

GENESIS: Genome Evolution Scenarios

Simon Gog, Martin Bader, and Enno Ohlebusch

Institute of Theoretical Computer Science, University of Ulm, D-89069 Ulm, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

ABSTRACT

Summary: We implemented a software tool called GENESIS for three different genome rearrangement problems: Sorting a unichromosomal genome by weighted reversals and transpositions (SwRT), sorting a multichromosomal genome by reversals, translocations, fusions, and fissions (SRTI), and sorting a multichromosomal genome by weighted reversals, translocations, fusions, fissions, and transpositions (SwRTTI).

Availability: Source code can be obtained by the authors, or use the web interface <http://www.uni-ulm.de/in/theo/research/genesis.html>.

Contact: {martin.bader, simon.gog, enno.ohlebusch}@uni-ulm.de

1 INTRODUCTION

During evolution, DNA molecules are subject to local and global mutations. Local mutations (point mutations) consist of the substitution, insertion, or deletion of single nucleotides, while global mutations (genome rearrangements) change the DNA molecules on a large scale. In unichromosomal genomes, the most common rearrangements are inversions (also called reversals in bioinformatics), where - from a mathematical point of view - a section of the genome is excised, reversed in orientation, and re-inserted. Biologically, inversions can be caused by replication errors. But also large-scale duplications, deletions (gene loss), insertions (e.g. horizontal gene transfer), and transpositions play a role. In a transposition, a section of the genome is excised and inserted at a new position in the genome; this may or may not also involve an inversion. In genomes with multiple chromosomes further genome rearrangements are translocations (in a reciprocal translocation, two non-homologous chromosomes break and exchange fragments), fusions (where two chromosomes fuse), and fissions (where a chromosome breaks into two parts).

Since the 1980s, computer scientists have developed several algorithms for reconstructing genome rearrangement scenarios that transform one genome into another genome (or equivalently, to sort a signed permutation into the identity permutation). These algorithms can be categorized according to the genome rearrangement operations they can deal with, and as to whether they take multiple chromosomes into account. For unichromosomal genomes, the following results are known. If only reversals are allowed, Hannenhalli and Pevzner's [7] algorithm yields an exact solution to the problem. The currently best algorithm for transpositions is a 1.375-approximation [4]. For equally weighted reversals and transpositions, Hartman and Sharan [8] devised a 1.5-approximation algorithm. Bader and Ohlebusch [1] extended their algorithm to a 1.5-approximation algorithm for any weight ratio between 1:1 and

1:2 (reversals:transpositions). Another program for the weighted case is DERANGE [3], but the authors of [3] did not provide a guaranteed approximation ratio.

For multichromosomal genomes, less results have been obtained so far. The most realistic solution to the problem was given by Hannenhalli and Pevzner [6]. Their algorithm takes reversals, translocations, fusions, and fissions into account and returns an exact solution. To the best of our knowledge, the new algorithm presented below is the first that augments their algorithm with transpositions.

We have implemented the following three algorithms:

1. The algorithm for SwRT by Bader and Ohlebusch [1] with quadratic running time. Moreover, the combination of this algorithm with a greedy strategy resulted in a practicable method. The price to be paid for this improvement is a cubic running time.
2. The algorithm for SRTI by Hannenhalli and Pevzner [6] with the improvements of Tesler [11] and of Ozery-Flato and Shamir [10]. The running time is quadratic.
3. An algorithm for SwRTTI, which is a combination of the two algorithms above. The running time of the algorithm is cubic. Our experiments show that it produces good results for biologically reasonable weights. To guarantee an approximation ratio of 2, we further combined the new algorithm with the strategy presented by Yancopoulos et al. [12], albeit for a different set of rearrangement operations.

GENESIS consists of two parts: The programs themselves and the web interface. The web interface is a simple front-end: One can choose the algorithm, set source and target genome, and gets the resulting rearrangement scenario. For the sake of readability, the web interface limits the size of permutations to 80, whereas the main programs can handle permutations of several thousands elements.

2 METHODS

All our algorithms work with the *reality-desire diagram* (also called *breakpoint graph*) as described in [2]. Let $c(\pi)$ denote the number of cycles in the reality-desire diagram, $c_{odd}(\pi)$ the number of cycles with an odd number of reality-edges (called *odd cycles*), and $c_{even}(\pi)$ the number of cycles with an even number of reality-edges (called *even cycles*). For the weighted algorithms, let w_r be the weight of reversals, translocations, fusions, and fissions, and let w_t be the weight of transpositions. We make the assumption that $w_r \leq w_t \leq 2w_r$.

The algorithm for SwRT is an implementation of the algorithm devised in [1]. The *score* of a permutation is defined by $\sigma(\pi) = c_{odd}(\pi) + (2 - 2 \cdot w_r/w_t)c_{even}(\pi)$. In each step, the algorithm

