Memory efficient calculation of phylogenetic distances for whole genome sequences

Dirk Willrodt[1], Bernhard Haubold[2], Stefan Kurtz[1]

[1] *Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany*

[2] *Max-Planck-Institute for Evolutionary Biology August-Thienemann-Str. 2, 24306 Plön, Germany*

15. August 2011

When comparing whole genomes one often needs to calculate a distance measure which is related to the number of mutations in the genomes. Traditionally, this is done by multiple sequence alignment methods, which however, do not scale for large genomes. For closely related organisms, alignment free methods can be applied. Several studies have shown that alignment free distance calculations well approximate alignment based methods and often lead to considerable improvements in calculation time [1].

One example of such an alignment free method is the Jukes-Cantor corrected distance measure K_r [2] which is based on the shortest unique substrings between pairs of genomes. K_r is, for example, implemented in the software *KR2* [3]. The latter implements the search for shortest unique substrings using enhanced suffix arrays [4]. These can become very large, often requiring server class machines when *KR2* is applied to large eukaryotic genomes. For example, to compare 12 *Drosophila* genomes [5], *KR2* requires 72 GB of RAM.

We modified the algorithm used in *KR2* such that random access to the enhanced suffix array is eliminated. The resulting algorithm streams the tables of the enhanced suffix array in sequential order, leading to a very small memory peak during the calculation. We implemented the modified algorithm in the *GenomeTools*-toolkit (`http://genometools.org`). This provides a scalable solution to computing the K_r-distance measure. For example, the 12 *Drosophila* genomes can be compared in about 3 hours using only 3 GB of RAM, which means a 20-fold space improvement over *KR2* and a comparable running time.

We will show that our approach allows to calculate phylogenetic distances on the basis of whole genome data for sets of whole drosophila and mouse genomes.

Contact E-mail: willrodt@zbh.uni-hamburg.de

Literatur

[1] S. Vinga and J. Almeida. Alignment-free sequence comparison-a review. *Bioinformatics*, 19:513–523, Mar 2003.

[2] B. Haubold, P. Pfaffelhuber, M. Domazet-Lošo, and T. Wiehe. Estimating mutation distances from unaligned genomes. *Journal of Computational Biology*, 16:1487–1500, 2009.

[3] M. Domazet-Lošo and B. Haubold. Efficient estimation of pairwise distances between genomes. *Bioinformatics*, 25:3221–3227, 2009.

[4] M.I. Abouelhoda, S. Kurtz, and E. Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, 2:53–86, 2004.

[5] Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–18, 2007.