Genome Rearrangement: A Planning Approach

Tansel Uras and **Esra Erdem**
Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey

Abstract

Evolutionary trees of species can be reconstructed by pairwise comparison of their entire genomes. Such a comparison can be quantified by determining the number of events that change the order of genes in a genome. Earlier Erdem and Tillier formulated the pairwise comparison of entire genomes as the problem of planning rearrangement events that transform one genome to the other. We reformulate this problem as a planning problem to extend its applicability to genomes with multiple copies of genes and with unequal gene content, and illustrate its applicability and effectiveness on three real datasets: mitochondrial genomes of *Metazoa*, chloroplast genomes of *Campanulaceae*, chloroplast genomes of various land plants and green algae.

Introduction

In biology, evolutionary trees (or phylogenies) can be reconstructed from the comparison of genomes of species (Sankoff and Blanchette 1998). One metric of evolutionary distance for this purpose is the number of rearrangement events such as transpositions and inversions to convert one genome to the other where a smaller number of such events implies a closer lineage. A rearrangement event is a genome-wide mutation that changes the order and/or orientations of genes (and sometimes their existence) in a genome. Finding the minimum number of these rearrangement events between genomes is called the genome rearrangement problem.

We consider the genome rearrangement problem as a planning problem as in (Erdem and Tillier 2005): one of the genomes is represented as the initial state and the other one as the goal state; the planner is prompted to find a sequence of at most k actions (rearrangement events) that leads the initial state to the goal state. As in (Erdem and Tillier 2005), we describe the genome rearrangement problem in ADL (Pednault 1989), and use TLPLAN (Bacchus and Kabanza 2000) to compute solutions. TLPLAN allows us to assign priorities/costs to actions and experiment with different search strategies. Our formulation of the genome rearrangement problem differs from that of (Erdem and Tillier 2005) in the following ways. First of all, it extends the descriptions of genomes to be able to handle duplicate genes—genes that occur multiple times in a single genome. Accordingly, it not only extends the descriptions of transpositions, inversions, inverted transpositions (transversions) but also introduces new operators for insertions and deletions. The temporal control information is described as preconditions of events. Also, the goal-check is done in a more computationally-efficient way by means of a biologically motivated measure, called the *breakpoint distance*, which does not require us to check the whole gene orders.

We have illustrated the applicability and the effectiveness of our planning-based approach to genome rearrangement on three sets of real data: mitochondrial genomes of Metazoa (animals with a nervous system, and muscles) (Blanchette, Kunisawa, and Sankoff 1999), chloroplast genomes of Campanulaceae (flowering plants) (Cosner et al. 2000), and chloroplast genomes of various land plants and green algae (Cui et al. 2006). Our results match the most recent and widely accepted results.

Considering that the genomes of many species (in particular, the chloroplast genomes) contain duplicate genes, and that no existing genome rearrangement software (e.g., GRAPPA (Moret et al. 2001), DERANGE 2 (Blanchette, Kunisawa, and Sankoff 1996)) can handle them, our methods provide a useful tool for experts.

Genome Rearrangement Problem

The *genome* of a single-chromosome organism can be represented by circular configurations of numbers $1, \ldots, n$, with a sign $+$ or $-$ assigned to each of them. Numbers $\pm 1, \ldots, \pm n$ will be called *labels*. Intuitively, a label corresponds to a gene, and its sign corresponds to the orientation of the gene. By $(l_1, \ldots, l_n)$ we denote the genome formed by the labels $l_1, \ldots, l_n$ ordered clockwise.

About genomes g, g' we say that g' is a *transposition* of g if, for some labels $l_1, \ldots, l_n$ and numbers k, m $(0 < k, m \leq n)$, $g = (l_1, \ldots, l_n)$ and $g' = (l_k, \ldots, l_m, l_1, \ldots, l_{k-1}, l_{m+1}, \ldots, l_n)$. Similarly, about genomes g, g', we say that g' can be obtained from g by a *deletion* (or g can be obtained from g' by an *insertion*) if, for some labels $l_1, \ldots, l_n$ and a number m $(0 < m \leq n)$, $g = (l_1, \ldots, l_n)$ and $g' = (l_1, \ldots, l_{m-1}, l_{m+1}, \ldots, l_n)$. Other events can be defined similarly.

We say that there is a *breakpoint* between two genomes if one of the genomes includes the pair l, l' and the other

genome includes neither the pair l, l' nor the pair $-l', -l$. For instance, there are 3 breakpoints between $(1, 2, 3, 4, 5)$ and $(1, 2, -5, -4, 3)$. The number of breakpoints between two genomes is called their *breakpoint distance*.

The *genome rearrangement problem* can be defined as follows: given two genomes g and g', and a positive integer k, decide whether g' can be obtained from g by at most k successive events. We view the genome rearrangement problem as a planning problem: given two genomes g and g', and a nonnegative integer k, find a sequence of at most k events that reduces the number of breakpoints between g and g' to 0. Note that this planning problem is different from the one described in (Erdem and Tillier 2005) in that both genomes are specified in the initial world, and that the goal is specified in terms of the number of breakpoints.

Representing the Planning Problem

We represent a genome by specifying the clockwise order of its labels, and by identifying which genes are duplicates. Both genomes are described in the initial state. To describe the goal, we introduce a functional fluent to denote the number of breakpoints: initially, it is counted; after that, at each step, it is decreased by the application of a rearrangement event. We introduce five actions to describe transpositions, inversions, transversions, insertions and deletions, and represent them as ADL-style operators in the language of TLPLAN. The definitions of these operators extend the ones in (Erdem and Tillier 2005) to handle duplicates.

Experimental Results

We have experimented with three sets of data using TLPLAN: *Metazoan* mitochondrial genomes (Blanchette, Kunisawa, and Sankoff 1999), *Campanulaceae* chloroplast genomes (Cosner et al. 2000), and chloroplast genomes of various land plants and green algae (Cui et al. 2006). Only in the last data set, genomes are of unequal content with duplicate genes. To analyze the accuracy of our approach, for each data set, first we have computed a small number of events for each pair of genomes and constructed a distance matrix, and then we have constructed a phylogeny using the distance matrix program NEIGHBOR (Felsenstein 2009). In all experiments, TLPLAN is run with the depth-best-first search strategy. The cost of each action is 1 (except the 0-cost action of swapping duplicates); so the goal is to find a plan with a small cost (rather than a shortest plan). The priorities of insertions, deletions, and swaps are much higher than the other events.

Mitochondrial genomes of *Metazoa*: Each one of these 11 genomes consists of 36 genes. The priorities of transpositions, inversions, transversions are specified as 2, 1, 1 respectively. All 45 plans (each with 1–26 events) are computed in less than 3 minutes. The phylogeny constructed by NEIGHBOR groups chordates and echinoderms together, arthropods, molluscs and annelids together; nematodes are a sister to these two groupings. These results match the results of (Nielsen 2001) based on morphological data, and the most widely accepted view of Metazoan Systematics and Tree of Life, based on the analysis of molecular data.

Chloroplast genomes of *Campanulaceae*: We consider 13 genomes, each with 105 genes. The priorities of transpositions, inversions, transversions are specified as 2, 3, 4 respectively (since inversions often occur in chloroplast genomes). All 66 plans (each with 1–12 events) are computed in less than 1 minute. According to the phylogeny constructed by NEIGHBOR, the groupings are identical to the ones in Fig. 4 of (Cosner et al. 2000).

Chloroplast genomes of land plants and green algae: These 7 genomes share 85 genes; each genome is of length 87–97. The priorities of transpositions, inversions, transversions are specified as 2, 3, 4 respectively. All 21 plans (each with 6–47 events) are computed in less than an hour. (The computation of a phylogeny for these species takes almost 25 days in (Cui et al. 2006).) The phylogeny constructed by NEIGHBOR groups *Nicotiana* and *Marchantia* with *Chaetosphaeridium*, and *Chlorella* and *Chlamydomonas* with *Nephroselmis*; *Mesostigma* is an outlier. These results conform with the biological evidence based on the analysis of proteins (Cui et al. 2006).

For further discussion please see (Uras and Erdem 2010).

References

Bacchus, F., and Kabanza, F. 2000. Using temporal logic to express search control knowledge for planning. *Artificial Intelligence* 116(1–2):123–191.

Blanchette, M.; Kunisawa, T.; and Sankoff, D. 1996. Parametric genome rearrangement. *Gene-Combis* 172:11–17.

Blanchette, M.; Kunisawa, T.; and Sankoff, D. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution* 49:193–203.

Cosner, M.; Jansen, R.; Moret, B.; Raubeson, L.; Wang, L.; Warnow, T.; and Wyman, S. 2000. An empirical comparison of phylogenetic methods on chloroplast gene order data in *Campanulaceae*. In Sankoff, D., and Nadeau, J., eds., *Comparative Genomics*. Kluwer. 99–122.

Cui, L.; Leebens-Mack, J.; Wang, L.; Tang, J.; Rymarquis, L.; Stern, D.; and dePamphilis, C. 2006. Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach. *BMC Evol. Bio.* 6:13.

Erdem, E., and Tillier, E. 2005. Genome rearrangement and planning. In *Proc. of AAAI*, 1139–1144.

Felsenstein, J. 2009. PHYLIP (phylogeny inference package) version 3.6. Distributed by the author.

Moret, B.; Wyman, S.; Bader, D.; Warnow, T.; and Yan, M. 2001. A new implementation and detailed study of breakpoint analysis. In *Proc. of PSB*, 583–594.

Nielsen, C. 2001. *Animal Evolution: Interrelationships of the Living Phyla*. Oxford University Press.

Pednault, E. 1989. ADL: Exploring the middle ground between STRIPS and the situation calculus. In *Proc. of KR*, 324–332.

Sankoff, D., and Blanchette, M. 1998. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology* 5:555–570.

Uras, T., and Erdem, E. 2010. Genome rearrangement and planning: Revisited. In *Proc. of ICAPS*.