Supplementary Material for LAMBDA - The Local Aligner for Massive Biological Data

Hannes Hauswedell, Jochen Singer and Knut Reinert
Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany

1 S1: Post-processing of candidate regions

In order to verify a candidate region the seed must be extended, which is computationally expensive. Therefore, one should try to remove extension candidates which will (a) lead to biologically insignificant alignments, or (b) produce alignments that have already been computed as the result of another hit.

Masking To identify candidates for the first sort of problem, so called *low complexity region filters* are frequently used, among them SEG [5] for protein sequences and DUST for nucleic acid sequences [see the supplementary data to 4]. Low complexity regions include repetitive or insignificant sequence patterns that can score well nevertheless – and hinder efficient searches [1]. Lambda can read masking files generated by `segmasker` (the SEG implementation provided by BLAST+ [2]) for the subject sequences and it implements soft-masking in that it filters out any seeds reaching into a masked region with at least half of their length.

Merging In order to address the second problem, overlapping seeds and seeds within a certain proximity are merged – since it is likely that if both are extended to valid alignments, it will be the same one. This is a very heuristic parameter, and its influence on results is not immediately intuitive, however a value equal to the seed length has shown good results, both in terms of sensitivity and performance.

Local alignment Another optimization is computing a local alignment of the candidate region using the original alphabet and discarding candidates that score below a certain threshold (without extending them). This is possible because the clustering of the alphabet reduction groups some amino acids together that are further related than others, and stretches of these distantly related amino acids are likely insignificant. For Blosum62 we have found that minimum scores of 3-3.5 times the seed length show good results on different datasets. Through these it is possible to avoid up to 85% of extensions with minimal impact on sensitivity.

As the seeds have no gaps, a custom local alignment function was developed for Lambda. It comes without the overhead of the general implementation and decreases running time by 22% over the generalized SeqaAn-implementation.

2 S2: MEGAN Analysis

In order togive an impression of the results reported by the tools, we conducted a cluster analysis with the help of MEGAN5[1], the newest version of the widely used program MEGAN [3]. In order to do so, we used the default settings of MEGAN, however we disabled the filters[2], such that the results are not biased by MEGAN manipulation.

Figure 1 shows the Neighbor Joining Tree (NJ-Tree) of the cluster analysis of the first dataset. As can be seen, UBLAST, RAPSearch2 and LAMBDA, in their default and slow configuration, show relatively small distances to each other and to BLAST+. In contrast, their fast configuration as well as PAUDA show much larger distances. Therefore, the cluster analysis conforms the similar quality of the results of RAPSearch2, UBlast and Lambda.

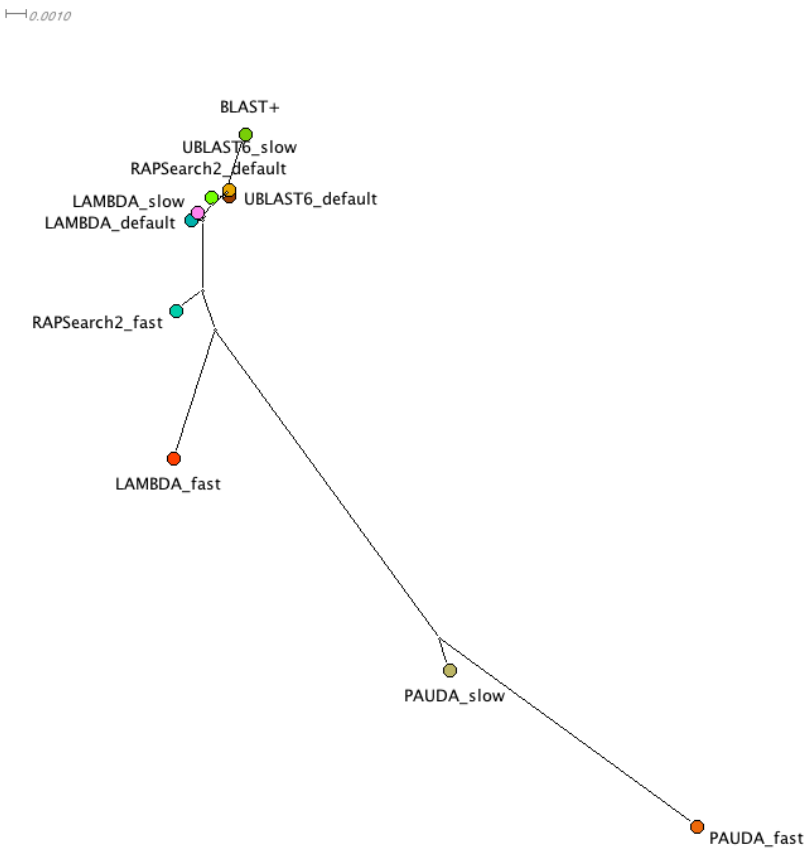

Figure 1: Result of the cluster analysis of the first dataset performed with MEGAN. The NJ-Tree of the data is shown.

As in Figure 1 Figure 2 shows very similar results for the second dataset. The major difference to Figure 1 is that the distance of the default and slow configuration of UBLAST, RAPSearch2 and LAMBDA to BLAST+ are much larger. However, the trend is the same as in Figure 1.

[1]http://ab.inf.uni-tuebingen.de/software/megan5/

[2]Min Support: 1, Min Score: 0.0, Max Expected: 1.0, Top Percent: 100, Min Complexity: 0.0

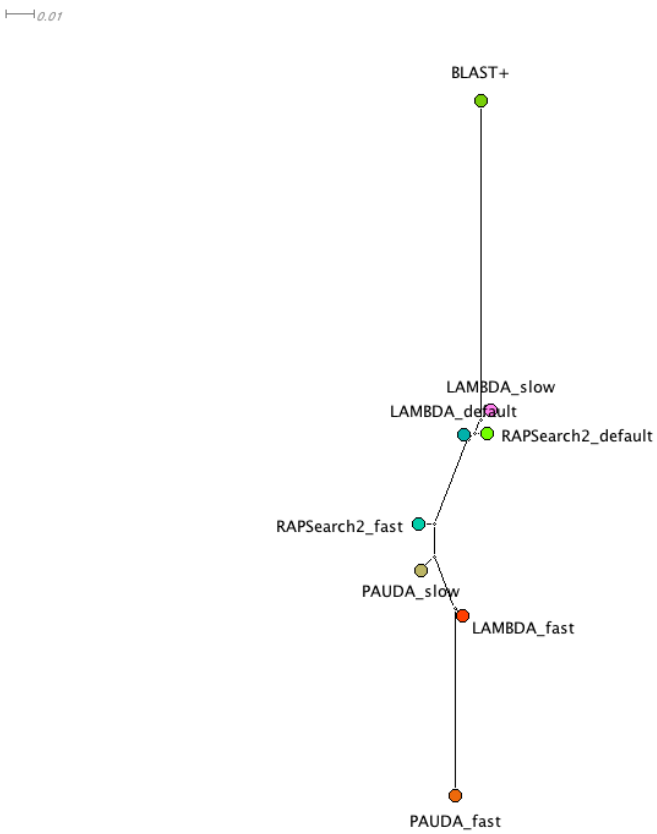

Figure 2: Result of the cluster analysis of the second dataset performed with MEGAN. The NJ-Tree of the data is shown.

References

[1] S. F. Altschul, M. S. Boguski, W. Gish, and J. C. Wootton. Issues in searching molecular sequence databases. *Nat. Genet.*, 6(2):119–129, Feb 1994.

[2] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason S. Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009.

[3] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. MEGAN analysis of metagenomic data. *Genome Res.*, 17(3):377–386, March 2007.

[4] Aleksandr Morgulis, E. Michael Gertz, Alejandro A. Schäffer, and Richa Agarwala. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, 22(2):134–141, 2006.

[5] John C. Wootton and Scott Federhen. Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. *Computers & Chemistry*, 17(2):149–163, 1993.