MPIPairwiseStatSig: Parallel Pairwise Statistical Significance Estimation of Local Sequence Alignment

Ankit Agrawal, Sanchit Misra, Daniel Honbo, Alok Choudhary
Dept. of Electrical Engg. and Computer Science
Northwestern University
2145 Sheridan Rd
Evanston, IL 60201
USA
{ankitag,smi539,dkh301,choudhar}@eecs.northwestern.edu

ABSTRACT

Sequence comparison is considered as a cornerstone application in bioinformatics, which forms the basis of many other applications. In particular, pairwise sequence alignment is a fundamental step in numerous sequence comparison based applications, where the typical purpose of pairwise sequence alignment step is homology detection, i.e., identifying related sequences. 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 superior than database statistical significance for homology detection (evaluated in terms of 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. In this paper, we present a parallel algorithm for pairwise statistical significance estimation, called MPIPairwiseStatSig, implemented in C using MPI. Distributing the most compute-intensive portions of the pairwise statistical significance estimation procedure across multiple processors has been shown to result in near-linear speed-ups for the application.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms

Experimentation

1. INTRODUCTION

With sequencing becoming easier and more affordable, a tremendous amount of biological sequence data is becoming easily available in the public domain. Analyzing this data and deriving meaningful information requires intelligently designed computational methodologies. High performance computing techniques are expected to play a key role in achieving this goal, and are already widely used in various applications in bioinformatics and computational biology. Pairwise sequence alignment is one of the most important computational problems in bioinformatics for analyzing and comparing DNA and protein sequences [29, 9, 10]. There exist many algorithms [32, 16, 15] and heuristic-based approaches [26, 10, 28, 19, 18] for local sequence alignment.

In bioinformatics, almost everything depends on the interrelationship between sequence, structure, and function, which makes sequence comparison a cornerstone application in bioinformatics for finding biologically related sequences (homologs). Pairwise alignment methods report an alignment score for an alignment of two sequences, and pairs of related sequences should, in general, have higher scores. But the alignment score is just a mathematical score which by itself does not tell anything about the relatedness of the sequences. For instance, two related sequences of length 100 can have an optimal alignment score of 50, and two unrelated sequences of length 500 can have an optimal alignment score of 200. Therefore, to comment on the relatedness of the two sequences being aligned, the statistical significance of the alignment score, which is the likelihood of that alignment score being produced by the alignment of two unrelated sequences of similar features, is commonly estimated. An alignment score is more statistically significant if it has a low probability of occurring by chance. Since the alignment score distribution depends on various factors like the alignment program, scoring scheme, sequence lengths, and sequence compositions [22], 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 [7, 25, 22, 20].

The statistical significance of hits (database sequences found to be similar to the query sequence) reported by popular database search programs like BLAST [10], FASTA [27, 28], SSEARCH (using full implementation of Smith-Waterman algorithm [32]), and PSI-BLAST [10, 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

HPDC'10, June 20–25, 2010, Chicago, Illinois, USA.

Table 1: Execution time break-up for different stages of pairwise statistical significance estimation

Seq. length	Shuffling	Alignment	Fitting
128	1.2%	98.6%	0.2%
256	0.7%	99.3%	0.0%
512	0.3%	99.7%	0.0%
1024	0.2%	99.8%	0.0%

BLAST and PSI-BLAST programs [31, 34, 35], which have been shown to improve search performance using composition-based statistics and other enhancements.

Recently, an alternate method to evaluate the statistical significance of an alignment was studied. Known as pairwise statistical significance [1, 2], it is specific to the sequence-pair being aligned, and independent of any database. Further studies on pairwise statistical significance using multiple parameter sets [4], and sequence-specific/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, the current implementations are quite slow for pairwise statistical significance to be used in many applications.

Message Passing Interface (MPI) [13] is a library specification for the message-passing model of parallel computation. In this computation model, independent processes share data and synchronize execution by sending and receiving messages. MPI is widely implemented and deployed on distributed memory systems, from small clusters to massively parallel supercomputers, making it a good choice for the implementation of distributed algorithms.

In this paper, we use MPI to parallelize the pairwise statistical significance estimation procedure. The pairwise statistical significance estimation procedure can be decomposed into three main tasks: shuffling, alignment, and fitting. As shown in Table 1, the alignment and shuffling tasks comprise the overwhelming majority of the total execution time. Decreasing the time taken by these two tasks will therefore result in significant performance benefits for the algorithm as a whole, which is confirmed by experimental results.

The rest of the paper is organized as follows: Section 2 presents a formal description of pairwise statistical significance estimation procedure, followed by the proposed parallel algorithm 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

The score distribution for ungapped local alignment is known to follow a Gumbel-type EVD [17], 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 [33, 8, 27, 21, 24, 15], 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 [20]. There exist some excellent reviews on statistical significance in sequence comparison [25, 30, 22, 20].

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 [21], PRSS [28], censored-maximum-likelihood fitting [11], linear regression fitting [15]. The maximum likelihood fitting with censoring left of peak (described as type-I censoring in [11]) was found to be 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.

3. PARALLEL PAIRWISE STATISTICAL SIGNIFICANCE ESTIMATION

As presented in the previous section, each shuffling operation during the pairwise statistical significance estimation procedure is independent, as is each alignment operation. Moreover, each alignment operation depends only on the output of a single shuffling operation. We can therefore express the procedure as N independent shuffled alignments, each of

Input: Sequence-pair ($Seq1$, $Seq2$), Substitution matrix S, Gap opening penalty p, Gap extension penalty r, Number of shuffles N.
Output: Pairwise statistical significance pss of pairwise alignment score pas of $Seq1$ and $Seq2$.

The number of processors is p.

for each processor p_i $0 < i < p-1$

1. Initialization
   (a) Load inputs
   (b) srand(time(NULL)+i*i) # give different random seeds for different processors
   (c) $N = ceil(N/p) * p$ # make N a multiple of p
2. $pas = SWAlignmentScore(Seq1, Seq2, S, p, r)$ # get pairwise alignment score using Smith-Waterman algorithm
3. Construct distributed empirical score distribution: *do* N/p times,
   (a) $Shuffle(Seq2)$
   (b) Add $SWAlignmentScore(Seq1, Seq2, S, p, r)$ to $localScores[]$ # get N/p scores by aligning $Seq1$ with shuffled versions of $Seq2$ using Smith-Waterman algorithm
4. $Gather(scores, localScores)$ # gather all $localScores$ at different processors to $scores$ at processor p_0
5. if $(i == 0)$ # Only p_0 does the fitting
   (a) $[K, \lambda] = EVDCensoredMLFit(scores)$ # get statistical parameters K and λ using censored maximum likelihood fitting
   (b) $pss = 1 - exp(-Kmn * \exp^{-\lambda \times pas})$ # pairwise statistical significance

Figure 1: MPIPairwiseStatSig psuedo-code.

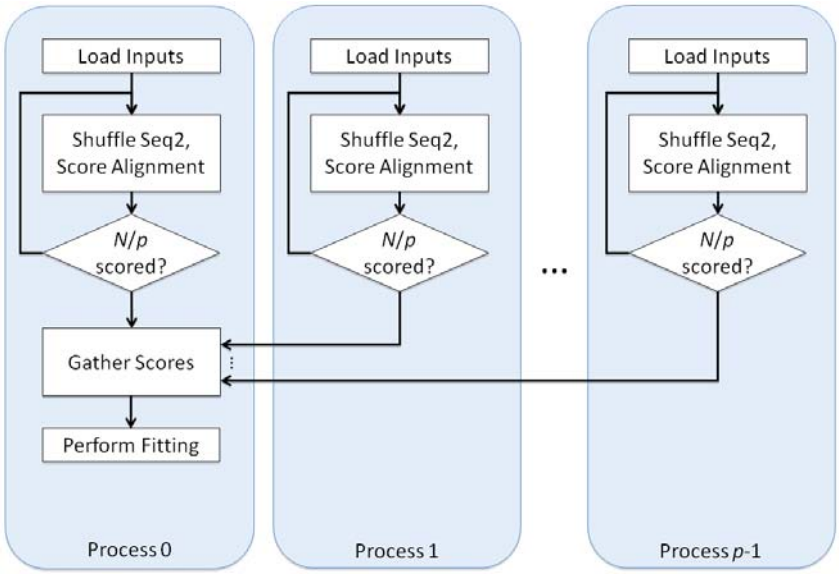

Figure 2: Flowchart depicting the distribution of workload for MPIPairwiseStatSig.

which shuffles $Seq2$ and aligns the shuffled sequence against $Seq1$, followed by the fitting task. The independent shuffled alignment tasks map very well to programming models capable of expressing task parallelism, like MPI. Figures 1 and 2 present the pseudo code and a simplified flowchart for the parallelized algorithm.

4. EXPERIMENTS AND RESULTS

In this section we present the timing results of MPIPairwiseStatSig. The MPIPairwiseStatSig program was used to estimate the pairwise statistical significance of five pairs of sequences of length 100, 200, 400, 800, and 1600, for 2, 4, 8, 16, 32, and 64 processors. The nodes of the cluster used for these experiments were 2.8 GHz dual Intel Xeon nodes. The BLOSUM50 substitution matrix was used for alignment with affine gap penalty of $10+2k$ for a gap of length k. The number of shuffles N was set to 1000, as used in previous studies on pairwise statistical significance [1, 2, 3, 4, 5]. The running times for each run were recorded and compared with the running times of the sequential version of the program. Fig. 3 presents the timing and speed-up results. All times are in seconds. Almost linear speed-ups are observed due to parallelizing the most task-parallel, compute-intensive step of the algorithm.

Fig. 4 presents the break-up of the time the algorithm spent during various stages. The shuffling and alignment task, which is distributed across the different processors, takes the maximum amount of time. As number of processors are increased, the shuffling and alignment time per processor reduces, thereby giving near-linear speed-ups.

MPIPairwiseStatSig is guaranteed to deliver the same performance as PairwiseStatSig [1, 2] in terms of statistical significance accuracy and retrieval accuracy, since the random shuffles across the different processors are seeded differently, which makes the shuffles equally pseudo-random as performing them all on a single processor.

For the database search application, pairwise statistical significance has earlier been shown to give significantly better results than popular database search programs like BLAST, PSI-BLAST, and SSEARCH, and hence, MPIPairwiseStatSig can now make pairwise statistical significance more readily usable for many sequence-comparison applications such as database search, for which using single processor would be extremely slow.

The MPIPairwiseStatSig program for pairwise statistical significance estimation is available for free academic use at www.cs.iastate.edu/ ∼ankitag/ MPIPairwiseStatSig.html

5. CONCLUSION AND FUTURE WORK

In this paper, the pairwise statistical significance estimation procedure is parallelized using Message Passing Interface (MPI) library, and a C implementation of the same is made available. By distributing the most heavily compute-intensive task of shuffling and alignment, near-linear speed-ups have been obtained, which is expected to be extremely useful in the wide variety of applications based on sequence comparison.

Future work includes further improving end-to-end application speed-up using data parallel methods for accelerating the alignment task. Apart from the task parallel implementation presented in this paper, there also exist data parallel methods for accelerating the alignment task, such as Streaming SIMD Extension (SSE) implementations [12], and FPGA implementations [23, 14], which exploit parallelism within each alignment operation. These approaches are complementary to the task parallel implementation presented here, in the sense that MPI can be used to parallelize the shuffled alignments across multiple nodes, and data parallel methods can be used to speed up the computations performed on each node. Combining the two approaches should yield better speed-ups.

Future work also includes using MPIPairwiseStatSig for sequence-comparison applications in bioinformatics. In particular, since pairwise statistical significance has been earlier shown to give superior retrieval accuracy than popular database search programs, a database search program can be designed based on pairwise statistical significance using MPIPairwiseStatSig. MPIPairwiseStatSig can also be used in conjunction with a fast database search program like 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 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] 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.

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

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and

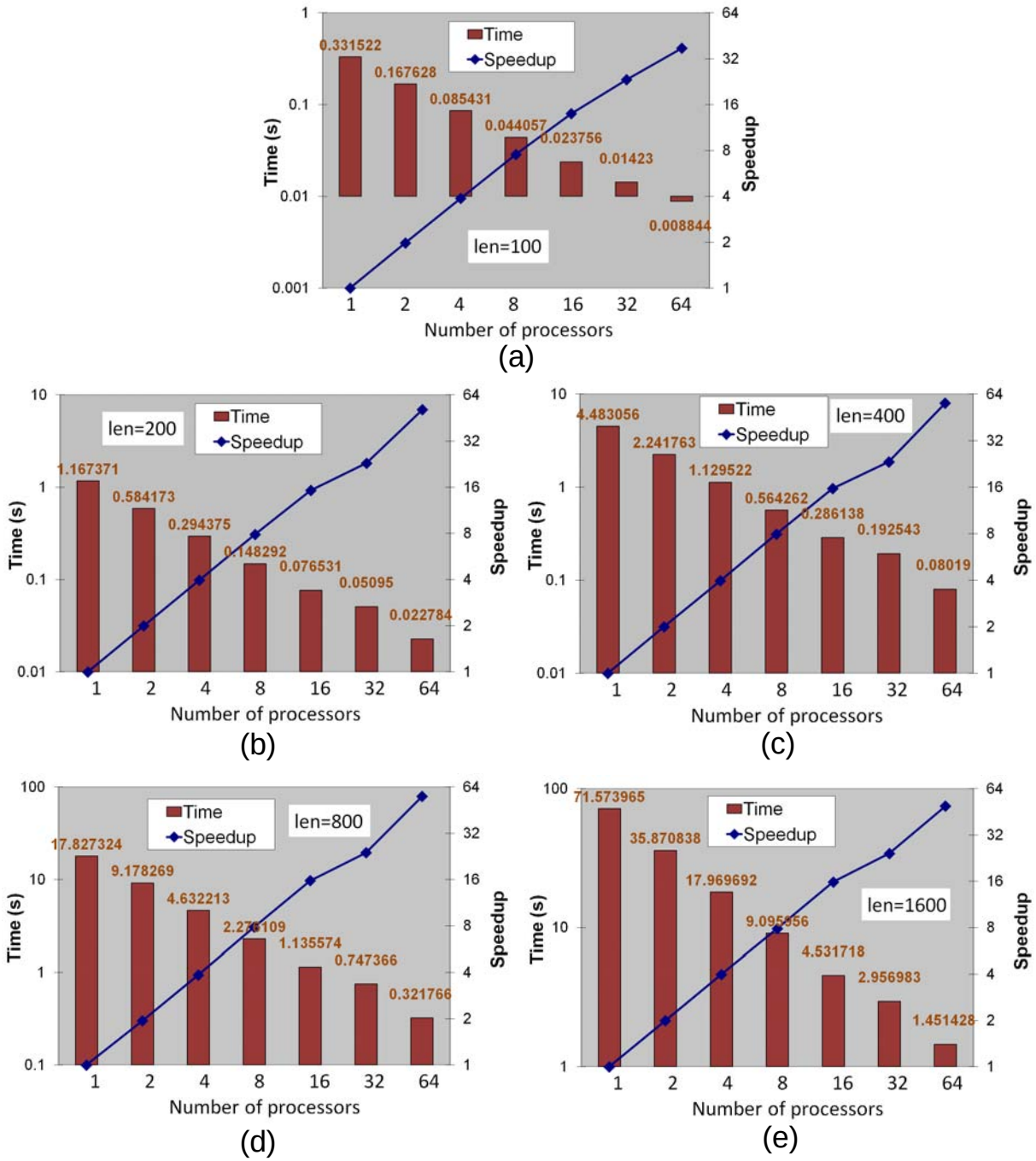

Figure 3: Timing and speed-up results using MPIPairwiseStatSig for sequence-pairs of length (a) 100, (b) 200, (c) 400, (d) 800, (e) 1600.

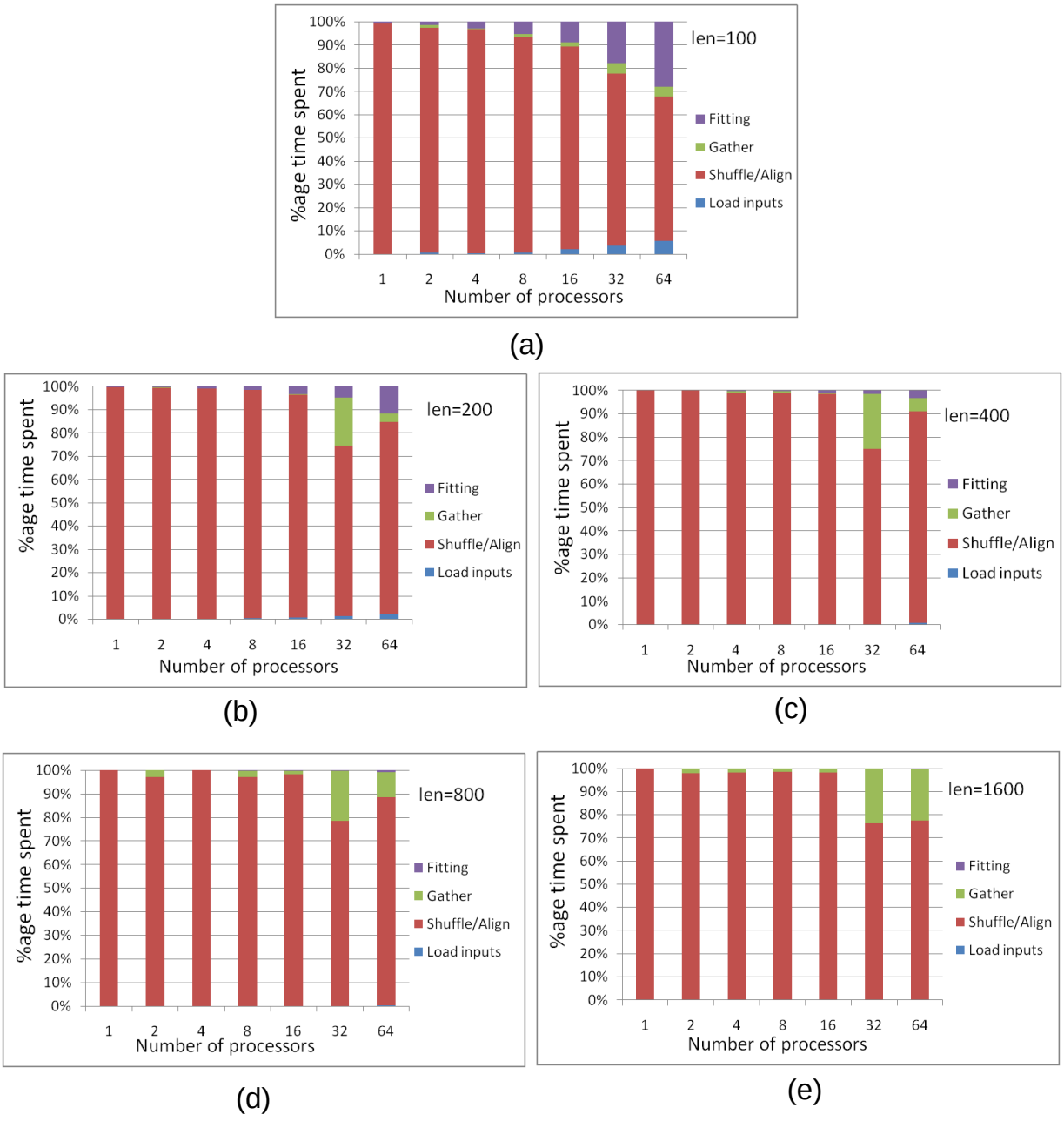

Figure 4: Percentage of time spent during different stages of the algorithm for sequence-pairs of length (a) 100, (b) 200, (c) 400, (d) 800, (e) 1600.

D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[10] 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.

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

[12] M. Farrar. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics*, 23(2):156–161, 2007.

[13] W. Gropp, E. Lusk, N. Doss, and A. Skjellum. A high-performance, portable implementation of the mpi message passing interface standard. *Parallel Comput.*, 22(6):789–828, 1996.

[14] D. Honbo, A. Agrawal, and A. Choudhary. Fpga-accelerated local sequence alignment for pairwise statistical significance estimation. 2010. to appear.

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

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

[17] 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.

[18] 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.

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

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

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

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

[23] T. Oliver, B. Schmidt, and D. Maskell. Reconfigurable architectures for bio-sequence database scanning on fpgas. *Circuits and Systems II: Express Briefs, IEEE Transactions on*, 52(12):851 – 855, dec. 2005.

[24] 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.

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

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

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

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

[29] 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.

[30] 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.

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

[33] 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.

[34] 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.

[35] 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.