Derived Distribution Points Heuristic for Fast Pairwise Statistical Significance Estimation

Ankit Agrawal, Alok Choudhary
Dept. of Electrical Engg. and Computer Science
Northwestern University
2145 Sheridan Rd
Evanston, IL 60201
USA
{ankitag,choudhar}@eecs.northwestern.edu

Xiaoqiu Huang
Dept. of Computer Science
Iowa State University
226 Atanasoff Hall
Ames, IA 50011
USA
xqhuang@cs.iastate.edu

ABSTRACT

Estimation of statistical significance of a pairwise sequence alignment is crucial in homology detection. A recent development in the field is the use of pairwise statistical significance as an alternative to database statistical significance. Although pairwise statistical significance has been shown to be potentially better than database statistical significance in terms of homology detection retrieval accuracy, currently it is much time consuming since it involves generating an empirical score distribution by aligning one sequence of the sequence-pair with N random shuffles of the other sequence. A high value of N produces (statistically and potentially biologically) accurate estimates, but also consumes more time. A low value of N leads to inaccurate fitting of the score distribution, and hence poor estimates of statistical significance. In this paper, we propose a simple heuristic, called the Derived Distribution Points (DDP) heuristic, which is designed taking into account the features of the pairwise statistical significance estimation procedure, and has shown to significantly improve the quality of pairwise statistical significance estimates (evaluated in terms of retrieval accuracy) even when using low values of N. Alternatively, it can be thought of as speeding-up pairwise statistical significance estimation using high values of N, where comparable performance is achieved by actually using a much lower number of random shuffles. Experiments indicate that a speed-up of up to 40 as compared to current implementations can be achieved without loss in retrieval accuracy.

1. INTRODUCTION

Biological sequence alignment is one of the most important computational problem in bioinformatics for analysis and comparison of DNA and protein sequences [28, 10, 11]. There exist classical algorithms for optimal local sequence alignment [33], based on which, many other algorithms [15, 14] have been proposed, which try to model sequence comparison in a more biologically relevant way. Several heuristics have also been proposed [25, 11, 27, 18, 17] which are extremely useful in database search application where it is impractical to use exact algorithms for sequence alignment.

The inter-relationship between sequence, structure, and function, which forms the basis of a vast number of applications in bioinformatics, motivates researchers to devise better methods for sequence alignment based sequence comparison. Pairwise alignment methods report an alignment score for an alignment of two sequences, and pairs of related sequences (known as homologs) should, in general have higher alignment score. But the alignment score itself does not reflect anything about the relatedness of the sequences. For instance, two related sequences of length 50 can have an optimal alignment score of 50, and two unrelated sequences of length 500 can have an optimal alignment score of 100. Therefore, to comment on the relatedness of the two sequences being aligned, a common practice is to estimate the statistical significance of the alignment score, which is an estimate of the likelihood of that alignment score being produced by the alignment of two unrelated sequences of similar features. An alignment score, therefore, is more 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 [21], 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. 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 [8, 24, 21, 19].

The statistical significance of hits (database sequences found to be similar to the query sequence) reported by popular database search programs like BLAST [11], FASTA [26, 27], SSEARCH (using full implementation of Smith-Waterman algorithm [33]), and PSI-BLAST [11, 31] is called database statistical significance, which is dependent on the size and composition of the database being searched. Over the last few years, there have been significant improvements to the BLAST and PSI-BLAST programs [31, 35, 36], which have been shown to improve search performance using composition-based statistics and other enhancements.

Recently, an alternative method to evaluate the statistical

ACM-BCB 2010, Niagara Falls, NY, USA

significance of an alignment was studied known as pairwise statistical significance [1, 2], which is specific to the sequence-pair being aligned, and independent of any database. Further studies on pairwise statistical significance using multiple parameter sets [4], sequence-specific and position-specific substitution matrices [5] have demonstrated it to be a promising method capable of producing much more biologically relevant estimates of statistical significance than database statistical significance. However, it is significantly slow since it involves the generation of sequence-specific empirical score distributions by aligning one sequence with N shuffled versions of the other sequence, and subsequent fitting of the distributions to estimate statistical significance.

In this paper, we propose a simple heuristic to speed-up the pairwise statistical significance procedure. Reducing the number of random shuffles N would in general reduce the running time at the cost of accuracy. The proposed Derived Distribution Points (DDP) heuristic is designed taking the features of pairwise statistical significance estimation procedure, where each alignment is made to contribute multiple relevant scores in the empirical distribution, thus giving comparable performance as with high values of N, even though effectively using a very low number of shuffles. Experiments indicate that a speed-up of up to 40 can be achieved without loss in retrieval 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 of the heuristic, which is 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.

2. PAIRWISE STATISTICAL SIGNIFICANCE

Score distribution for ungapped local alignment is known to follow a Gumbel-type EVD [16], with analytically calculable 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

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

and m and n are the lengths of the two sequences being aligned.

For gapped alignment score distribution, no perfect statistical theory has yet been developed, although there is ample empirical evidence that it also closely follows Gumbel-type EVD [34, 9, 26, 20, 22, 14], even when using multiple parameter sets [4] and position-specific substitution matrices, as used by PSI-BLAST. 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 [19]. There exist some excellent reviews on statistical significance in sequence comparison [24, 29, 21, 19].

Pairwise statistical significance is an attempt to make the statistical significance estimation process more specific to the sequence pair being compared. A study of pairwise statistical significance and its comparison with database statistical significance [1, 2] compared various approaches to estimate pairwise statistical significance like ARIADNE [20], PRSS [27], censored-maximum-likelihood fitting [12], linear regression fitting [14] to find that maximum likelihood fitting with censoring left of peak (described as type-I censoring in [12]) is the most accurate method for estimating pairwise statistical significance.

Pairwise statistical significance described in [1, 2] can be thought of as being 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. The function $PairwiseStatSig$, therefore, generates a score distribution by aligning $Seq1$ with N shuffled versions of $Seq2$, fits the distribution to an EVD 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. The scoring scheme SC can be extended to use sequence-pair-specific distanced substitution matrices or multiple parameter sets, as used in [3] and [4] respectively. Further, a sequence-specific/position-specific scoring scheme SC_1 specific to one of the sequences (say $Seq1$) can be used to estimate pairwise statistical significance using sequence-specific/position-specific substitution matrices [5]. Pairwise statistical significance has also been used to reorder the hits from a fast database search program like PSI-BLAST [6]. However, since estimation of pairwise statistical significance for a single pair involves N alignments, it is very time consuming and can be impractical for estimating pairwise statistical significance of a large number of sequence pairs.

It is easy to see that the number of shuffles N has an immediate effect on statistical significance accuracy and execution time. 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 [36], 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 since we are censoring the scores left of peak (about half of the low scores in the distribution).

2. The region of interest (right half of the distribution) is non-increasing.

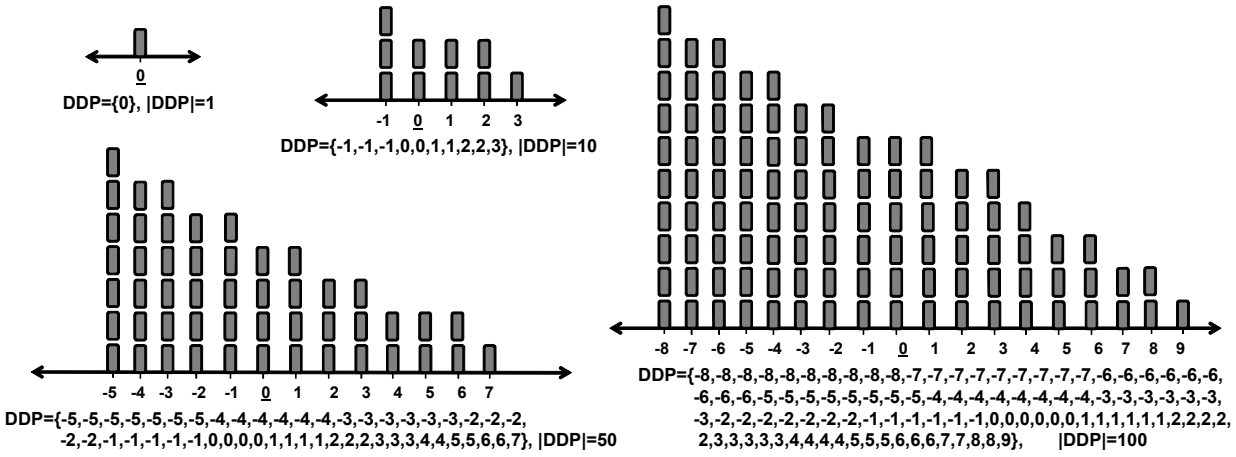

Figure 1: Four DDP sets used in this work. Each alignment score contributes $|DDP|$ scores to the histogram around itself thereby adversely affecting the score distribution only left of peak but not right of peak. The special case of $DDP = \{0\}$ essentially disables the DDP heuristic.

3. It should be possible to improve retrieval accuracy even though compromising on statistical significance accuracy.

3. PROPOSED HEURISTIC

Based on the observations made in the previous section, here we present the Derived Distribution Points (DDP) heuristic to speed up the pairwise statistical significance estimation process.

3.1 Derived Distribution Points

This heuristic derives multiple alignment scores from each actual alignment score obtained by aligning $Seq1$ with a shuffled version of $Seq2$. It attempts to do so without adversely affecting the right tail of the distribution. Given sequences $Seq1$, $Seq2$, the number of shuffles N, and a derived distribution points set $DDP = \{DDP_1, DDP_2, \ldots, DDP_{N_{ddp}}\}$ with $N_{ddp} = |DDP|$, this heuristic derives N_{ddp} scores from each of the N actual scores obtained by N shuffles. 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 $N \times N_{ddp}$ alignment scores. Once the modified score distribution is thus obtained, a censored-maximum-likelihood fitting is performed (type-I censoring) to estimate the statistical parameters K and λ. Pairwise statistical significance of the alignment score of $Seq1$ and $Seq2$ is subsequently estimated by the P-value formula.

The choice of the set DDP is such that it contributes a non-increasing mini-histogram for every alignment score s centered around s, which adversely affects the left tail of the distribution but not the right tail (the left tail of the distribution is anyway censored before fitting [1, 2]). Still, since the slope of the distribution is not the same throughout the right-half of the distribution, this introduces some error in the distribution which could degrade statistical significance accuracy.

The use of DDP heuristic can possibly be justified on the grounds that if a score s is obtained by aligning $Seq1$ and a shuffled version of $Seq2$, it is possible that the same score s and scores around s can also be obtained by aligning $Seq1$ and other valid shuffles of $Seq2$. For example, if we swap two amino acids from $Seq2$ from a region which did not participate in the optimal local alignment for score s, it is possible (although not guaranteed) that the new optimal local alignment and the alignment score remains unchanged. Of course, the existence of scores around s is an assumption to support the use of DDP heuristic and is certainly not correct for all 'real' s. Thus, it is expected that this methodology would introduce some error in statistical significance accuracy. But experiments and results presented in the next section demonstrate that the more important retrieval accuracy is maintained.

Fig. 1 shows four DDP sets used in this work, including the special case of $DDP = \{0\}$, which effectively disables the DDP heuristic.

4. EXPERIMENTS AND RESULTS

In this section we present the retrieval accuracy, statistical significance accuracy, and timing results of fast pairwise statistical significance using the proposed DDP heuristic.

4.1 Retrieval accuracy evaluation

To evaluate the performance of the proposed heuristic in terms of retrieval accuracy, we used the same experiment setup as used in [32], and later in [2, 3, 4, 5]. A non-redundant subset of the CATH 2.3 database (Class, Architecture, Topology, and Hierarchy, [23]) available at ftp://ftp.ebi.ac.uk/ pub/software/ unix/fasta/prot_sci_04/ was selected in [32] to evaluate seven structure comparison programs and two sequence comparison programs. As described in [32], 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 [30].

For each of the 86×2771 comparisons, pairwise statistical significance was estimated, and the retrieval accuracy across

different methods methods is visually compared using Error per Query (EPQ) versus Coverage plots. 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.

Fig. 2 presents the retrieval accuracy performance on the benchmark datasets using pairwise statistical significance with different values for the parameter N (number of shuffles). Clearly, retrieval accuracy is very poor for low N, and improves as N is increased, and more or less saturates after $N = 500$. Previous works on pairwise statistical significance [2, 3, 4, 5] have used a value of 1000 for N. Fig. 2 also supports the finding reported in [36] that improving statistical significance accuracy may not necessarily improve retrieval accuracy, in the sense that for this dataset, retrieval accuracy in our experiments saturates after $N = 500$, although a higher value of N would statistically give more accurate estimates of P-value. Nonetheless, the number of shuffles N cannot be made too low since retrieval accuracy would also drop by that.

For the proposed DDP heuristic, we used three DDP sets with $|DDP|$ as 10, 50, and 100 as shown in Fig. 1. Note that the special case of $DDP = \{0\}$ corresponds to not using the DDP heuristic for pairwise statistical significance estimation. Figures 3, 4, and 5 present the EPQ vs. Coverage curves for pairwise statistical significance estimation using the above mentioned DDP sets for $N = 25$, $N = 50$, and $N = 100$ respectively. For comparison with earlier works on pairwise statistical significance using $N = 1000$, the corresponding curve (named 'base') is also shown in these figures. It is clear from these curves that the retrieval accuracy performance can be significantly enhanced using the proposed heuristic, and performance comparable to the base case (and even better in some cases) can be obtained by using a larger DDP set, even with a small number of shuffles. Again, this figure supports the finding presented in [36] that improving statistical significance accuracy may not improve retrieval accuracy, in the sense that by using a larger DDP set, we are introducing more error in the distribution which would degrade statistical significance accuracy, but it results in an enhancement in retrieval accuracy.

The enhancement in retrieval accuracy performance can be attributed to the fact the proposed heuristic has been designed taking into consideration the features of the pairwise statistical significance estimation procedure, as described in Section 2. For validating our choice of DDP sets proposed in this work, we also performed additional experiments with different DDP sets for $N = 25$ and $|DDP| = 100$. The following sets were used:

1. An non-decreasing DDP-set, as opposed to the non-increasing DDP set as proposed in this work.
2. A random DDP set with the same spread as proposed DDP set around the obtained 'real' score.
3. A random DDP set with the spread twice as compared to that of the proposed DDP set around the obtained 'real' score.
4. A random DDP set with the spread four times as compared to that of the proposed DDP set around the obtained 'real' score.
5. A constant DDP set with no spread, which simply adds the obtained 'real' score |DDP| times to the distribution.

Fig. 6 shows the retrieval accuracy performance in the above described four scenarios, along with the curve with proposed non-increasing DDP set. The figure shows that when the spread of the DDP sets is same as that of the proposed non-increasing DDP sets, the retrieval accuracy performance is also almost comparable to using proposed non-increasing DDP sets, but the retrieval accuracy with proposed non-increasing DDP sets is at least as good as or better than with other DDP sets. As the spread increases, the performance significantly degrades.

4.2 Statistical significance accuracy

Although retrieval accuracy is more important than statistical significance accuracy for bioinformatics applications, we also evaluate the proposed method in terms of statistical significance accuracy to further validate our choice of the DDP sets proposed in this work. We used the method earlier used in [1, 2] for this purpose. For evaluating in terms of statistical significance accuracy, we would need to know the actual alignment score distribution for a sequence-pair. Since the true theoretical distribution is unknown for gapped alignment, we constructed a score distribution with a million shuffles for a sequence-pair (sequences 1qktA0 and 2lbd00 in the CATH database), and considered it to be the closest representative of the underlying score distribution. Subsequently, the task of evaluating a sequence comparison method in terms of statistical significance accuracy reduces to comparing the predicted P-values (using estimated K and λ) against the normalized E-values (normalized alignment score distribution; also known as complementary distribution in terms of statistics) of the million-shuffle distribution. Since a good sequence comparison method is expected to accurately predict the P-values in the right tail region, we looked at distribution of scores for which normalized E-value was less than 0.01, i.e., top 1% alignment scores.

Fig. 7 shows the comparison of the sum of squares of differences (SSD) between the predicted P-values and the actual normalized E-values for the five cases outlined in the previous subsection, along with the SSD for the base case (N=1000,$|DDP| = 1$) and for using the proposed non-increasing DDP set. The y-axis of the bar-chart shows the SSD, which can be thought of as the error in predicted P-values. As expected, the SSD using the proposed non-increasing DDP set increases as compared to the base case, since adding artificial scores would lead to a loss in statistical significance

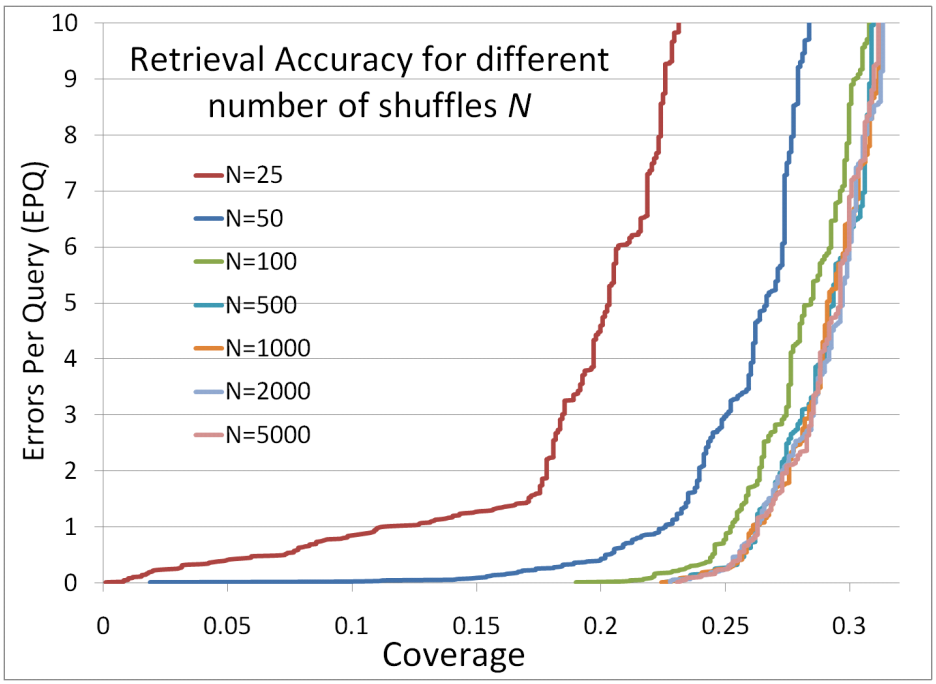

Figure 2: EPQ vs. Coverage plots for pairwise statistical significance with different number of shuffles. A very low value of N gives poor retrieval accuracy performance, which improves with increase in N. The alteration in retrieval accuracy beyond $N = 500$ is not very significant.

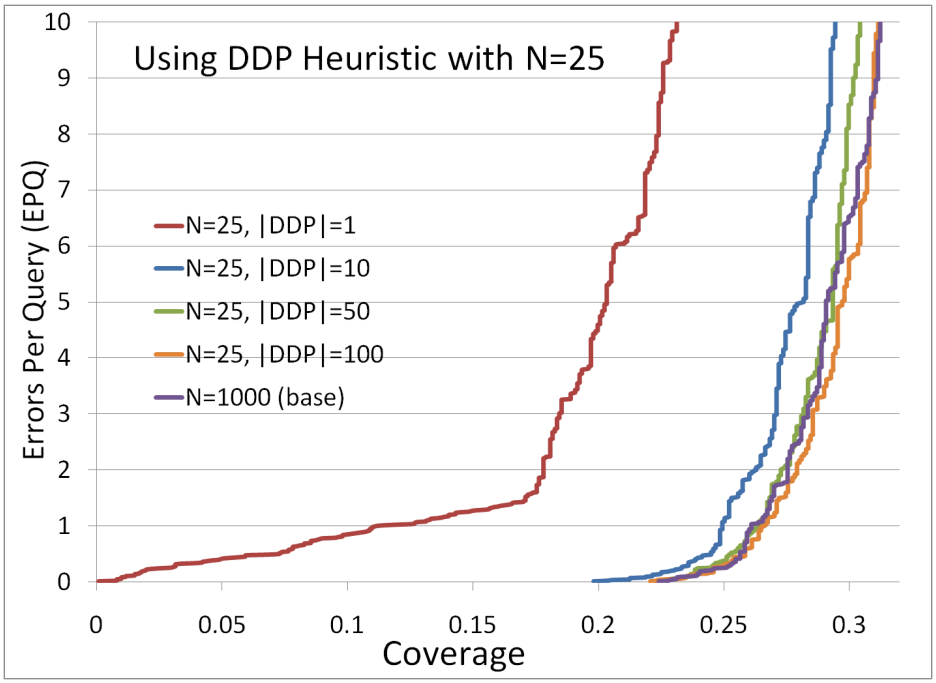

Figure 3: EPQ vs. Coverage plots for pairwise statistical significance with number of shuffles $N = 25$ using DDP heuristic. $|DDP|$=1 corresponds to normal pairwise statistical significance with the proposed heuristic disabled. Using DDP heuristic significantly enhances retrieval accuracy. Retrieval accuracy performance with the base case ($N = 1000$) is also shown. Results with $N = 25, |DDP| = 100$ is at least comparable and sometimes better than the base case.

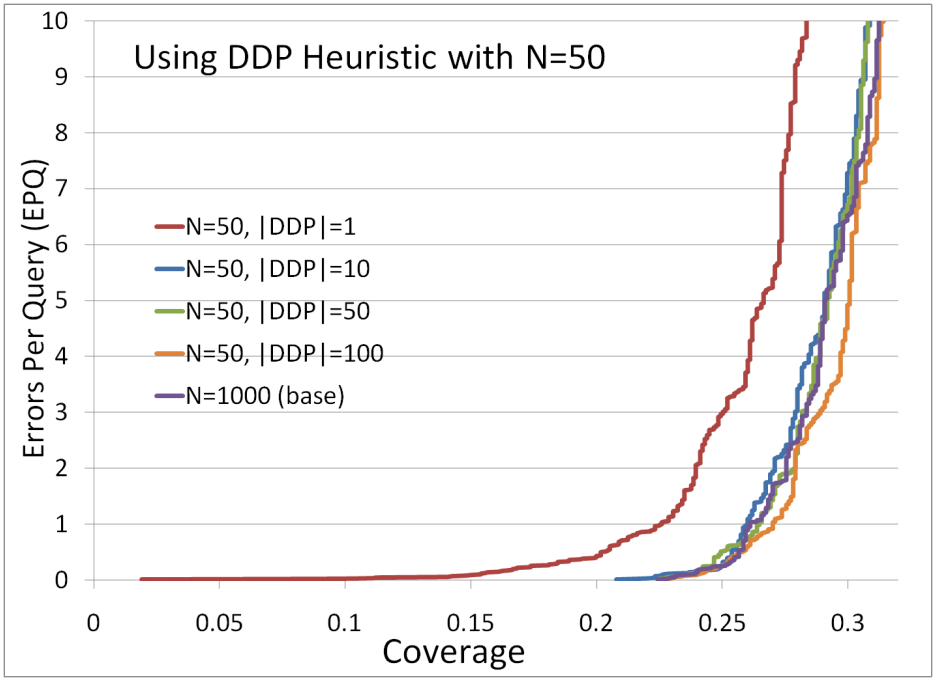

Figure 4: EPQ vs. Coverage plots for pairwise statistical significance with number of shuffles $N = 50$ using DDP heuristic. Results with all experimented DDP sets are comparable to the base case.

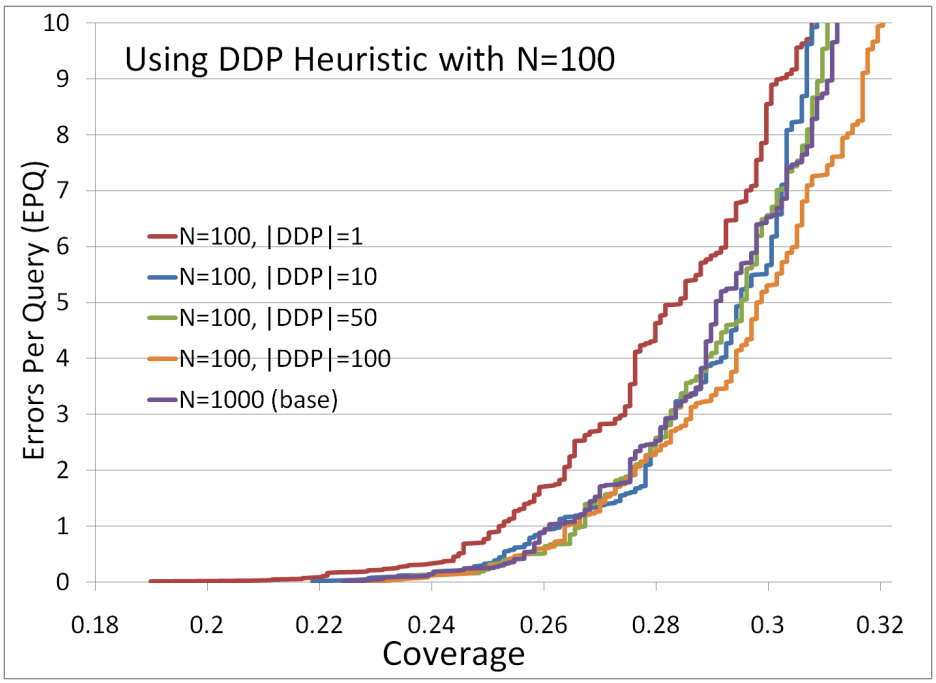

Figure 5: EPQ vs. Coverage plots for pairwise statistical significance with number of shuffles $N = 100$ using DDP heuristic. Results with all experimented DDP sets are comparable to the base case.

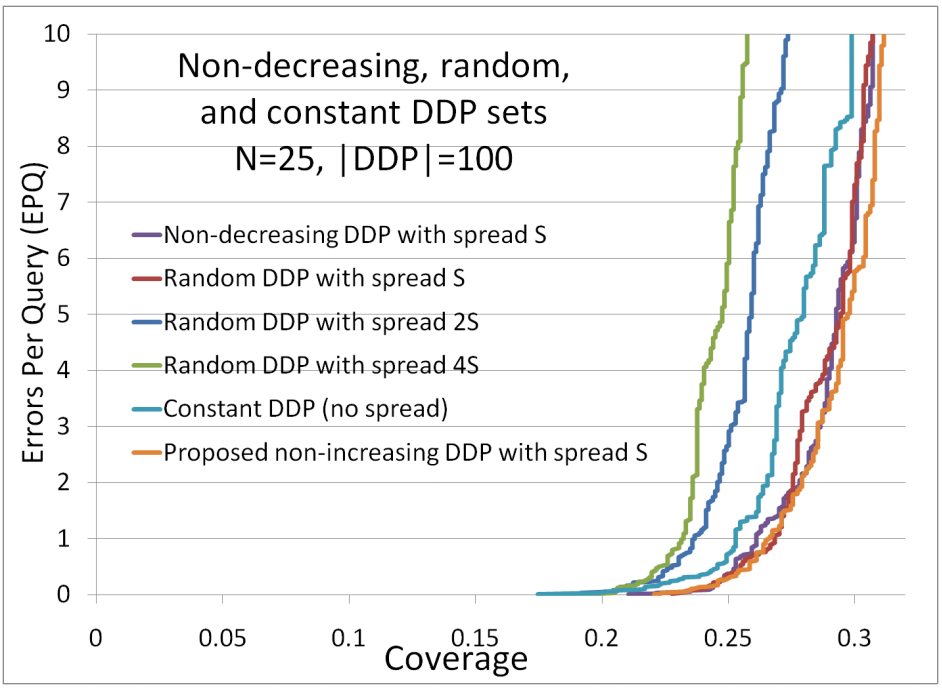

Figure 6: Retrieval accuracy comparison for pairwise statistical significance with non-decreasing, random, constant, and proposed non-increasing DDP sets.

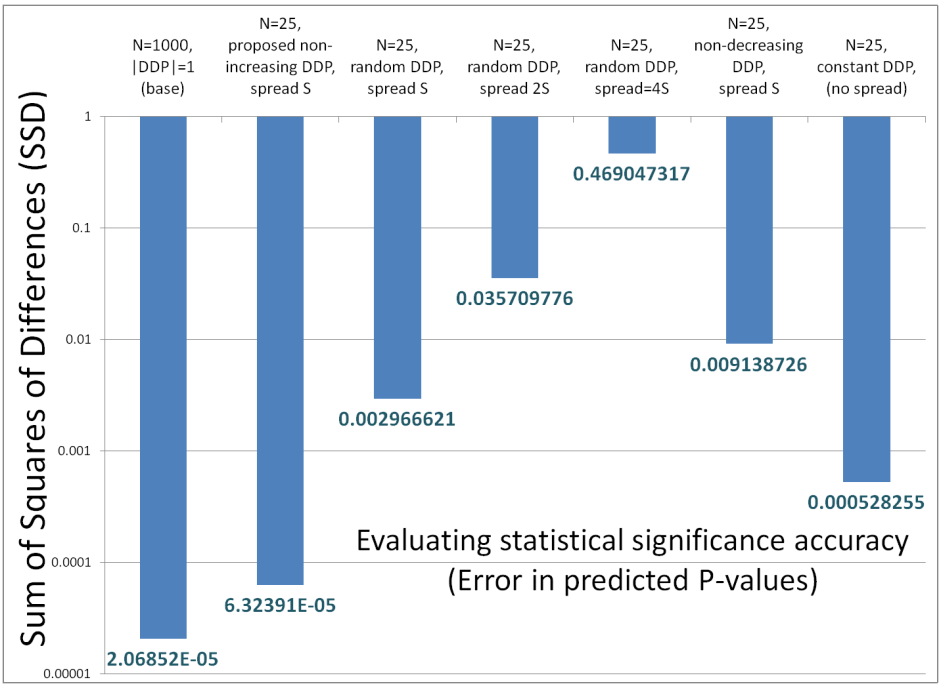

Figure 7: Statistical significance accuracy comparison for pairwise statistical significance with non-decreasing, random, constant, and proposed non-increasing DDP sets, and without using the DDP heuristic (base case). The SSD bars represent the error in predicted P-values. Lower the SSD, more accurate the method is in terms of statistical significance accuracy.

Table 1: Execution time and Speed-up with DDP heuristic. $|DDP|$=1 corresponds to normal pairwise statistical significance. One ATU (Alignment Time Unit) is defined as the time required to align two sequences of length around 250 using Smith-Waterman algorithm.

N	$\|DDP\|$	Time (s)	Time (ATU)	Speedup
1000	1	3.277	1001.65	1
100	10	0.329	100.56	9.96
100	50	0.33	100.87	9.93
100	100	0.33	100.87	9.93
50	10	0.165	50.43	19.86
50	50	0.166	50.74	19.74
50	100	0.165	50.43	19.86
25	100	0.083	25.37	39.48

accuracy. However, the SSD for other DDP sets is significantly worse, thereby experimentally justifying the careful choice of proposed DDP sets used in this work.

4.3 Timing results

Since many combinations of N and DDP values give comparable or better results than the base case ($N = 1000$), we can define speed-up as the time improvement to get at least comparable biologically relevant estimates of pairwise statistical significance. Table 1 and Fig. 8 presents the timing and speed-up results. All times are in seconds, and represent the time taken to estimate the pairwise statistical significance of alignment score of two average length ($\sim$250 aa) protein sequences. In addition to reporting the time in seconds, the execution time is also reported in Alignment Time Units (ATUs) to better visualize the speed-up independent of underlying processor used. One ATU is defined as the time required to align two average length sequences using Smith-Waterman algorithm [33]. In our experiments, 1 ATU = 0.0032716 s.

All experiments were performed on an Intel 2.8GHz processor. The substitution matrix, gap opening, and gap extension penalties used were BLOSUM62, 11, and 1 respectively (default used in BLAST).

Since the bulk of time in pairwise statistical significance estimation is spent for aligning sequences, the observed speed-ups are almost same as the factor reduction in number of shuffles, as can be seen from Table 1. This can also be seen in light of time complexity analysis. With the proposed heuristic, the time complexity changes from $O(Nmn)$ to $O(\frac{N}{|DDP|}mn)$, which is a speed-up of $|DDP|$. The proposed DDP heuristic as used in this work has been shown to give speed-ups of up to 40 without degrading retrieval accuracy performance, which can be extremely useful in many applications requiring the analysis of sequence-pairs. The method can now easily be used for small database searches, for which pairwise statistical significance has earlier been shown to give significantly better results than popular database search programs like BLAST, PSI-BLAST, and SSEARCH.

The implementation of the proposed heuristic for fast pairwise statistical significance estimation is available for free academic use at:
www.cs.iastate.edu/$\sim$ankitag/PairwiseStatSig_DDP.html

5. CONCLUSION AND FUTURE WORK

In this paper, we propose and implement the Derived Distribution Points heuristic for fast pairwise statistical significance estimation, which takes advantage of the nature of pairwise statistical significance estimation process. Carefully chosen DDP sets have been shown to give significant speed-ups and at the same time maintain retrieval accuracy and statistical significance accuracy, as compared to not adding artificial scores at all, or adding scores in a fashion different than the proposed way. The proposed heuristic has been shown to give significant speed-up of up to 40 without loss of retrieval accuracy, which is expected to be extremely useful in the wide variety of applications based on sequence comparison.

This work provides a lot of scope for future work. It would be interesting to explore more with other values of the parameters used in this work, like number of shuffles N, and the set DDP. It may be possible to make the DDP sets more sequence-specific by making them dependent on the alignment parameters (since the alignment score distribution depends on various factors like alignment program, scoring scheme, sequence lengths, sequence compositions [21]).

Also, considering that the speed-up found is obtained on a single processor, it may be combined with hardware acceleration using multi-processors [7, 13], which can possibly make it an extremely useful tool in bioinformatics, given the all-pervading application of sequence comparison in bioinformatics. Future work also includes the application of the proposed heuristics for estimating pairwise statistical significance using position-specific substitution matrices. It may also be used in conjunction with BLAST to recover the homologs missed by BLAST.

6. REFERENCES

[1] A. Agrawal, V. Brendel, and X. Huang. Pairwise statistical significance versus database statistical significance for local alignment of protein sequences. In *Bioinformatics Research and Applications*, volume 4983 of *LNCS(LNBI)*, pages 50–61. Springer Berlin/Heidelberg, 2008.

[2] A. Agrawal, 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. 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.

[4] 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*, 10(Suppl 3):S1, 2009.

[5] A. Agrawal and X. Huang. Pairwise statistical

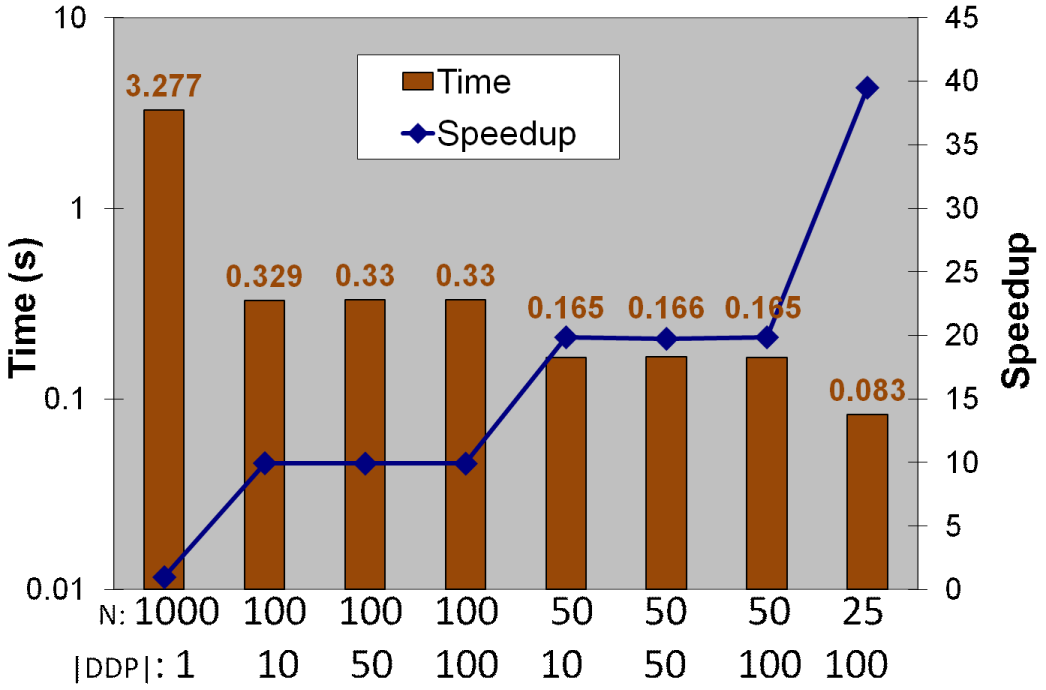

Figure 8: Execution times and speed-ups with DDP heuristic. Only those combinations of N and DDP are used which gave comparable retrieval accuracy performance to the base case (with DDP heuristic disabled)

significance of local sequence alignment using sequence-specific and position-specific substitution matrices. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2009. 25 Sept. 2009.

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

[7] A. Agrawal, S. Misra, D. Honbo, and A. Choudhary. Mpipairwisestatsig: Parallel pairwise statistical significance estimation of local sequence alignment. In *ECMLS proceedings of HPDC 2010*, 2010. to appear.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] D. Honbo, A. Agrawal, and A. Choudhary. Fpga-accelerated local sequence alignment for pairwise statistical significance estimation. In *Proceedings of BIOCOMP 2010*, 2010. to appear.

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

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

[16] 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.

[17] 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.

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

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

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

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

[22] 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.

[23] 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.

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

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

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

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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.

[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*, 21(7):902–911, 2005.

[36] 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.