Simultaneous Solution to Synteny Blocks Construction and Genome Rearrangements Problems

Shuai Jiang and Max A. Alekseyev*

University of South Carolina, Columbia, SC, U.S.A.

One of the key computational problems in comparative genomics is reconstruction of the ancestral genomes from genomes of living species and the sequence of evolutionary events (*evolutionary history*), such as genome rearrangements as well as gene duplications and deletions, between the genomes. Traditional approaches addressing this problem solve the following sub-problems in order:

Synteny block construction. *Synteny blocks*, which are fundamental components in revealing evolutionary history, are formed by conserved sequences of homologous genes in different genomes. To enable rearrangement analysis of given genomes, they are first represented as sequences of non-overlapping synteny blocks (*synteny block construction problem*). Recent approaches for synteny blocks construction [3, 5] employ A-Bruijn graphs that were initially introduced for repeat classification problem [4]. These rather universal approaches do not however take into account information about the nature of structural changes in one genome as compared to the others.

Genome rearrangements problem. *Genome rearrangements* (such as *reversals*, *translocations*, *fusions*, and *fissions*) are evolutionary events that shuffle genomic material without altering it otherwise. When two genomes are represented as permutations of the same synteny blocks, it becomes possible to define the *evolutionary distance* as the minimal number of genome rearrangements required to transform one genome into the other. This definition further inspires the most parsimonious *genome rearrangements problem* asking for reconstruction of ancestral genomes (at the internal nodes of the phylogenetic tree) composed of the same synteny blocks as the given genomes, such that the total evolutionary distance along the branches of the phylogenetic tree is minimized. Traditionally, these computational problems are addressed using graph-theoretical models (such as *breakpoint graphs*) and algorithms. There exist a number of ancestral genomes reconstruction tools, including our own *Multiple Genome Rearrangements and Ancestors* (MGRA) tool [1].

*Corresponding author. Email: maxal@cse.sc.edu

However, most of such tools do not deal with gene duplications and deletions, while others account for these evolutionary events but are limited to the case of two closely related genomes [2, 6].

We show that synteny blocks construction and genome rearrangements problems can be solved simultaneously, which also improves quality of the resulting reconstruction of the ancestral genomes. Our approach is based on the notion of *noisy breakpoint graph* that is constructed directly from the homologs, similarly how conventional breakpoint graphs are constructed from the synteny blocks. We classify and analyze *artifacts* that are specific to noisy breakpoint graphs as compared to breakpoint graphs. We identify sources of these artifacts (such as erroneous homologs or duplications in the genomes) and eliminate them with appropriate graph "cleaning" operations. Combining these cleaning operations with traditional genome rearrangements analysis (as in MGRA), we efficiently transform the noisy breakpoint graph into the identity breakpoint graph on the set of true (non-erroneous) homologs. This corresponds to a transformation of each of the given genomes into the same ancestral genome along the branches of their phylogenetic tree. This transformation can be further used to recover ancestral genomes at the internal nodes of the phylogenetic tree [1] as well as to identify conserved (synteny) blocks of true homologs in the given genomes.

This work is supported by the National Science Foundation under Grant No. IIS-1253614.

[1] Alekseyev, M., and Pevzner, P. A. Breakpoint graphs and ancestral genome reconstructions. *Genome Research 19* (2009), 943–957.

[2] Bader, M. Sorting by reversals, block interchanges, tandem duplications, and deletions. *BMC Bioinformatics 10*, Suppl 1 (2009), S9.

[3] Peng, Q., Alekseyev, M., Tesler, G., and Pevzner, P. Decoding the Genomic Architecture of Mammalian and Plant Genomes: Synteny Blocks and Large-Scale Duplications. *Communications in Information and Systems 10*, 1 (2010), 1–22.

[4] Pevzner, P., Tang, H., and Tesler, G. De Novo Repeat Classification and Fragment Assembly. *Genome Research 14* (2004), 1786–1796.

[5] Pham, S. K., and Pevzner, P. A. DRIMM-Synteny: Decomposing Genomes into Evolutionary Conserved Segments. *Bioinformatics 26*, 20 (2010), 2509–2516.

[6] Yancopoulos, S., and Friedberg, R. Sorting genomes with insertions, deletions and duplications by dcj. *Lecture Notes in Computer Science 5267* (2008), 170–183.