

# ISMB 2002 Tutorial on Comparative Genomics — OUTLINE

David Sankoff

## Introduction

Chromosome breakage and mistakes in repair, errors in replication or recombination due to repetitive sequence, along with a number of other processes, give rise to changes in gene order. These have important consequences for the cell, the organism, the population, and for the evolution of species.

This field has become of increasing interest with the availability of genome sequences from many prokaryotes, eukaryote organelles and, more recently, eukaryote nuclei. Models and methods have been developed that are of importance in understanding the many unexpected similarities and divergences in the large-scale organization of these genomes.

This tutorial focuses on analytical methods for the characterization of genome rearrangements, for inferences about them, and their use in phylogenetic analysis. It should be comprehensible to anyone with basic knowledge of genetics or molecular biology and with some familiarity with the language of mathematical modelling, probability and statistics, and/or algorithmics. Common biological and genetic terminology will be used as well as light mathematical notation, but lengthy derivations and proofs will be avoided. The objective will be to survey a broad spectrum of topics and approaches rather than to delve into the details of particular mathematical results. Though some programs and packages will be mentioned or referenced, this will not be an demonstration of how to use bioinformatics software.

**1. Overview** During biological evolution, inter- and intrachromosomal exchanges of chromosomal fragments disrupt the order of genes on a chromosome and, for multichromosomal genomes, the partition of genes among these chromosomes. Any (maximal) contiguous region of the genome in which gene content and order have been conserved in two species is a conserved segment. Adjacent conserved segments are separated by breakpoints. Conserved segments become shorter and more numerous over time as they are disrupted by new events creating new breakpoints. The number of conserved segments (or breakpoints) measures the genomic distance between two species.

Genome rearrangements have been studied within many distinct disciplines. We sketch an overview of genome rearrangement from the perspectives of cytological genetics, clinical genetics, population biology, comparative mapping and phylogeny.

The duality between breakpoints and conserved segments is mirrored in the two research traditions for reconstructing genomic history based on gene orders in two or more genomes. The genome rearrangements approach, making use of combinatorial optimization techniques, attempts to infer a most economical sequence of rearrangement events to account for the differences among the genomes. The comparative mapping approach focuses on the statistics of genes per segment to infer the amount of evolutionary divergence.

**2. Genomic Distances.** The algorithmic study of comparative genomics tries to explain differences in gene orders in two or more genomes in terms of a limited number of rearrangement processes. For single-chromosome genomes, this requires the calculations of an edit distance between two linear orders on the same set of objects, representing the ordering of homologous genes in two genomes. In the "signed" version of the problem, a plus or minus is associated with each gene, representing the direction of transcription. One edit operation consists of the inversion, or reversal, of any number of consecutive terms in the ordered set, which, in the case of signed orders, also reverses the polarity of each term within the scope of the inversion. The calculation of the distance for unsigned genomes with inversions only is NP-hard; for signed problem it is of polynomial complexity. For multi-chromosome genomes, another important edit operation is reciprocal translocation, representing the exchange of terminal fragments between two chromosomes. Some formulations of the distance problem for translocation are of polynomial complexity. We sketch the Hannenhalli-Pevzner algorithms and available software for inversions and for translocations.

**3. Gene-order Phylogenies.** Though distance matrix methods are applicable to genome edit distances, the optimisation of a given tree, including the reconstruction of ancestral gene orders, is NP-hard,

even with only three input genomes. The number of breakpoints is an alternate, easily computed, genomic distance which, however, is also theoretically hard to generalise to the phylogenetic context. Nevertheless, it can be transformed into the Traveling Salesman Problem, for which efficient software exists for moderate-sized input. We illustrate the application of this methodology to the phylogeny of mitochondrial genomes in animals and in protists.

**4. Genome Rearrangement with Gene Families.** The above methods based on permutations of gene order are seriously compromised for larger genomes where several copies of the same gene, or several highly homologous (paralogous) genes may be scattered across the genome. One approach to overcoming this difficulty is based on exemplars: for each genome, an exemplar sequence is constructed by deleting all but one occurrence of each gene family. The exemplar distance between two genomes is the minimum, over all their exemplar strings, of the distance between their exemplars. We describe algorithms and software for exemplar breakpoint distance and exemplar inversions distance. We show how to extend exemplar analysis to phylogenetic analysis, using output from gene-tree/species-tree algorithms.

**5. Hybridization and genome duplication** Several types of biological processes, widespread in the plant kingdom, give rise to hybrids. A variety of mathematical problems arising in trying to model these processes. These problems differ according to the process responsible for hybridisation, the kinds of data available, and the aim of the evolutionary reconstruction. For example, one form of hybridization of two karyotypically distinct species sees the fusion of two genomes followed by a series of chromosomal rearrangement events until the hybrid genome is finally stabilised as a diploid. We seek to infer an ancestral hybrid genome, using data on which hybrid genes originated from each of the parent species, and/or additional data (synteny and gene order) from the modern purebred descendants of these parents. A number of algorithms are available for these problems. Hybrids may also be formed through the exceptional fertilisation of two distinct though related species, the parent species possibly differing from each other and from the modern hybrid by numerous genome rearrangements. This case may be analyzed using the three-genome generalisation discussed above.

The effects of genome doubling show up across the eukaryote spectrum, e.g., in yeast, plants and vertebrates. We propose a suite of genome halving problems, and associated algorithms, for reconstructing the ancestral, pre-duplication genome, each problem depending on the level of detail of the data and the desired reconstruction. In each case the idea is to find a genome consisting of pairs of identical chromosomes, representing the original tetraploid, such that the number of translocations (including chromosome fusions and fissions) required to transform it to the modern genome is minimised.

**6. Comparative Mapping.** The simplest model of genomic divergence, deriving from a 1984 study by Nadeau and Taylor, assumes the spatial homogeneity of both breakpoint and gene distributions along the chromosomes. The main focus has been the severe underestimation of the number of segments in comparisons where there are relatively few genes common to the data sets for a pair of species. To this end, we study the marginal probability that a segment contain  $r$  genes, as well as the probability of observing a non-empty segments if there are  $m$  genes and  $n$  breakpoints. We show that  $a$  is a sufficient statistic for the number of breakpoints, and provide estimators for  $n$  given  $m$  and  $a$ . We discuss how to relax the homogeneity assumptions of the model.

As map data accumulate, it becomes increasingly difficult to find segments in which gene content and order are strictly parallel in two genomes, due in part to experimental error, but also to high rates of inversion of small regions of chromosomes. We introduce a method for estimating the configuration of conserved segments resulting from the evolutionary history of reciprocal translocations, by minimizing a weighted sum of mapping error and rearrangement costs. This involves a variant of single link stepwise cluster analysis performed simultaneously on all conserved syntenies, with the interim results from each cluster analysis affecting the current state of all the others.

**7. Relationship with genome rearrangements in cancer and infertility:** In this section, we analyze data on rearrangement breakpoints resulting from individual real-time cytogenetic events in order to help understand the distribution of multiple breakpoints in comparative maps. We compare breakpoint positions from four different databases, on reciprocal translocations, inversions and deletions in neoplasms (the Mitelman database), reciprocal translocations and inversions in families carrying rearrangements (the HC Forum database) and the human-mouse comparative map. For each set of positions we construct

breakpoint distributions for as many as possible of the the 44 autosomal arms. We identify and interpret four main types of distribution:

- the uniform distribution associated both with families carrying translocations or inversions, and with the comparative map,
- telomerically skewed distributions of translocations or inversions detected consequent to births with malformations,
- medially clustered distributions of translocation and deletion breakpoints in tumor karyotypes, and
- bimodal translocation breakpoint distributions for chromosome arms containing telomeric proto-oncogenes.

**8. Modelling eukaryote-prokaryote differences:** Two independent sets of recent observations on newly sequenced microbial genomes pertain to the prevalence of short inversion as a gene order rearrangement process and to the lack of conservation of gene order within conserved gene clusters. We develop a model of inversion where the key parameter is the length of the inverted fragment. We show that there is a *qualitative* difference in the pattern of evolution when the inversion length is small with respect to the cluster size and when it is large. This suggests an explanation of the lack of parallel gene order in conserved clusters and raises questions about the statistical validity of putative functionally selected gene clusters if these have only been tested against inappropriate null hypotheses.

**9. Tests for gene clusters:** Comparing chromosomal gene order in two or more related species is an important approach to studying the forces that guide genome organization and evolution. Linked clusters of similar genes found in related genomes are often used to support arguments of evolutionary relatedness or functional selection. However, as the gene order and the gene complement of sister genomes diverge progressively due to large scale rearrangements, horizontal gene transfer, gene duplication and gene loss, it becomes increasingly difficult to determine whether observed similarities in local genomic structure are indeed remnants of common ancestral gene order, or are merely coincidences.

A rigorous comparative genomics requires principled methods for distinguishing chance commonalities, within or between genomes, from genuine historical or functional relationships. We show how to construct tests for significant groupings against null hypotheses of random gene order, taking incomplete clusters, multiple genomes and gene families into account. We consider both the significance of individual clusters of pre-specified genes, and the overall degree of clustering in whole genomes.