The Effect Read Length has on the Performance of Adaptive Seeds for Sequence Alignment

Raymond Wan[1] r.wan@aist.go.jp
Szymon M. Kiełbasa[2] szymon.kielbasa@molgen.mpg.de
Paul Horton[1] horton-p@aist.go.jp
Martin C. Frith[1] martin@cbrc.jp

[1] Computational Biology Research Center, AIST, 2-4-7, Aomi, Koto-ku, Tokyo, 135-0064, Japan
[2] Department of Computational Biology, Max Planck Institute for Molecular Genetics, Ihnestr 63-73, 14195 Berlin, Germany

Keywords: local sequence alignment, seed-and-extend heuristics, adaptive seeds

1 Introduction

Efficient alignment algorithms are needed to keep up with the continued growth in sequencing technologies. While Smith-Waterman-based methods continue to improve, many practitioners still prefer faster heuristic methods despite the lack of a guarantee in optimality. Seed-and-extend heuristics (as used by BLAST) employ fixed-size seeds (or words) as starting points for local alignment.

In this work, we consider *adaptive seeds* which change in length based on the characteristics of the query and target sequences. We investigated the performance of adaptive seeds for reads whose lengths are typically found in sequencing data. In addition to real data, synthetic data with query sequence lengths longer than current standard read lengths were also considered.

2 Method and Results

Local sequence alignment aligns the similar parts (if present) of a sequence s to their matching regions in a target sequence t. Seed-and-extend heuristics create seeds of length l starting from every position in s to find *exact* matches in t. At every exact match, the seed is extended and local alignment is performed.

In contrast to fixed-size seeds, adaptive ones are associated with a frequency f instead of a length. They increase in length until the frequency of the seed in t is less than or equal to f. The main advantage of adaptive seeds is the ability to adjust the length of the seed based on the repetitiveness of the genome region matching the query. The extension phase of both types of seeds are identical.

Support for both types of seeds using a suffix array has been implemented in a publicly-available system called `LAST`. Further details about `LAST` can be found at its web site [1].

We compared the performance of adaptive seeds against fixed-size seeds by measuring the relationship between accuracy and elapsed running time, as l and f are varied. For each query, we recorded the maximum score attained by any method (the *pool*) and scored all methods attaining that maximum as correct for that query. A particular method is accurate if many of its query sequences achieve the same scores as the pool. The following alignment parameters were chosen: match score of 1, mismatch scores of -1, gap existence cost of 2, and a gap extension cost of 1. Alignments with scores less than 30 were discarded. All experiments were executed on a 2.0 GHz Dual-Core AMD Opteron Processor 246 with 6 GB of RAM.

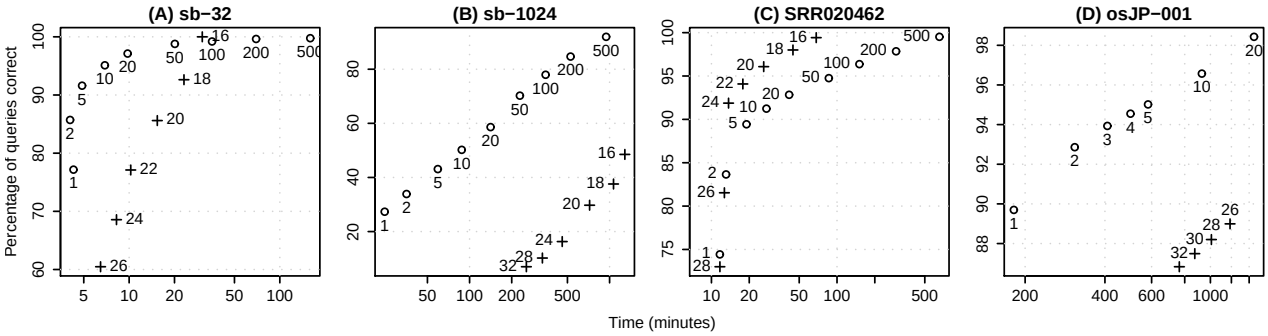

Figure 1: Performance of adaptive seeds "○" compared to fixed-size seeds "+" for synthetic data sets of (A) 32 nts and (B) 1,024 nts; (C) next-generation sequences (Illumina Genome Analyzer II, SRR020462) [35 nts]; (D) trace data (*O. sativa*, japonica cultivar group accession 001) [median of 753 nts].

Data sets were obtained from the genome, trace, and sequence read repositories of NCBI [2]. We downloaded the rice (*O. sativa*, japonica cultivar group) genome and one data set from each of the trace and sequence read archives. Synthetic query sets were created by taking random sequences of another grain (*S. bicolor*[1]). We used 1,048,576 and 32,768 queries of lengths 32 and 1,024 respectively. Duplicate sequences were removed for all data sets.

Figure 1 presents our results, beginning with the synthetic data sets ((A) and (B)). The results show that adaptive seeds out-perform fixed-size seeds. For example, for synthetic data of 1,024 nts (B), in 500 minutes, adaptive and fixed-size seeds achieve sensitivities of 80% and 20%, respectively. Additional experiments (not shown) indicate that this trend also holds for intermediate sequence lengths. Experiments with real data also show that adaptive seeds perform better as query sequence lengths increase ((C) and (D)).

Interestingly, the results of (C) show that fixed size seeds perform better for next-generation sequences, even though the read lengths are similar to those of (A). We believe the reason for this is that the performance of these two types of seeds depend on at least two other factors as well: repetitiveness of the genome and similarity of the reads to the genome. While the same target genome was used throughout our experiments, our synthetic data sets use *S. bicolor* instead of *O. sativa*.

3 Discussions

Our experiments have shown a general trend of adaptive seeds out-performing fixed-size seeds as the length of the sequences increase. While this is true for both synthetic and real data sets, our work has also shown that the performance of adaptive and fixed-size seeds depend on other factors as well. Consideration of how all of these factors interact with each other remains a topic for future work.

Acknowledgements This research was supported by funding from INTEC Systems Institute, Inc.

References

[1] LAST. `http://last.cbrc.jp/`.

[2] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37(Database issue):D5–D15, January 2009.

[1] This data set has not yet been curated by NCBI and, therefore, is available from `ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/`.