

UniMoG—a unifying framework for genomic distance calculation and sorting based on DCJ

Rolf Hilker^{1,2,*}, Corinna Sickinger², Christian N.S. Pedersen³ and Jens Stoye^{2,4}

¹Computational Genomics, Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany, ²Institute for Bioinformatics, Center for Biotechnology (CeBiTec), Bielefeld University, 33615 Bielefeld, Germany, ³Bioinformatics Research Center (BiRC), Aarhus University, 8000 Aarhus, Denmark and ⁴Genome Informatics, Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany

Associate Editor: David Posada

ABSTRACT

Summary: UniMoG is a software combining five genome rearrangement models: double cut and join (DCJ), restricted DCJ, Hannenhalli and Pevzner (HP), inversion and translocation. It can compute the pairwise genomic distances and a corresponding optimal sorting scenario for an arbitrary number of genomes. All five models can be unified through the DCJ model, thus the implementation is based on DCJ and, where reasonable, uses the most efficient existing algorithms for each distance and sorting problem. Both textual and graphical output is possible for visualizing the operations.

Availability and implementation: The software is available through the Bielefeld University Bioinformatics Web Server at <http://bibiserv.techfak.uni-bielefeld.de/dcj> with instructions and example data.

Contact: rhilker@cebitec.uni-bielefeld.de

Received on July 5, 2011; revised on July 5, 2012; accepted on July 6, 2012

1 INTRODUCTION

Genome rearrangements describe the dynamics of evolution at an abstracted genomic level, in contrast to local mutations of single DNA base pairs. Very little is known about the exact procedure of rearrangement events and how and when they are triggered. More detailed knowledge of evolution could help to improve the understanding of the mechanisms important for survival and development of species. The evolutionary distance between at least two organisms with shared gene content can be estimated by solving the combinatorial problem of finding a possible sequence of rearrangement operations among their shared genes under the aspect of parsimony. Thus, all genes unique to one of the genomes are ignored and only one representative among duplicated genes is chosen for the comparison.

In recent years, large amounts of genomic data have become available and genome comparison has become a routine task. For example, the Chimpanzee Sequencing and Analysis Consortium (2005) compared chimpanzee and human genomes and developed a catalogue of genetic differences. Since both are closely related, only one fusion of two chromosomes and several inversions were identified. Another example is the comparison of human and mouse genomes by Pevzner and Tesler (2003).

Among other methods they used GRIMM (Tesler, 2002b) for the analysis, because utilizing automated methods allows for easier and faster analyses, no matter how divergent the investigated organisms are. GRIMM is based on the Hannenhalli and Pevzner (HP) model (Hannenhalli and Pevzner, 1995), thus its set of rearrangement operations comprises inversions, translocations, fusions and fissions of linear genomes. However, one can investigate the phylogenetic distance under different aspects and the HP model is only one of the common models. Besides the HP model we consider four additional models. The inversion model (Hannenhalli and Pevzner, 1999) allows for inversions of internal genomic regions in linear, uni-chromosomal genomes, while the translocation model (Hannenhalli, 1996) comprises the exchange of two linear chromosome ends. As already mentioned, HP combines both models and adds fusions and fissions of two chromosomes to the repertoire of rearrangement operations. Among the included models, the most general is the double cut and join (DCJ) model (Bergeron *et al.* 2006b; Yancopoulos *et al.* 2005), which allows for all common rearrangement operations: inversions, translocations, fusions, fissions, circularizations and decircularizations. Besides these operations, block interchanges, which describe the exchange of two DNA segments, can be mimicked through two operations by all models, except the inversion model. Finally, the restricted DCJ model (Kováč *et al.*, 2011) allows the same operations as the DCJ model, but constricts it by requiring immediate decircularization in the next step for emerging circular chromosomes.

In our software, UniMoG, the DCJ Adjacency Graph data structure (Bergeron *et al.*, 2006b), serves as basis for all calculations, and in contrast to GRIMM it implements, based on DCJ, all of these five distance models and is able to return either the desired distance or the distance and a corresponding optimal sorting scenario. For fast comparisons between the different distances, it is also possible to calculate all five distances and sorting scenarios at once, if applicable. Another advantage is that the input is neither limited to two genomes at a time nor can genes only be represented by integers. Instead, gene names are converted to integers for the internal representation. In the case of multiple input genomes, all of them are compared pairwise with each other. The distance results are then returned in a matrix, which is also provided in PHYLIP format (Fig. 1, inset), and can further be fed into distance-based phylogenetic tree reconstruction methods, possibly after applying distance

*To whom correspondence should be addressed.

correction models like the ones presented by Lin and Moret (2008).

Because of the efficient implementation, UniMoG can handle large genomes and was tested with genomes up to 32 500 genes without encountering any problems. For further improvement of the computational performance, regions with identical gene order can be merged into larger syntenic blocks, since none of the considered models can break up conserved blocks.

Conflict of Interest: none declared.

Bergeron, A. *et al.* (2005) The inversion distance problem. In Gascuel, O. (ed.) *Mathematics of Evolution and Phylogeny*. Oxford University Press, Oxford, UK, pp. 262–290.

Bergeron, A. *et al.* (2006a) On sorting by translocations. *J. Comput. Biol.*, **13**, 567–578.

Bergeron, A. *et al.* (2006b) A unifying view of genome rearrangements. In *Proceedings of WABI 2006*. LNBI 4175. Springer Verlag, Heidelberg, pp. 163–173.

Erdős, P.L. *et al.* (2011) Balanced vertices in trees and a simpler algorithm to compute the genomic distance. *Appl. Math. Lett.*, **24**, 82–86.

Hannenhalli, S. (1996) Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Appl. Math.*, **71**, 137–151.

Hannenhalli, S. and Pevzner, P.A. (1995) Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of FOCS 1995*. IEEE Computer Society Press, Los Alamitos, CA, pp. 581–592.

Hannenhalli, S. and Pevzner, P.A. (1999) Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *J. ACM*, **46**, 1–27.

- Jean, G. and Nikolski, M. (2007) Genome rearrangements: a correct algorithm for optimal capping. *Inform. Process. Lett.*, **104**, 14–20.
- Kováč, J. *et al.* (2011) Restricted DCJ model: rearrangement problems with chromosome reincorporation. *J. Comput. Biol.*, **18**, 1231–1241.
- Lin, Y. and Moret, B.M.E. (2008) Estimating true evolutionary distances under the DCJ model. *Bioinformatics*, **24**, 114–122.
- Pevzner, P.A. and Tesler, G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.*, **13**, 37–45.
- Tannier, E. *et al.* (2007) Advances on sorting by reversals. *Discrete Appl. Math.*, **155**, 881–888.
- Tesler, G. (2002a) Efficient algorithms for multichromosomal genome rearrangements. *J. Comput. Systems Sci.*, **65**, 587–609.
- Tesler, G. (2002b) GRIMM: genome rearrangements web server. *Bioinformatics*, **18**, 492–493.
- The Chimpanzee Sequencing and Analysis Consortium Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
- Yancopoulos, S. *et al.* (2005) Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, **21**, 3340–3346.