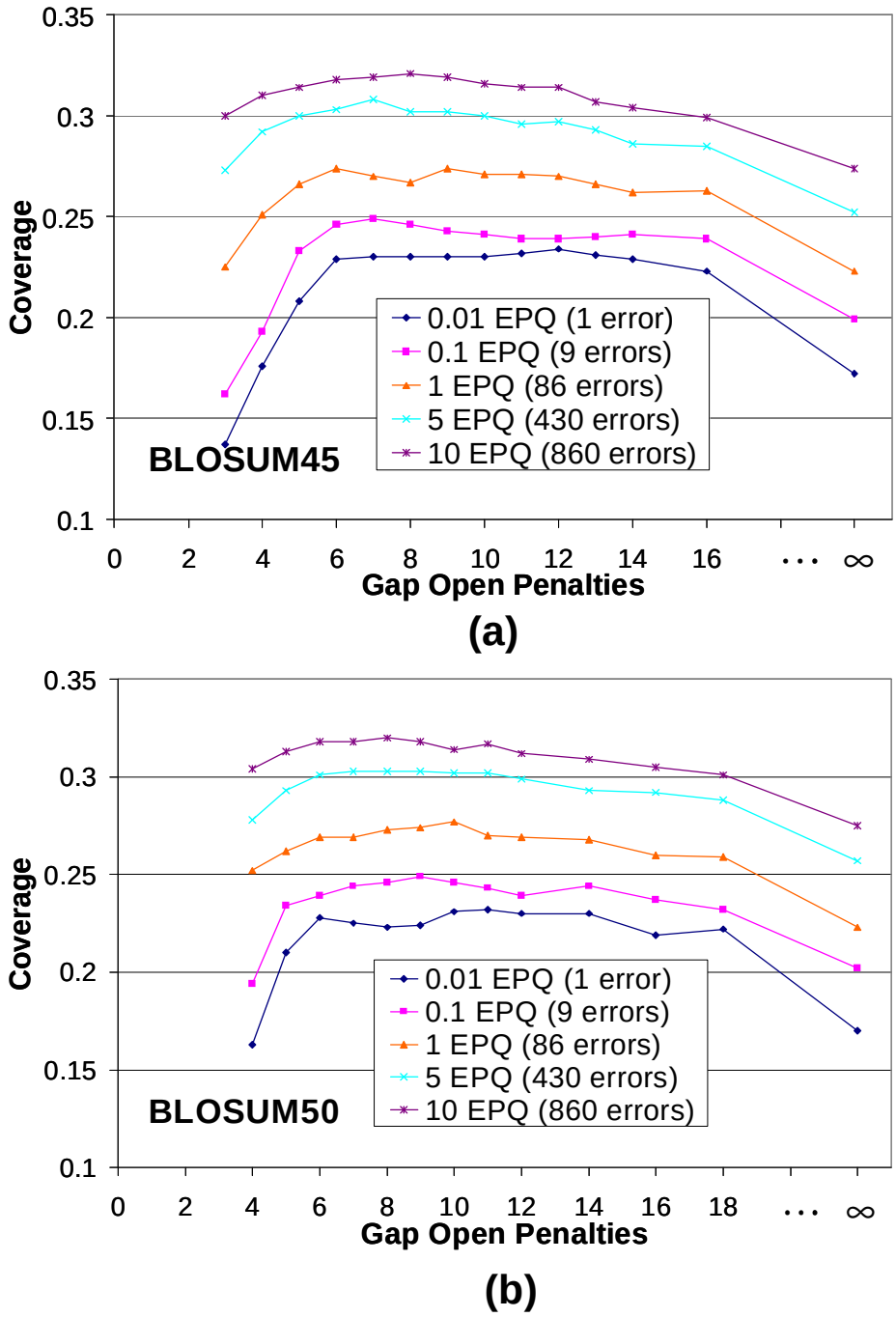

Figure 2.3 Coverage vs. Gap Opening Penalty plots using PairwiseStatSig at different errors per query for BLOSUM matrices: (a) BLOSUM45, (b) BLOSUM50. The curves show the expected concave downward behavior, with poor coverage at very low and very high gap opening penalty values. These graphs can be used to determine the range of best gap opening penalties for each substitution matrix.

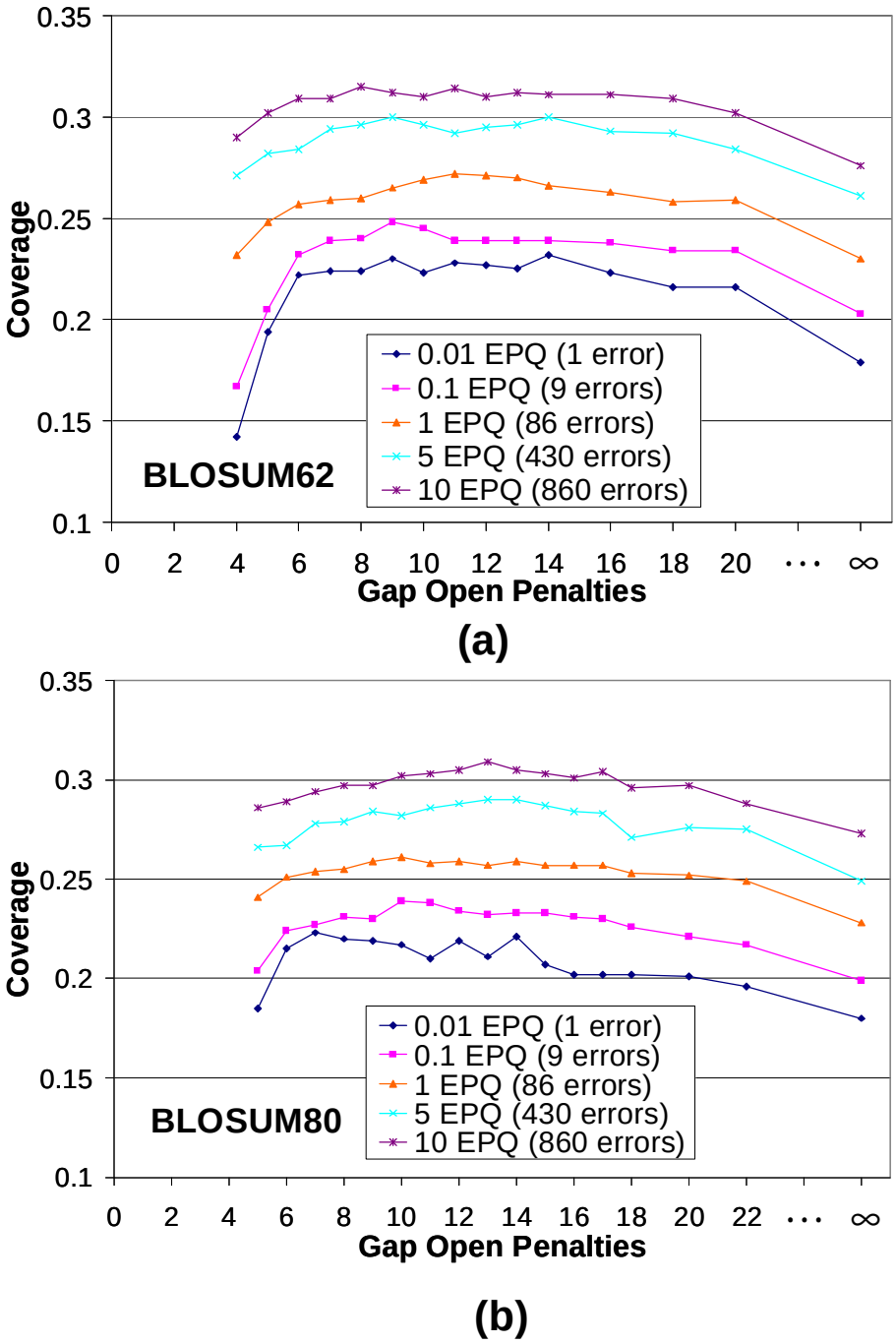

Figure 2.4 Coverage vs. Gap Opening Penalty plots using PairwiseStatSig at different errors per query for BLOSUM matrices: (a) BLOSUM62, (b) BLOSUM80. The curves show the expected concave downward behavior, with poor coverage at very low and very high gap opening penalty values. These graphs can be used to determine the range of best gap opening penalties for each substitution matrix.

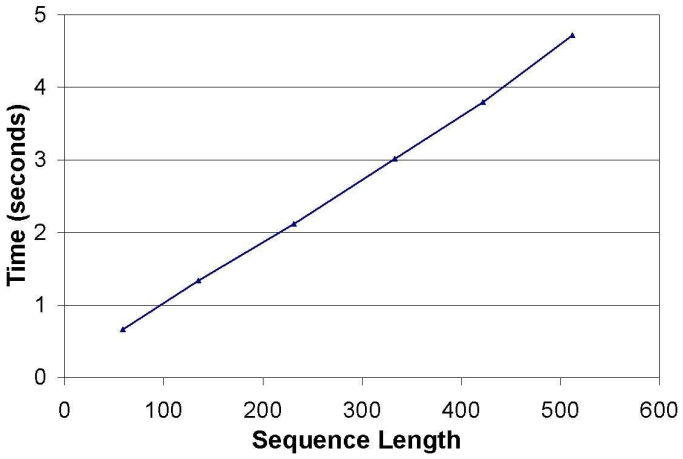

Figure 2.5 Time per comparison vs. Sequence length plot using PairwiseStatSig program with sequences of different length.

3. PAIRWISE STATISTICAL SIGNIFICANCE OF LOCAL SEQUENCE ALIGNMENT USING MULTIPLE PARAMETER SETS AND EMPIRICAL JUSTIFICATION OF PARAMETER SET CHANGE PENALTY

A paper published in BMC Bioinformatics

Ankit Agrawal and Xiaoqiu Huang

Abstract

Background: Accurate estimation of statistical significance of a pairwise alignment is an important problem in sequence comparison. Recently, a comparative study of pairwise statistical significance with database statistical significance was conducted. In this paper, we extend the earlier work on pairwise statistical significance by incorporating with it the use of multiple parameter sets.

Results: Results for a knowledge discovery application of homology detection reveal that using multiple parameter sets for pairwise statistical significance estimates gives better coverage than using a single parameter set, at least at some error levels. Further, the results of pairwise statistical significance using multiple parameter sets are shown to be significantly better than database statistical significance estimates reported by BLAST and PSI-BLAST, and comparable and at times significantly better than SSEARCH. Using non-zero parameter set change penalty values give better performance than zero penalty.

Conclusions: The fact that the homology detection performance does not degrade when using multiple parameter sets is a strong evidence for the validity of the assumption that the alignment score distribution follows an extreme value distribution even when using multiple parameter sets. Parameter set change penalty is a useful parameter for alignment using multiple parameter sets. Pairwise statistical significance using multiple parameter sets can be effectively used to determine the relatedness of a (or a few) pair(s) of sequences without performing a time-consuming database search.

Background

Local sequence alignment plays a major role in the analysis of DNA and protein sequences [52, 12, 13]. It is the basic step of many other applications like detecting homology, finding protein structure and function, deciphering evolutionary relationships, etc. There exist several local sequence alignment programs that use well-known algorithms [63, 58] or their heuristic versions [13, 49, 51]. Database search is a special case of pairwise local sequence alignment where the second sequence is a database in itself consisting of many sequences. Recently, there have been many enhancements in alignment program features [32, 31] using difference blocks and multiple scoring matrices, in an attempt to incorporate more biological features in the alignment algorithm.

Why statistical significance?

The local sequence alignment programs report alignment scores for the alignments constructed, and related (homologous) sequences will have *higher* alignment scores. But the definition of *high* depends strongly on the alignment score distribution, which gives importance to the concept of statistical significance. 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 [42], it implies that 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,

instead of using the alignment score for detecting homology, the statistical significance of an alignment score is more widely accepted as a metric to comment on the relatedness of the two sequences being aligned. Of course, it is important to emphasize here that although statistical significance is a good preliminary indicator of biological significance, it does not necessarily imply biological significance [9, 42].

Accurate statistical theory for the ungapped alignment score distribution is available [34]. However, no precise statistical theory currently exists for the gapped alignment score distribution and for score distributions from alignment programs using additional features. Accurate estimation of statistical significance of gapped sequence alignment scores has attracted a lot of attention in the recent years [65, 11, 50, 43, 41, 10, 57, 59, 68]. Although there exists some understanding of the statistics of gapped alignment score distributions for simple scoring schemes [36, 25], but a complete mathematical description of the optimal score distribution remains far from reach [25]. There exist some excellent reviews on statistical significance in sequence comparison in the literature [48, 53, 42, 40].

Database statistical significance versus pairwise statistical significance

Recently, a study of pairwise statistical significance and its comparison with database statistical significance [1] was conducted. In summary, the database statistical significance reported by most database search programs like SSEARCH, FASTA, PSI-BLAST is database-dependent, and hence, the same alignment of two sequences with the same alignment score can be evaluated as having different significance values in database searches with different databases, and even with the same database at different times, as the database size can be variable. On the other hand, pairwise statistical significance is specific to the sequence pair being aligned, and is database-independent. 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, this method was compared with database statistical significance in a homology detection experiment to find that pairwise statistical significance performs comparably to and sometimes

significantly better than database statistical significance.

Accurate statistical significance estimates for 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, it can also be used to compare different combination of alignment parameters - like the alignment program itself, substitution matrices, gap costs. A comparison of different gap opening penalties for four commonly used BLOSUM matrices using pairwise statistical significance was presented in [2]. In addition to the standard local alignment algorithms [63, 58], some recent algorithms have been developed [32, 31] that take into account other desirable biological features in addition to gaps - like difference blocks or the use of multiple parameter sets (substitution matrices, gap penalties). These features of the alignment programs enhance the sequence alignment of real sequences by better suiting to different conservation rates at different spatial locations of the sequences. As pointed out earlier, accurate 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 can be very useful. Further, quick and accurate estimates of pairwise statistical significance can also be helpful for applications like multiple sequence alignment and phylogenetic tree construction which are based on pairwise sequence alignment to select most related pairs of sequences, for example, in a progressive multiple sequence alignment.

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) [35]. This is an important and useful fact, because in principle 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 [34]. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP (High-scoring Segment Pairs which correspond to ungapped local alignments) scores are characterized by two parameters, K and λ. The probability (P-value) that the optimal local alignment score S exceeds x is estimated by:

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

where E is the E-value and is given by

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

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 [34], and the parameters K and λ can be computed analytically from the substitution scores and sequence compositions. For gapped alignments, no perfect statistical theory has yet been developed, although there exist some good starting points for the problem as mentioned before [36, 25]. Recently, researchers have also looked closely at the low probability tail distribution, and the work in [66] applied a rare-event sampling technique earlier used in [27] 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 [44, 65, 11, 50, 41, 46, 31, 1].

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

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

In theory, this normalized E-value is same as the P-value (for large N).

Contributions

In this paper, we extend the existing work on pairwise statistical significance [1] to incorporate in it the use of multiple parameter sets, and evaluate it on an important knowledge discovery application - homology detection. We conducted similar experiments as reported in [62], and later in [1] on a subset of the CATH 2.3 database to compare pairwise statistical significance with single and multiple parameter sets. This benchmark database was earlier created in [62] to evaluate seven protein structure comparison methods and two sequence comparison programs: SSEARCH and PSI-BLAST. SSEARCH uses the original Smith-Waterman algorithm [63], and is considered the most sensitive algorithm in terms of retrieval accuracy, better than the heuristic versions like BLAST and FASTA [15, 17]. PSI-BLAST is a modification to the BLAST program, where position specific scoring matrices are constructed over multiple iterations of BLAST algorithm. Comparison of pairwise statistical significance results using multiple parameter sets with pairwise statistical significance using a single parameter set shows that at least for some error levels, using multiple parameter sets is significantly better than using a single parameter set. This is because sequences can have different conservation rates at different spatial locations, which can be better aligned using multiple parameter sets (substitution matrices, gap penalties, etc.). Comparison with database statistical significance results show that pairwise statistical significance with multiple parameter sets gives significantly better performance than the statistical significance estimates reported by BLAST and PSI-BLAST, and comparable and at times significantly better performance than the SSEARCH program. Further, the results also give concrete evidence that for the practical application of homology detection, the score distribution from alignment program using multiple parameter sets can also be assumed to follow an extreme value distribution. This is an important and useful finding since it is in general difficult to accurately determine statistics of alignment scores from enhanced alignment programs. Finally, experiments with different values of parameter set change penalties indicate that it is indeed important to use a non-zero parameter set change penalty while performing alignment using multiple parameter sets.

Methods

Pairwise statistical significance estimation

Consider the pairwise statistical significance described 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 (substitution matrix, gap penalties), 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 the 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 in the P-value formula. More details on pairwise statistical significance can be found in [1]. In this paper, we dynamically use multiple parameter sets instead of a single scoring scheme SC for estimation of pairwise statistical significance.

Dynamic use of multiple parameter sets in sequence alignment

Usually, pairwise sequence alignment is done with a single parameter set (substitution matrix, gap penalties). But to suit the different levels of conservation between sequences, there exists an algorithm [31] which can dynamically use multiple parameter sets and generate a single optimal alignment with possibly different parameter sets used in different regions of the alignment. The algorithm is implemented in a program named GAP4. The algorithm uses a dynamic programming approach as explained in [31]. Consider alignment of two sequences $A = a_1, a_2, ..., a_m$ and $B = b_1, b_2, ...b_n$ using p parameter sets $P_1, P_2, ..., P_p$. Let A_i and B_j be the subsequences $a_1, a_2, ...a_i$ and $b_1, b_2, ..., b_j$ respectively. For each alignment position (i, j) and each parameter set P_k, the algorithm keeps track of the optimal alignment score of the subsequences A_i and B_j where the last component (substitution, gap, or difference block) is scored using P_k. Dynamic programming is used to get optimal alignment for progressive alignment positions, until i becomes m and j becomes n. Appropriate modification of the algorithm also allows it to calculate the optimal local alignment. More details about using

multiple parameter sets for pairwise sequence alignment can be found in [31].

Evaluation methodology

To evaluate the performance of pairwise statistical significance using multiple parameter sets, we used a non-redundant subset of the CATH 2.3 database (Class, Architecture, Topology, and Hierarchy, [47]) provided by [62] and available at ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/ prot_sci_04/. This database was selected in [62] to evaluate seven structure comparison programs and two sequence comparison programs. As described in [62], this dataset consists of 2771 domain sequences and includes 86 selected test query sequences. This domain set is considered as a valid benchmark for testing protein comparison algorithms [56].

We used this database and query set for experimenting with pairwise statistical significance using multiple parameter sets. For each of the 86×2771 comparisons, we used the maximum likelihood method with censoring left of peak with 1000 shuffles to fit the score distribution from the GAP4 program with substitution matrices BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, and their all possible combinations ($2^4 - 1 = 15$ in number). All matrices were used in 1/3 bit scale. The gap opening penalties for each of these matrices was set to the values empirically determined to be the best for this database in [2]. These are listed in Table 3.1. The gap extension penalties were set to 2 for all the four matrices.

Table 3.1 Effective gap opening penalties for commonly used BLOSUM matrices determined for the benchmark database used

BLOSUM matrix	**Gap opening penalty**
BLOSUM45	7
BLOSUM50	9
BLOSUM62	11
BLOSUM80	13

Following [62, 1], Error per Query (EPQ) versus Coverage plots were used to present the results. To create these plots, the list of pairwise comparisons was sorted based on decreasing statistical significance (increasing P-values). Going down the list, the coverage count is increased by one if the two sequences of the pair are homologs, and the error count is increased

by one if they are not. At a given point in the list, Errors Per Query (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. In the ideal case, the curve would go from 0% to 100% coverage, without incurring any errors, which would correspond to a straight line on the x-axis. Therefore, the more the curve is towards the right, the better the curve is.

Just as gap opening and gap extension penalties are dynamically charged during the alignment process whenever a gap is inserted and extended respectively, the GAP4 [31] program allows the use of a parameter set change penalty, which is dynamically charged whenever the parameter set mapping is changed during the alignment process. To see the effect of parameter set change penalty on the coverage performance, we conducted a series of homology detection experiments with one of the substitution matrix combinations (BLOSUM45 and BLOSUM62) with different parameter set change penalties. Coverage vs. parameter set change penalty curves were plotted at different error levels to find the usefulness of the parameter set change penalty, as reported in the next section.

Results

Comparison with pairwise statistical significance using single parameter set

Out of the 15 substitution matrix combinations, 4 are using single parameter sets, 6 are using two parameter sets, 4 are using three parameter sets, and 1 is using all four parameter sets. The EPQ vs. Coverage curves using pairwise statistical significance with two, three, and four parameter sets are presented in Figures 3.1,3.2, and 3.3 respectively. For comparison purposes, the EPQ vs. Coverage curves using corresponding single parameter sets are also presented in the same figures. The y-axis (error-axis) in all these graphs is in log-scale, and hence there is more information in the upper part of the graphs. These figures suggest that pairwise statistical significance using multiple parameter sets performs comparably to and sometimes significantly better than pairwise statistical significance using a single parameter set for most instances of using a single parameter set, and at most error levels.

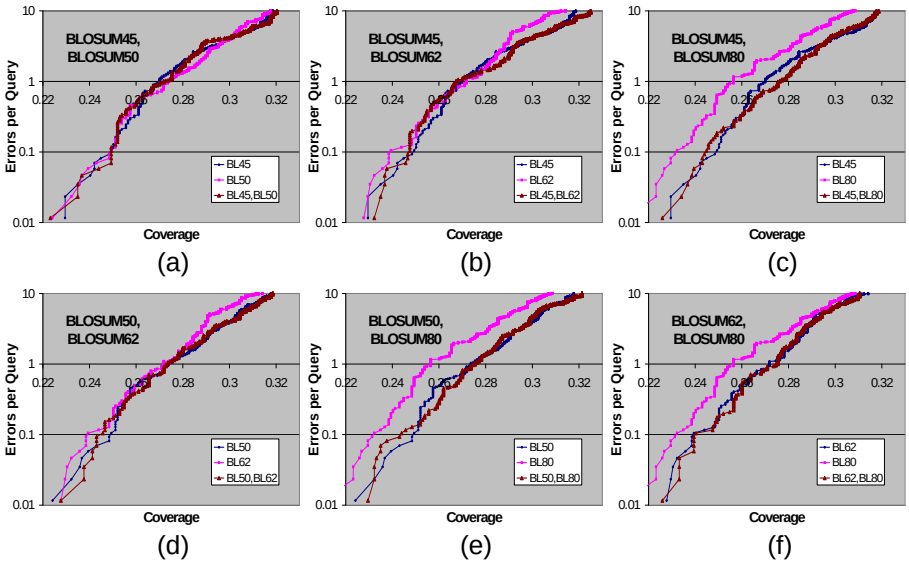

Figure 3.1 Pairwise statistical significance using two parameter sets. Errors per Query vs. Coverage plot for pairwise statistical significance using two parameter sets, along with the curves using corresponding single parameter sets. (a) BLOSUM45, BLOSUM50; (b) BLOSUM45, BLOSUM62; (c) BLOSUM45, BLOSUM80; (d) BLOSUM50, BLOSUM62; (e) BLOSUM50, BLOSUM80; (f) BLOSUM62, BLOSUM80. In 5 panels (b) through (f) out of 6, using two parameter sets leads to better coverage than using single parameter set at most error levels.

Comparison with database 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) [62], to compare the performance of pairwise statistical significance using multiple parameter sets with database statistical significance, we examined the performance of the methods with individual queries, following the work in [62]. 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 sequence comparison method to be evaluated. Fig. 4.3 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. 4.3 is derived from the Fig.

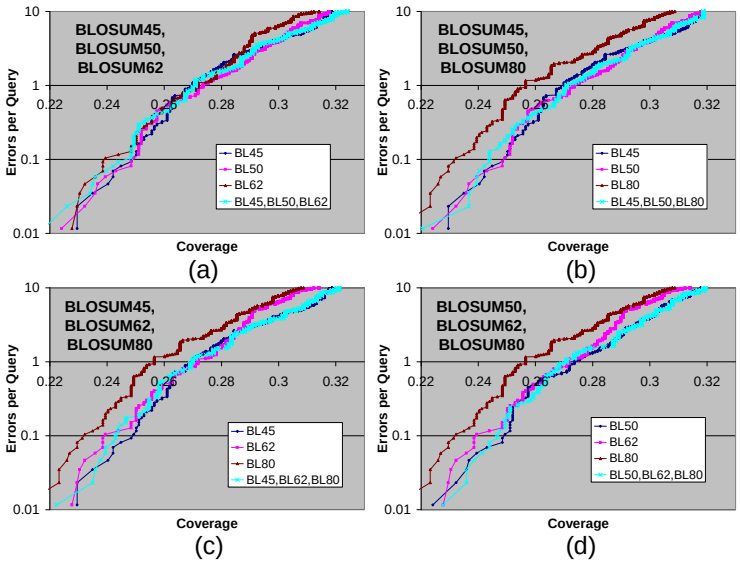

Figure 3.2 Pairwise statistical significance using three parameter sets. Errors per Query vs. Coverage plot for pairwise statistical significance using three parameter sets, along with the curves using corresponding single parameter sets. (a) BLOSUM45, BLOSUM50, BLOSUM62; (b) BLOSUM45, BLOSUM50, BLOSUM80; (c) BLOSUM45, BLOSUM62, BLOSUM80; (d) BLOSUM50, BLOSUM62, BLOSUM80. In all 4 panels, using three parameter sets leads to better coverage than using single parameter set at most error levels for at least two instances of using single parameter set.

2A in [62]. The curves for BLAST and PSI-BLAST were obtained by experimentation. It is clear that the proposed pairwise statistical significance using multiple parameter sets performs significantly better than BLAST and PSI-BLAST at all error levels, comparable to SSEARCH at low error levels, and significantly better than SSEARCH at higher error levels.

Empirical justification of parameter set change penalty

The coverage vs. parameter set change penalty plot for the substitution matrix combination of BLOSUM45 and BLOSUM62 is illustrated in Fig. 3.5. The curve shows a poor coverage performance for the case when the parameter set change penalty is not charged, i.e., when the alignment algorithm is freely allowed to change the parameter set during alignment without charging any penalty. This can be explained by the fact that the algorithm would try to

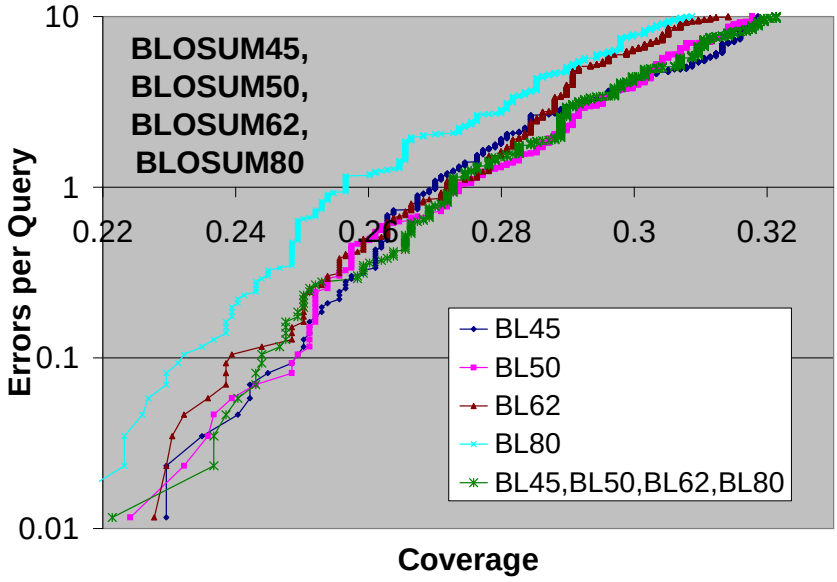

Figure 3.3 Pairwise statistical significance using four parameter sets. Errors per Query vs. Coverage plot for pairwise statistical significance using four parameter sets BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80 along with the curves using corresponding single parameter sets. Using four parameter sets results in better coverage than using single parameter set at most error levels for at least three instances of using single parameter set.

mathematically maximize the alignment score by changing the parameter set as frequently as possible, which may produce more biologically irrelevant alignments. A similar phenomenon is also observed when very low gap penalty is used [2]. The coverage performance clearly improves for non-zero values of parameter set change penalty, which provides its empirical justification.

Discussion

As pointed out earlier, SSEARCH employs the original Smith-Waterman algorithm for alignment, and is considered more sensitive than its heuristic implementations like BLAST and FASTA. PSI-BLAST uses an iterative approach with query-specific substitution matrices, and its performance mainly depends on the quality of position-specific scoring matrices

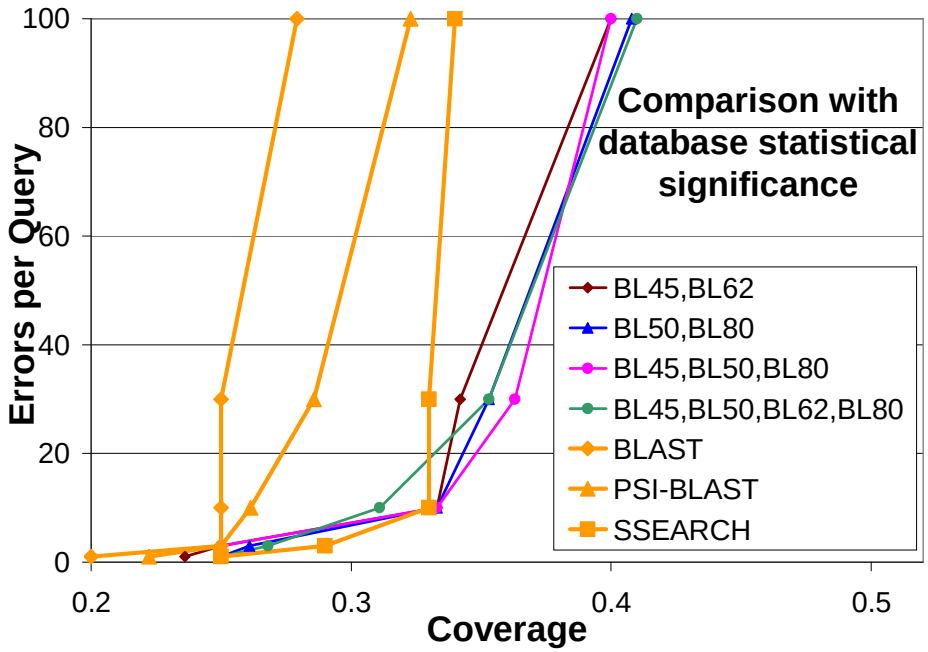

Figure 3.4 Comparison with database statistical significance. Comparison of pairwise statistical significance (using multiple parameter sets) and database statistical significance. All parameter combinations are significantly better than database statistical significance estimates reported by BLAST and PSI-BLAST at all error levels, and better than SSEARCH especially at higher error levels.

(PSSMs) constructed iteratively. The results show that PSI-BLAST gave poorer performance than pairwise statistical significance using multiple parameter sets, even with PSSMs constructed against the benchmark CATH database used in our experiments. However, using PSSMs derived against BLAST non-redundant protein database has been shown to give better results [62] as it uses much more information than just a pair of sequences. Comparable and at times significantly better results than SSEARCH using pairwise statistical significance with multiple parameter sets implies that statistical significance estimates at least as good as database statistical significance can be obtained by pairwise statistical significance using multiple parameter sets without having to do a time-consuming database search. This can be very useful to estimate accurate pairwise statistical significance of two (or a few) sequences, which is a common scenario in many pairwise alignment based applications like phylogenetic

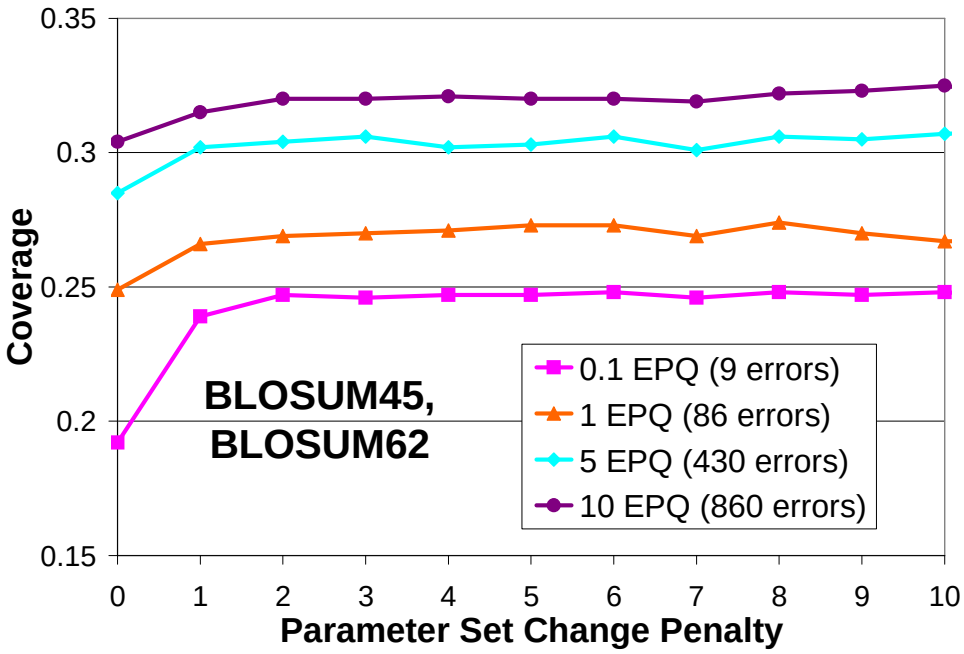

Figure 3.5 Empirical justification of parameter set change penalty. Coverage vs. Parameter Set Change Penalty plots at different errors per query for the substitution matrix combination of BLOSUM45 and BLOSUM62. Poor coverage is obtained if the parameter set change penalty is zero. The coverage is better and steady for non-zero values of parameter set change penalty.

tree construction, progressive multiple sequence alignment.

It is important to note that the proposed method is not a database search method but statistical significance estimation method for pairwise local alignment, and the comparison with database search programs like BLAST, SSEARCH, and PSI-BLAST is of their statistical significance estimation strategies. The proposed method, as of now is not scalable to a database search, but can be used to refine the results from a fast database search program like BLAST.

Since pairwise alignment using multiple parameter sets takes more computational time than using a single parameter set, pairwise statistical significance estimation using multiple parameter sets also takes more time than pairwise statistical significance estimation using a single parameter set. In general, using k parameter sets increases the computation time by a factor little more than k. Therefore, faster methods for significance estimation can be very helpful.

Conclusions

This paper extends the work on pairwise statistical significance by incorporating in it the use of multiple parameter sets (substitution matrices, gap penalties, etc.), and compares it with database statistical significance for the knowledge discovery application of homology detection. The results show that pairwise statistical significance using multiple parameter sets performs better than pairwise statistical significance using a single parameter set. It also performs significantly better than database statistical significance using BLAST and PSI-BLAST, and comparable and at times significantly better than database statistical significance using SSEARCH. Further, an empirical justification of the use of parameter set change penalty is provided.

Since PSI-BLAST results can be improved by using better quality PSSMs derived from larger protein databases, we believe that the performance of pairwise statistical significance can also be improved using sequence-specific/position-specific substitution matrices, which is a significant part of our future work. Another important contribution can be to estimate the pairwise statistical significance accurately in less time, since using multiple parameter sets increases the significance estimation time. Faster methods for determining pairwise statistical significance would be very useful.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

AA conceived the study based on the initial idea given by XH. Both AA and XH did the programming, AA carried out the experiments and analysis, and drafted the initial manuscript. Both AA and XH read and approved the final version of the manuscript.

Acknowledgements

The authors would like to thank Dr. Volker Brendel for helpful discussions and providing links to the data.

4. CONSERVATIVE, NON-CONSERVATIVE AND AVERAGE PAIRWISE STATISTICAL SIGNIFICANCE OF LOCAL SEQUENCE ALIGNMENT

A paper published in the Proceedings of IEEE International Conference on Bioinformatics and Biomedicine 2008

Ankit Agrawal and Xiaoqiu Huang

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.

Introduction

Statistical Significance of Sequence Alignment Scores

Sequence alignment is an underlying application in the comparison of DNA and protein sequences [13]. There exist programs for sequence alignment that use popular algorithms [63] or their heuristic versions [13, 51]. 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 [42], 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.

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.

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 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.

Experiments and Results

To evaluate the performance of the proposed significance measures, we used the experiment setup used earlier in [62], and subsequently in [1]. A non-redundant subset of the CATH 2.3

database (Class, Architecture, Topology, and Hierarchy, [47]) available at ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/prot_sci_04/ was selected in [62] 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 [62], 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. 4.1. 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. 4.1, the curve for original 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. 4.2 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 [5]. Fig. 4.2 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.

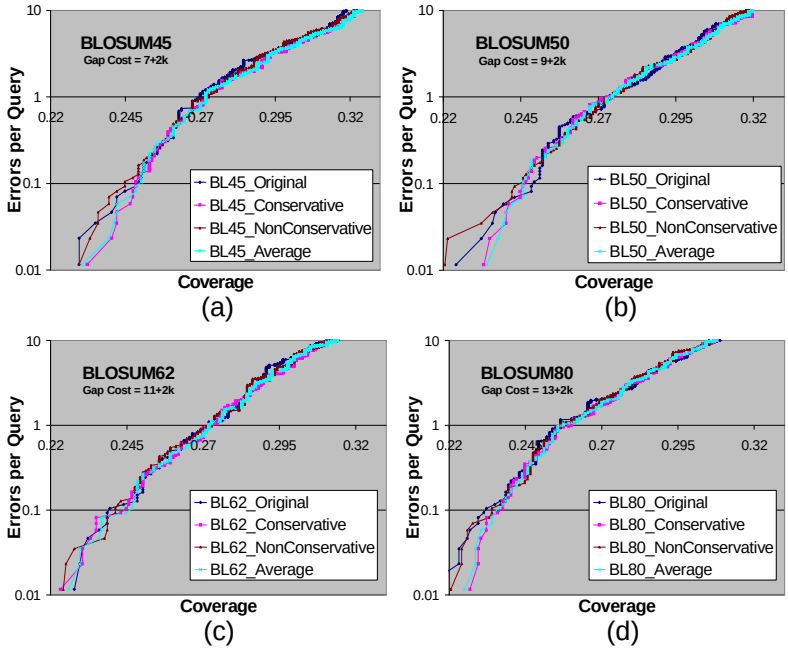

Figure 4.1 EPQ vs. Coverage plot for original, conservative, non-conservative, and average pairwise statistical significance using four substitution matrices. (a) BLOSUM45; (b) BLOSUM50; (c) BLOSUM62; (d) BLOSUM80. Although the curves are very close to each other, in all the four figures, the curve for original pairwise statistical significance is towards the left for most error levels.

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) [62], 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 [62]. 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. Fig. 4.3 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. 4.3 is derived from the figure 2A in [62]. All other curves were obtained by experimentation. The curves suggest that using the proposed variants gives significantly better

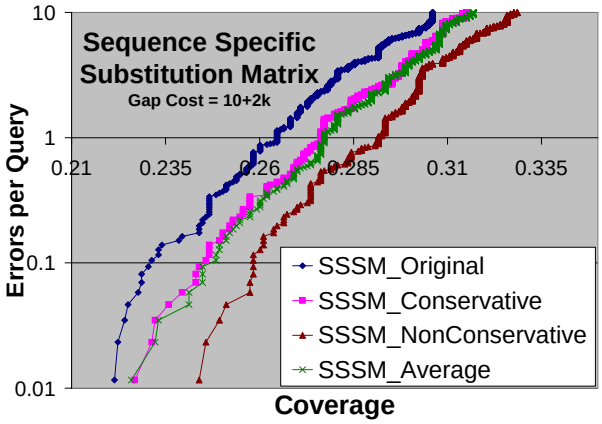

Figure 4.2 EPQ vs. Coverage plot for different kinds of pairwise statistical significance using sequence specific substitution matrices. Non–conservative pairwise statistical significance outperforms other variants of pairwise statistical significance. All the three variants proposed in this paper are better than original pairwise statistical significance.

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 [62], 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.

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

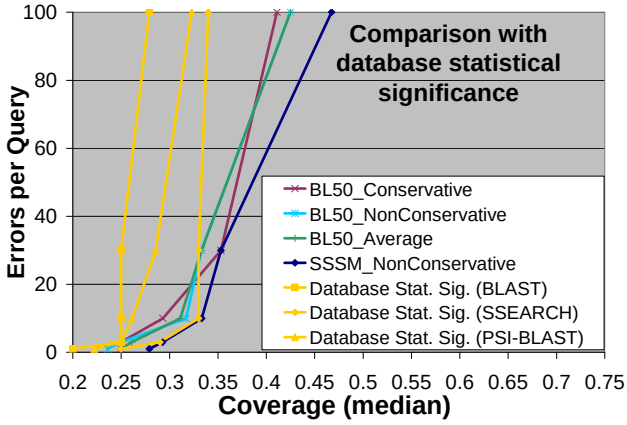

Figure 4.3 Comparison of proposed significance measures with database statistical significance using BLAST, PSI-BLAST and SSEARCH. The proposed measures are significantly better than BLAST and PSI-BLAST. With general substitution matrices, the performance of the proposed measures is significantly better 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.

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.

5. PAIRWISE STATISTICAL SIGNIFICANCE OF LOCAL SEQUENCE ALIGNMENT USING SEQUENCE-SPECIFIC AND POSITION-SPECIFIC SUBSTITUTION MATRICES

A paper submitted to IEEE Transactions on Computational Biology and Bioinformatics

Ankit Agrawal and Xiaoqiu Huang

Abstract

Pairwise sequence alignment is a central problem in bioinformatics which forms the basis of many other applications. Two related sequences are expected to have a high alignment score, but relatedness is usually judged by statistical significance rather than by alignment score. Recently, it was shown that pairwise statistical significance is better and quicker than database statistical significance for getting individual significance estimates of pairwise alignment scores. The improvement was mainly attributed to making the statistical significance estimation process sequence-specific and database-independent. In this paper, we use sequence-specific and position-specific substitution matrices to derive the estimates of pairwise statistical significance, which is expected to use more sequence-specific information in estimating pairwise statistical significance. Experiments with sequence-specific substitution matrices at different levels of sequence-specific contribution were conducted, and results confirm that using sequence-specific substitution matrices for estimating pairwise statistical significance is significantly better than using a standard matrix like BLOSUM62, and than database statistical significance estimates reported by popular database search programs like BLAST, PSI-BLAST and SSEARCH on a benchmark database, but PSI-BLAST results can be significantly improved by using pre-trained PSSMs. Further, using position-specific substitution matrices for estimating pairwise

statistical significance gives significantly better results even than PSI-BLAST using pre-trained PSSMs.

Introduction

Sequence alignment is an underlying application in the analysis and comparison of DNA and protein sequences [52, 12, 13]. Although a computational problem, its primary application in bioinformatics is homology detection, i.e., identifying sequences evolved from a common ancestor, generally known as homologs or related sequences. Homology detection further forms the key step of many other bioinformatics applications making various high level inferences about the DNA and protein sequences - like finding protein function, protein structure, deciphering evolutionary relationships, drug design, etc. There exist several programs for sequence alignment that use popular algorithms [63, 58] or their heuristic versions [49, 13, 51, 39, 38]. The heuristic implementations of sequence alignment are especially useful for database search application, where one sequence is the query sequence, and the other sequence is a database. A lot of enhancements in alignment program features are also available [19, 32, 31] using difference blocks and multiple scoring matrices, in an attempt to capture some more biological features in the alignment algorithm.

Why Statistical Significance?

Sequence alignment programs invariably report alignment scores for the alignments constructed, and related (homologous) sequences will have *higher* alignment scores. But the threshold score above which the score can be considered *high* depends on the alignment score distribution, and hence estimating statistical significance of an alignment score is very useful in sequence comparison. An alignment score is considered statistically significant if it has a low probability of occurring by chance. The alignment score distribution depends on various factors like alignment program, scoring scheme, sequence lengths, sequence compositions [42], which means that it is possible that two sequence pairs have optimal alignment scores x and y with $x < y$, but x is more statistically significant than y. Therefore, instead of using the align-

ment score alone as the metric for homology, it is a common practice to estimate the statistical significance of an alignment score to comment on the relatedness of the two sequences being aligned. Of course, it is important to note here that although statistical significance may be a good preliminary indicator of biological significance, it does not necessarily imply biological significance [9, 42].

The knowledge of accurate statistics for score distribution of ungapped alignments is available [34]. However, till now there is no precise statistical theory for the gapped alignment score distribution and for score distributions from enhanced alignment programs using additional features like difference blocks [32] or multiple parameter sets [31]. Accurate estimation of statistical significance of gapped sequence alignment has attracted a lot of attention in the recent years [65, 11, 50, 43, 41, 18, 10, 57, 59, 54, 68, 1, 3]. There exist a couple of good starting points for statistically describing gapped alignment score distributions for simple scoring schemes [36, 25], but a complete mathematical description of the optimal score distribution remains far from reach [25]. There exist many excellent reviews on statistical significance in sequence comparison in the literature [48, 53, 42, 40].

Database statistical significance versus pairwise statistical significance

Recently, a thorough study of pairwise statistical significance and its comparison with database statistical significance was conducted [1, 2]. In summary, the database statistical significance which is commonly reported by most database search programs like BLAST, FASTA, SSEARCH, PSI-BLAST is dependent on the database, and hence the same pairwise alignment with same alignment score can be assessed different significance values in different database searches, and even with the same database at different times, since the size of the database keeps on changing. Pairwise statistical significance, on the other hand is specific to the sequence pair being aligned, and is independent of any database. In [1, 2], various approaches to estimate pairwise statistical significance like ARIADNE [41], PRSS [51], censored-maximum-likelihood fitting [22], linear regression fitting [31] were compared to find that maximum likelihood fitting with censoring left of peak (described as type-I censoring in [22]) is the most accurate

method for estimating pairwise statistical significance. Further, this method was compared with database statistical significance in a homology detection experiment to find that pairwise statistical significance performs better than database statistical significance using BLAST and PSI-BLAST on a benchmark database and comparable to SSEARCH, but PSI-BLAST results can be significantly improved by using pre-trained PSSMs (position-specific scoring matrices). In another related work [3], a simple extension of pairwise statistical significance was shown to be better than ordinary pairwise statistical significance, where the concept of conservative, non-conservative, and average pairwise statistical significance was introduced. This corresponds to estimating two different values of pairwise statistical significance for the same pair of sequences by shuffling each sequence independently to generate separate score distributions, and reporting the final pairwise statistical significance estimate as the maximum, minimum, and average of the two values respectively. In [3], non-conservative pairwise statistical significance was shown to perform better than the other two variants of pairwise statistical significance and original pairwise statistical significance. Pairwise statistical significance using multiple parameter sets [4] and sequence-pair-specific distanced substitution matrices [6] has also been explored, which give slightly better results than original pairwise statistical significance, but not comparable to the methods described in this paper.

Relevance

Accurate statistical significance estimates for pairwise alignments can be very useful to comment on the relatedness of a pair of sequences independent of any database. Further, it can also be used to compare different combination of alignment parameters - like the alignment program, substitution matrices, gap costs. A comparison of different gap opening penalties for four commonly used BLOSUM matrices using pairwise statistical significance was presented in [2]. There has been a lot of recent development in alignment programs [32, 31] taking into account other desirable biological features of a sequence alignment 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

better suiting to different rates of conservation at different spatial locations of the sequences. As pointed out earlier, accurate 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 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 can be extremely useful. Further, accurate estimates of pairwise statistical significance can also be highly valuable for many other pairwise-alignment-based applications - like multiple sequence alignment (in particular progressive MSA), phylogenetic tree construction, etc. With the all-pervasive use of sequence alignment methods in bioinformatics making use of ever-increasing sequence data, and with development of more and more sophisticated alignment methods with unknown statistics, we believe that computational and statistical approaches for accurate estimation of statistical significance of pairwise alignment scores would prove to be very useful for computational biologists and bioinformatics community.

Contributions

In this paper, we explore the use of sequence-specific and position-specific substitution matrices with pairwise statistical significance, which is expected to be still more specific to the sequence-pair being aligned, and hence yield better performance. To evaluate the results of using sequence-specific substitution matrices with pairwise statistical significance, we conducted similar experiments as reported in [62], and later in [1, 3] on a subset of the CATH 2.3 database. [62] had earlier created this database to evaluate seven protein structure comparison methods and the two sequence comparison programs. In the current work, experiments were conducted with different levels of sequence-specific contribution (using sequence-specific substitution matrices with different levels of sequence-specific contribution), and the results confirm that deriving more sequence-specific information by using sequence-specific substitution matrices gives better coverage performance than using a standard substitution matrix like BLOSUM62. A sequence-specific substitution matrix for a given sequence is derived using the

alignments obtained from a BLAST search with the given sequence as the query. More details are provided later. Further, the optimal level of sequence-specific contribution was identified for the benchmark database used for the experiments (a subset of CATH 2.3 database). Also, the comparison with database statistical significance shows that using sequence-specific substitution matrices with pairwise statistical significance gives significantly better estimates of statistical significance than database statistical significance estimates using BLAST, PSI-BLAST and even SSEARCH, which is considered most sensitive as it uses original implementation of Smith-Waterman algorithm (and subsequently takes large amount of time for database search). Although using sequence-specific substitution matrices for estimating pairwise statistical significance gives significantly better performance than PSI-BLAST on the benchmark database, PSI-BLAST results can be significantly improved by using pre-trained PSSMs (position-specific scoring matrices). To fairly compare database statistical significance using PSI-BLAST with pairwise statistical significance, we also conducted experiments with pairwise statistical significance using the same pre-trained PSSMs used with PSI-BLAST, which indeed gives significantly better performance than PSI-BLAST. It is important to note that the methods proposed in this paper are for estimating pairwise statistical significance and not for general database search application, and the comparison with database search methods like BLAST, PSI-BLAST, SSEARCH is of their statistical significance estimation strategies.

The rest of the paper is organized as follows: In Section 2, an introduction to the extreme value distribution in the context of estimating statistical significance for gapped and ungapped alignments is presented, followed by the description of the methods used to create sequence-specific substitution matrices and modifying the Smith-Waterman algorithm to use position-specific substitution matrices in Section 3. Experiments and results are reported in Section 4, and finally the conclusion and future work is presented in Section 5.

The Extreme Value Distribution for Ungapped and Gapped Alignments

It is a well-known fact that 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). Similarly, the distribution of the maximum of a large number of i.i.d. random variables tends to an extreme value distribution (EVD) [35]. The distribution of Smith-Waterman local alignment score between random, unrelated sequences is known to follow a Gumbel-type EVD [34]. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP (High-scoring Segment Pairs which correspond to ungapped local alignment) scores is characterized by two parameters, K and λ. The probability that the optimal local alignment score S exceeds x is given by the P-value, which is defined as:

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

where E is the E-value and is given by

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

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

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

In theory, this normalized E-value is same as the P-value (for large N). 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 [34], and the parameters K and λ can be computed analytically from the substitution scores and sequence compositions. 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 [36, 25]. Recently, researchers have also looked closely at the low probability tail distribution, and the work in [66] applied a rare-event sampling technique earlier used in [27] 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 [65, 11, 50, 41, 46, 44, 31, 1, 3].

Methods

Creating Sequence-Specific Substitution Matrix for a Given Sequence

The entries of a typical substitution matrix like BLOSUM62 are essentially log-odds scores. The score $s(a, b)$ for aligning two residues a and b is:

$$s(a, b) = c \times \log_2 \frac{p(a, b)}{\pi(a)\pi(b)}$$

where $p(a, b)$ denotes the probability that the residues a and b are correlated because they are homologous, $\pi(a)$ is the equilibrium probability of residue a, and c is the scaling factor. Therefore, $p(a, b)$ is the target frequency: the probability of observing residues a and b aligned in homologous sequence alignments, and $\pi(a)\pi(b)$ is the probability that the two residues are uncorrelated and unrelated, occurring independently. The resulting substitution matrix is said to be in $1/c$ bit units. An excellent introduction to fundamental concepts of substitution matrices is provided in [23].

Further, the probabilities $p(a, b)$ and $\pi(a)$ can be easily estimated from a count matrix C, where the entry $C(a, b)$ gives the count of the number of times residue a was seen aligned to b in a set of alignments (both pairwise or multiple sequence alignments) of homologous sequences. Usually, the count matrix is added to its transpose to ensure symmetry, and hence, $C(a, b) = C(b, a)$. Then,

$$p(a, b) = \frac{C(a, b)}{\sum_c \sum_d C(c, d)}$$

$$\pi(a) = \frac{\sum_b C(a, b)}{\sum_c \sum_d C(c, d)}$$

Therefore, the task of generating sequence-specific substitution matrices reduces to obtaining sequence specific count matrices. For a given sequence S, a sequence-specific count matrix can be obtained using the simple procedure as follows: Run BLAST program with S as the query sequence against non-redundant protein database (provided with the BLAST suite of programs) with a sufficiently high e-value threshold so that more alignments can be obtained. The entries of the sequence-specific count matrix C_S can be obtained by counting the number of times residue a is aligned with b. Subsequently, C_S is added to its transpose to ensure symmetry.

Just as a count matrix can be used to get the substitution matrix, one can also back-calculate the count matrix for a given substitution matrix and equilibrium frequencies. Calculating the probabilities $p(a, b)$ from scores $s(a, b)$ and equilibrium frequencies $\pi(a)$ involves solving for a non-zero λ in $\sum_{ab} \pi(a)\pi(b)e^{\lambda s(a,b)} = 1$, and a C implementation of this procedure is available in the supplementary notes of [23]. Subsequently, these probabilities can be multiplied by a suitably large integer to get a representative count matrix C. Let the count matrix this obtained for the BLOSUM62 matrix be C_{BL62}.

This can be used to derive sequence-specific substitution matrices with different levels of sequence-specific contribution. We define $\alpha \in [0, 1]$ as the sequence-specific contribution. Then, for a given sequence S, sequence-specific count matrix with sequence-specific contribution α can be obtained as follows:

$$C_{S,\alpha} = \alpha C_S + (1 - \alpha)C_{BL62}$$

which can be subsequently used to obtain a sequence-specific substitution matrix for sequence S at sequence-specific contribution α using the procedure described earlier in this section. This is one of the many possible approaches to get a sequence-specific substitution matrix that we tried for our experiments. Results presented in the next section demonstrate the potential of the approach.

Using Position-Specific Substitution Matrices with Smith-Waterman algorithm

The Smith-Waterman algorithm [63] produces an optimal local alignment of two sequences. In its original form, it is designed to work for a standard substitution matrix with substitution scores for all possible pairs of residue substitutions. The algorithm can be trivially modified to work with position-specific substitution matrix for one of the two sequences being aligned.

Let $A = a_1, a_2, ..., a_m$ and $B = b_1, b_2, ..., b_n$ be two sequences of length m and n. Without loss of generality, let the position-specific scoring matrix be available for sequence A, given by a $m \times |\Sigma|$ matrix S, where $|\Sigma|$ is the number of different residues in the alphabet (e.g. 20 for protein sequences). $S(i, b)$ represents the substitution score of aligning the i^{th} residue of A (i.e. a_i) with residue b. Let q be the non-negative gap opening penalty, and r be the non-negative

gap extension penalty, so that the score of a gap of length k is $-(q + k \times r)$. The optimal local alignment of A and B is the global alignment of the subsequences α of A and β of B, whose similarity is maximal, i.e., which has maximum alignment score for the given scoring scheme (substitution matrix and gap penalties).

A local alignment of A and B consists of two types of configurations: substitutions and gaps. A substitution associates a residue of A with a residue of B. A gap consists only of residues from one sequence with each residue associated with the symbol $-$. There are two kinds of gaps. A deletion gap with respect to sequence A consists only of residues from A and an insertion gap with respect to sequence A consists only of residues from B.

Let $A_i = a_1, a_2, ...a_i$ and $B_j = b_1, b_2, ..., b_j$ be initial segments of A and B of length i and j respectively. Define $V(i, j)$ to be the score of the optimal local alignment of A_i and B_j. Define $G(i, j)$ to be the score of the optimal local alignment of A_i and B_j where a_i and b_j are aligned with each other. Define $I(i, j)$ to be the score of the optimal local alignment of A_i and B_j that end with an insertion gap with respect to A. Similarly, define $D(i, j)$ to be the score of the optimal local alignment of A_i and B_j that end with an deletion gap with respect to A. Then, the following recurrences are used to calculate the optimal local alignment:

Base Conditions:

$$\begin{aligned} V(0,0) &= 0 \\ V(i,0) &= 0 \quad \forall i \\ V(0,j) &= 0 \quad \forall j \\ I(i,0) &= -q \quad \forall i \geq 0 \\ D(0,j) &= -q \quad \forall j \geq 0 \end{aligned}$$

Recurrence relations:

$$\begin{aligned} V(i,j) &= max\left\{G(i,j), I(i,j), D(i,j), 0\right\} \\ G(i,j) &= max\left\{V(i-1,j-1) + S(i,b_j)\right\} \end{aligned}$$

$$I(i,j) = max\{I(i,j-1)-r, V(i,j-1)-q-r\}$$
$$D(i,j) = max\{D(i-1,j)-r, V(i-1,j)-q-r\}$$

The score of the optimal local alignment of A and B is given by $V(i',j') = max_{1\leq i\leq m, 1\leq j\leq n} V(i,j)$.

Note that the only difference in the above algorithm and the standard Smith-Waterman algorithm adapted to work with affine gap penalties (provided in Appendix) is in the recursion for matrix G. Both time and space complexity of the algorithm is $O(mn)$. The actual alignment can be calculated by following a trace-back procedure from $V(i',j')$ as described in [63]. The space-complexity can be reduced to $O(min\{m,n\})$ using a divide-and-conquer strategy developed by Hirschberg [29] after identifying the starting and ending indices of the optimal local alignment.

Experiments and Results

We use sequence-specific and position-specific substitution matrices to estimate the pairwise statistical significance of a pairwise alignment score, which is expected to be more specific to the sequence pair being aligned, and hence give better performance. Statistical significance estimates can be evaluated in terms of statistical significance accuracy or retrieval accuracy. Statistical significance accuracy is a measure of how accurate the significance estimates are, in relation to the true score distribution. Retrieval accuracy is a measure of the ability of significance estimates to distinguish between true homologs and false homologs, which is practically very useful for the primary application of local sequence alignment, which is homology detection. A good sequence comparison strategy, therefore, should assign higher significance values (lower P-values) to true homolog pairs, and lower significance values (higher P-values) to false homolog pairs. [1] examined the statistical significance methods both in terms of statistical significance accuracy and retrieval accuracy. Here, we evaluate the performance of the proposed methods and existing methods in terms of retrieval accuracy because firstly, it is far more important than statistical significance accuracy and secondly, here we use the same

method that was found in [1] to be most accurate in terms of statistical significance accuracy (maximum-likelihood fitting of score distribution censored left of peak).

To evaluate the performance of using sequence-specific and position-specific substitution matrices for estimating pairwise statistical significance in terms of retrieval accuracy and compare it with using a general matrix for pairwise statistical significance and with database statistical significance, we used the same experiment setup as used in [62], and later in [1, 3]. A non-redundant subset of the CATH 2.3 database (Class, Architecture, Topology, and Hierarchy, [47]) available at ftp://ftp.ebi.ac.uk/ pub/software/ unix/fasta/prot_sci_04/ was selected in [62] to evaluate seven structure comparison programs and two sequence comparison programs. As described in [62], this dataset consists of 2771 domain sequences and includes 86 query sequences. This domain set is considered as a valid benchmark for testing protein comparison algorithms [56].

Sequence-specific substitution matrices were obtained for each of the 2771 sequences in the database using the method described in the previous section. We used to BLAST program (version 2.2.17) with a relatively high e-value threshold of 1000 (-e 1000) so that we can collect enough alignments for filling the count matrix. Further, to view 1000 best alignments in the output, the '-b 1000' option was used. The BLAST alignments were used to generate the count matrix, and subsequently the substitution matrix for different values of sequence-specific contribution α. The scaling factor c was chosen to be 3, and hence, all substitution matrices were generated in 1/3-bit scale. We used these sequence-specific substitution matrices for estimating pairwise statistical significance [1]. Here, we used non-conservative pairwise statistical significance, which has been shown to be more effective compared to other variants of pairwise statistical significance [3]. Non-conservative pairwise statistical significance involves estimation of two different pairwise statistical significance estimates obtained by independent shuffling each of the two sequences in the sequence pair being aligned, and the final significance reported is the minimum of the two values.

For each of the 86×2771 comparisons, we estimated the non-conservative pairwise statistical significance using the sequence-specific substitution matrices at different levels of sequence-

specific contribution. The number of shuffles to generate the empirical distribution was set to 1000. The gap opening and gap extension penalties were set to 10 and 2 respectively.

Following [62, 1, 3], Error per Query (EPQ) versus Coverage plots were used to visualize and compare the results. To create these plots, the list of pairwise comparisons was sorted based on decreasing statistical significance (increasing P-values). While traversing the sorted list from top to bottom, the coverage count 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. The 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 better curve is one which is more to the right.

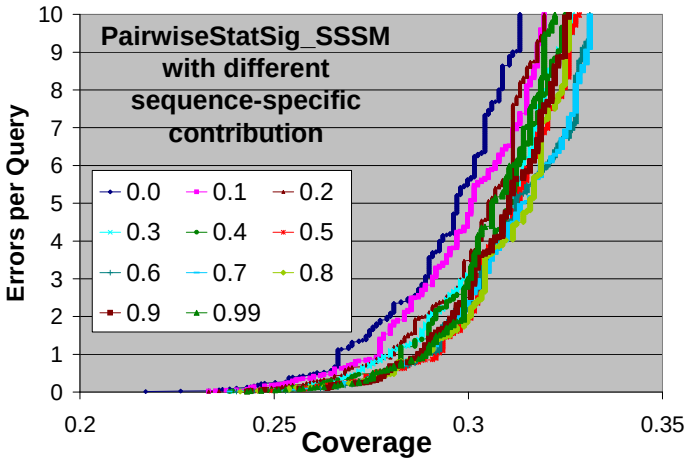

Figure 5.1 EPQ vs. Coverage plot for different levels of sequence-specific contribution α. The left-most curve is for $\alpha = 0$, i.e., 0% sequence-specific contribution, which corresponds to using a general substitution matrix (BLOSUM62). For all values of $\alpha > 0$, the coverage performance is significantly better than the performance with $\alpha = 0$, suggesting that using sequence-specific substitution matrices for estimating pairwise statistical significance is significantly better than using general substitution matrices.

The EPQ vs. Coverage curves for different levels of sequence-specific contribution α are presented in Fig. 5.1. The left-most curve is for $\alpha = 0$, i.e., 0% sequence-specific contribution, which corresponds to using a general substitution matrix (BLOSUM62). For all values of

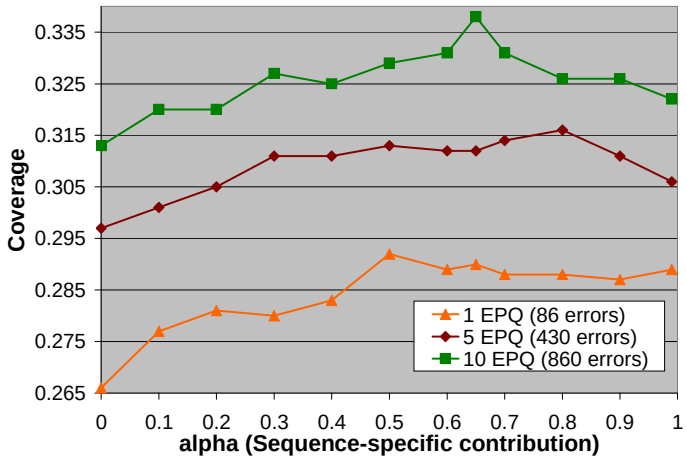

Figure 5.2 Coverage vs. sequence-specific contribution (α) plot for three different error levels. The coverage performs increases as α increases, reaches a maximum, and decreases a little for high values of α. $\alpha = 0.65$ is identified to be the best value for the benchmark dataset used.

$\alpha > 0$, the coverage performance is significantly better than the performance with $\alpha = 0$, suggesting that using sequence-specific substitution matrices for estimating pairwise statistical significance is significantly better than using general substitution matrices. The curves are quite close to each other, and it is difficult to determine the best value of α for this dataset from this graph. Therefore, we further use Coverage vs. Sequence-specific contribution plots at different error levels to determine the optimal value of α for this dataset, as presented in Fig. 5.2. It shows the coverage values at three different error levels for different values of α. There is a clear improvement in coverage performance as α increases from 0. But for values of α close to 1.0, the coverage performance decreases slightly, which is expected since some sequences in the database may not get sufficient hits in the BLAST search, which would leave the count matrix very sparse, and without sufficiently filled count matrix, the corresponding substitution matrix would not be of good quality, which would affect the coverage performance. From Fig. 5.2, we can determine a range of α values which gives the best performance. Clearly, for this dataset it can be safely considered to be $[0.5, 0.8]$. Further, within this range, $\alpha = 0.65$ is visually identified to be the best value for this dataset. It is important to note here that these results are obtained on the subset of CATH 2.3 database which is a benchmark database for

protein comparison, but the results and best value of α may not be generalized to all databases.

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) [62], for comparing the performance across different comparison methods, we examine the performance of the methods with individual queries, following the work in [62]. The coverage of each of the 86 queries at the 1st, 3rd, 10th, 30th, and 100th error was recorded, and percentile analysis was done for each error level across the 86 queries. A comparison of pairwise statistical significance using sequence-specific substitution matrices (PairwiseStatSig_SSSM) and database statistical significance reported by BLAST, PSI-BLAST and SSEARCH is presented in Fig. 5.3. Fig. 5.3(a) shows the 25th percentile coverage level at the 1st, 3rd, 10th, 30th, and 100th false positive for homologs (i.e. 21 of the queries have worse coverage, and 65 have better coverage), Fig. 5.3(b) shows the same results for 50th percentile of coverage, i.e. the median coverage (43 queries performed better, 43 worse), and Fig. 5.3(c) shows the same results for 75th percentile of coverage (i.e. 65 of the queries have worse coverage, and 21 have better coverage). The curves for SSEARCH in Fig. 5.3(a) and Fig. 5.3(b) are derived from the figures 2A and 2B in [62]. The results for SSEARCH corresponding to Fig. 5.3(c) were not available in [62].

The curves suggest that using more sequence-specific information for statistical significance estimation (by using sequence-specific substitution matrices) gives significantly better results than database statistical significance using BLAST, PSI-BLAST and even SSEARCH.

In the above described experiments, only the benchmark database was used to construct the PSSMs (position-specific scoring matrices) over a maximum of 5 iterations. Since PSI-BLAST allows the use of pre-constructed PSSMs for the query sequence, we derived PSSMs for all the 86 test queries against the non-redundant protein database (provided along with the BLAST package) over a maximum of 5 iterations. Subsequently, these pre-trained PSSMs were used as starting PSSMs for PSI-BLAST searches of each of the 86 queries against the benchmark database, further refined for a maximum of 5 iterations. Using better quality pre-trained PSSMs in this way is expected to give superior performance for PSI-BLAST. For a fair

comparison of pairwise statistical significance with PSI-BLAST using pre-trained PSSMs, we also conducted experiments with pairwise statistical significance using the same pre-trained PSSMs used as starting PSSMs for PSI-BLAST searches on the benchmark database. For this purpose, the popular Smith-Waterman algorithm [63] was trivially modified to calculate the optimal local alignment using a position-specific substitution matrix instead of a general substitution matrix, as described in the previous section. The implementation of the GAP3 program [32] was suitably modified to get the optimal alignment score of a pairwise alignment using position-specific substitution matrix. Again, the number of shuffles was set to 1000. Gap opening and gap extension penalties were set to 11 and 1 respectively, since these were the default values using which the PSI-BLAST PSSMs were constructed. A comparison of pairwise statistical significance using position-specific scoring matrices (PairwiseStatSig_PSSM) and PSI-BLAST is presented in Fig. 5.4. There are two comparisons: one using the PSSMs derived against the benchmark database (a subset of CATH), and the other using pre-trained PSSMs derived against the non-redundant protein database (NRP) provided with the BLAST package. Fig. 5.4(a), Fig. 5.4(b), and Fig. 5.4(c) show the 25th percentile, 50th percentile, and 75th percentile coverage level at the 1st, 3rd, 10th, 30th, and 100th false positive for homologs in a Errors per Query vs. Coverage plot. As is clear from these figures, using position-specific substitution matrices for estimating pairwise statistical significance is significantly better than database statistical significance using PSI-BLAST, for both kinds of PSSMs.

Finally, in Fig. 5.5, we present the combined comparison results of Fig. 5.3 and Fig. 5.4, with sub-figures Fig. 5.5(a), Fig. 5.5(b), and Fig. 5.5(c) showing the 25th percentile, 50th percentile, and 75th percentile coverage level as shown in earlier figures. There are three observations that can be made from the figures: Firstly, for all relevant comparisons, pairwise statistical significance performs at least comparable or significantly better than database statistical significance. Secondly, in general, position-specific sequence comparison is superior to sequence-specific analysis, which is better than sequence-independent analysis (using general substitution matrix), which is expected. Thirdly, depending on the quality of sequence-specific and position-specific substitution matrices, there are some exceptions to the second

observation. For example, using PSI-BLAST on the benchmark database gives inferior performance than using sequence-specific substitution matrices with pairwise statistical significance, although PSI-BLAST uses position-specific substitution matrices. Also, pairwise statistical significance using sequence-specific substitution matrices (derived from BLAST searches against non-redundant protein database) performs comparable to pairwise statistical significance using position-specific substitution matrices (derived from PSI-BLAST searches against the benchmark database).

As mentioned earlier, SSEARCH employs the original Smith-Waterman algorithm for alignment, and is considered more sensitive than its heuristic implementations like BLAST and FASTA. PSI-BLAST uses an iterative approach with position-specific scoring matrices (PSSMs), and its accuracy depends heavily on the quality of PSSMs. BLAST, PSI-BLAST and SSEARCH are database search methods which report database statistical significance. Significantly better results than these database search methods by using sequence-specific and position-specific substitution matrices, at least on one benchmark database implies that statistical significance estimates significantly better than database statistical significance can be obtained by using sequence-specific substitution matrices with pairwise statistical significance. This can be very useful to estimate accurate pairwise statistical significance of a (or a few) pair of sequences, which is a common situation in many pairwise alignment based applications like phylogenetic tree construction, progressive multiple sequence alignment.

Since the computation time for finding an optimal local sequence alignment is more or less the same for the cases of using a standard substitution matrix, sequence-specific substitution matrix, and position-specific substitution matrix, it is highly recommended to use sequence-specific and position-specific substitution matrices for estimating pairwise statistical significance, if they are available. Further, since the significant improvement of results using the proposed methods is mainly due to the use of sequence-specific and position specific substitution matrices, this research is also expected to motivate researchers to develop better quality sequence-specific and position-specific substitution matrices.

Another important contribution of this work is the evidence of the applicability of Karlin-

Altschul statistics for local alignment scores using sequence-specific and position-specific substitution matrices, where the statistical parameters K and λ can be estimated by fitting an Gumbel-type extreme value distribution to the observed score distribution. As discussed earlier, accurate statistics are available only for ungapped local sequence alignment scores [34], which is theoretically applicable only when using a single standard substitution matrix and infinite length sequences; and accurate statistics of newer and more sophisticated alignment methods is not expected to be straightforward. The results presented in this work support the assumption that the score distribution from the newer alignment methods considered in this paper also follows extreme value distribution, wherein the statistical significance of an alignment score and be accurately estimated for a practical application of separating true homologs from false homologs (homology detection - measured in terms of retrieval accuracy).

It is important to note that the methods described in this paper are for estimating pairwise statistical significance and not for general database search application, and the comparison with database search methods like BLAST, PSI-BLAST, SSEARCH is of their statistical significance estimation strategies. The proposed method can be used to estimate the pairwise statistical significance of a pair (or few pairs) of sequences quickly, but will take impractically long time for all pairwise comparisons in a large database search. Since the performance of pairwise statistical significance is shown to be superior than database statistical significance, it can be used in conjunction with a fast database search program like BLAST or PSI-BLAST to refine their results. This can be especially useful since no extraneous computation would be required to get the BLAST output file or PSI-BLAST PSSM file.

An implementation of the proposed method and related programs in C are available for free academic use at www.cs.iastate.edu/∼ankitag/PairwiseStatSig_SSSM.html and www.cs.iastate.edu/∼ankitag/PairwiseStatSig_PSSM.html

Conclusion

This paper extends the work on pairwise statistical significance by exploring the use of sequence-specific and position-specific substitution matrices for estimating pairwise statistical

significance, and compares them with database statistical significance in a homology detection experiment. Results indicate using sequence-specific substitution matrices performs significantly better than using general substitution matrices with pairwise statistical significance, and also significantly better than database statistical significance (using BLAST, PSI-BLAST and SSEARCH), but the accuracy of PSI-BLAST can be improved using pre-trained position-specific scoring matrices (PSSMs). Pairwise statistical significance using position-specific substitution matrices is significantly better than PSI-BLAST using pre-trained PSSMs.

Although pairwise statistical significance has been shown to give significantly better results than database statistical significance in terms of retrieval accuracy, the pairwise statistical significance estimation methods described in this paper can be used only for estimating the pairwise statistical significance of a few pairs of sequences in a reasonable time, and hence, cannot be used as a method for all pairwise comparisons in a large database search. It, however can be used in conjunction with fast heuristic-based database search programs like BLAST and PSI-BLAST to refine their results.

The current work provides for a lot of scope for future work. Significant improvement in retrieval accuracy with pairwise statistical significance using sequence-specific and position-specific substitution matrices underscores the influence of substitution matrices in sequence comparison. Hence, better quality sequence-specific and position-specific substitution matrices can be extremely useful. Also, faster methods for pairwise statistical significance estimation will be quite helpful.

Acknowledgment

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.

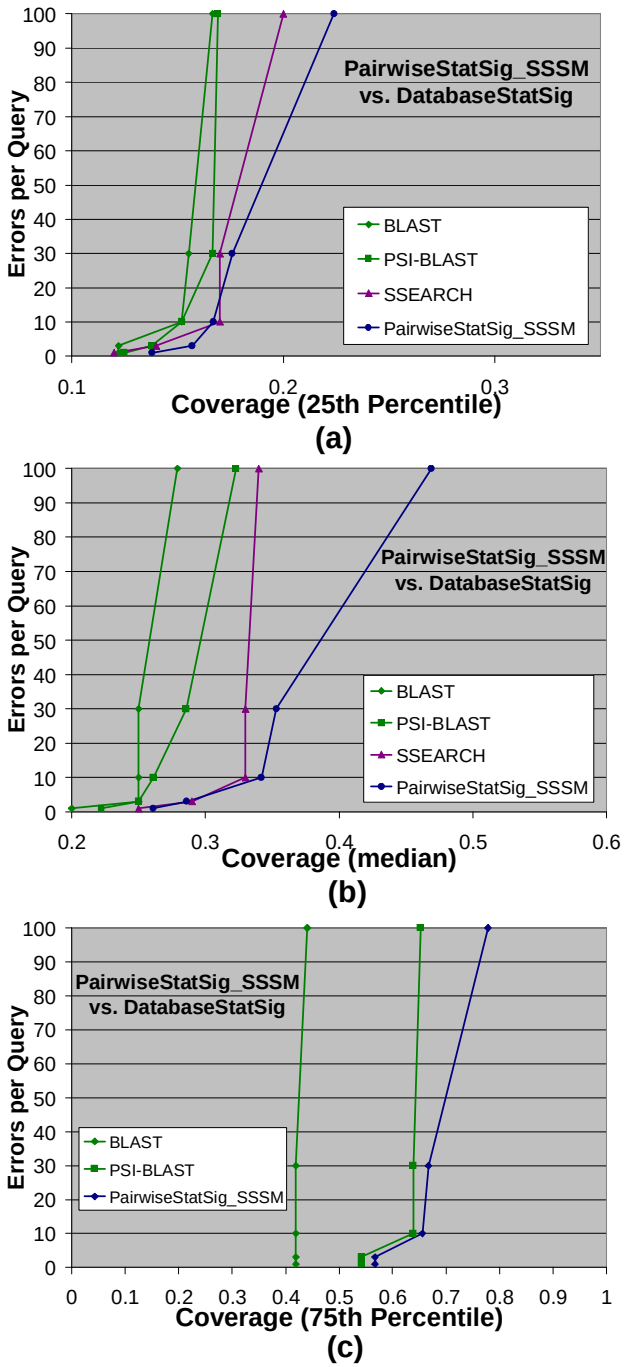

Figure 5.3 Comparison of using sequence-specific substitution matrices for estimating pairwise statistical significance with database statistical significance. Using sequence-specific substitution matrices for estimating pairwise statistical significance is significantly better than database statistical significance using BLAST, PSI-BLAST and even SSEARCH, on the benchmark database.

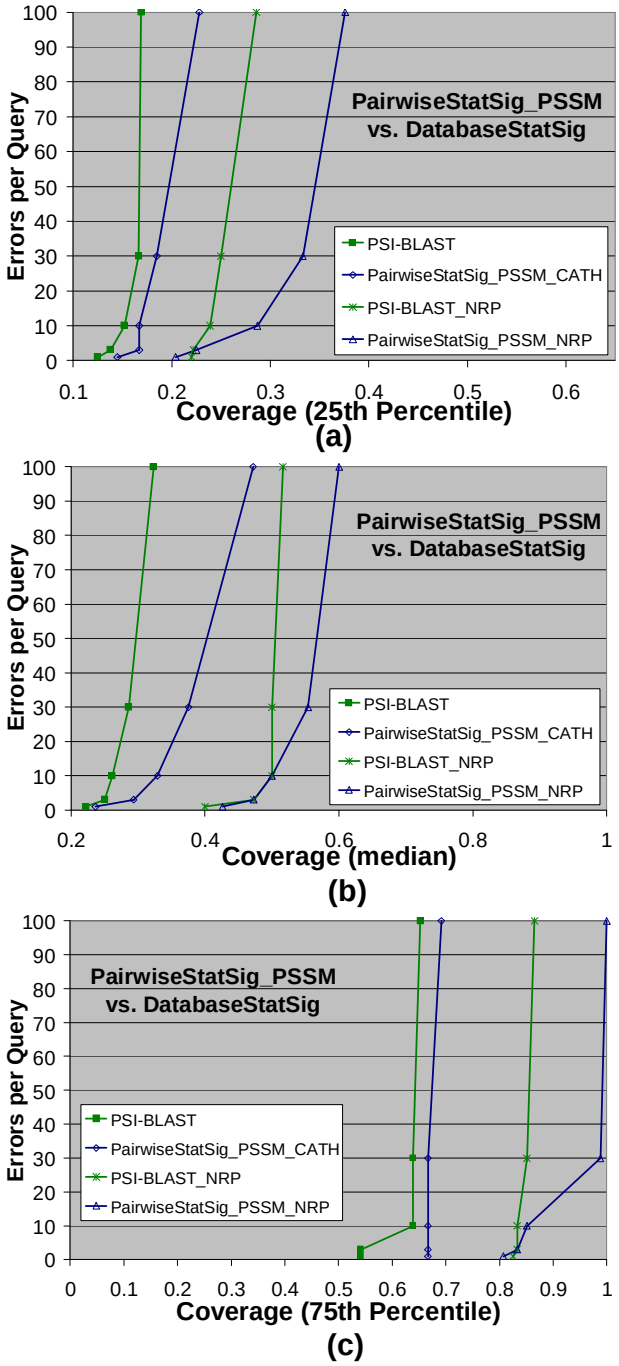

Figure 5.4 Comparison of using position-specific substitution matrices for estimating pairwise statistical significance with database statistical significance using PSI-BLAST. Using position-specific substitution matrices for estimating pairwise statistical significance is significantly better than database statistical significance using PSI-BLAST, for both types of PSSMs (derived against the benchmark database and against non-redundant protein database provided with the BLAST package).

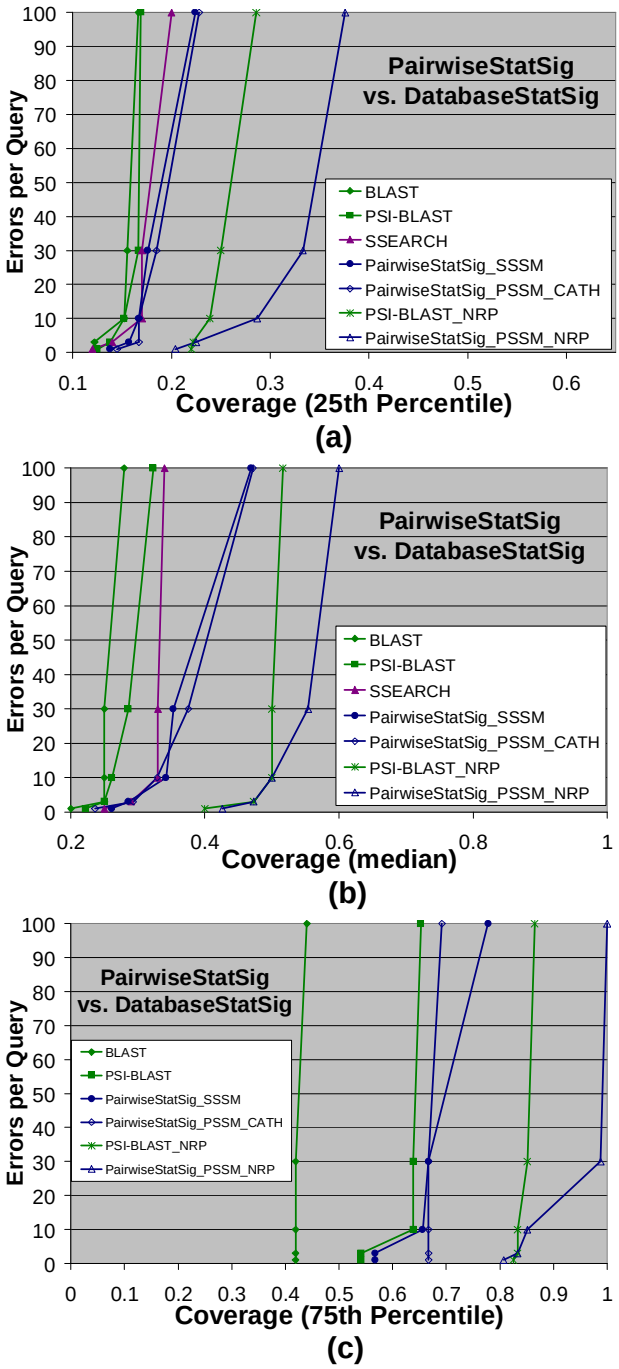

Figure 5.5 Comparison of using sequence-specific and position-specific substitution matrices for estimating pairwise statistical significance with database statistical significance. For all relevant comparisons, pairwise statistical significance performs significantly better than database statistical significance using BLAST, PSI-BLAST and SSEARCH.

6. PSIBLAST_PairwiseStatSig: REORDERING PSI-BLAST HITS USING PAIRWISE STATISTICAL SIGNIFICANCE

A paper published in Bioinformatics

Ankit Agrawal and Xiaoqiu Huang

Abstract

Summary: We present an add-on to BLAST and PSI-BLAST programs to reorder their hits using pairwise statistical significance. Using position-specific substitution matrices to estimate pairwise statistical significance has been recently shown to give promising results in terms of retrieval accuracy, which motivates its use to refine PSI-BLAST results, since PSI-BLAST also constructs a position-specific substitution matrix for the query sequence during the search. The obvious advantage of the approach is more accurate estimates of statistical significance because of pairwise statistical significance, along with the advantage of BLAST/PSI-BLAST in terms of speed.

Availability: The implementation as a C library is freely available at www.cs.iastate.edu/~ankitag/PSIBLAST_PairwiseStatSig.html

Contact: ankitag@cs.iastate.edu

Introduction

Database search is one of the most important applications of pairwise sequence alignment. The most popular heuristic-based methods for database search are the BLAST and PSI-BLAST

programs [13]. PSI-BLAST uses an iterative approach to BLAST using position-specific substitution matrices which are refined with every iteration, and its performance can be significantly better than BLAST. Another slightly slower but more accurate database search program than BLAST is FASTA [51], which also employs heuristics to obtain a sub-optimal alignment. There also exists the SSEARCH program, which uses the full implementation of the Smith-Waterman algorithm [63]. Although more accurate, it can take many hours to days for a modest database search.

The hits of a database search are ranked according to statistical significance of the alignment scores rather than by alignment score themselves. An alignment score is considered statistically significant if it has a low probability of occurring by chance. The alignment score distribution (and hence statistical significance) depends on various factors like alignment program, scoring scheme, sequence lengths, sequence compositions [42]. Accurate estimation of statistical significance of alignment scores is an important aspect of sequence comparison.

The methods to estimate the statistical significance of a pairwise alignment can be categorized into two primary methods. The statistical significance of the hits reported by database search programs is called database statistical significance, which is in general dependent on the size and composition of the database being searched. An alternative method to estimate statistical significance of a pairwise alignment independent of any database is to estimate pairwise statistical significance, which uses statistical parameters specific to the sequence-pair to estimate statistical significance.

In the last few years there have been considerable improvements to the BLAST and PSI-BLAST programs [57, 67, 68], which have been shown to improve database search performance by using composition-based statistics and substitution matrix rescaling techniques, together with pre-computed statistical parameters for a wide range of alignment parameters. Recently, a study of pairwise statistical significance was conducted [2]. It compared various approaches to find that maximum likelihood fitting of an empirical distribution with censoring left of peak is most accurate for estimating pairwise statistical significance. Further, using position-specific substitution matrices to estimate pairwise statistical significance [5] gives the best results in

terms of retrieval accuracy since it uses maximal sequence-specific information. Relevant details on pairwise statistical significance can be found in the supplementary notes (Appendix B).

Proposed Approach

The advantage of using position-specific substitution matrices (PSSMs) with pairwise statistical significance strongly motivates its use to refine PSI-BLAST results, since PSI-BLAST naturally constructs a PSSM for the query sequence, which can be used for estimating pairwise statistical significance. In this application note, we present an add-on to the BLAST and PSI-BLAST programs to refine their results using pairwise statistical significance. The proposed approach is implemented as a program named PSIBLAST_PairwiseStatSig which takes a query sequence, a database, the PSI-BLAST output file, and the PSI-BLAST constructed PSSM (if available), and gives the new pairwise statistical significance estimates.

To evaluate PSIBLAST_PairwiseStatSig, we used the same benchmark database (a non-redundant subset of CATH2.3 database of 2771 sequences, and its subset of 86 query sequences) as earlier used in [62, 2, 5]. For refining BLAST results, the BLSOUM62 matrix was used for the alignments as it is the default substitution matrix for the BLAST program. For refining PSI-BLAST results, the PSI-BLAST constructed PSSM was used. To further take advantage of PSSMs, non-conservative pairwise statistical significance was also estimated (see supplementary notes (Appendix B)). Note that for non-conservative pairwise statistical significance estimation with PSSMs, we would need PSSMs for both the sequences being aligned. But in general, after a PSI-BLAST run, we get a PSSM for only the query sequence, and not for the hits obtained. Therefore, here we use the standard substitution matrix BLOSUM62 instead of PSSM for second sequence, hoping that the PSSM for query sequence is significantly different from BLOSUM62 to take advantage of non-conservative pairwise statistical significance. The number of shuffles N was set to 1000.

The two evaluation methodologies used to compare the results are explained in detail in the supplementary notes (Appendix B). Here we only present the results using the standard methodology (earlier used in [16, 62]) due to limited space. Fig. 6.1 shows the Er-

ror Per Query vs. Coverage curves for BLAST and PSI-BLAST with and without reordering their hits using pairwise statistical significance, depicting the improvement in performance using PSIBLAST_PairwiseStatSig (a curve more towards the right is better). The PSIBLAST_PairwiseStatSig program is tested to work with BLAST/PSI-BLAST output files for BLAST 2.2.17, but is expected to work for other versions as well.

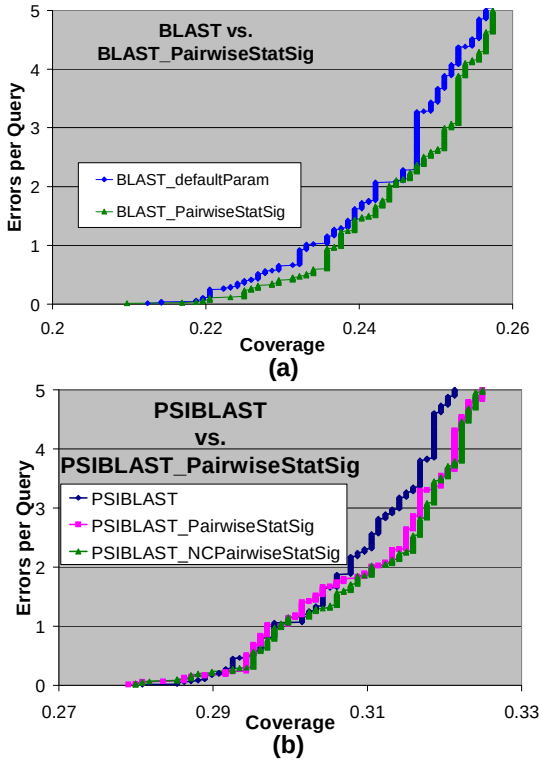

Figure 6.1 Errors per Query vs. Coverage plots comparing the performance of (a) BLAST and BLAST_PairwiseStatSig (reordering BLAST results using pairwise statistical significance), and (b) PSIBLAST, PSIBLAST_PairwiseStatSig, and PSI-BLAST_NCPairwiseStatSig (reordering PSI-BLAST results using non-conservative pairwise statistical significance). Reordering BLAST/PSI-BLAST hits using pairwise statistical significance leads to superior performance in terms of retrieval accuracy.

Assuming that BLAST/PSI-BLAST output file is already available, the running time of the proposed method is dependent on the number of hits given by BLAST/PSI-BLAST. For a single sequence-pair, the pairwise statistical significance estimation time depends on the

length of the two sequences. For typical protein sequence lengths (248 and 255), it took 0.45s to estimate pairwise statistical significance on a 2.8 GHz Intel processor.

An obvious disadvantage of the proposed approach is that its performance is upper-bounded by the number of true homologs detected by BLAST/PSI-BLAST. It can only reorder the hits with an attempt to rank the true homologs higher, but cannot recover any more homologs. However, considering this limitation, PSIBLAST_PairwiseStatSig has been demonstrated to give better results than BLAST and PSI-BLAST just by reordering the hits using pairwise statistical significance. It is also important to note that the proposed method is studied in context of protein database searches and not DNA, which may require substantial modification considering the arbitrary lengths of the DNA sequences.

Acknowledgement

The authors would like to thank Dr. Volker Brendel for helpful discussions and providing links to the data. Special thanks are due to the anonymous reviewers for their insightful comments, which made the manuscript stronger.

7. FAST PAIRWISE STATISTICAL SIGNIFICANCE ESTIMATION USING DERIVED DISTRIBUTION POINTS AND DATABASE SEARCH HEURISTICS

A paper to be submitted

Ankit Agrawal and Xiaoqiu Huang

Abstract

Evaluation of statistical significance of a pairwise sequence alignment is crucial in homology detection. A major recent development in the field is the use of pairwise statistical significance as an alternate to database statistical significance. Although pairwise statistical significance has been shown to be comparable and at times significantly better than database statistical significance in terms of homology detection retrieval accuracy, it is also much time consuming since it involves generating an empirical score distribution by alignment of one sequence with random shuffles of the other sequence. In this paper, we devise heuristics to speed up pairwise statistical significance estimation taking advantage of the nature of the estimation procedure. Both the proposed derived distribution points heuristic and a specific application of database search heuristic have been individually shown to give significant speedup compared to normal pairwise statistical significance with negligible loss of accuracy. Using both the heuristics in conjunction can give a speed up of more than 200, without significant loss of accuracy.

Introduction

Sequence alignment is an underlying application in the analysis and comparison of DNA and protein sequences [52, 12, 13], which forms the basis of numerous applications in bioinformatics,

beginning with homology detection, i.e., identifying sequences evolved from a common ancestor, generally known as homologs or related sequences. Homology detection further forms the key step of many other bioinformatics applications making various high level inferences about the DNA and protein sequences - like finding protein function, protein structure, deciphering evolutionary relationships, drug design, etc. There exist classical algorithms for optimal local sequence alignment [63], based on which, there exist many other algorithms [32, 31], which try to model sequence comparison in a better way by incorporating more biological features, like different conservation level along the length of the sequence, etc. Several heuristics have also been proposed [49, 13, 51, 39, 38] which are extremely useful in database search application where it is impractical to use exact algorithms for sequence alignment.

Since the chief application of sequence alignment is homology detection, sequence alignment methods can be compared in terms of their ability to distinguish between pairs of related and unrelated sequences. Although homologous pairs (pairs of related sequences) are expected to have high alignment score, the potential relatedness of two sequences is judged by statistical significance rather than by alignment score alone. 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 [42], it is possible that two sequence pairs have optimal alignments with scores x and y with $x < y$, but x more statistically significant than y. Of course, it is important to note here that although statistical significance may be a good preliminary indicator of biological significance, it does not necessarily imply biological significance [9, 42].

A good sequence comparison strategy should assign lower probabilities (higher statistical significance) for related sequence pairs, and higher probabilities (lower statistical significance) for unrelated sequence pairs. Recently, a study of pairwise statistical significance and its comparison with database statistical significance was conducted [1, 2], and its use with multiple parameter sets [7] and sequence-specific/position-specific substitution matrices [5] was explored, which depicted a clear advantage of pairwise statistical significance as compared to database statistical significance. Pairwise statistical significance has also been used to reorder

the hits from a fast database search program like PSI-BLAST [8]. However, since estimation of pairwise statistical significance involves generating a score distribution specific to the sequence-pair by aligning one sequence by multiple shuffles of the other sequence, it is very time consuming and can be impractical for estimating pairwise statistical significance of a large number of sequence pairs.

In this paper, we devise suitable heuristics to speed up pairwise statistical significance estimation, taking advantage of the nature of the estimation procedure. Both the proposed derived distribution points heuristic and a specific application of the database search heuristic have been individually shown to give significant speedup with negligible loss of accuracy. When used together, these heuristics can give a speedup of more than 200 without significant loss of accuracy.

The rest of the paper is organized as follows: Section 2 presents a description of the features of pairwise statistical significance estimation that motivated the design and application of the heuristics, which are discussed in Section 3 of the paper. Experiments and results are presented in Section 4, followed by the conclusion and future work in Section 5.

Pairwise Statistical Significance

As mentioned earlier, statistical significance of alignment score is more commonly used to comment on the relatedness of sequence-pairs than alignment score. The distribution of Smith-Waterman local alignment score between random, unrelated sequences is approximately a Gumbel-type EVD [34]. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP (High-scoring Segment Pairs which correspond to 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:

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

where E is the E-value and is given by

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

The above formulae are theoretically valid only for ungapped alignment, and the corresponding statistical parameters K and λ can be analytically determined for given sequence lengths, compositions and scoring scheme. For gapped alignments, however, no precise theory exists, but the gapped score distribution has been widely observed to follow the same distribution, even when using multiple parameter sets [7] and position-specific substitution matrices, as used by PSI-BLAST. Thus, the corresponding statistical parameters K and λ can be estimated by fitting an EVD to experimentally generated score distribution.

Pairwise statistical significance is an attempt to make the statistical significance estimation process more specific to the sequence pair being compared. Recently a study of pairwise statistical significance and its comparison with database statistical significance was conducted [1, 2]. It compared eight different schemes of estimating pairwise statistical significance in terms of statistical significance accuracy, i.e., their ability to predict the P-value in the extreme right tail of the distribution. It was found that maximum likelihood fitting of the censored score distribution (censored left of peak/fitting only the right tail) is the most accurate method. 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.

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 (substitution matrix, gap penalties), 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].

It is easy to see that the number of shuffles used to generate the empirical score distribution has an obvious effect on statistical significance accuracy. Higher the number of shuffles,

smoother the empirical distribution obtained, better the maximum-likelihood fitting, and hence better the statistical significance accuracy. However, it has been reported that improving the statistical significance accuracy may not necessarily improve retrieval accuracy [68], which is clearly more important for bioinformatics applications.

Since the estimation of pairwise statistical significance of the optimal alignment score of two sequences of length m and n involves computing N alignment scores, where N is the number of shuffles, the time complexity of the estimation procedure is $O(Nmn)$.

These features of pairwise statistical significance estimation strategy lead to the following observations which can help in speeding up the estimation process:

1. Scores in the right tail are more important than those in the left tail.
2. It should be possible to reduce number of shuffles without losing too much on retrieval accuracy.
3. Rather than using the time consuming dynamic programming algorithm to get the scores for the score distribution, fast heuristic-based methods can be used.

Proposed Heuristics

Based on the observations made in the previous section, two heuristics are described in this section to speed up the pairwise statistical significance estimation process.

Derived Distribution Points

This heuristic attempts to generate a score distribution faster by reducing the number of effective shuffles without adversely affecting the right tail of the distribution. Given the number of shuffles N and a derived distribution points set $DDP = \left\{DDP_1, DDP_2, \ldots, DDP_{N_{ddp}}\right\}$ with $N_{ddp} = |DDP|$, this heuristic reduces the number of shuffles from N to N/N_{ddp}, and the alignment score s from each actual shuffle contributes N_{ddp} alignment scores ($s + DDP_i$, $1 \leq i \leq N_{ddp}$) in the histogram, making a total of N alignment scores. The choice of the set DDP is such that it contributes a decreasing mini-histogram for every alignment score s

centered around s, which adversely affects the left tail of the distribution but not the right tail. This heuristic is expected to give a speedup of N_{ddp}. Figure 7.1 shows three DDP sets used in this work, including the special case of $DDP = \{0\}$, which effectively disables the DDP heuristic.

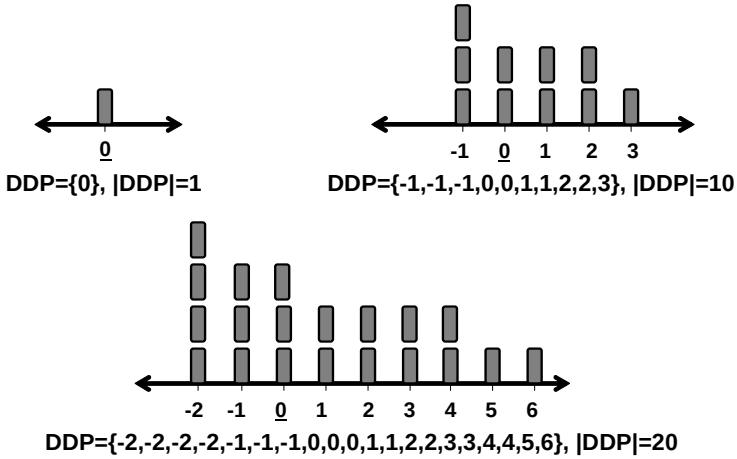

Figure 7.1 Three DDP sets used in this work. Each alignment score contributes $|DDP|$ scores to the histogram around itself thereby adversely affecting the score distribution only left of peak but not right of peak. The special case of $DDP = \{0\}$ essentially disables the DDP heuristic.

Database Search Heuristic

To construct the empirical distribution to estimate $PairwiseStatSig(Seq1, Seq2, SC, N)$, we propose to do a database search with $Seq1$ as the query against a database constructed of N random shuffles of $Seq2$. This approach is expected to give high scores quickly, which would constitute the right tail of the score distribution which we are interested to fit.

A basic heuristic-based database search approach can be outlined as follows:

1. Find short exact matches of length w between the query and the database.

2. Extend the matches in either direction allowing for mismatches in an attempt increase the score to get segments of score at least ic.

3. Combine the segments allowing for gaps by chaining segments across different diagonals in an attempt to increase the score to get chains of score at least fc.

The commonly used database search program BLAST [13] uses the above basic approach interweaved with much more intricate heuristics. The time required to get the score distribution for statistical significance estimation depends primarily on the parameter w, and also on ic. Smaller the word length w, more the number of exact matches, and hence higher the execution time. Similarly, lower the parameter ic, more the segments, and higher the execution time.

BLAST uses a default value of 3 for the word size (for protein comparisons). Instead of the cutoff ic on the score of the segments, BLAST places a cutoff on statistical significance of segment scores (known as E-value cutoff, with default value 10). It is known that the average expected score of optimal pairwise alignment of two sequences increases proportional to the logarithm of the product of the sequences, and hence, for the purpose of estimating pairwise statistical significance, we make the parameter ic dependent on the lengths of the sequences accordingly.

The above approach can generally give more than one hit corresponding to a single sequence in the database. To be consistent with the earlier definition of $PairwiseStatSig$ function, we select the best (highest) score corresponding to each sequence in the database. With proper values of the parameters w and ic, it is expected that at least right half of the distribution is obtained.

Algorithm for Fast Pairwise Statistical Estimation

The two heuristics discussed in the previous section are independent of each other, and can be used both individually and collectively. Here we outline the proposed algorithm for fast pairwise statistical significance estimation using these heuristics.

Input: Sequence 1 $Seq1$ with length $len1$, Sequence 2 $Seq2$ with length $len2$, Substitution matrix S with entropy H and bit-scale c, Gap opening penalty p, Gap extension penalty r, Number of shuffles N, Derived distribution points set DDP with $|DDP| = N_{ddp}$.

Output: Pairwise statistical significance pss of pairwise alignment score pas of $Seq1$ and $Seq2$.

1. Initialization
   (a) $minlen = min(len1, len2)$
   (b) $\mu = c \times \log_2(len1 \times len2)$
   (c) $fc = min(0.5 \times minlen \times H, 0.75 \times \mu)$ #minimum chain score
   (d) $ic = fc/2$ #minimum segment score
   (e) $w = (minlen < 50)?2 : 3$ #word size
2. $pas = SWAlignmentScore(Seq1, Seq2, S, p, r)$ #get pairwise alignment score using Smith-Waterman algorithm
3. Construct empirical score distribution
   (a) $N_{effective} = N/N_{ddp}$
   (b) $DB = Seq2DB(Seq2, N_{effective})$ #create database of N shuffles of $Seq2$
   (c) $scores[\,] = (useDBSearchHeuristic)\,?$
   $DBSearch(Seq1, DB, S, p, r, w, ic, fc) : SWAlignmentScores(Seq1, DB, S, p, r)$ #get up to $N_{effective}$ scores either by searching $Seq1$ against DB with scoring scheme SC, word size w, segment score cutoff ic and chain score cutoff fc, or by aligning $Seq1$ with all $N_{effective}$ sequences in database DB using Smith-Waterman algorithm
   (d) $Hist = createHistogram(scores, DDP)$ #for each available score (up to $N_{effective}$), add N_{ddp} alignment scores ($s + DDP_i$, $1 \le i \le N_{ddp}$) to the histogram
4. Perform a censored-maximum likelihood fitting of the histogram
   (a) $peak = Peak(Hist)$ #find the peak of the histogram
   (b) $peak = peak - min_i\{DDP_i\}$ $1 \le i \le N_{ddp}$ #correct for error in peak due to DDP
   (c) if $nRightScores(Hist, peak) < 100$, $\{N = 2 \times N;$ *goto step* $3\}$ #increase N if have too less scores to fit
   (d) $[K, \lambda] = EVDCensoredMLFit(Hist, peak)$ #get statistical parameters K and λ
5. $pss = 1 - exp(-Kmn * \exp^{-\lambda \times pas})$ #pairwise statistical significance

Experiments and Results

In this section we present the timing and retrieval accuracy results of fast pairwise statistical significance. For the derived distribution points heuristic, we used two DDP sets with $|DDP|$ as 10 and 20, as shown in 7.1. Note that the special case of $DDP = \{0\}$ corresponds to not using the DDP heuristic for pairwise statistical significance estimation. For the database search heuristic, we used a variation of the DPS (DNA-Protein Search) program [30], which we would refer to as PPS (Protein-Protein Search). The PPS program can be thought of as a basic implementation of the fundamental database search heuristic described in the previous section, which although much less complex than BLAST, is significantly faster than currently available versions of BLAST, and is good enough for our purposes of getting the right half of the score distribution. We present results for both the heuristics both when used individually and when used together.

All experiments were performed on an Intel 2.8GHz processor. The timing and speedup results using FastPairwiseStatSig are presented in Table 7.1 and Fig. 7.2. The first row in Table 7.1 with $|DDP|$=1, PPS=0 corresponds to normal pairwise statistical significance, with both proposed heuristics disabled. The times represent the time taken to estimate the optimal pairwise alignment score of two sequences of length around 250 and its pairwise statistical significance. The substitution matrix, gap opening, and gap extension penalties used were BLOSUM62, 11, and 1 respectively (default used in BLAST), and the number of shuffles N was set to 1000. In addition to reporting the time in seconds, the execution time is also reported in Alignment Time Units (ATUs) to better visualize the speedup independent of underlying processor used. 1 ATU is defined as the time required to align two sequences of length around 250 using Smith-Waterman algorithm. When used alone, the DDP heuristic gives the expected speedup of close to $|DDP|$ (10 and 20), and the database search heuristic (hereafter referred to as PPS heuristic) gives a speedup of more than 25. When both heuristics are used together, the overall speedup is more than 200.

To evaluate the performance of FastPairwiseStatSig in terms of retrieval accuracy and compare it with PairwiseStatSig, we used the same experiment setup as used in [62], and later

Table 7.1 Execution time and Speedup with FastPairwiseStatSig. $|DDP|$=1,PPS=0 corresponds to normal pairwise statistical significance. One ATU (Alignment Time Unit) is defined as the time required to align two sequences of length 250 using Smith-Waterman algorithm

$\|DDP\|$	**PPS**	**Time (s)**	**Time (ATU)**	**Speedup**
1	0	3.405	1000	1
10	0	0.345	101.32	9.87
20	0	0.172	50.51	19.80
1	1	0.12	35.24	28.38
10	1	0.017	4.99	200.29
20	1	0.012	3.52	283.75

in [2]. A non-redundant subset of the CATH 2.3 database (Class, Architecture, Topology, and Hierarchy, [47]) available at ftp://ftp.ebi.ac.uk/ pub/software/ unix/fasta/prot_sci_04/ was selected in [62] to evaluate seven structure comparison programs and two sequence comparison programs. As described in [62], this dataset consists of 2771 domain sequences and includes 86 query sequences. This domain set is considered as a valid benchmark for testing protein comparison algorithms [56].

For each of the 86×2771 comparisons, we used all the above compared methods to estimate pairwise statistical significance. Following [62, 1, 5], Error per Query (EPQ) versus Coverage plots were used to visualize and compare the results. To create these plots, the list of pairwise comparisons was sorted based on decreasing statistical significance (increasing P-values). While traversing the sorted list from top to bottom, the coverage count 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. The 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 better curve is one which is more to the right.

The advantage of the speedup using FastPairwiseStatSig is expected to incur a loss of retrieval accuracy. In Fig. 7.3, we present the comparison results of fast pairwise statistical significance and normal pairwise statistical significance in terms of retrieval accuracy. All

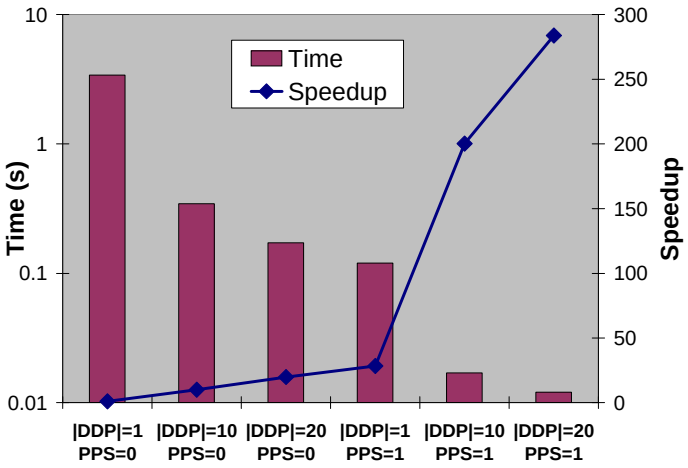

Figure 7.2 Execution time (s) and Speedup with Fast Pairwise Statistical Significance. $|DDP|$=1,PPS=0 corresponds to normal pairwise statistical significance with both the proposed heuristics disabled. A speedup of more than 200 is achieved by using both the heuristics together.

the curves are quite close to each other (except the curve for $|DDP|$=20,PPS=1). This indicates that the heuristics proposed in this work speed up the pairwise statistical significance estimation without a significant loss of retrieval accuracy.

Conclusion and Future Work

In this paper, we propose, implement and incorporate two independent heuristics for fast pairwise statistical significance estimation, which takes advantage of the nature of pairwise statistical significance estimation process. The two heuristics have been shown to give significant speedup of up to 200 without significant loss of retrieval accuracy, which is expected to be extremely useful in the wide variety of applications based on sequence comparison.

Future work includes application of the proposed heuristics for estimating pairwise statistical significance using position-specific substitution matrices, and improvement of the proposed heuristics to further speedup the pairwise statistical significance estimation process using intelligent methods to select heuristic parameters, especially the set DDP, and word size w. Fast pairwise statistical significance can also be used to design a database search method to

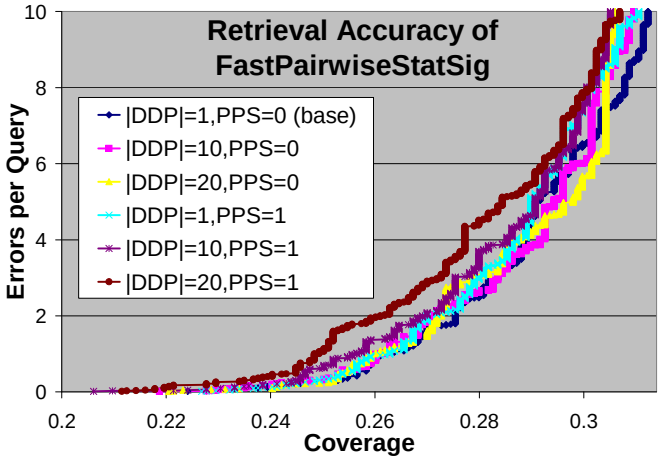

Figure 7.3 Comparison of FastPairwiseStatSig and PairwiseStatSig in terms of retrieval accuracy. $|DDP|$=1,PPS=0 corresponds to normal pairwise statistical significance with both the proposed heuristics disabled. All the curves are quite close to each other except the curve for $|DDP|$=20, PPS=1 which gives the maximum speedup. FastPairwiseStatSig performs comparable to PairwiseStatSig at least up to $|DDP|$=10,PPS=1 which gives a speedup of 200.

recover the hits missed by BLAST.

Acknowledgment

The authors would like to thank Dr. Volker Brendel for helpful discussions and providing links to the data.

8. CONCLUSIONS

In this research, we have made significant contributions to "accurately and quickly estimate the statistical significance of pairwise local sequence alignment for the purpose of identifying related sequences by using computational, statistical, and heuristic methods". This has been done using sequence-specific strategies for pairwise sequence alignment and pairwise statistical significance estimation. Sequence-specific sequence comparison indeed improves retrieval accuracy as it is evident from the fact that retrieval accuracy increases as the sequence comparison process is made more and more specific to the sequence pair being compared. Using pairwise statistical significance to refine the results of a fast database search program like PSI-BLAST, and the design of suitable heuristics to make the estimation process faster also makes it practical for large number of sequence pairs.

Given the all-pervasive utility of sequence comparison in many bioinformatics applications like database search, protein structure and function identification, multiple sequence alignment, phylogenetic tree construction, etc., this research has opened up multiple avenues of applications for future work. Apart from improving upon the methods described in this work, this research is expected to motivate researchers to develop efficient methods for constructing more accurate position-specific substitution matrices, which can greatly aid in sequence comparison applications. Fast pairwise statistical significance estimation can be used to develop an efficient database search method. Pairwise statistical significance can also be used to construct the guide tree in phylogenetic tree construction and for progressive multiple sequence alignment.

LIST OF PUBLICATIONS

Journal Publications:

1. Ankit Agrawal and Xiaoqiu Huang, PSIBLAST_PairwiseStatSig: Reordering PSI-BLAST Hits Using Pairwise Statistical Significance, Bioinformatics, 2009 25(8):1082-1083.

2. Ankit Agrawal and Xiaoqiu Huang, Pairwise Statistical Significance of Local Sequence Alignment Using Multiple Parameter Sets and Empirical Justification of Parameter Set Change Penalty, BMC Bioinformatics 2009, 10 (Suppl 3): S1.

3. Ankit Agrawal, Volker P. Brendel and Xiaoqiu Huang, Pairwise Statistical Significance and Empirical Determination of Effective Gap Opening Penalties for Protein Local Sequence Alignment, IJCBDD, 2008, 1(4):347-367.

Peer-reviewed Conference Publications:

1. Ankit Agrawal and Xiaoqiu Huang, Conservative, Non-Conservative and Average Pairwise Statistical Significance of Local Sequence Alignment, IEEE BIBM 2008, Philadelphia, PA, USA, Nov. 3-5, 2008 pp. 433-436.

2. Ankit Agrawal and Xiaoqiu Huang, Pairwise Statistical Significance of Local Sequence Alignment Using Multiple Parameter Sets, ACM DTMBIO 2008, Napa Valley, CA, USA, Oct. 30, 2008, pp. 53-60.

3. Ankit Agrawal, Volker Brendel and Xiaoqiu Huang, Pairwise Statistical Significance Versus Database Statistical Significance for Local Alignment of Protein Sequences, ISBRA 2008, Atlanta, GA, May 7, 2008, LNCS (LNBI), Springer, Heidelberg, vol. 4983, pp. 50-61.

Under Review/In Preparation:

1. Ankit Agrawal and Xiaoqiu Huang, Fast Pairwise Statistical Significance Estimation Using Derived Distribution Points and Database Search Heuristics, in preparation, 2009.

2. Ankit Agrawal and Xiaoqiu Huang, Pairwise Statistical Significance of Local Sequence Alignment Using Sequence-Specific and Position-Specific Substitution Matrices, IEEE TCBB, under review, 2008.

APPENDIX A. SMITH-WATERMAN ALGORITHM FOR PAIRWISE LOCAL SEQUENCE ALIGNMENT

The Smith-Waterman algorithm [63] is a popular algorithm for performing local sequence alignment of two nucleotide or protein sequences for a given gap penalty function. Under an affine gap penalty model, the algorithm can be described as follows.

Let $A = a_1, a_2, ..., a_m$ and $B = b_1, b_2, ..., b_n$ be two sequences of length m and n. The optimal local alignment of A and B is the global alignment of the subsequences α of A and β of B, whose similarity is maximal, in terms of a scoring scheme (pairwise substitution scores and gap penalties).

A local alignment of A and B consists of two types of configurations: substitutions and gaps. A substitution associates a residue of A with a residue of B. A gap consists only of residues from one sequence with each residue associated with the symbol -. There are two kinds of gaps. A deletion gap with respect to sequence A consists only of residues from A and an insertion gap with respect to sequence A consists only of residues from B. Let the scoring scheme be as follows. Substitution matrix S is a square matrix consisting of scores for each possible substitution between residue pairs, with $S(a, b)$ representing the substitution score of aligning residue a with residue b. Let q be the non-negative gap opening penalty, and r be the non-negative gap extension penalty, so that the score of a gap of length k is $-(q + k \times r)$.

Let $A_i = a_1, a_2, ...a_i$ and $B_j = b_1, b_2, ..., b_j$ be initial segments of A and B of length i and j respectively. Define $V(i, j)$ to be the score of the optimal local alignment of A_i and B_j. Define $G(i, j)$ to be the score of the optimal local alignment of A_i and B_j where a_i and b_j are aligned with each other. Define $I(i, j)$ to be the score of the optimal local alignment of A_i and B_j that end with an insertion gap with respect to A. Similarly, define $D(i, j)$ to

be the score of the optimal local alignment of A_i and B_j that end with an deletion gap with respect to A. Then, the following recurrences are used to calculate the optimal local alignment:

Base Conditions:

$$\begin{aligned} V(0,0) &= 0 \\ V(i,0) &= 0 \quad \forall i \\ V(0,j) &= 0 \quad \forall j \\ I(i,0) &= -q \quad \forall i \geq 0 \\ D(0,j) &= -q \quad \forall j \geq 0 \end{aligned}$$

Recurrence relations:

$$\begin{aligned} V(i,j) &= max\left\{G(i,j), I(i,j), D(i,j), 0\right\} \\ G(i,j) &= max\left\{V(i-1,j-1) + S(a_i, b_j)\right\} \\ I(i,j) &= max\left\{I(i,j-1) - r, V(i,j-1) - q - r\right\} \\ D(i,j) &= max\left\{D(i-1,j) - r, V(i-1,j) - q - r\right\} \end{aligned}$$

The score of the optimal local alignment of A and B is given by $V(i',j') = max_{1 \leq i \leq m, 1 \leq j \leq n} V(i,j)$.

Both time and space complexity of the algorithm is $O(mn)$. The actual alignment can be calculated by following a trace-back procedure from $V(i',j')$ as described in [63]. The space-complexity can be reduced to $O(min\{m,n\})$ using a divide-and-conquer strategy developed by Hirschberg [29] after identifying the starting and ending indices of the optimal local alignment.

APPENDIX B. Supplementary Notes for PSIBLAST_PairwiseStatSig: REORDERING PSI-BLAST HITS USING PAIRWISE STATISTICAL SIGNIFICANCE

Pairwise Statistical Significance

Consider the pairwise statistical significance described in [2] 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 the 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. Using this function two times with different ordering of sequence inputs, non-conservative pairwise statistical significance was introduced in [3]. Let

$$S1 = PairwiseStatSig(Seq1, Seq2, SC_1, N)$$
$$S2 = PairwiseStatSig(Seq2, Seq1, SC_2, N)$$

Then, non-conservative pairwise statistical significance is defined as $minS1, S2$. SC_1 and SC_2 signifies that a scoring scheme specific to $Seq1$ and $Seq2$ can be used to estimate $S1$ and $S2$ respectively, since $Seq1(Seq2)$ is not shuffled during estimation of $S1(S2)$. Pairwise statistical significance using sequence-specific and position-specific substitution matrices [5] indicates that best results are obtained by using position-specific substitution matrices as it uses maximal sequence-specific information.

Evaluation methodology

Errors per Query vs. Coverage curves

Plotting Errors per Query vs. Coverage curves to represent and compare the results of a homology detection experiment is one of the standard methods as used in [16, 62]. For plotting such curves, the list of all pairwise comparisons are sorted in decreasing order of statistical significance (increasing order of E-values/P-values). Subsequently, the list is traversed from top to bottom and the count of true homologs detected and errors incurred so far is kept track of at every point. Finally coverage (fraction of true homologs detected) and errors per query (number of errors divided by total number of queries) are plotted on x and y axis respectively. Ideally, all true homologs should be at the top of the list, which would correspond to a straight line on the x-axis, as 100% coverage is achieved at 0 EPQ. Therefore, the more the curve is towards the right, the better it is. The comparison results with this evaluation strategy are presented in the main manuscript.

Average Error Rate vs. Coverage curves

Rather than constructing a single list of comparisons combining the results from all the queries, another approach is to analyze the list of each query separately, and aggregate the results. For this approach, the number of errors incurred for each query at different levels of attained coverage was calculated, and the average number of errors incurred at different coverage levels was plotted. To the best of our knowledge, this method has not been used in this form in any previous work to compare performance of sequence-comparison/database search programs.

Fig. B.1 (a) gives the avg. error rate vs. coverage curves comparing the results of BLAST and BLAST_PairwiseStatSig, and Fig B.1 (b) gives the corresponding curves comparing the results for PSI-BLAST, PSIBLAST_PairwiseStatSig, and PSIBLAST_NCPairwiseStatSig (re-ordering PSI-BLAST results using non-conservative pairwise statistical significance). The curves are not non-decreasing since most of the queries saturate at a certain coverage level, i.e., their maximum possible coverage is achieved. This is because as explained in the main manuscript, the results of all the approaches compared here are upper-bounded by the number of true homologs detected by BLAST/PSI-BLAST. For instance, suppose a query has10

homologs in the searched database, and the PSI-BLAST search gives 10 hits, out of which 5 are true homologs and 5 are erroneous hits. Therefore, the maximum possible coverage for this query is 50%, and reordering the hits by pairwise statistical significance can only alter how soon this maximum coverage is attained. Therefore, at each coverage level, only the errors from unsaturated queries are used to compute the average error. Thus, because of different denominators used to calculate the average error, the curves are not necessarily supposed to be non-decreasing Since the average error rate is plotted on the y-axis, the lower the curve, the better it is. It is important to note that such a comparison is valid in this case since all the methods compared on a single plot have exactly same saturated queries at different coverage levels. This is because the maximum coverage for a given query is determined by the number of true homologs detected in the BLAST/PSI-BLAST searches. But if a method were to have the ability to extract more true homologs missed by BLAST/PSI-BLAST, then this comparison methodology would not have been fair.

As expected, these curves show more irregularity than EPQ vs. Coverage curves with the proposed method doing better than BLAST/PSI-BLAST at certain coverage levels and worse at other coverage levels. However, for most coverage levels (0.01 to 0.5, 0.68 to 0.87 for BLAST_PairwiseStatSig, and 0.01 to 0.4, 0.5 to 1.0 for PSIBLAST_PairwiseStatSig), the proposed method performs comparable to or better than BLAST/PSI-BLAST.

As mentioned earlier, at each coverage level, the average of error counts is calculated only across those queries which are not saturated, which means that different denominators are used at different coverage levels to calculate the average error. Thus, some coverage levels may have very few unsaturated queries, and the resulting curve may be highly susceptible to noise and may not show a definitive trend, as found here. Although the results from this evaluation methodology by and large support the results from the earlier evaluation methodology, these curves do not seem to be as conclusive of the superiority of one method over the other as EPQ vs. Coverage curves.

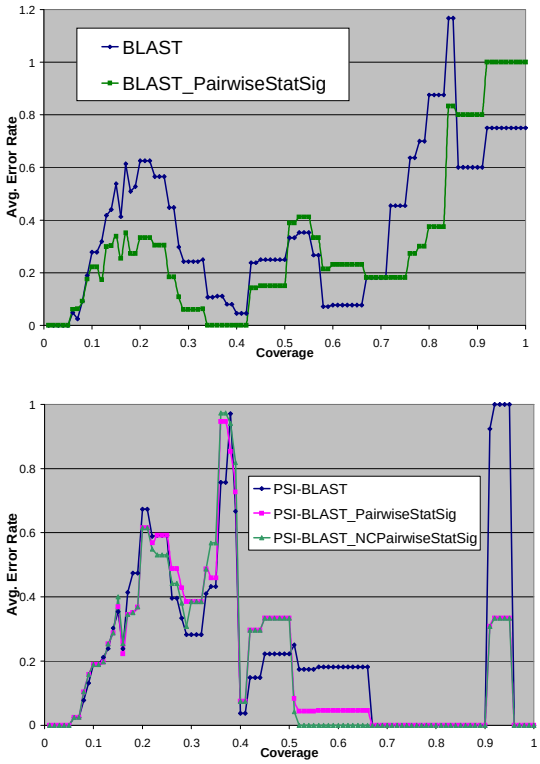

Figure B.1 Avg. Error Rate vs. Coverage curves.

Bibliography

[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, V. P. Brendel, and X. Huang. Pairwise Statistical Significance and Empirical Determination of Effective Gap Opening Penalties for Protein Local Sequence Alignment. *International Journal of Computational Biology and Drug Design*, 1(4):347–367, 2008.

[3] A. Agrawal and X. Huang. Conservative, non-conservative and average pairwise statistical significance of local sequence alignment. In *Proc. of IEEE Intl. Conf. on Bioinformatics and Biomedicine, BIBM*, pages 433–436, 2008.

[4] A. Agrawal and X. Huang. Pairwise statistical significance of local sequence alignment using multiple parameter sets. In *Proc. of ACM 2nd Intl. Workshop on Data and Text Mining in Bioinformatics, DTMBIO*, pages 53–60, 2008.

[5] A. Agrawal and X. Huang. Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. 2008. under review.

[6] A. Agrawal and X. Huang. Pairwise statistical significance of local sequence alignment using substitution matrices with sequence-pair-specific distance. In *Proc. of Intl. Conf. on Information Technology, ICIT*, pages 94–99, 2008.

[7] A. Agrawal and X. Huang. Pairwise Statistical Significance of Local Sequence Align-

ment Using Multiple Parameter Sets and Empirical Justification of Parameter Set Change Penalty. *BMC Bioinformatics*, 10(Suppl 3):S1, 2009.

[8] A. Agrawal and X. Huang. PSIBLAST_PairwiseStatSig: reordering PSI-BLAST hits using pairwise statistical significance. *Bioinformatics*, 25(8):1082–1083, 2009.

[9] S. F. Altschul, M. S. Boguski, W. Gish, and J. C. Wootton. Issues in searching molecular sequence databases. *Nature Genetics*, 6(2):119–129, 1994.

[10] S. F. Altschul, R. Bundschuh, R. Olsen, and T. Hwa. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Research*, 29(2):351–361, 2001.

[11] S. F. Altschul and W. Gish. Local Alignment Statistics. *Methods in Enzymology*, 266:460–80, 1996.

[12] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[13] 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.

[14] T. L. Bailey and M. Gribskov. Estimating and Evaluating the Statistics of Gapped Local-Alignment Scores. *Journal of Computational Biology*, 9(3):575–93, 2002.

[15] S. E. Brenner. Practical database searching. *Trends in Biotechnology*, 16(1):9–12, 1998.

[16] S. E. Brenner, C. Chothia, and T. J. P. Hubbard. Assessing Sequence Comparison Methods with Reliable Structurally Identified Distant Evolutionary Relationships. *Proceedings of the National Academy of Sciences, USA*, 95(11):6073–6078, 1998.

[17] P. Bucher and K. Hofmann. A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 44–51. AAAI Press, 1996.

[18] R. Bundschuh. Rapid Significance Estimation in Local Sequence Alignment with Gaps. In *RECOMB '01: Proceedings of the fifth annual International Conference on Computational biology*, pages 77–85, New York, NY, USA, 2001. ACM.

[19] K.-M. Chao. Calign: aligning sequences with restricted affine gap penalties. *Bioinformatics*, 15(4):298–304, 1999.

[20] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, pages 345–352. National Biomedical Research Foundation., Washington DC, 1978.

[21] S. R. Eddy. Multiple Alignment Using Hidden Markov Models. In C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 114–120. AAAI Press, Menlo Park, 1995.

[22] S. R. Eddy. Maximum likelihood fitting of extreme value distributions. 1997. unpublished work.

[23] S. R. Eddy. Where did the blosum62 alignment score matrix come from? *Nat Biotechnol*, 22(8):1035–1036, August 2004.

[24] P. Fariselli, I. Rossi, E. Capriotti, and R. Casadio. The WWWH of remote homolog detection: The state of the art. *Brief Bioinform*, 8(2):78–87, 2007.

[25] S. Grossmann and B. Yakir. 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.

[26] E. J. Gumbel. *Statistics of Extremes*. Columbia University Press, New York, 1958.

[27] A. K. Hartmann. Sampling Rare Events: Statistics of Local Sequence Alignments. *Physical Review E*, 65(5):056102, 2002.

[28] S. Henikoff and J. Henikoff. Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[29] D. S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, 18(6):341–343, 1975.

[30] X. Huang. Fast comparison of a dna sequence with a protein sequence database. *Microbial & Comparative Genomics*, 1:281–291, 1996.

[31] X. Huang and D. L. Brutlag. Dynamic Use of Multiple Parameter Sets in Sequence Alignment. *Nucleic Acids Research*, 35(2):678–686, 2007.

[32] X. Huang and K.-M. Chao. A Generalized Global Alignment Algorithm. *Bioinformatics*, 19(2):228–233, 2003.

[33] X. Huang and W. Miller. A Time-efficient Linear-space Local Similarity Algorithm. *Advances in Applied Mathematics*, 12(3):337–357, 1991.

[34] S. Karlin and S. F. Altschul. 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.

[35] S. Kotz and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*, chapter 1, pages 3–4. Imperial College Press, London, UK, 2000.

[36] M. Kschischo, M. Lässig, and Y.-K. Yuc. Toward an Accurate Statistics of Gapped Alignments. *Bulletin of Mathematical Biology*, 67:169–191, 2004.

[37] E. L. Lehmann. *Testing Statistical Hypotheses*. Wiley, New York, 1986.

[38] M. Li, B. Ma, D. Kisman, and J. Tromp. PatternHunter II: Highly Sensitive and Fast Homology Search. *Journal of Bioinformatics and Computational Biology*, 2(3):417–439, 2004. Early version in GIW 2003.

[39] B. Ma, J. Tromp, and M. Li. PatternHunter: faster and more sensitive homology search . *Bioinformatics*, 18(3):440–445, 2002.

[40] A. Y. Mitrophanov and M. Borodovsky. Statistical Significance in Biological Sequence Analysis. *Briefings in Bioinformatics*, 7(1):2–24, 2006.

[41] R. Mott. Accurate Formula for P-values of Gapped Local Sequence and Profile Alignments. *Journal of Molecular Biology*, 300:649–659, 2000.

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

[43] R. Mott and R. Tribe. Approximate Statistics of Gapped Alignments. *Journal of Computational Biology*, 6(1):91–112, 1999.

[44] R. F. Mott. Maximum-likelihood Estimation of the Statistical Distribution of SmithWaterman Local Sequence Similarity Scores. *Bulletin of Mathematical Biology*, 54:59–75, 1992.

[45] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, March 1970.

[46] R. Olsen, R. Bundschuh, and T. Hwa. Rapid assessment of extremal statistics for gapped local alignment. In *Proc. of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 211–222. AAAI Press, 1999.

[47] 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.

[48] M. Pagni and C. V. Jongeneel. Making Sense of Score Statistics for Sequence Alignments. *Briefings in Bioinformatics*, 2(1):51–67, 2001.

[49] W. R. Pearson. Effective Protein Sequence Comparison. *Methods in Enzymology*, 266:227–259, 1996.

[50] W. R. Pearson. Empirical Statistical Estimates for Sequence Similarity Searches. *Journal of Molecular Biology*, 276:71–84, 1998.

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

[52] W. R. Pearson and D. J. Lipman. Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Sciences, USA*, 85(8):2444–2448, 1988.

[53] W. R. Pearson and T. C. Wood. Statistical Significance in Biological Sequence Comparison. In D. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 39–66. Chichester, UK: Wiley, 2001.

[54] A. Poleksic, J. F. Danzer, K. Hambly, and D. A. Debe. Convergent Island Statistics: A Fast Method for Determining Local Alignment Score Significance. *Bioinformatics*, 21(12):2827–2831, 2005.

[55] J. T. Reese and W. R. Pearson. Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics*, 18:1500–1507(8), November 2002.

[56] J. Rocha, F. Rosselló, and J. Segura. Compression Ratios Based on the Universal Similarity Metric Still Yield Protein Distances far from CATH Distances. *CoRR*, abs/q-bio/0603007, 2006.

[57] A. A. Schäffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-based Statistics and Other Refinements. *Nucleic Acids Research*, 29(14):2994–3005, 2001.

[58] P. H. Sellers. Pattern Recognition in Genetic Sequences by Mismatch Density. *Bulletin of Mathematical Biology*, 46(4):501–514, 1984.

[59] S. Sheetlin, Y. Park, and J. L. Spouge. The Gumbel Pre-factor k for Gapped Local Alignment can be Estimated From Simulations of Global Alignment. *Nucleic Acids Research*, 33(15):4987–4994, 2005.

[60] D. Siegmund and B. Yakir. *Approximate P-values for Local Sequence Alignments*, 2000.

[61] D. Siegmund and B. Yakir. *Correction: approximate P-values for Local Sequence Alignments (vol 28, pg 657, 2000)*, 2003.

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

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

[64] T. A. Tatusova and T. L. Madden. Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.*, 174:247–250, 1999.

[65] M. S. Waterman and M. Vingron. Rapid and Accurate Estimates of Statistical Significance for Sequence Database Searches. *Proceedings of the National Academy of Sciences, USA*, 91(11):4625–4628, 1994.

[66] S. Wolfsheimer, B. Burghardt, and A. K. Hartmann. Local Sequence Alignments Statistics: Deviations from Gumbel Statistics in the Rare-event Tail. *Algorithms for Molecular Biology*, 2(9), 2007.

[67] Y.-K. Yu and S. F. Altschul. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, 21(7):902–911, 2005.

[68] Y.-K. Yu, E. M. Gertz, R. Agarwala, A. A. Schäffer, and S. F. Altschul. Retrieval Accuracy, Statistical Significance and Compositional Similarity in Protein Sequence Database Searches. *Nucleic Acids Research*, 34(20):5966–5973, 2006.