Pairwise Statistical Significance of Local Sequence Alignment Using Sequence-Specific and Position-Specific Substitution Matrices

Ankit Agrawal and Xiaoqiu Huang

Abstract—Pairwise sequence alignment is a central problem in bioinformatics, which forms the basis of various 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 gives promising results as an alternative to database statistical significance for getting individual significance estimates of pairwise alignment scores. The improvement was mainly attributed to making the statistical significance estimation process more 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 on a benchmark database 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 (without pretrained PSSMs), and SSEARCH on a benchmark database, but with pretrained PSSMs, PSI-BLAST results are significantly better. Further, using position-specific substitution matrices for estimating pairwise statistical significance gives significantly better results even than PSI-BLAST using pretrained PSSMs.

Index Terms—Database statistical significance, homologs, pairwise statistical significance, position-specific scoring matrices (PSSMs), sequence alignment, substitution matrices.

1 Introduction

Sequence alignment is an underlying application in the analysis and comparison of DNA and protein sequences [1], [2], [3]. 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, etc. There exist several programs for sequence alignment that use popular algorithms [4], [5], [6] , or their heuristic versions [7], [3], [8], [9], [10]. A lot of enhancements in alignment program features are also available [11], [12], [13] using difference blocks and multiple scoring matrices, in an attempt to capture some more biological features in the alignment algorithm.

1.1 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, and sequence compositions [14]. This implies that it is possible to have two scores x and y with $x < y$, but x more statistically significant than y. For instance, two sequences of length 50 may produce a highly statistically significant score of 75, whereas another two sequences each of length 250 may have an optimal alignment score of 100, which may not be statistically significant. Therefore, instead of using the alignment 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 [14], [15].

The knowledge of accurate statistics for score distribution of ungapped alignments is available [16]. However, there is no rigorous statistical theory for the gapped alignment score distribution yet and for score distributions from enhanced alignment programs using additional features like difference blocks [12] or multiple parameter sets [13]. Accurate estimation of statistical significance of gapped sequence alignment has attracted a lot of attention in the recent years [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. There exist a couple of good

• *The authors are with the Department of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011-1041. E-mail: {ankitag, xqhuang}@iastate.edu.*

Manuscript received 29 Sept. 2008; revised 1 July 2009; accepted 9 Aug. 2009; published online 25 Sept. 2009.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-09-0171. Digital Object Identifier no. 10.1109/TCBB.2009.69.

starting points for statistically describing gapped alignment score distributions for simple scoring schemes [30], [31], but a complete mathematical description of the optimal score distribution remains far from reach [31]. There exist many excellent reviews on statistical significance in sequence comparison in the literature [32], [33], [14], [34].

1.2 Database Statistical Significance versus Pairwise Statistical Significance

The hits reported by common database search programs like BLAST, FASTA, SSEARCH, and PSI-BLAST are evaluated by database statistical significance, which is in general dependent on the size and composition of the database being searched. In the last few years, there have been considerable improvements to the BLAST and PSI-BLAST programs [24], [35], [27], which have been shown to improve retrieval accuracy of database searches by using composition-based statistics and other enhancements.

An alternative method to estimate statistical significance of a pairwise alignment is to estimate pairwise statistical significance, which is database-independent and sequence-specific. Recently, a study of pairwise statistical significance and its comparison with database statistical significance was conducted [28] wherein various approaches to estimate pairwise statistical significance like ARIADNE [21], PRSS [8], censored-maximum-likelihood fitting [36], and linear regression fitting [13] were compared to find that maximum likelihood fitting with censoring left of peak (described as type-I censoring in [36]) 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 (without pretrained PSSMs) on a benchmark database and comparable to SSEARCH, but PSI-BLAST gives significantly better results by using pretrained PSSMs (position-specific scoring matrices). In another related work [29], a simple extension of pairwise statistical significance was shown to be better than ordinary pairwise statistical significance, where the concept of nonconservative pairwise statistical significance was introduced. Pairwise statistical significance using multiple parameter sets [37] and sequence-pair-specific distanced substitution matrices [38] has also been explored, which, in some cases, give slightly better results than original pairwise statistical significance, but not comparable to the methods described in this paper. The brief background of the relevant details on pairwise statistical significance is presented in the next section.

1.3 Relevance

It is a well-known fact that almost everything in bioinformatics depends on the interrelationship between sequence, structure, and function (all encapsulated in the term "relatedness"), which is far from being well-understood. With sequencing becoming more and more easy and affordable, there is an increasing deluge of sequence data in the public domain, for the analysis of which, computational sequence comparison techniques would have to play a key role. In the progressive march toward this goal, 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, which, in turn, can be very useful in any application, which employs sequence alignment for examining the relatedness of sequence pairs. 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 more sophisticated alignment programs, therefore, is not expected to be straightforward. For comparing the performance of newer alignment programs, which try to incorporate more biological aspects of sequence alignment, accurate estimates of pairwise statistical significance can be extremely useful. 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.

1.4 Contributions

Earlier work on pairwise statistical significance has shown it to be a promising alternative to database statistical significance when used with standard substitution matrices owing to it being more specific to the sequence-pair being aligned. In this paper, we attempt to make the sequence comparison more sequence-specific by using sequence-specific and position-specific substitution matrices (SSSMs and PSSMs). To construct an SSSM for a given sequence, we used a simple intuitive method aimed at constructing substitution matrices with some sequence-specific contribution, since we could not find any such related program in the public domain. To construct a PSSM for a given sequence, we used the popularly used database search program PSI-BLAST to search the given sequence against a large database (nr protein database provided by BLAST), which naturally constructs a PSSM for the query sequence. To effectively separate the influence of pairwise statistical significance and SSSMs/PSSMs, we also conducted homology detection experiments with SSSMs/PSSMs using both pairwise statistical significance and database statistical significance as reported by popular database search programs BLAST, PSI-BLAST, and SSEARCH. Results provide clear empirical evidence that as sequence comparison is made more and more sequence-specific by using SSSMs and PSSMs, pairwise statistical significance provides significantly better estimates of statistical significance as compared to database statistical significance for the biologically relevant and practical application of homology detection, which can be measured in terms of retrieval accuracy.

Although the comparison results in this paper show that the proposed methods outperform the traditional database search programs like BLAST, PSI-BLAST (pairwise statistical significance with SSSMs performs better than PSI-BLAST used directly on benchmark test database (a subset of CATH database first used in [39]) with default parameters (e-value threshold of 10 and maximum five rounds), but with pretrained PSI-BLAST PSSMs (obtained from database search on a larger database (nr)), PSI-BLAST performs significantly better; pairwise statistical significance with same PSSMs performs significantly better than PSI-BLAST

with pretrained PSSMs), and SSEARCH, the proposed methods, nevertheless, are currently much computationally expensive to be directly used in a large database search, since it involves generation of sequence-specific empirical score distributions and subsequent curve-fitting. The intended goal of this work was to make sequence comparison more effective by making it more sequence-specific; and the central question in sequence comparison is—"How closely the given two sequences are evolutionary related?." The proposed method, as of now, can be used for any application requiring an examination of relatedness of a few sequence pairs, like small database searches, refining the results of database searches, creating distance matrices for phylogenetic tree construction, etc. Since the essential quality of a good sequence comparison strategy in all such applications would be its ability to order the sequence pairs according to biological significance (relatedness), we chose to evaluate the proposed method in terms of retrieval accuracy, which targets biological significance and not just statistical precision. As far as statistical significance accuracy is concerned, we here use the same method (censored maximum likelihood fitting) to get statistical parameters from score distributions, which was earlier found to be the best in terms of statistical significance accuracy [28], and hence do not evaluate the proposed method again in terms of statistical precision.

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, along with a brief description of relevant details on pairwise statistical significance. Section 3 presents the methods used to create sequence-specific substitution matrices and estimating pairwise statistical significance using sequence-specific and position-specific substitution matrices. Experiments and results are reported in Section 4, and finally, the conclusion and future work are presented in Section 5.

2 Background

2.1 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) [40]. The distribution of Smith-Waterman local alignment score between random, unrelated sequences is known to follow a Gumbel-type EVD [16]. 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:

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

where E is the E-value and is given by

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

For E-values less than 0.01, both E-value and P-values are very close to each other. The above formulas are valid for ungapped alignments [16], and the parameters K and λ can be computed analytically from the substitution scores and sequence compositions. For the gapped alignment, no rigorous statistical theory has yet been developed, although there exist some good starting points for the problem as mentioned before [30], [31]. Recently, researchers have also looked closely at the low probability tail distribution, and the work in [41] applied a rare-event sampling technique earlier used in [42] 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 [17], [18], [19], [21], [43], [44], [13], [28], [29].

2.2 Pairwise Statistical Significance

The pairwise statistical significance described in [28] can be understood 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 opening penalty, and gap extension penalty), 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. A simple extension of the $PairwiseStatSig$ function was presented in [29], wherein the function was used two times with different ordering of sequence inputs, and nonconservative pairwise statistical significance was introduced. Let

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

Then, nonconservative pairwise statistical significance is defined as $min\{S1, S2\}$.

3 Methods

3.1 Creating Sequence-Specific Substitution Matrix for a Given Sequence

In this section, we outline a simple method for constructing a 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 [45].

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 a large database (nr database used in our experiments) with a relatively high e-value threshold (1,000 used in our experiments) so that enough alignments can be obtained to fill up the count matrix. 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 nonzero 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 [45]. Subsequently, these probabilities can be multiplied by a suitably large integer to get a representative count matrix C. Let the count matrix thus 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. Both C_S and C_{BL62} are individually normalized to have a constant matrix sum (1,000,000 used in our experiments), so that they are compatible for addition. 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.

The simple method described above is one of the many possible approaches to get a sequence-specific substitution matrix. Although simple, it currently has several shortcomings, some of which are as follows: A general criticism of the method is that this approach assumes that the $p(a,b)$ are the probabilities of amino acids a and b being aligned within correct alignments of homologous sequences. Even though we use a large database (nr) for the BLAST search to fill up the count matrix, we are not guaranteed to always have correct homologous alignments, not only because BLAST is not optimal search method, but also because of the high e-value threshold (1,000), which may return a large number of false positives. We would still like to use a relatively high e-value threshold to collect sufficient alignments to fill up the count matrix (we got as low as 5 hits for a query (1dp5B0) even with an e-value threshold of 1,000 in our experiments). At the same time, we would not like to raise the e-value cutoff too much, since it would add to the BLAST search time. Hence, with the proposed method, we can expect BLAST to return of the order of 1,000 false positives, although it would rarely happen, since we collect only top 1,000 alignments from the BLAST search to fill up the count matrix (details presented in the Results section). Of course, to the extent that the assumed statistical model of [16] is applicable to gapped alignments, the resulting $p(a,b)$ from the false positives will converge on the $p(a,b)$ implicit in the substitution matrix used, i.e., BLOSUM62; so this simply adds additional weight to the default matrix. Nevertheless, the procedure clearly gives much greater weight to the default matrix for queries with fewer homologs in the database than to sequences with more homologs. Another aspect of the uneven distribution of hits returned by BLAST in this method is that for some "popular" query sequences, most hits would be its close homologs, which would overwhelm the amino-acid pair counts. Again, the provision of having contributions from both sequence-specific count matrix C_S and the default count matrix C_{BL62} allows for controlled sequence-specific contribution, but the issue of over-representation from very similar sequences can possibly be addressed using several techniques as will be discussed later in Section 5. These issues are not addressed here since the goal here was not to design a very effective method for constructing SSSMs, but to construct substitution matrices with some sequence-specific contribution for experimentation with pairwise statistical significance to investigate for any performance enhancement in the absence of an easily available program in the public domain to do so. Experiments were conducted using the resulting SSSMs with both pairwise statistical significance and database statistical significance, and our strategy for constructing SSSMs, even though not free from defects, is expected to influence both methods of estimating statistical significance in a similar way. The comparison results presented in the next section demonstrate the potential of the approach. However, addressing the potential problems discussed above can possibly result in an improved performance, both for pairwise statistical significance and for database statistical significance.

3.2 Pairwise Statistical Significance Using Sequence-Specific and Position-Specific Substitution Matrices

As described in the earlier section, pairwise statistical significance can be understood 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 opening penalty, gap extension penalty), and N is the number of shuffles. Since only $Seq2$ is shuffled during the significance estimation procedure, it is possible to easily use a scoring scheme specific to $Seq1$ to estimate pairwise statistical significance. Therefore, pairwise statistical significance of a pairwise alignment using sequence-specific/position-specific substitution matrix for one of the two sequences (let that be $Seq1$) can be estimated by using it in SC. Let the sequence-specific/position-specific scoring scheme specific to $Seq1$ be thus denoted by SC_1. Then, $PairwiseStatSig(Seq1, Seq2, SC_1, N)$ would denote the pairwise statistical significance estimate using sequence-specific/position-specific substitution matrix, depending on whether SC_1 is sequence-specific or position-specific.

If, however, sequence-specific/position-specific substitution matrices are available for both the sequences being compared, we can use the concept of nonconservative pairwise statistical significance to make the estimation process more specific to the sequences being aligned

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

SC_1 and SC_2 represents a scoring scheme specific to $Seq1$ and $Seq2$, respectively, which can be sequence-specific or position-specific depending on the substitution matrix. Nonconservative pairwise statistical significance using sequence-specific/position-specific substitution matrix is thus given by $min\{S1, S2\}$.

4 Experiments and Results

To evaluate the proposed approach, we used the same experiment setup as used in [39], and later in [28], [29]. A nonredundant subset of the CATH 2.3 database (Class, Architecture, Topology, and Hierarchy, [46]) available at ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/prot_sci_04/ was selected in [39] to evaluate seven structure comparison programs and two sequence comparison programs. This data set consists of 2,771 domain sequences and includes 86 query sequences, and is considered as a valid benchmark for testing protein comparison algorithms [47].

Following [39], [28], [29], 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 percent 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.

4.1 Pairwise Statistical Significance Using SSSMs

Sequence-specific substitution matrices were obtained for each of the 2,771 sequences in the database using the method described in the previous section. We used the BLAST program (version 2.2.17) to query the 2,771 sequences against the nonredundant protein database (nr) provided with BLAST programs with a relatively high e-value threshold of 1,000 ($-\mathrm{e}\ 1000$) so that we can collect enough alignments for filling the count matrix. Further, to view 1,000 best alignments in the output, the "$-\mathrm{b}\ 1000$" 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. The number of shuffles to generate the empirical distribution was set to 1,000. The gap opening and gap extension penalties were set to 10 and 2, respectively, which are the default in FASTA and SSEARCH programs, that also use substitution matrices in 1/3-bit scale.

4.1.1 Using SSSMs at Different Levels of Sequence-Specific Contribution

We experimented with sequence-specific substitution matrices at different levels of sequence-specific contribution. Here, we used nonconservative pairwise statistical significance, which has been shown to be more effective compared to other variants of pairwise statistical significance [29], since SSSMs for all the sequences were available.

The EPQ versus Coverage curves for different levels of sequence-specific contribution α are presented in Fig. 1. The left-most curve is for $\alpha = 0$, i.e., 0 percent 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 curves are quite close to each other, and it is difficult to determine the best value of α for this data set from this graph. Therefore, we further use Coverage versus Sequence-specific contribution plots at different error levels to determine the optimal value of α for this data set, as presented in Fig. 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. For example, in our experiments, we got as low as 5 hits for a query (1dp5B0) even with an e-value threshold of 1,000. Thus, some contribution of amino-acid pair counts from the default count matrix is helpful. From Fig. 2, we can determine a range of α values, which gives the best performance. Clearly, for this data set, 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 data set. 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

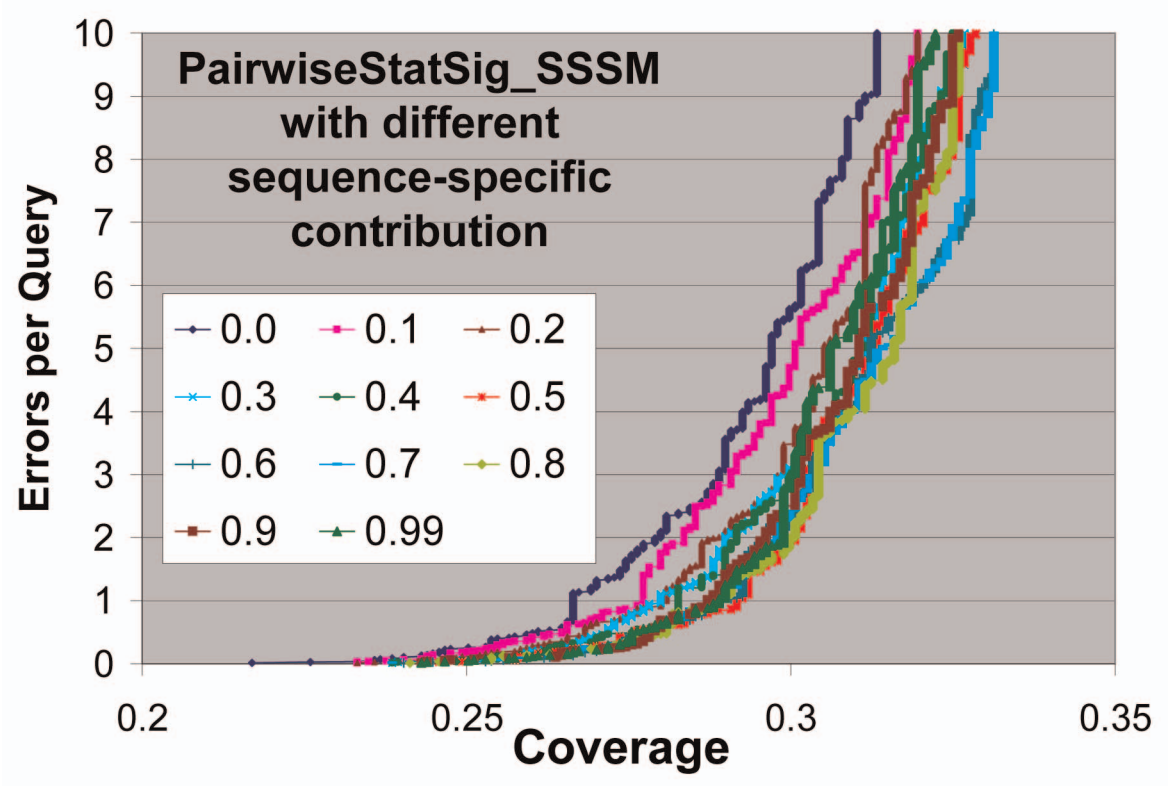

Fig. 1. EPQ versus coverage plot for different levels of sequence-specific contribution α. The leftmost curve is for $\alpha = 0$, i.e., 0 percent sequence-specific contribution, which corresponds to using a general substitution matrix (BLOSUM62). For all values of $\alpha > 0$, the coverage performance 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.

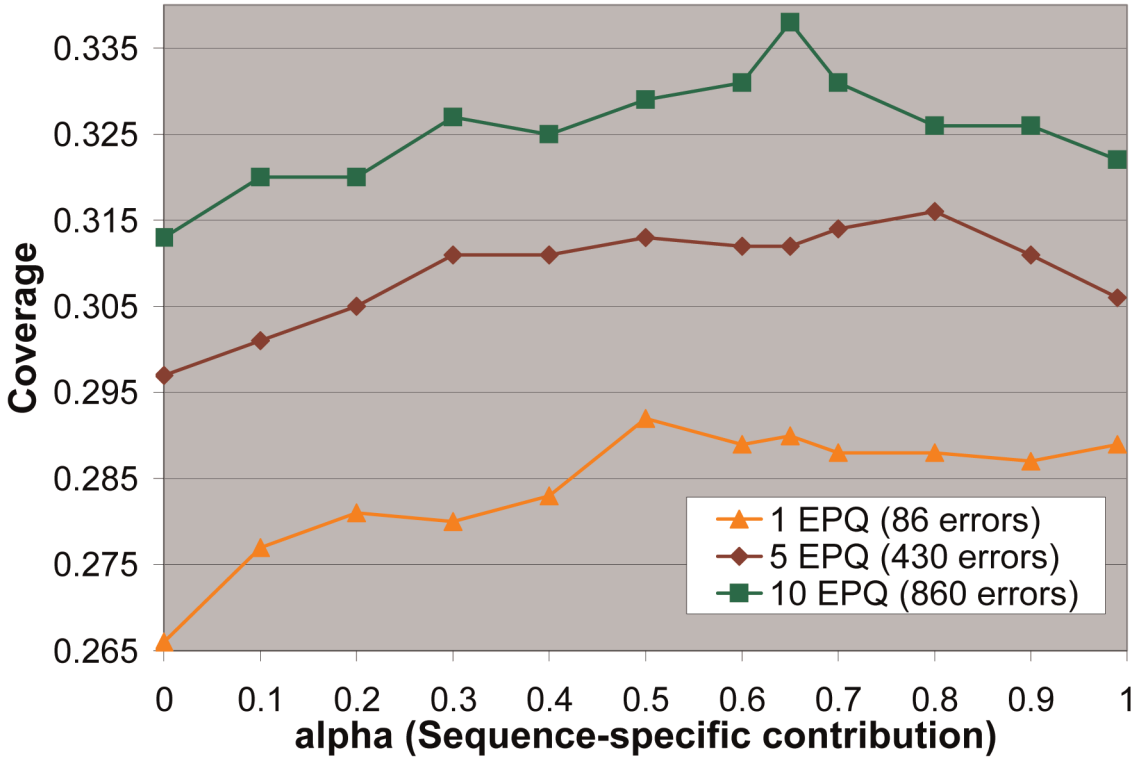

Fig. 2. Coverage versus sequence-specific contribution (α) plot for three different error levels. The coverage performs increases as α increases, reaches a maximum, and decreases a little for high values of α. $\alpha = 0.65$ is identified to be the best value for the benchmark data set used.

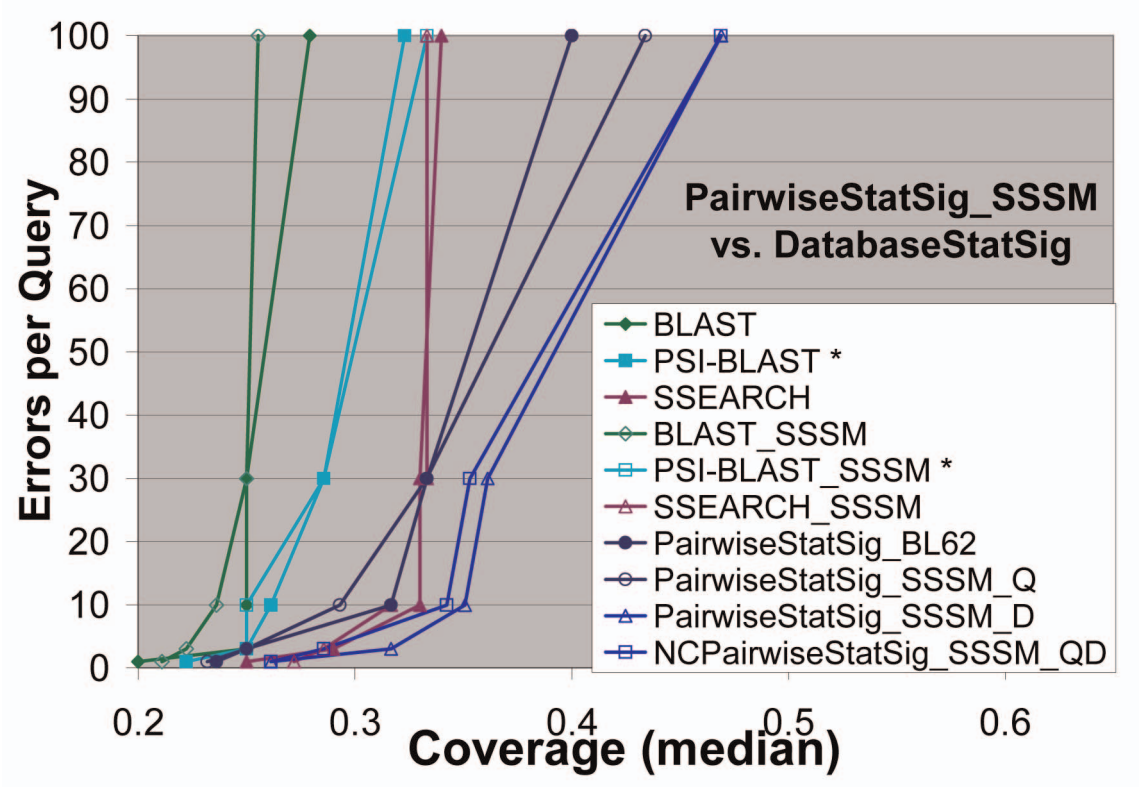

Fig. 3. Pairwise statistical significance versus database statistical significance using both standard and sequence-specific substitution matrices. PSI-BLAST* and PSI-BLAST_SSSM* denote that these results were obtained by using PSI-BLAST directly on the test database with default BLOSUM62 matrix and SSSMs for the queries, respectively, as starting points and not pretrained PSSMs. The last three curves are corresponding to using pairwise statistical significance using SSSMs only for query sequences; only for database sequences; and both for query and database sequences (using nonconservative pairwise statistical significance). Using SSSMs for estimating pairwise statistical significance is significantly better than database statistical significance using BLAST, PSI-BLAST, and SSEARCH on the benchmark database.

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

4.1.2 Comparison with Database Statistical Significance

We compare the results of using pairwise statistical significance with that of using database statistical significance (as reported by popular database search programs like BLAST, PSI-BLAST, and SSEARCH), both with and without using SSSMs. Here, the PSI-BLAST results are obtained by directly running it on the test database with default parameters without using pretrained PSSMs, and are presented just to make the analysis more "complete." This, of course, is not the best use of PSI-BLAST, and a more proper use of PSI-BLAST using pretrained PSSMs is presented in the next section to compare it with pairwise statistical significance using PSSMs.

For estimating pairwise statistical significance using SSSMs, SSSMs can be used for either or both of the sequences being compared, and here we report the experimental results for all three cases: 1) using SSSMs for query sequences only; 2) using SSSMs for database sequences only; and 3) using SSSMs for both query and database sequences (using nonconservative pairwise statistical significance). Further, to effectively isolate the influence of pairwise statistical significance and the use of SSSMs, we also report the results of pairwise statistical significance using standard substitution matrix (BLOSUM 62). For the same reason, we also conducted experiments with BLAST, PSI-BLAST, and FASTA with the SSSMs that were used for experiments with pairwise statistical significance. BLAST/PSI-BLAST use a 1/2-bit scaling of substitution matrices instead of 1/3-bit scaling, and hence, the SSSMs were appropriately rescaled for use with BLAST/PSI-BLAST.

Since the EPQ versus Coverage curves on the complete data set can be distorted due to poor performance by one or two queries (if those queries produce many errors at low coverage levels) [39], for comparing the performance across different comparison methods, we examine the performance of the methods with individual queries, following the work in [39]. The coverage of each of the 86 queries at the 1st, 3rd, 10th, 30th, and 100th error was recorded, and the median coverage at each error level was compared across different sequence-comparison methods. A comparison of pairwise statistical significance using sequence-specific substitution matrices (PairwiseStatSig_SSSM) and database statistical significance reported by BLAST, PSI-BLAST (used directly on test database without pretrained PSSMs) and SSEARCH is presented in Fig. 3. Fig. 3 shows the median coverage level at the 1st, 3rd, 10th, 30th, and 100th false positive for homologs (i.e., 43 of the queries have worse coverage, and 43 have better coverage).

The curves present several interesting experimental findings. First, the overall trend of increasing performance is - BLAST $<$ PSI-BLAST (without pretrained PSSMs) $<$ SSEARCH $\sim$ PairwiseStatSig_BL62 $<$ PairwiseStatSig_SSSM, which is not very surprising. Second, using

SSSMs with BLAST, PSI-BLAST, and SSEARCH does not significantly affect their performance. Although little surprising, it may be justified on grounds that our SSSMs may not be good enough, which is also reflected in the fact that when SSSMs are used only for the queries, even pairwise statistical significance shows only marginal improvement over using standard substitution matrices. However, the most surprising observation is that using SSSMs only for database sequences gives very similar performance to nonconservative pairwise statistical significance, which uses SSSMs for both query and database sequences. Although this does not provide a definitive proof of the superiority of using SSSMs, but gives some empirical evidence of the fact that using more sequence-specific information improves the performance of sequence comparison. A further implication of these results is that if this method were to be used for a database search application, we would only need SSSMs for the database sequences, which can be precomputed, and save the search time by not constructing SSSMs for the query sequences. On the other hand, the obvious drawback is that it would require much precomputation. A final important point about the use of SSSMs with BLAST, PSI-BLAST, and SSEARCH is that these methods do not currently support the use of SSSMs for the database sequences and, hence, in their experiments, only the SSSMs for queries were used. Specifically, for BLAST/PSI-BLAST, using SSSMs for different database sequences is not possible because it treats the entire database as one big sequence and scores the hits using a single substitution matrix. Although SSEARCH compares the query sequence with each database sequence independently, it still does not support using different substitution matrices for each database sequence, since it creates an empirical distribution from the scores obtained by comparing the query sequence with the database sequences, and if different substitution matrices are used for comparison with different database sequences, the Karlin-Altschul statistics [16] would not be applicable to estimate the statistical significance of the hits.

4.2 Pairwise Statistical Significance Using PSSMs

In the experiments with PSI-BLAST described in the previous section, only the benchmark database was used to construct the PSSMs over a maximum of five iterations. Since PSI-BLAST allows the use of preconstructed PSSMs for the query sequence, we derived PSSMs for all the 86 test queries against the nonredundant protein database (provided along with the BLAST package) over a maximum of five iterations and with other default parameters. Subsequently, these pretrained PSSMs were used as starting PSSMs for PSI-BLAST searches against the benchmark database, further refined for a maximum of five iterations. Using better quality pretrained 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 pretrained PSSMs, we also conducted experiments with pairwise statistical significance using the same pretrained PSSMs used as starting PSSMs for PSI-BLAST searches on the benchmark database. For this purpose, the popular Gotoh-Smith-Waterman algorithm [4], [5] was trivially modified to calculate the optimal local alignment using a position-specific substitution matrix instead of a general substitution matrix. The time and space complexity of the algorithm is quadratic w.r.t. sequence lengths, but space complexity can be reduced to linear using a divide-and-conquer strategy developed by Hirschberg [48] after identifying the starting and ending indices of the optimal local alignment. The actual alignment can be calculated by following a trace-back procedure as described in [49]. The implementation of the GAP3 program [12] was suitably modified to get the optimal alignment score of a pairwise alignment using a PSSM. Again, the number of shuffles was set to 1,000. 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 and PSI-BLAST, is presented in Fig. 4. There are two comparisons: one using the PSSMs derived against the benchmark database (a subset of CATH), and the other using pretrained PSSMs derived against the nonredundant protein database (NRP) provided with the BLAST package. 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.

4.3 Overall Picture

Finally, in Fig. 5, we present all relevant distinct comparison results together from Figs. 3 and 4. There are three observations that can be made from Fig. 5: First, for all relevant comparisons, pairwise statistical significance performs at least comparable or significantly better than database statistical significance. Second, 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. Third, 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 nonredundant protein database) performs comparable to pairwise statistical significance using position-specific substitution matrices (derived from PSI-BLAST searches against the benchmark database). These are visually summarized in Fig. 6.

In all our experiments with BLAST, PSI-BLAST, and SSEARCH, we used the default parameters, unless otherwise stated. In particular, experiments with PSI-BLAST runs for getting PSSMs on nr database and for retrieval accuracy comparison experiments on the test database were conducted with the default parameters of an e-value threshold (expected chance similarities) of 10.0, gap opening/extension penalty of 11/1, e-value threshold for inclusion in next round of search as 0.002 for a maximum of five iterations, using composition-based statistics as described in [24].

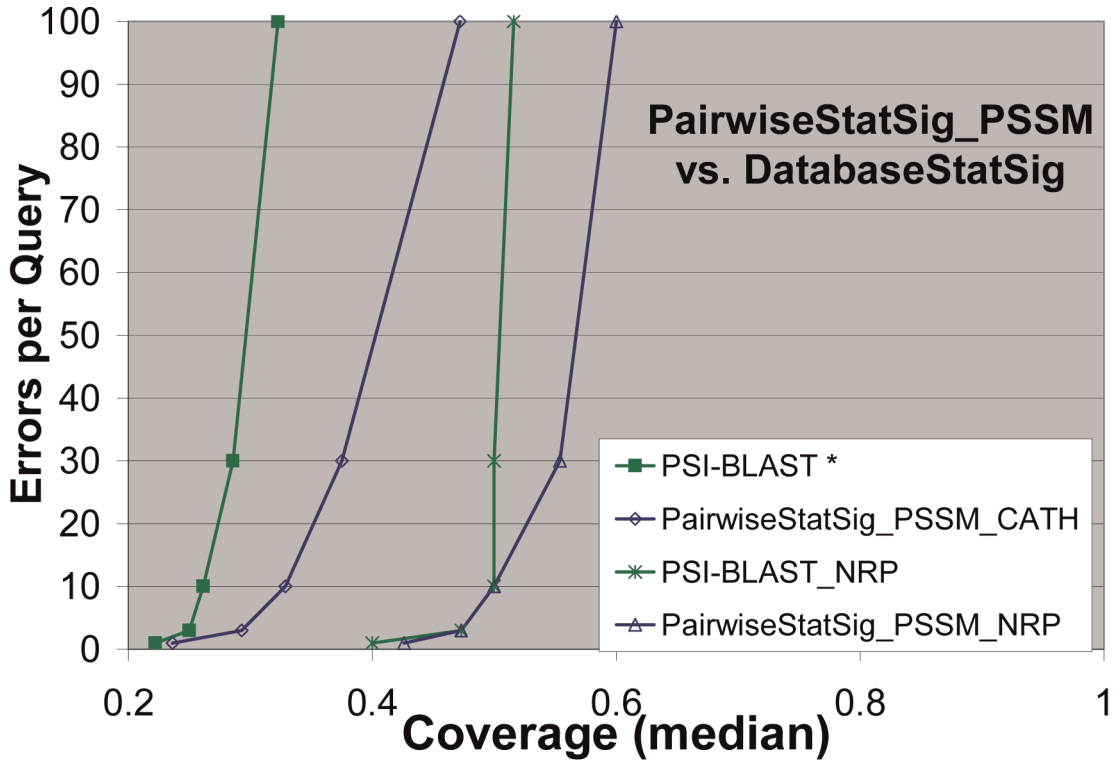

Fig. 4. Pairwise statistical significance versus database statistical significance using position-specific substitution matrices. PSI-BLAST* denotes that these results were obtained by using PSI-BLAST directly on the test database without using pretrained PSSMs. PSI-BLAST_NRP denotes the results obtained by PSI-BLAST on the test database using the pretrained PSSMs derived against nonredundant protein (nr) database. The results for pairwise statistical significance using both kinds of PSSMs are also shown. 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.

For average length protein sequences (200-250), the estimation of pairwise statistical significance takes about two seconds on an Intel 2.8 GHz processor. 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.

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

5 Conclusion and Future Work

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. The results provide clear empirical evidence that sequence-comparison performance improves as the sequence-comparison process is made more and more sequence-specific. 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 pretrained position-specific scoring matrices (PSSMs). Pairwise statistical significance using position-specific substitution matrices is significantly better than PSI-BLAST using pretrained PSSMs.

Although the comparison results in this paper show that the proposed methods outperform the traditional database search programs like BLAST, PSI-BLAST (pairwise statistical significance with SSSMs performs better than PSI-BLAST used directly on benchmark test database, but with pretrained PSSMs, PSI-BLAST performs significantly better; pairwise statistical significance with same PSSMs performs significantly better than PSI-BLAST with pretrained PSSMs), and SSEARCH, the proposed methods, nevertheless, are currently much computationally expensive to be directly used in a large database search since it involves generation of sequence-specific empirical score distributions and subsequent curve-fitting. However, its demonstrated ability to provide biologically more relevant estimates of "relatedness" of sequence pairs, as evaluated in terms of retrieval accuracy makes it a useful tool for many bioinformatics applications relying on sequence-comparison to estimate relatedness of a few sequence-pairs like small database

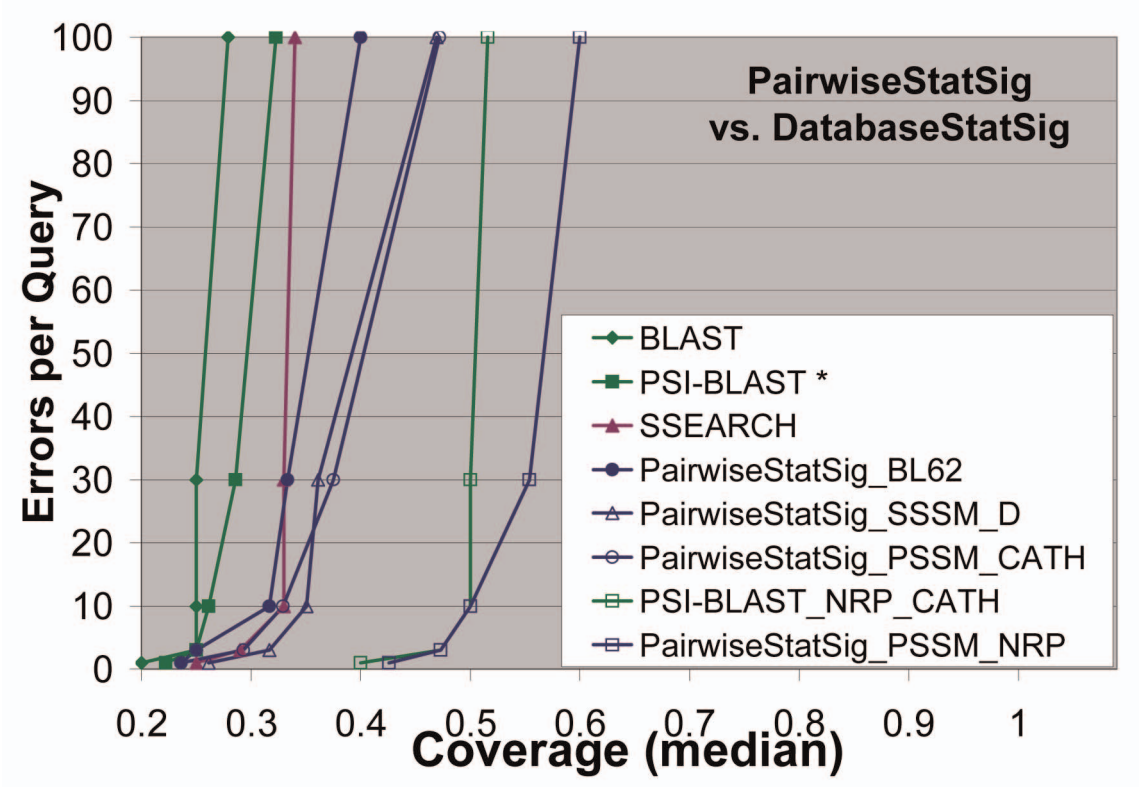

Fig. 5. Comparison of using sequence-specific and position-specific substitution matrices for estimating pairwise statistical significance with database statistical significance. PSI-BLAST* denotes that these results were obtained by using PSI-BLAST directly on the test database without using pretrained PSSMs. PSI-BLAST_NRP denotes the results obtained by PSI-BLAST on the test database using the pretrained PSSMs derived against nonredundant protein (nr) database. For all relevant comparisons, pairwise statistical significance performs significantly better than database statistical significance using BLAST, PSI-BLAST and SSEARCH.

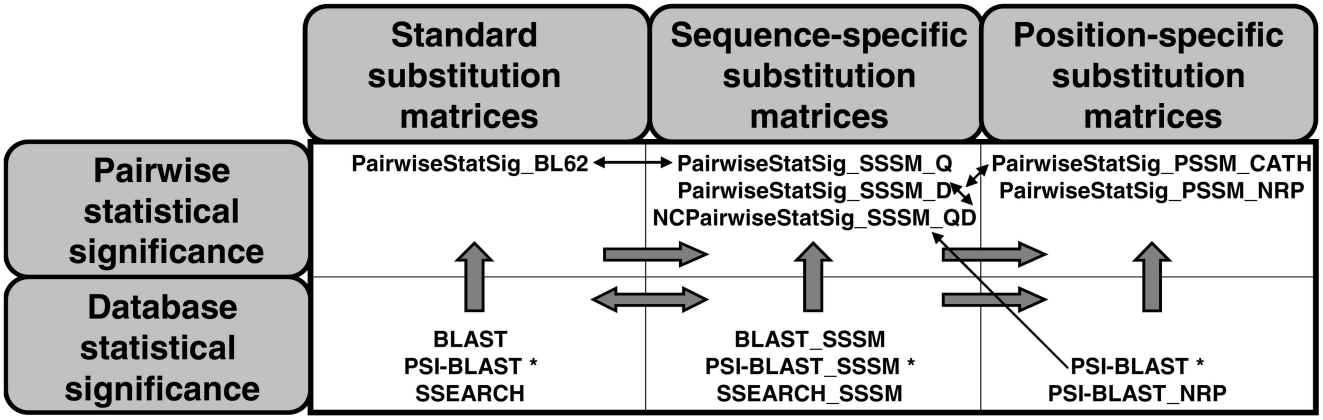

Fig. 6. Summary of the results using pairwise statistical significance and database statistical significance with standard, sequence-specific, and position-specific substitution matrices. The titles in each of the six blocks are the same as the EPQ versus Coverage curve representations in previous figures. PSI-BLAST* is placed both in database statistical significance using standard and position-specific substitution matrices because it begins with a standard matrix and constructs PSSMs during the search against the test database. Arrows are pointing toward the combination which performed better in terms of retrieval accuracy in our experiments. Block arrows show the general trend, and thin arrows indicate the exceptions.

searches, progressive multiple sequence alignment, creating distance matrices for phylogenetic tree construction, etc.

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. Further enhancement in the performance of using pairwise statistical significance with SSSMs may be possible by addressing some of the weaknesses of the SSSM construction approach used in this work, as outlined earlier. In particular, to address the issue of over-representation from very similar sequences in the BLAST search to collect homologous alignments to fill up the count matrix, several approaches are possible. All hits with similarity, more than a cutoff (e.g., 62 percent), may be represented by a single sequence while filling the count matrix, which is similar to the approach of creating BLOSUM matrices. Another possibility is to give

weights to each hit while constructing the count matrices, which is something like what PSI-BLAST uses to construct PSSMs. Combining information from known sequence families may also be helpful in constructing better quality SSSMs/PSSMs.

In this work, we have used PSSMs for only one sequence (query sequences) to estimate pairwise statistical significance. Using PSSMs for database sequences also is expected to further improve retrieval accuracy, as found in the case of SSSMs. Another very important aspect of future work is to develop faster methods to estimate pairwise statistical significance without sacrificing much on the front of retrieval accuracy.

ACKNOWLEDGMENTS

The authors thank Dr. Sean Eddy for making the C 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. Special thanks are due to the anonymous reviewers, whose insightful comments made this manuscript stronger.

REFERENCES

[1] W.R. Pearson and D.J. Lipman, "Improved Tools for Biological Sequence Comparison," *Proc. Nat'l Academy of Sciences USA* vol. 85, no. 8, pp. 2444-2448, http://www.pnas.org/cgi/content/abstract/85/8/2444, 1988.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic Local Alignment Search Tool.," *J. Molecular Biology,* vol. 215, no. 3, pp. 403-410, http://dx.doi.org/10.1006/jmbi.1990.9999, 1990.

[3] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research,* vol. 25, no. 17, pp. 3389-3402, http://dx.doi.org/10.1093/nar/25.17.3389, 1997.

[4] T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences," *J. Molecular Biology,* vol. 147, no. 1, pp. 195-197, http://view.ncbi.nlm.nih.gov/pubmed/7265238, 1981.

[5] O. Gotoh, "An Improved Algorithm for Matching Biological Sequences," *J. Molecular Biology,* vol. 162, no. 3, pp. 705-708, Dec. 1982.

[6] P.H. Sellers, "Pattern Recognition in Genetic Sequences by Mismatch Density," *Bull. of Math. Biology,* vol. 46, no. 4, pp. 501-514, http://www.springerlink.com/content/2v4477481102w030, 1984.

[7] W.R. Pearson, "Effective Protein Sequence Comparison," *Methods in Enzymology,* vol. 266, pp. 227-259, 1996.

[8] W.R. Pearson, "Flexible Sequence Similarity Searching with the FASTA3 Program Package," *Methods in Molecular Biology,* vol. 132, pp. 185-219, 2000.

[9] B. Ma, J. Tromp, and M. Li, "PatternHunter: Faster and More Sensitive Homology Search," *Bioinformatics,* vol. 18, no. 3, pp. 440-445, 2002.

[10] M. Li, B. Ma, D. Kisman, and J. Tromp, "PatternHunter II: Highly Sensitive and Fast Homology Search," *J. Bioinformatics and Computational Biology,* vol. 2, no. 3, pp. 417-439, 2004.

[11] K.-M. Chao, "Calign: Aligning Sequences with Restricted Affine Gap Penalties," *Bioinformatics,* vol. 15, no. 4, pp. 298-304, 1999.

[12] X. Huang and K.-M. Chao, "A Generalized Global Alignment Algorithm," *Bioinformatics,* vol. 19, no. 2, pp. 228-233, 2003.

[13] X. Huang and D.L. Brutlag, "Dynamic Use of Multiple Parameter Sets in Sequence Alignment," *Nucleic Acids Research,* vol. 35, no. 2, pp. 678-686, http://nar.oxfordjournals.org/cgi/content/abstract/35/2/678, 2007.

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

[15] S.F. Altschul, M.S. Boguski, W. Gish, and J.C. Wootton, "Issues in Searching Molecular Sequence Databases," *Nature Genetics,* vol. 6, no. 2, pp. 119-129, 1994.

[16] S. Karlin and S.F. Altschul, "Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes," *Proc. Nat'l Academy of Sciences USA,* vol. 87, no. 6, pp. 2264-2268, http://www.pnas.org/cgi/content/abstract/87/6/2264, 1990.

[17] M.S. Waterman and M. Vingron, "Rapid, Accurate Estimates of Statistical Significance for Sequence Database Searches," *Proc. Nat'l Academy of Sciences USA,* vol. 91, no. 11, pp. 4625-4628, http://www.pnas.org/cgi/content/abstract/91/11/4625, 1994.

[18] S.F. Altschul and W. Gish, "Local Alignment Statistics," *Methods in Enzymology,* vol. 266, pp. 460-80, 1996.

[19] W.R. Pearson, "Empirical Statistical Estimates for Sequence Similarity Searches," *J. Molecular Biology,* vol. 276, pp. 71-84, 1998.

[20] R. Mott and R. Tribe, "Approximate Statistics of Gapped Alignments," *J. Computational Biology,* vol. 6, no. 1, pp. 91-112, 1999.

[21] R. Mott, "Accurate Formula for P-Values of Gapped Local Sequence and Profile Alignments," *J. Molecular Biology,* vol. 300, pp. 649-659, 2000.

[22] R. Bundschuh, "Rapid Significance Estimation in Local Sequence Alignment with Gaps," *Proc. Fifth Ann. Int'l Conf. Research in Computational Molecular Biology (RECOMB '01),* pp. 77-85, 2001.

[23] S.F. Altschul, R. Bundschuh, R. Olsen, and T. Hwa, "The Estimation of Statistical Parameters for Local Alignment Score Distributions," *Nucleic Acids Research,* vol. 29, no. 2, pp. 351-361, 2001.

[24] 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,* vol. 29, no. 14, pp. 2994-3005, 2001.

[25] 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,* vol. 33, no. 15, pp. 4987-4994, 2005.

[26] A. Poleksic, J.F. Danzer, K. Hambly, and D.A. Debe, "Convergent Island Statistics: A Fast Method for Determining Local Alignment Score Significance," *Bioinformatics,* vol. 21, no. 12, pp. 2827-2831, 2005.

[27] 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,* vol. 34, no. 20, pp. 5966-5973, 2006.

[28] A. Agrawal, V.P. Brendel, and X. Huang, "Pairwise Statistical Significance and Empirical Determination of Effective Gap Opening Penalties for Protein Local Sequence Alignment," *Int'l J. Computational Biology and Drug Design,* vol. 1, no. 4, pp. 347-367, 2008.

[29] A. Agrawal and X. Huang, "Conservative, Non-Conservative and Average Pairwise Statistical Significance of Local Sequence Alignment," *Proc. IEEE Int'l Conf. Bioinformatics and Biomedicine,* pp. 433-436, 2008.

[30] M. Kschischo, M. Lässig, and Y.-K. Yu, "Toward an Accurate Statistics of Gapped Alignments," *Bull. of Math. Biology,* vol. 67, pp. 169-191, 2004.

[31] 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,* vol. 10, no. 5, pp. 829-845, 2004.

[32] M. Pagni and C.V. Jongeneel, "Making Sense of Score Statistics for Sequence Alignments," *Briefings in Bioinformatics,* vol. 2, no. 1, pp. 51-67, 2001.

[33] W.R. Pearson and T.C. Wood, "Statistical Significance in Biological Sequence Comparison," *Handbook of Statistical Genetics,* D. J. Balding, M. Bishop, and C. Cannings, eds., pp. 39-66, Wiley, 2001.

[34] A.Y. Mitrophanov and M. Borodovsky, "Statistical Significance in Biological Sequence Analysis," *Briefings in Bioinformatics,* vol. 7, no. 1, pp. 2-24, 2006.

[35] Y.-K. Yu and S.F. Altschul, "The Construction of Amino Acid Substitution Matrices for the Comparison of Proteins with Non-Standard Compositions," *Bioinformatics,* vol. 21, no. 7 pp. 902-911, 2005.
[36] S.R. Eddy, "Maximum Likelihood Fitting of Extreme Value Distributions," unpublished work, citeseer.ist.psu.edu/370503.html, 1997.
[37] A. Agrawal and X. Huang, "Pairwise Statistical Significance of Local Sequence Alignment Using Multiple Parameter Sets and Empirical Justification of Parameter Set Change Penalty," *BMC Bioinformatics,* vol. 10, suppl. 3, p. S1, 2009.
[38] A. Agrawal and X. Huang, "Pairwise Statistical Significance of Local Sequence Alignment Using Substitution Matrices with Sequence-Pair-Specific Distance," *Proc. Int'l Conf. Information Technology, (ICIT '08),* pp. 94-99, 2008.
[39] M.L. Sierk and W.R. Pearson, "Sensitivity and Selectivity in Protein Structure Comparison," *Protein Science,* vol. 13, no. 3, pp. 773-785, 2004.
[40] S. Kotz and S. Nadarajah, *Extreme Value Distributions: Theory and Applications,* ch. 1, pp. 3-4. Imperial College Press, 2000.
[41] 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,* vol. 2, p. 9, 2007.
[42] A.K. Hartmann, "Sampling Rare Events: Statistics of Local Sequence Alignments," *Physical Rev. E,* vol. 65, no. 5, p. 056102, 2002.
[43] R. Olsen, R. Bundschuh, and T. Hwa, "Rapid Assessment of Extremal Statistics for Gapped Local Alignment," *Proc. Seventh Int'l Conf. Intelligent Systems for Molecular Biology,* pp. 211-222, 1999.
[44] R.F. Mott, "Maximum-Likelihood Estimation of the Statistical Distribution of Smith Waterman Local Sequence Similarity Scores," *Bull. of Math. Biology,* vol. 54, pp. 59-75, 1992.
[45] S.R. Eddy, "Where did the Blosum62 Alignment Score Matrix Come from?," *Nature Biotechnology,* vol. 22, no. 8, pp. 1035-1036, Aug. 2004.
[46] 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,* vol. 28, no. 1, pp. 1093-1108, 1997.
[47] J. Rocha, F. Rosselló, and J. Segura, "Compression Ratios Based on the Universal Similarity Metric Still Yield Protein Distances Far from CATH Distances," *CoRR,* vol. abs/q-bio/0603007, 2006.
[48] D.S. Hirschberg, "A Linear Space Algorithm for Computing Maximal Common Subsequences," *Comm. ACM,* vol. 18, no. 6, pp. 341-343, 1975.
[49] S. Altschul and B. Erickson, "Optimal Sequence Alignment Using Affine Gap Costs," *Bull. of Math. Biology,* vol. 48, no. 5, pp. 603-616, Sept. 1986.

Ankit Agrawal received the BTech degree in computer science and engineering from the Indian Institute of Technology, Roorkee. He was the graduating topper of his batch. At present, he is working toward the PhD degree in computer science at Iowa State University. His research interests include bioinformatics, fuzzy logic, data mining, and signal processing.

Xiaoqiu Huang received the PhD degree in computer science from Pennsylvania State University in 1990. He is a professor in the Department of Computer Science at Iowa State University. He is the author of a widely used CAP3 assembly program. His research interests are in bioinformatics. He and his collaborators have recently developed a whole-genome assembly program, named PCAP. PCAP has been used by Washington University Genome Center in chimpanzee and chicken genome projects.