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

Abstract

Sequence alignment calculates the similarity between two sequences to assess whether or not they are homologous to each other. Efficient sequence alignment remains an important topic due to rapidly advancing sequencing technologies.

*While seed-and-extend heuristics (i.e., BLAST) do not guarantee optimality in comparison to Smith-Waterman-based methods, they offer a good compromise between efficiency and effectiveness. We have developed a seed-and-extend heuristic that employs **adaptive seeds**. In this work, we:*

1) Assess their performance against fixed-length seeds for next generation sequence data.
2) Conduct experiments with both real and synthetic data that represent current and future trends in read lengths.

1. Fixed-length vs Adaptive seeds

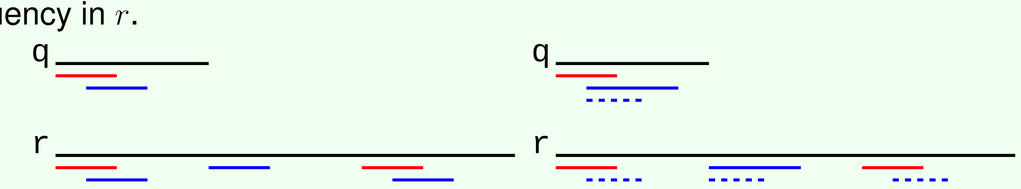
Local sequence alignment aligns a query q (of a query set Q) against a genome r. Seed-and-extend methods:

1) Take sub-sequences from q (i.e., **seeds**) to find exact matches in r.
2) And then perform Smith-Waterman alignments at these locations.

The difference between fixed-length and adaptive seeds is:

- Fixed-length seeds are seeds of a pre-determined length l.
- Adaptive seeds change in length until the seed occurs $\leq$ some frequency threshold f in r.

Adaptive seeds are better suited for the repetitive regions of a genome because the seed's length will increase as necessary to maintain a nearly constant frequency in r.



(a) fixed-length seeds (b) adaptive seeds ($f = 2$)

2. Software

We developed LAST which incorporates both fixed-length and adaptive seeds. Further details about LAST and adaptive seeds is described elsewhere [in preparation]. Here, we only give an overview and report on the performance of adaptive seeds relative to read length.

Source code to LAST: http://last.cbrc.jp/

3. Pooling

Our experiments use pooling to evaluate and compare the two types of seeds. Alignment results for a query set Q is a list of queries with their maximum alignment score.

- For both fixed-length and adaptive seeds, we consider several parameters for l and f (respectively).
- These results are merged into a single result set (the pool) by recording the maximum alignment score (if any) for each $q \in Q$.
- We then report the performance of each method relative to this pool as a percentage.

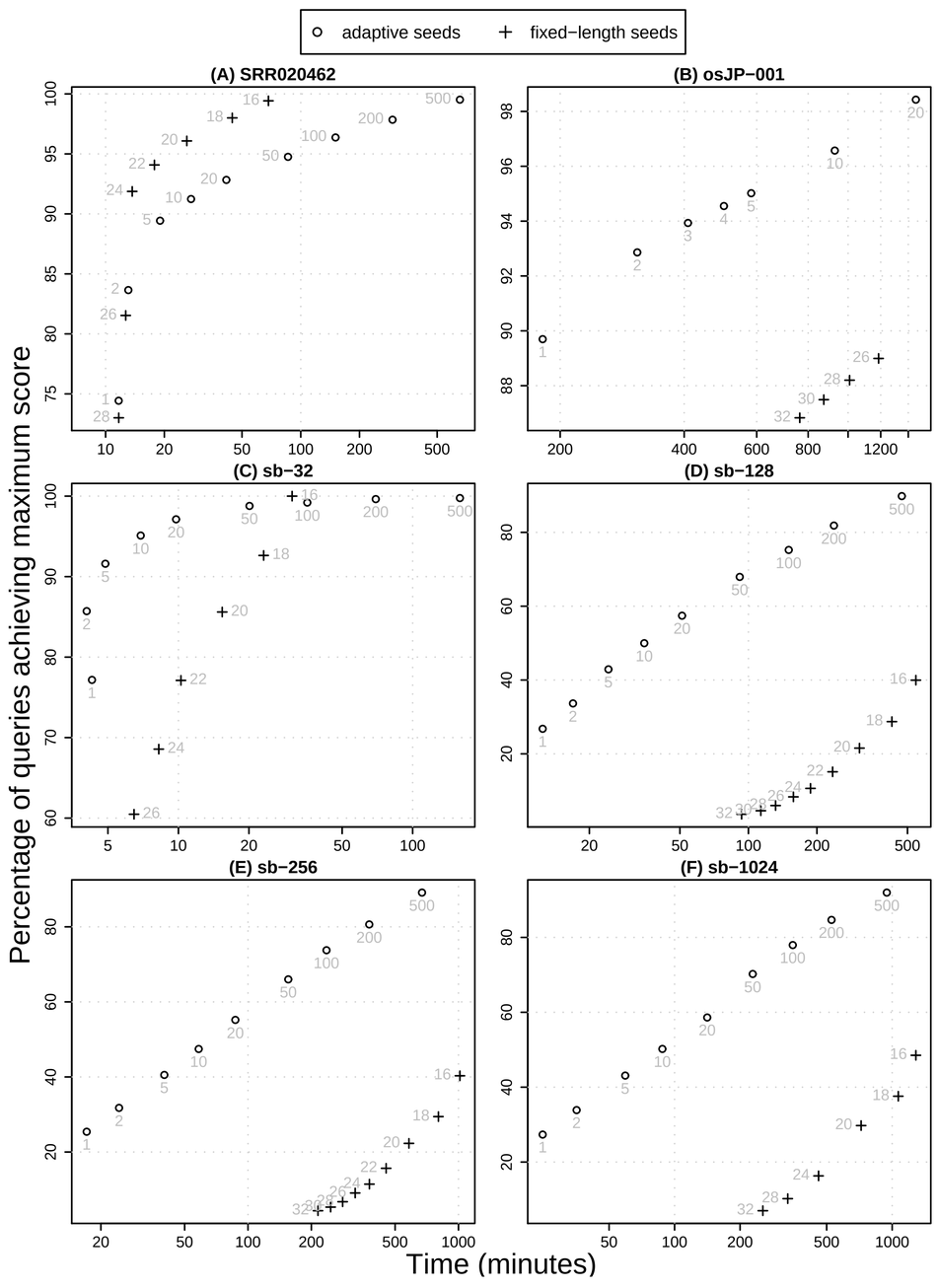
4. Experiments

Experiments were performed under these conditions and using these data sets:

- Match score: 1 Gap existence cost: 2
  Mismatch score: -1 Gap extension cost: 1
- Alignment scores: ≥ 30
- CPU: 2.0 GHz Dual-Core AMD Opteron Processor 246 with 6 GB of RAM

Data type	Species	NCBI source	Accession	Number of queries	Median length	Figs.
Illumina Genome Analyzer II	O. sativa	Short read archive	SRR020462	2,983,140	35	(A)
Trace data	O. sativa	Trace archive	osJP-001	499,922	753	(B)
Synthetic data (random)	S. bicolor	Genome ftp site	–	1,048,576 to 32,768	32 to 1,024	(C–F)
Genome	O. sativa	Genome ftp site	–	–	–	(A–F)

5. Results



6. Discussion

The results show the following.

- Adaptive seeds always out-perform fixed-length seeds.
- The exception is SRR020462 [Fig. A], but the difference between the seed types is not as large as the other cases [5-10% versus at most 60%].
- As the read lengths increase the adaptive seeds perform better than fixed-length ones.

Other factors that affect the performance of seed-and-extend methods include:

- repetitiveness of the genome,
- compositional bias of the genome,
- similarity of the reads to the genome (i.e., same or different species).

We believe that considering these factors along with read length will further explain our results above.

The 2009 Annual Meeting of the Computational Biology Research Center (**CBRC 2009**), December 3-4, 2009, Tokyo, Japan; The 20th International Conference on Genome Informatics (**GIW 2009**), December, 14-16, 2009, Yokohama, Japan