Genome Rearrangement

David Sankoff Nadia El-Mabrouk

1 Introduction

The difference between genome rearrangement theory and other approaches to comparative genomics, and indeed most other topics in computational molecular biology, is that it is not directly based on macromolecular sequences, either nucleic acids or proteins. Rather like classical genetics, its building blocks are genes, and the structures of interest are chromosomes, abstracted in terms of the linear order of the genes they contain. Of course genes and their RNA and protein products are macromolecules, but here we do not focus on the internal structure of genes and assume that the problems of determining the identity of each gene, and its homologs in other genomes, have been solved, so that a gene is simply labeled by a symbol indicating the class of orthologs to which it belongs. Moreover, the linearity of chromosomal structure does not evolve by a nucleotide substitution process in the way DNA does, or even by the same type of insertion/deletion processes, but by a number of very different rearrangement processes which are non-local, i.e. their scope may involve an arbitrarily large proportion of a chromosome. As a consequence, the formal analysis of rearrangements bears little resemblance in detail to DNA or protein comparison algorithms.

Nevertheless, in analogy with sequence comparison, the study of genome rearrangements has focused on inferring the most economical explanation for observed differences in gene orders in two or more species, as represented by their genomes, in terms of a small number of elementary processes. After first formalizing in Section 2 the notion of a genome as a set of chromosomes, each consisting of an ordered set of genes, we will proceed in Section 3 to a survey of genomic distance problems. More detail on the Hannehanlli-Pevzner theory for "signed" distances follows in Section 4. Section 5 will be devoted to phylogenetic extensions, and Section 6 to problems of gene and genome duplication and their implications for genomic distance and genome-based phylogeny.

2 The formal representation of the genome.

As a first approximation, a genome can be thought of as a set containing on the order of 10^3 (some bacteria) to 10^5 (human) distinct elements called genes. In more realistic analyses, it may be necessary to consider that some genes occur with multiplicity two or higher in a genome, which cannot be captured in a set formulation. The latter situation will be explored in Section 6.

2.1 Synteny

The genes in plants, animals, yeasts and other eukaryotes are partitioned among a number of chromosomes, generally between 10 and 100 in number, though it can be as low as 2 or 3 (Jackson, 1957; Lima-de Faria, 1980), or much higher than 100. Two genes located on the same chromosome in a genome are said to be *syntenic* in that genome.

Some genome rearrangements involve parts of one chromosome being relocated to another chromosome. Syntenic structure is generally different between different species and usually identical among all the members of a single species. A few species tolerate population "heterogeneity" involving small differences in syntenic structure, where heterokaryotypic individuals are not only viable, but fertile (McAllister, 2000).

In prokaryotic genomes, comprising both eubacteria and archaebacteria, the genome typically resides on a single chromosome. Organelles, such as the mitochondria found in most eukaryotes and the chloroplasts in plants and algae, also have relatively small single-chromosome genomes, containing less than 100 (mitochondria) or 250 (chloroplasts) genes, and are believed to be the highly reduced descendants of prokaryotic endosymbionts.

2.2 Order and polarity

Syntenic structure,as we shall see in Section 3.6.1, suffices to initiate the study of genome rearrangements. Two additional levels of chromosomal structure, when they are available, add valuable information about rearrangement. The first is gene order. The genes on each chromosome are have a linear order that is characteristic of each genome. Note that although our discussion in this paper is phrased in terms of the order of genes along a chromosome, the key aspect for mathematical purposes is the order and not the fact that the entities in the order are genes. They could as well be blocks of genes contiguous in the two (or N) species being compared, conserved chromosomal segments in comparative genetic maps (cf. Nadeau and Sankoff (1998)) or, indeed, the results of any decomposition of the chromosome into disjoint ordered fragments, each identifiable in the two (or in all N) genomes.

The next level of structure is the transcription direction associated with each gene. In the double-stranded DNA of a genome, typically some genes are found on one strand and are read in the direction associated with that strand, while other genes are on the complementary strand which is read in the opposite direction. To capture this distinction in the mathematical notation for a genome, the genes on one strand are designated as of positive polarity and those on the other as negative. The latter are written with a minus sign preceding the gene label, and genomes and genome distance problems where this level of structure is known and taken into account are called “signed” in contrast to the situation where no directional information is used, the “unsigned” case.

2.3 Linearity *versus* circularity

In eukaryotes such as yeast, amoeba, or humans, the genes on a chromosome are ordered linearly. There is no natural left-to-right order; i.e. there is no structural asymmetry or polarity between one end of a chromosome and the other. Biologists distinguish between the short and long “arms” of a chromosome for nomenclatural purposes, and while we shall see in Section 2.4 that this has a structural basis, there is no biological reason to order the long arm before the short arm, or vice-versa.

In prokaryotes and in organelles, the single chromosome is generally circular. This leads to terminological and notational adjustments – the arbitrariness of of left-to-right order becomes the arbitrariness of clockwise versus counterclockwise ordering, and the notion of one gene appearing in the order somewhere before another is no longer meaningful. Most computational problems in genome comparison are no more difficult for circular genomes than linear ones, though there is one clear exception where the circular problem is much harder, as described in Section 3.1.

2.4 Centromeres and telomeres

Two structural aspects of eukaryote chromosomes are especially pertinent to genome rearrangements. The centromere is a structurally specialized noncoding region of the DNA, situated somewhere along the length of the chromosome, associated with specific proteins, which plays a key role in assuring the proper allocation of chromosomes among the daughter cells during cell division. The centromere divides the chromosome into two arms, both of which normally contain genes. The end of each arm is the telomere, also consisting of non-coding DNA in association with particular proteins.

Because the telomere “protects” the end of the chromosome and is also necessary in cell division, as is the centromere, genome rearrangements usually do not involve the telomere and do not entail the creation of a chromosome without a centromere or with more than one centromere, though on the evolutionary time scale there are exceptions. New centromeres occasionally emerge remote from existing centromeres and take over the role of the latter, which quickly lose their erstwhile function. Chromosomes sometimes fuse in an end-to-end manner, involving the loss of two telomeres; while the opposite process, fission, is another possibility.

2.5 Multigene families

Implicit in the rearrangements literature is that both genomes being compared contain an identical set of genes and the one-to-one homologies (orthologies) between all pairs of corresponding genes in the two genomes have previously been established. While this hypothesis of *unique genes* may be appropriate for some small genomes, e.g. viruses and mitochondria, it is clearly unwarranted for divergent species where several copies of the same gene, or several homologous (paralogous) genes — a *multigene family*, may be scattered across a genome.

2.5.1 The pertinence of sequence comparison

We stressed at the outset that genome rearrangement analysis is usually carried out separately from, and subsequent to, gene homology assessments. A partial exception to this must be made in the study of multigene familes, where we must take into account *degrees* of homology, so that the input data are more subtle than the binary distinction between homologous genes and unrelated genes.

3 Operations and distances

There are many ways of comparing two linear (or circular) orders on a set of objects. In Subsection 3.1, we first discuss one which is not based on any biologically-motivated model. In Subsection 3.2, we introduce a distance which is motivated by general characteristics of genome rearrangements. In the remainder of this section, we review the many edit distances which are based on particular types of rearrangement.

3.1 Alignment traces

One of the earliest suggestions for comparing genomes was to adapt concepts of alignment in sequence comparison, in particular the notion of the trace of an alignment. In its graphic version, this requires displaying the n genes in each of the two genomes, ordered from left to right, one genome above the other, and connecting each of the n pairs of homologous genes with a line. The number of intersections between pairs of lines is a measure of how much one genome is scrambled with respect to the other (Sankoff and Goldstein, 1989). (In a classical sequence alignment, there are no intersections.) For linear orders, this measure is easily calculated and analytical tests are available for detecting non-random similarities in order; the circular case is much more difficult. The problem has to do with the optimal alignment of the two genomes, where one circular genome is superimposed on the other and rotated in such a way as to minimize the number of intersections between trace lines connecting genes in the two genomes (Sankoff et al., 1990; Bafna et al., 2000).

3.2 Breakpoints

Since genome rearrangements generally involve incorrectly repaired breaks between adjacent genes, it seems appropriate to focus on adjacencies when comparing rearranged genomes. For two genomes X and Y, we define $b(X, Y)$ to be the number of pairs of genes that are adjacent in genome X but not in Y. The easily calculated measure b is and was first defined in the context of genome rearrangements by Watterson et al. (1982), but was already implicit much earlier in cytogenetic assessments of chromosomal evolution. For signed genomes, the notion of adjacency requires that the configuration of transcription directions be conserved, so that if genome X contains two genes ordered as gh, then these two genes are adjacent in Y only if they occur as gh or as $-h-g$.

The breakpoint distance can be extended to apply to two genomes X and Y which do not contain identical sets of genes. Here we create two smaller genomes X' and Y' by simply deleting those genes which are only in one of the genomes. Then the "induced breakpoint" distance $b_I(X, Y)$ between X and Y is defined to be $b(X', Y')$. For multiple comparisons, as in phylogenetic applications, it is preferable to use the normalized measure $b_\nu(X, Y) = b_I(X, Y)/l$, where l is the number of genes in X' and Y'.

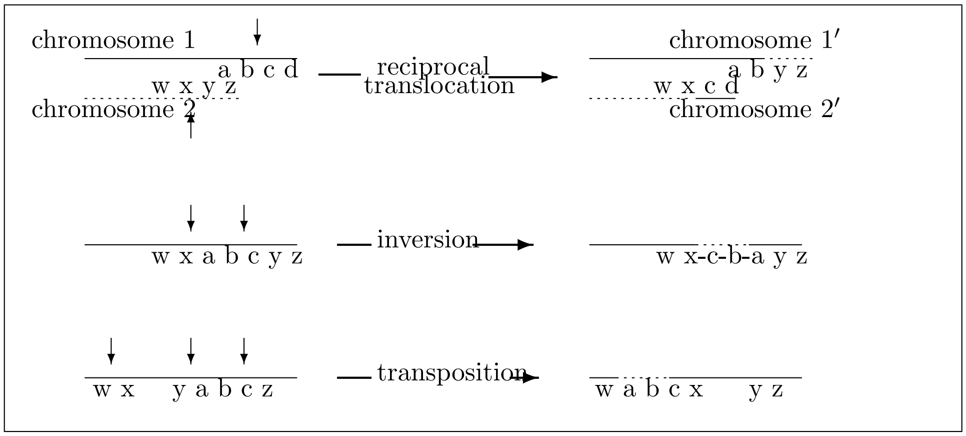

Figure 1: Schematic view of genome rearrangement processes. Letters represent positions of genes. Vertical arrows at left indicate breakpoints introduced into original genome. Reciprocal translocation exchanges end segments of two chromosomes. Reversal (or inversion) reverses the order and sign of genes between two breakpoints (dotted segment). Transposition removes a segment defined by two breakpoints and inserts it at another breakpoint (dotted segment), in the same chromosome or another.

3.3 Edit distances

Distance problems motivated by particular types of rearrangement processes require calculating an edit distance between two linear or circular orders on the same set of objects, representing the ordering of homologous genes in two genomes. The elementary edit operations may include one or more of the processes depicted in Figure 1.

3.4 Reversal distances

Reversal, or **inversion**, reverses the order of any number of consecutive terms in the ordered set, which, in the case of signed orders, also changes the sign of each term within the scope of the reversal. Kececioglu and Sankoff (1995) re-introduced the problem—earlier posed by Watterson et al. (1982), and even earlier in the genetics literature, e.g. Sturtevant and Novitski (1941)—of computing the minimum reversal distance between two given permutations in the unsigned case, and gave approximation algorithms and an exact algorithm feasible for moderately long permutations. Bafna and Pevzner (1996) gave improved approximation algorithms and Caprara (1997) showed this problem to be NP-complete. Kececioglu and Sankoff (1994) also found tight lower and upper bounds for the signed case and implemented an exact algorithm which worked rapidly for long permutations. Indeed, Hannenhalli and Pevzner (1999) showed that the signed problem is only of polynomial complexity, and improvements to their algorithm were given by Berman and Hannenhalli (1996) and by Kaplan et al. (2000). We will return to the Hannenhalli-Pevzner approach in Sections 4 and 6.

3.5 Transposition distance

A **transposition** moves any number of consecutive terms from their position in the order to a new position between any other pair of consecutive terms. Computation of the transposition distance between two permutations was considered by Bafna and Pevzner (1998), but its NP-completeness has not yet been confirmed. This has been more difficult to analyze than the reversals distance problem (Meidanis and Dias, 2000).

3.6 Translocation distance

Kececioglu and Ravi (1995) began the investigation of translocation distances, and Hannenhalli (1996) showed that the problem is of polynomial complexity, using methods similar to the reversals distance algorithm.

3.6.1 Syntenic distance

Ferretti et al. (1996) proposed a relaxed form of translocation distance applicable when chromosomal assignment of genes, but not their order, is known. Let A and B be two chromosomes, considered to be sets of genes. A translocation then transforms A and B into $(A - A') \cup B'$ and $(B - B' \cup A')$, respectively, where at least one of A' and B' is a proper subset of A or B. A fusion occurs when, e.g. $A' = A$ and $B' =$ the null set, and a fission when either A or B is replaced by the null set, in this formulation.

Then the syntenic distance between two genomes G and H, considered as two different partitions of the same set into subsets (chromosomes), is defined to be the minimum number of translocations necessary to transform G into H. The complexity of its calculation was shown to be NP-complete by DasGupta et al. (1998) and its structure was further investigated by Liben-Nowell (1999); Kleinberg and Liben-Nowell (2000).

3.7 Combined distances

Distances based on single operations may be of mathmatical interest and are appropriate starting points for investigating genomic rearrangements, but realistic models must allow for several types of operation. Several studies have attempted this. The most successful is the extension of the Hannenhalli-Pevzner theory to cover the case where both translocation and reversal operations are considered (Hannenhalli and Pevzner, 1995).

Another exact polynomial algorithm extending the Hannenhalli-Pevzner theory applies to two genomes which do not have the identical set of genes. This requires the calculation the the minimum number of reversals, and insertions or deletions of contiguous segments of the chromosome necessary to convert one genome into another (El-Mabrouk, 2000).

There have also been a number of studies combining transposition and reversals (Gu et al., 1997; Walter et al., 1998).

An edit distance which is a weighted combination of inversions, transpositions and deletions has been studied by Sankoff (1992), Sankoff et al. (1992) and Blanchette et al. (1996). Dalevi et al. (2000) have developed a simulation-based method for determining appropriate weighting parameters in the context of prokaryotic evolution, and applied this to the divergence of of *Chlamydia trachomatis* and *Chlamydia pneumoniae*. (See also Andersson and Eriksson (2000).) Their results quantify the propensity for shorter rather than longer inversions.

4 The Hannenhalli-Pevzner theory

In this section, we introduce the structures necessary to understand the results of the three polynomial-time algorithms devised by Hannenhalli and Pevzner. In particular, we sketch how they calculate the edit distance between two genomes, although we do not enter into the details of how they recover the actual operations which convert one of the genomes into the other.

Given two genomes H_1 and H_2 containing the same genes, where each gene appears exactly once in each genome, the genome rearrangement problem is to find the minimum number of rearrangement operations necessary to transform H_1 into H_2 (or H_2 into H_1). Polynomial algorithms were designed for the reversals-only version of the problem (in the case of single-chromosome genomes) (Hannenhalli and Pevzner, 1999), the translocations-only version (Hannenhalli, 1996), and the version with both reversals and translocations (Hannenhalli and Pevzner, 1995) (the latter two for multichromosomal genomes). The two methods allowing translocations require that the genomes H_1 and H_2 share the same set of chromosomal endpoints, but this can be taken care of by means of the addition of dummy endpoints, if necessary.

The algorithms all depend on a bicoloured graph $\mathcal{G}$ constructed from H_1 and H_2. The details of this construction vary from model to model, due to the different ways chromosomal endpoints must be handled, but the general character of the graph is the same and may be summarized as follows.

Graph $\mathcal{G}$: If gene x of H_1 has positive sign, replace it by the pair $x^t x^h$, and if it is negative, by $x^h x^t$. Then the vertices of $\mathcal{G}$ are just the x^t and the x^h for all genes x. Any two vertices which are adjacent in some chromosome in H_1, other than x^t and x^h from the same x, are connected by a black edge, and any two adjacent in H_2, by a gray edge. In the case of a single chromosome, the black edges may be displayed linearly according to the order of the genes in the chromosome. For a genome containing N chromosomes, N such linear orders are required; in the model allowing both reversals and translocations, however, the N orders are concatenated in each of the two genomes, so that we are again left with a single linear order.

Now, each vertex is incident to exactly one black and one gray edge, so that there is a unique decomposition of $\mathcal{G}$ into c disjoint cycles of alternating edge colours. By the **size of a cycle** we mean the number of black edges it contains. Note that c is maximized when $H_1 = H_2$, in which case each cycle has one black edge and one gray edge.

A rearrangement operation ρ, either a reversal or a translocation, is determined by the two black edges e and f where it "cuts" the current genome. Rearrangement operations may change the number of cycles, so that minimizing the number of operations can be seen in terms of increasing the number of cycles as fast as possible. Let $\mathcal{G}$ be a cycle graph, ρ a rearrangement operation, and $\Delta(c)$ the difference between the number of cycles before and after applying the operation ρ. Hannenhalli and Pevzner showed that $\Delta(c)$ may take on values 1, 0 or -1, in which cases they called ρ **proper, improper** or **bad**, respectively. Roughly, an operation determined by two black edges in two different cycles will be bad, while one acting on two black edges within the same cycle may be proper or improper, depending on the type of cycle and the type of edges considered.

Key to the Hannenhalli-Pevzner approach are the graph components. Two cycles, say Cycles 1 and 2, all of whose black edges are related by the same linear order (i.e. are on the same line), and containing gray edges that "cross", e.g., gene i linked to gene j by a black edge (i.e. in H_1) in Cycle 1, gene k linked to gene t by a black edge in Cycle 2, but ordered i, k, j, t in H_2, are connected. A component of $\mathcal{G}$ is a subset of the cycles (not consisting of a single cycle of size 1), built recursively from any of its cycle, at each step adding all the remaining cycles connected to any of those already in the construction. A component is termed **good** if it can be transformed to a set of cycles of size 1 by a series of proper operations, and **bad** otherwise. Bad components are called *subpermutations* in the translocations-only model, *hurdles* in the reversals-only model, and *knots* in the combined model. This property may be readily ascertained for each component by means of simple tests.

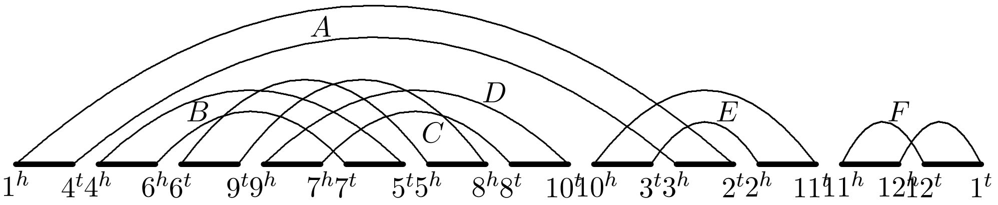
A
B
C
D
E
F
1^h $4^t 4^h$ $6^h 6^t$ $9^t 9^h$ $7^h 7^t$ $5^t 5^h$ $8^h 8^t$ $10^t 10^h$ $3^t 3^h$ $2^t 2^h$ $11^t 11^h$ $12^t 12^t$ 1^t

Figure 2: Graph $\mathcal{G}$ corresponding to circular genomes (i.e. first gene is adjacent to last gene) $H_1 = +1 +4 -6 +9 -7 +5 -8 +10 +3 +2 +11 -12$ and $H_2 = +1 +2 +3 \cdots +12$. A, B, C, D, E and F are the 6 cycles of $\mathcal{G}$. $\{A, E\}, \{B, C, D\}$ and $\{F\}$ are the three components of $\mathcal{G}$.

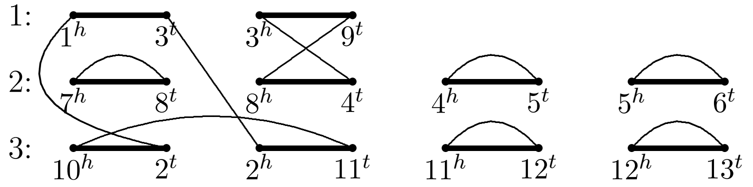
1:
2:
3:
1^h 3^t 3^h 9^t
7^h 8^t 8^h 4^t 4^h 5^t 5^h 6^t
10^h 2^t 2^h 11^t 11^h 12^t 12^h 13^t

Figure 3: Graph $\mathcal{G}$ corresponding to genomes H_1, H_2, both with 3 chromosomes, where $H_1 = \{1 : 1\ 3\ 9\ ;\ 2 : 7\ 8\ 4\ 5\ 6\ ;\ 3 : 10\ 2\ 11\ 12\ 13\}$ and $H_2 = \{1 : 1\ 2\ 3\ 4\ 5\ 6\ ;\ 2 : 7\ 8\ 9\ ;\ 3 : 10\ 11\ 12\ 13\}$. All genes are signed '+'. The edges which are on the same horizontal row of the graph corresponds to a chromosome of H_1. 7 cycles are present. As no cycle of size > 1 is contained on one row, $\mathcal{G}$ does not contain any component. Both genomes have the same set of endpoints, so we can omit the first vertices (x^t for initial genes and x^h for terminal genes).

The Hannenhalli-Pevzner formulae for all three models may be summarized as follows:

$$d(H_1, H_2) = n(\mathcal{G}) - c(\mathcal{G}) + m(\mathcal{G}) + f(\mathcal{G})$$

where $d(H_1, H_2)$ is the minimum number of rearrangement operations (reversals and/or translocations) $n(\mathcal{G})$ is the number of black edges of $\mathcal{G}$, $c(\mathcal{G})$ is the number of cycles, $m(\mathcal{G})$ is the number of bad components, and $f(\mathcal{G})$ is a correction of size 0, 1 or 2 depending on the set of bad components.

5 Phylogenetic analyses

Reconstruction of phylogeny may be approached through the application of generic methods (neighbour-joining, least-squares fitting, agglomerative clustering, etc.) to a distance matrix, independent of the nature of the data giving rise to the summary distances, or through ancestral inference methods (maximum likelihood, parsimony, etc.), where the tree shape is optimized simultaneously with the reconstruction of ancestral forms associated with non-terminal nodes, analogous to the input data associated with the terminal nodes. Distance matrices based on genomic distances can and have been used in traditional ways for phylogenetic reconstruction (Sankoff et al., 1992, 2000b), but approaches involving ancestral inference pose new analytical problems.

The problem of inferring ancestors may be decomposed into two aspects which must be solved simultaneously – finding the optimal shape, or topology, of the tree, and optimizing the ancestral reconstruction at each non-terminal node. Again, there are traditional search methods for optimal trees, but the reconstruction of ancestral genomes, given a fixed topology, is a new type of task, and it is on this question that we focus in this section.

5.1 The median problem

The solution of the *median problem* is of key importance in inferring the ancestral states in a phylogenetic tree. Given a distance d and three genomes A, B and C, the median is a genome $M \in \mathcal{S}$, the set of all possible genomes, such that the sum $d(A,X)+d(B,X)+d(C,X)$ is minimal over $\mathcal{S}$ for $X = M$. Algorithms for finding the median can be used to reconstruct ancestors in a given phylogeny through the process of *steinerization*. Unfortunately, the median problem is NP-hard, even in the case of unique genes, for all known rearrangement distances d including signed inversion distance. Even heuristic approaches to this problem work well only for very small instances (cf Hannenhalli et al. (1995); Sankoff et al. (1996)).

5.1.1 Reversals

Recall that reversal distance on signed genomes can be calculated in polynomial time; indeed, in only quadratic time. Can polynomial efficiency be extended to the median problem? The answer is no, as proved by Caprara (1999). Moreover, no reasonably effective heuristics have been tested for this problem.

5.1.2 Breakpoints

For the breakpoint distance d, where $d(Y, Z)$ is the number of pairs of genes that are adjacent in genome Y but not in Z, the median problem is also NP-hard (Pe'er and Shamir, 1998; Bryant, 1998). Nevertheless, it can be solved in a relatively simple manner for three genomes A, B and C having the same gene content. Indeed, in this case, the problem can be reduced to the Traveling Salesman Problem (TSP) (Sankoff and Blanchette, 1997).

For unsigned genomes, consider the complete graph Γ whose vertices are all the genes. For each edge gh, let $u(gh)$ be the number of times g and h are adjacent in the three genomes A, B and C. Set $w(gh) = 3 - u(gh)$. Then the solution to TSP on (Γ, w) traces out an optimal genome M, since if g and h are adjacent in M, but not in A, for example, then they form a breakpoint in M.

For signed genomes, the reduction of the median problem to TSP must be somewhat different to take into account that we must specify for the median genome whether it contains $x^t x^h$ or $x^h x^t$, in the notation of Section 4. Let Γ be a complete graph whose vertices include x^t and x^h for each gene x. For each pair of distinct genes x and y , let $u(xy)$ be the number of times x^h and y^t are adjacent in the genomes A, B and C and $w(xy) = 3 - u(xy)$. We also set $w(x^t x^h) = -Z$, where Z is large enough to assure that a minimum weight cycle must contain the edge $x^t x^h$.

Although the TSP is also NP-hard, there are a number of algorithms and software packages applicable in particular contexts (Reinelt, 1991). These allow us to find the median of three genomes of size $n = 100$ in a matter of minutes Sankoff and Blanchette (1998). Recently, we have developed a heuristic for this problem in the much more difficult case where the genomes do not have the same set of genes (Sankoff et al., 2000a).

Further work on these problems was done by Bryant (2000) and Pe'er and Shamir (2000).

5.2 Steinerization algorithm

An optimal tree is one where the sum of the edge lengths is minimal, the length being defined as the number of breakpoints when the genomes associated with the endpoints are compared. A binary unrooted tree may be decomposed into a number of overlapping median configurations. Each median consists of a non-terminal node together with its three colinear nodes, terminal or non-terminal, and the three edges which join them. In an optimal tree, the genome reconstructed at each non-terminal node will be a solution to the median problem defined by its three neighbours. We exploit this fact to reconstruct the ancestral genomes, starting with some reasonable initialization, and iterating the median algorithm on the list of non-terminal nodes until no improvement is found with any node. This may result in a local optimum, but sufficient repeated trials of the whole algorithm, with somewhat different initializations, should eventually indicate the best possible solution. Blanchette et al. (1999) applied this method to animal mitochondrial genomes.

5.3 Probability-based models

The development of likelihood or other probability-based methods for phylogenetic inference from gene order data requires the prior probabilization of genome rearrangement models, which is much more difficult than modeling sequence divergence according the Jukes-Cantor, Kimura or the many other available parametrizations for nucleotide or amino acid residue substitutions, or even models allowing gaps. Sankoff and Blanchette (1999a,b) gave a complete characterization of the evolution of gene adjacency probabilities for random reversals on unsigned circular genomes as well as a recurrence in the case of reversals on signed genomes. Concepts from the theory of invariants developed for the phylogenetics of homologous gene sequences were used to derive a complete set of linear invariants for unsigned reversals, as well as for a mixed rearrangement model for signed genomes, though not for pure transposition or pure signed reversal models. The invariants are based on an extended Jukes-Cantor semigroup. They illustrated the use of these invariants to relate mitochondrial genomes from a number of invertebrate animals.

5.4 Reducing Gene Order Data to 'Characters'

Gene adjacencies may be treated as characters in inferring a parsimony, maximum likelihood, or compatability tree from gene order data (cf Gallut et al. (2000); Cosner et al. (2000)). The advantage of this is that it allows the use of existing phylogenetic software. The disadvantage is that the character sets it reconstructs at the ancestor nodes are generally incompatible with any gene order.

6 Gene copies, gene families

There are a number of different ways in which duplicate genes can arise: tandem repeat through slippage during recombination, gene conversion, horizontal transfer and other transposition, hybridization and whole genome duplication.

Analytical methods for genome rearrangement, predicated on the hypothesis that the gene order of two genomes are basically permutations of each other, eventually run into the problem of duplicate genes. It is no longer clear how to obtain the basic datum for rearrangement analysis: *caba* is not a permutation of *abc*. Complicating the situation further is the process of sequence divergence, so that duplicate genes gradually become structurally and functionally differentiated; at some point they are no longer duplicates, but members of a gene family sharing some functional similarities as well as homology. Duplicate copies are also particularly prone to be lost, either by physical deletion or by becoming pseudogenes (non-functional ex-genes) through rapid sequence divergence. It is in these contexts that the study of gene order is often forced to take account of the degree of similarity among different genes, and not to rely on a binary distinction between homology and unrelaed.

This section is structured according to the mechanism giving rise to duplicate genes. First, we discuss the doubling of the whole genome and the hybridization through fusion of two distinct genomes, and then the processes of individual gene duplication.

6.1 Genome doubling

There is a difference between the duplication of single genes and processes which result in the doubling of large portions of a chromosome or even of the entire genome. In the latter case, not only is one copy of each gene free to evolve its own function (or to lose function, becoming a pseudogene and mutating randomly, eventually beyond recognition), but it can evolve in concert with any subset of the hundreds or thousands of other extra gene copies. Whole new physiological pathways may emerge, involving novel functions for many of these genes.

Evidence for the effects of genome duplication can be seen across the eukaryote spectrum, though it is always controversial (Ohno et al., 1968; Wolfe and Shields, 1997; Postlethwait et al., 1998; Skrabanek and Wolfe, 1998; Hughes, 1999; Smith et al., 1999)). Genome duplication and other mechanisms for combining two genomes (hybridization, allotetraploidization) are particularly prevalent in plants (Devos, 2000; Parkin, 2000; Paterson et al., 2000).

From the analytical point of view, partial or total genome duplication differs from mechanisms of duplication such as duplication-transposition, gene conversion or horizontal transfer in that it conserves gene order within conserved segments, and this can facilitate the analysis of genomes descended from a duplicated genomes.

A duplicated genome contains two identical copies of each chromosome, but through genome rearrangement parallel linkage patterns between the two copies are disrupted. Even after a considerable time, however, we can hope to detect a number of scattered chromosome segments, each of which has one apparent double, so that the two segments contain a certain number of paralogous genes in a parallel order. Similarly patterns should be visible after hybridization through allotetraploidization Sankoff and El-Mabrouk (1999). The main methodological question addressed in this field is: how can we reconstruct some or most of the original gene order at the time of genome duplication or hybridization, based on traces conserved in the ordering of those duplicate genes still identifiable? Some of the contributions to this methodology include Skrabanek and Wolfe (1998); El-Mabrouk et al. (1998, 1999); El-Mabrouk and Sankoff (1999), the latter applicable to single, circular chromosomal genomes, i.e., typical prokaryotes.

6.2 Multigene families and exemplar distances

Implicit in definitions of rearrangement distances is that both genomes contain an identical set of genes and the one-to-one homologies (orthologies) between all pairs of corresponding genes in the two genomes have previously been established. As we have stressed, while this hypothesis of *unique genes* may be appropriate for some small genomes, e.g. viruses and mitochondria, it is clearly unwarranted for divergent species where several copies of the same gene, or several homologous (paralogous) genes–a *multigene family*, may be scattered across a genome.

In a recent publication (Sankoff, 1999), we formulated a generalized version of the genomic rearrangement problem, where each gene may be present in a number of copies in the same genome. The central idea, based on a model of gene copy movement, is the deletion of all but one member of each gene family–its *exemplar*–in each of the two genomes being compared, so as to minimize some rearrangement distance d between the two reduced genomes thus derived. Thus the exemplar distance between two genomes X and Y is $e_d(X,Y) = \min d(X',Y')$ where the minimum is taken over all pairs of reduced genomes X' and Y' obtained by deleting all but one member of each gene family.

6.3 Duplication, Rearrangement, Reconciliation

The notion of exemplar distance takes on particular relevance in the phylogenetic context. Sankoff and El-Mabrouk (2000) investigated the problem of inferring ancestral genomes when the data genomes contain multigene families. We define a gene tree as a phylogenetic tree built from the sequences (according to some given method) of all copies of a gene g or all members of a gene family in all the genomes in the study. There are a number of techniques for inferring gene duplication events and gene loss events by projecting a gene tree T_g onto a 'true' species tree T; this is known as *reconciliation* (e.g. Page and Cotton (2000)).

We ask: Given

- a phylogenetic tree $\mathcal{T}$ on N species;
- their N genomes: strings of symbols belonging to an alphabet of size F;
- F gene trees, each T_g relating all occurrences of one symbol g in the N genomes;
- a distance d between two gene orders containing only unique genes,

the problem is to find, in each ancestral genome (internal node) of $\mathcal{T}$,

- its set of genes, as well as
- their relationships with respect to genes in the immediate ancestor,
- the order of these genes in the genome, and
- among each set of sibling genes (offspring of the same copy in the immediate ancestor),one gene, designated as the exemplar,

such that the sum of the branch lengths of the tree $\mathcal{T}$ is minimal. The length of the branch connecting a genome G to its immediate ancestor A is $e_d(G', A)$, where G' is the genome built from G by deleting all but the exemplar from each family.

Acknowldgments

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. DS is a Fellow, and NE-M a Scholar, of the Program in Evolutionary Biology of the Canadian Institute for Advanced Research.

References

Andersson, S. G. E. and Eriksson, K. (2000). Dynamics of gene order stuctures and genomic architectures. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 267–280, Dordrecht, NL. Kluwer Academic Press.

Bafna, V., Beaver, D., Fürer, M., and Pevzner, P. A. (2000). Circular permutations and genome shuffling. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 199–206, Dordrecht, NL. Kluwer Academic Press.

Bafna, V. and Pevzner, P. A. (1996). Genome rearrangements and sorting by reversals. *SIAM Journal on Computing*, 25(2):272–289.

Bafna, V. and Pevzner, P. A. (1998). Sorting by transpositions. *SIAM Journal on Discrete Mathematics*, 11(2):224–240.

Berman, P. and Hannenhalli, S. (1996). Fast sorting by reversals. In Hirschberg, D. and Myers, G., editors, *Proceedings of the Seventh Annual Symposium on Combinatorial Pattern Matching (CPM '96)*, volume 1075 of *Lecture Notes in Computer Science*.

Blanchette, M., Kunisawa, T., and Sankoff, D. (1996). Parametric genome rearrangement. *Gene*, 172:GC11–GC17.

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

Bryant, D. (1998). The complexity of the breakpoint median problem. Technical Report CRM-2579, Centre de recherches mathématiques, Université de Montréal.

Bryant, D. (2000). A lower bound for the breakpoint phylogeny problem. In Giancarlo, R. and Sankoff, D., editors, *Proceedings of the Eleventh Annual Symposium on Combinatorial Pattern Matching (CPM 2000)*, volume 1848 of *Lecture Notes in Computer Science*, pages 235–247.

Caprara, A. (1997). Sorting by reversals is difficult. In *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97)*, pages 75–83, New York. ACM.

Caprara, A. (1999). Formulations and hardness of multiple sorting by reversals. In Istrail, S., Pevzner, P. A., and Waterman, M. S., editors, *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, pages 84–93, New York. ACM.

Cosner, M. E., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., Wang, L.-S., 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. H., editors, *Comparative Genomics*, pages 99–121, Dordrecht, NL. Kluwer Academic Press.

Dalevi, D., Eriksen, N., Eriksson, K., and Andersson, S. (2000). Genome comparison: The number of evolutionary events separating *C. pneumoniae* and *C. trachomatis*. Technical report, University of Uppsala.

DasGupta, B., Jiang, T., Kannan, S., Li, M., and Sweedyk, E. (1998). On the complexity and approximation of syntenic distance. *Discrete Applied Mathematics*, 88(1–3):59–82.

Devos, K. M. (2000). Comparative genetics: from hexaploid wheat to arabidopsis. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 411–423, Dordrecht, NL. Kluwer Academic Press.

El-Mabrouk, N. (2000). Genome rearrangement by reversals and insertions/deletions of contiguous segments. In Giancarlo, R. and Sankoff, D., editors, *Proceedings of the Eleventh Annual Symposium on Combinatorial Pattern Matching (CPM 2000)*, volume 1848 of *Lecture Notes in Computer Science*, pages 222–234.

El-Mabrouk, N., Bryant, B., and Sankoff, D. (1999). Reconstructing the predoubling genome. In Istrail, S., Pevzner, P. A., and Waterman, M. S., editors, *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB'99)*, pages 154–163, New York. ACM.

El-Mabrouk, N., Nadeau, J. H., and Sankoff, D. (1998). Genome halving. In Farach-Colton, M., editor, *Proceedings of the Ninth Annual Symposium on Combinatorial Pattern Matching (CPM '98)*, volume 1448 of *Lecture Notes in Computer Science*, pages 235–250, Heidelberg. Springer Verlag.

El-Mabrouk, N. and Sankoff, D. (1999). On the reconstruction of ancient doubled circular genomes using minimum reversals. In Asai, K., Miyano, S., and Takagi, T., editors, *Genome Informatics 1999*, pages 83–93. Universal Academy Press, Tokyo.

Ferretti, V., Nadeau, J. H., and Sankoff, D. (1996). Original synteny. In Hirschberg, D. and Myers, G., editors, *Proceedings of the Seventh Annual Symposium on Combinatorial Pattern Matching (CPM '96)*, volume 1075 of *Lecture Notes in Computer Science*, pages 159–167, Heidelberg. Springer.

Gallut, C., Barriel, V., and Vignes, R. (2000). Gene order and phylogenetic information. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 123–132, Dordrecht, NL. Kluwer Academic Press.

Gu, Q.-P., Iwata, K., Peng, S., and Chen, Q.-M. (1997). A heuristic algorithm for genome rearrangements. In Miyano, S. and Takagi, T., editors, *Proceedings of the Eighth Workshop on Genome Informatics 1997*, pages 268–269, Tokyo. Universal Academy Press.

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

Hannenhalli, S., Chappey, C., Koonin, E. V., and Pevzner, P. A. (1995). Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics*, 30:299–311.

Hannenhalli, S. and Pevzner, P. (1995). Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science*, pages 581–592.

Hannenhalli, S. and Pevzner, P. A. (1999). Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Journal of the ACM*, 48:1–27.

Hughes, A. L. (1999). *Adaptive Evolution of Genes and Genomes.* Oxford University Press, New York.

Jackson, R. (1957). New low chromosome number for plants. *Science*, 126:1115–1116.

Kaplan, H., Shamir, R., and Tarjan, R. E. (2000). A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, 29:880–892.

Kececioglu, J. and Sankoff, D. (1994). Efficient bounds for oriented chromosome inversion distance. In M.Crochemore and D.Gusfield, editors, *Proceedings of the Fifth Annual Symposium on Combinatorial Pattern Match-*

ing (CPM '94), volume 807 of *Lecture Notes in Computer Science*, pages 162–176, Heidelberg. Springer.

Kececioglu, J. and Sankoff, D. (1995). Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*, 13:180–210.

Kececioglu, J. D. and Ravi, R. (1995). Of mice and men: algorithms for evolutionary distance between genomes with translocations. In *Proceedings of Sixth ACM-SIAM Symposium on Discrete Algorithms*, pages 604–613.

Kleinberg, J. and Liben-Nowell, D. (2000). The syntenic diameter of the space of n-chromosome genomes. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 185–197, Dordrecht, NL. Kluwer Academic Press.

Liben-Nowell, D. (1999). On the structure of syntenic distance. In Crochemore, M. and Paterson, M., editors, *Proceedings of the Tenth Annual Symposium on Combinatorial Pattern Matching (CPM '99)*, volume 1645 of *Lecture Notes in Computer Science*, pages 43–56, Heidelberg. Springer.

Lima-de Faria, A. (1980). How to produce a human with 3 chromosomes and 1000 primary genes. *Hereditas*, 93:47–73.

McAllister, B. F. (2000). Fixation of chromosomal rearrangements. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 19–27, Dordrecht, NL. Kluwer Academic Press.

Meidanis, J. and Dias, Z. (2000). An alternative algebraic formalism for genome rearrangements. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 213–233, Dordrecht, NL. Kluwer Academic Press.

Nadeau, J. H. and Sankoff, D. (1998). Counting on comparative maps. *Trends in Genetics*, 14:495–501.

Ohno, S., Wolf, U., and Atkin, N. B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas*, 59:169–187.

Page, R. D. M. and Cotton, J. A. (2000). Genetree: A tool for exploring gene family evolution. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 525–536, Dordrecht, NL. Kluwer Academic Press.

Parkin, I. (2000). Unraveling crucifer genomes through comparative mapping. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 425–537, Dordrecht, NL. Kluwer Academic Press.

Paterson, A. H., Bowers, J. E., Burow, M. D., Draye, X., Elsik, C. G., xiao Jiang, C., Katsar, C. S., Lan, T.-H., Lin, Y.-R., Ming, R., and Wright, R. J. (2000). Comparative genomics of plant chromosomes. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 439–457, Dordrecht, NL. Kluwer Academic Press.

Pe'er, I. and Shamir, R. (1998). The median problems for breakpoints are NP-complete. Electronic Colloquium on Computational Complexity Technical Report 98-071. `http://www.eccc.uni-trier.de/eccc`.

Pe'er, I. and Shamir, R. (2000). Approximation algorithms for the median problem in the breakpoint model. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 225–241, Dordrecht, NL. Kluwer Academic Press.

Postlethwait, J. H., Yan, Y.-L., Gates, M. A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E. S., Force, A., Gong, Z., Goutel, C., Fritz, A., Kelsh, R., Knapik, E., Liao, E., Paw, B., Ransom, D., Singer, A., Thomson, T., Abduljabbar, T. S., Yelick, P., Beier, D., Joly, J.-S., Larhammar, D., Rosa, F., Westerfield, M., Zon, L. I., and Talbot, W. S. (1998). Vertebrate genome evolution and the zebrafish gene map. *Nature Genetics*, 18:345–349.

Reinelt, G. (1991). *The traveling salesman - computational solutions for TSP applications.* Springer Verlag, Berlin.

Sankoff, D. (1992). Edit distance for genome comparison based on nonlocal operations. In Apostolico, A., Crochemore, M., Galil, Z., and Manber, U., editors, *Proceedings of the Third Annual Symposium on Combinatorial Pattern Matching (CPM '92)*, volume 644 of *Lecture Notes in Computer Science*, pages 121–135, Heidelberg. Springer.

Sankoff, D. (1999). Genome rearrangements with gene families. *Bioinformatics*, 15:909–917.

Sankoff, D. and Blanchette, M. (1997). The median problem for breakpoints in comparative genomics. In Jiang, T. and Lee, D. T., editors, *Computing and Combinatorics, Proceeedings of COCOON '97*, volume 1276 of *Lecture Notes in Computer Science*, pages 251–263. Springer, Berlin.

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

Sankoff, D. and Blanchette, M. (1999a). Comparative genomics via phylogenetic invariants for Jukes-Cantor semigroups. In Gorostiza, L. and Ivanoff, G., editors, *Proceedings of the International Conference on Stochastic Models*, Conference Proceedings series. Canadian Mathematical Society.

Sankoff, D. and Blanchette, M. (1999b). Phylogenetic invariants for genome rearrangements. *Journal of Computational Biology*, 6:431–445.

Sankoff, D., Bryant, D., Deneault, M., Lang, B. F., and Burger, G. (2000a). Early eukaryote evolution based on mitochondrial gene order breakpoints. In Shamir, R., Miyano, S., Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, pages 254–262. ACM, New York.

Sankoff, D., Cedergren, R., and Abel, Y. (1990). Genomic divergence through gene rearrangement. In Doolittle, R. F., editor, *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, volume 183 of *Methods in Enzymology*, pages 428–438. Academic Press.

Sankoff, D., Deneault, M., Bryant, D., Lemieux, C., and Turmel, M. (2000b). Chloroplast gene order and the divergence of plants and algae, from the normalized number of induced breakpoints. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map alignment and the Evolution of Gene Families*, volume 1 of *Series in Computational Biology*, Dordrecht, NL. Kluwer Academic Press.

Sankoff, D. and El-Mabrouk, E. (1999). Hybridization and genome rearrangement. In Crochemore, M. and Paterson, M., editors, *Combinatorial Pattern Matching. Tenth Annual Symposium*, volume 1645 of *Lecture Notes in Computer Science*, pages 78–87. Springer Verlag, Berlin.

Sankoff, D. and El-Mabrouk, N. (2000). Duplication, rearrangement and reconciliation. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map alignment and the Evolution of Gene Families*, volume 1 of *Series in Computational Biology*, Dordrecht, NL. Kluwer Academic Press.

Sankoff, D. and Goldstein, M. (1989). Probabilistic models of genome shuffing. *Bulletin of Mathematical Biology*, 51:117–124.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F., and Cedergren, R. J. (1992). Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences USA*, 89(14):6575–6579.

Sankoff, D., Sundaram, G., and Kececioglu, J. (1996). Steiner points in the space of genome rearrangements. *International Journal of the Foundations of Computer Science*, 7:1–9.

Skrabanek, L. and Wolfe, K. H. (1998). Eukaryote genome duplication—where's the evidence? *Current Opinion in Genetics and Development*, 8:694–700.

Smith, N. G. C., Knight, R., and Hurst, L. D. (1999). Vertebrate genome evolution: a slow shuffle or a big bang? *BioEssays*, 21:697–703.

Sturtevant, A. H. and Novitski, E. (1941). The homologies of chromosome elements in the genus drosophila. *Genetics*, 26:517–541.

Walter, M. E., Dias, Z., and Meidanis, J. (1998). Reversal and transposition distance of linear chromosomes. In *String Processing and Information Retrieval: A South American Symposium (SPIRE '98)*. Submitted to Journal of Computational Biology.

Watterson, G., Ewens, W., Hall, T., and Morgan, A. (1982). The chromosome inversion problem. *Journal of Theoretical Biology*, 99:1–7.

Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713.