Comparative Mapping and Genome Rearrangement

(Also published on-line in AgBiotechNet at URL *http://agbio.cabweb.org*)

DAVID SANKOFF

Centre de recherches mathématiques
Université de Montréal, Québec, Canada

Comparative mapping of two species allows the observation of chromosomal segments conserved in both genomes since divergence from a common ancestor. Statistical analysis of the distribution of genes in these segments leads to an estimation of the number of unobserved segments, i.e., those containing no homologous genes identified as yet in the two genomes, possibly far more numerous than the observed ones. The total number of segments, observed and unobserved, is a measure of the evolutionary divergence of the two species. The algorithmic study of genome rearrangements focuses on the combinatorics of which segments are adjacent to which others in each genome and estimates the number of translocations and/or inversions necessary to account for the observed configuration. This has extensions to the study of gene families, genome duplication and hybridization. We review recent developments in both fields and assess the potential for an integrated approach.

Introduction

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. When comparing two evolutionarily diverging species, any (maximal) contiguous region of the genome in which gene content and order have been conserved in both species is called a *conserved segment.* Between any two adjacent conserved segments is a *breakpoint*. The number of conserved segments increases as they are disrupted by new events, creating new breakpoints, so that they tend to become shorter over time. The number of chromosomal segments conserved during the divergence of two species, or equivalently, the number of breakpoints, can be used as a rough measure of their genomic distance. Note that neither conserved segments nor breakpoints have any physical existence within a single genome. They are purely analytical constructs based on the comparison of two or more genomes, though they do reflect historical events.

The duality between breakpoints and conserved segments is mirrored in the two research traditions that have focused on the reconstruction of genomic history based on the comparison of chromosomal gene content and order 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. This is based only on comparing which conserved segments are adjacent at the breakpoints in the two genomes, without detailed attention to the contents of segments. The *comparative mapping* approach, pioneered by Nadeau and Taylor (1984), is essentially that of probability modeling and statistical analysis. It is not concerned with the details of rearrangement history, though it assumes that a stochastic translocation process accounts for the differences in chromosomal gene content and order.

The comparative mapping approach is more data-oriented and is lacking in analytical methodology. Rearrangement theory, on the other hand, has focused on the mathematics of a range of analytical questions, with little attention paid to application to real data, except as illustrative examples, with no interpretation. And there are more profound differences between the two. In this paper, we review recent results obtained through the two approaches and discuss the possibility of integrating them. Our framework is that of mathematical and statistical modeling, our subject is the formulae and algorithms that have been discovered, but our concern is the eventual applicability of the results in real contexts.

Comparative Mapping

The simplest formulation of the Nadeau-Taylor model of genomic divergence (Nadeau and Sankoff, 1998b; Sankoff *et al.*, 1997b) assumes that each reciprocal translocation breaks chromosomes at random points on two randomly chosen chromosomes. As a consequence, when we compare two divergent genomes, the endpoints of the conserved segments making up each chromosome are uniformly and independently distributed along its length (spatial homogeneity of breakpoints). We also assume that which genes of a genome are discovered and mapped first does not depend on their position on the chromosome (spatial homogeneity of gene distribution), nor on their proximity to each other (independence of map positions).

The Marginal Probability of r-gene Segments

In trying to count the number of conserved segments for the quantification of evolution, the main focus has been the underestimation due to conserved segments in which genes have not yet been identified in one or both species. There are two in Figure 1: one from chromosome 4 of Genome 2 and the other from chromosome 17. This is particularly important if there are relatively few genes common to the data sets for a pair of species, so that many or most of the conserved segments are not represented in the comparison, and genomic distance may be severely underestimated. Nadeau and Taylor (1984) could only detect 13 segments out of the almost 200 now known to exist.

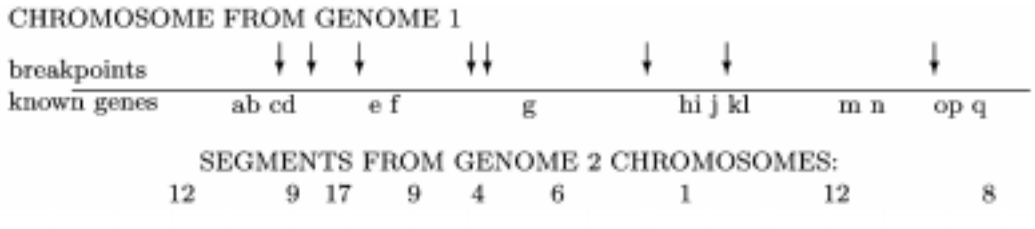

FIGURE 1. Fictitious example of conserved segments indicated on a chromosome from genome 1, with each segment labeled between its endpoints (adjacent arrows) as to which chromosome it is found on in genome 2. Homologous genes that have been discovered to date are indicated with letters.

We model the genome as a single long unit broken at n random breakpoints into $n+1$ segments, within each of which gene order has been conserved with reference to some other genome. Little is lost in not distinguishing between breakpoints and concatenation boundaries separating two successive chromosomes (Sankoff and Ferretti, 1996). The marginal probability that a segment contains r genes, $0 \leq r \leq m$, has been shown (Sankoff and Nadeau, 1996) to be:

$$\Pi(r) = \frac{n}{n+m} \binom{m}{r} \Big/ \binom{n+m-1}{r}.$$

We cannot compare the theoretical distribution $\mathbf{\Pi}(r)$ with n_r, the number of segments observed to contain r genes, since we cannot observe n_0, the number of segments containing no identified genes. We can, however, compare the relative frequencies $f(r)=n_r/\mathbf{\Sigma}_{r>0}\, n_r$ with the conditional probabilities $Q(r)=\mathbf{\Pi}(r|r>0)$, as in Figure 2, based on m=1423 human-mouse homologies documented in late 1996. The largest discrepancy is the comparison between $f(1)$ and $Q(1)$, most likely a reflection of error in the identification of homologous genes or other experimental error in chromosome assignment. When this source of error is removed, there is evidence that allowing inhomogeneity in breakpoint and gene distributions offers a closer fit to the data.

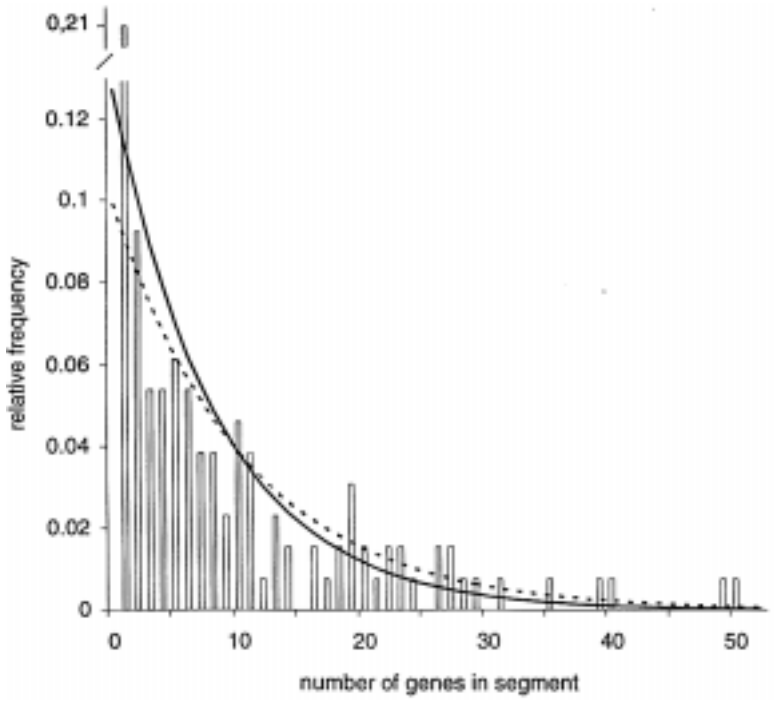

FIGURE 2. Comparison of relative frequencies $n_r/\mathbf{\Sigma}_{r>0}\, n_r$ of segments containing r genes, with predictions of the Nadeau-Taylor model. The value of n in the formula for Q is taken to be 141 (dotted curve) or 181 (uninterrupted curve), as estimated by maximum likelihood or a Kolmogorov-Smirnov best-fit method (Sankoff *et al.*, 1997a), respectively. Extrapolated $Q(0)$ permits comparison of the estimated number $(n+1)Q(0)$ of unobserved (empty) segments with the predictions $(n+1)Q(r)$ for $r \geq 1$. Three data points are off-scale, with r=55, 65 and 83, and the vertical axis is interrupted to allow an expanded scale.

The Inference Problem

It might seem that the number of segments n_r observed to contain r genes, for r=1,2,...m, would be useful data for inference about the Nadeau-Taylor model, in particular about n, the unknown number of breakpoints. It is remarkable, then, that to estimate n from m and n_r, only the number of non-empty segments $a=\mathbf{\Sigma}_{r>0}\, n_r$ is important, since it is a sufficient statistic for the estimation of n (Parent, 1999; Sankoff *et al.*, 1997a).

To estimate *n*, we study *P(a,m,n)*, the probability of observing *a* non-empty segments if there are *m* genes and *n* breakpoints. Combinatorial arguments give:

$$P(a,m,n) = \frac{\binom{m-1}{a-1}\binom{n+1}{a}}{\binom{n+m}{m}}$$

After observing *m* and *a*, it is an easy matter to find the value of *n* which maximizes *P*; Parent (1999) has shown that the computing formula *n'=m(A*-1*)/(m-A)* finds the maximum likelihood estimator over the entire range of *m* and *n* likely to be experimentally interesting, as long as a certain proportion of segments contain at least two genes.

Identification of Conserved Segments

Strictly speaking, conserved segments are regions of chromosomes in two related species in which both gene content and gene order are parallel, as in Figure 3(a). As map data accumulate, however, it becomes increasingly difficult to find segments that satisfy the criteria of content and order perfectly. This can be attributed in part to experimental error — either gross mistakes in chromosomal assignment of genes or quantitative errors in map positions affecting apparent gene order. Moreover, in the comparison of multi-chromosomal species such as humans and mice, we may wish to consider the segment structure to be that produced by translocation, and to consider as "noise", the effects of high rates of inversion or transpositions of small regions of chromosomes.

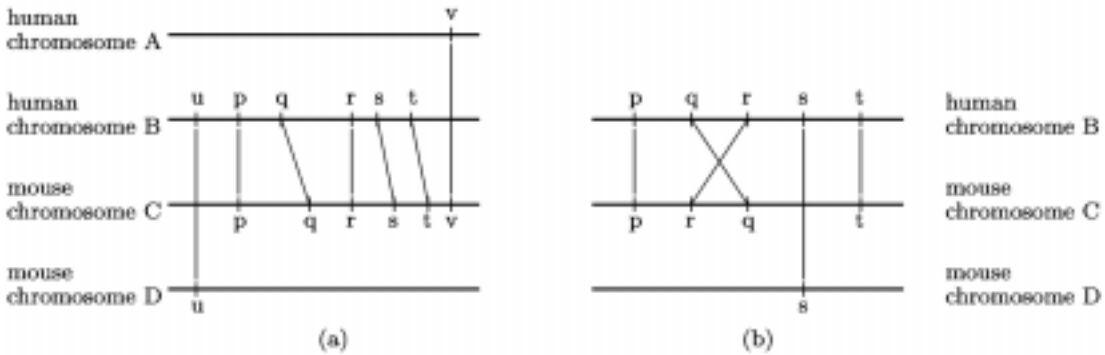

FIGURE 3. (a) Schematic example of a conserved segment in a human chromosome (B) and a mouse chromosome (C). Genes u and v have homologues elsewhere in the mouse and human genomes, respectively, and thus limit the leftward and rightward extension of the segment.
(b) Experimental mistake in the chromosomal assignment of s to mouse chromosome D, quantitative error in the assignment of q and/or r in the human or mouse map, or inversion of qr or transposition of q or r, results in the erroneous identification of three segments, p,qr,t, instead of just one, in human chromosome B and mouse chromosome C, and an additional one, s, in human chromosome B and mouse chromosome D.

We can estimate the configuration of conserved segments resulting from the evolutionary history of reciprocal translocations, and thus account for the gross differences between the genomes, by minimizing appropriately weighted mapping error plus rearrangement costs. For an appropriate choice of weighting parameters α, β, γ, we wish to find the subgroupings of conserved syntenic genes (i.e., all on a single chromosome in both species) into *a* segments, so as to minimize $D=\Sigma_{\{i=1,\ldots,a\}}D_i$, where D_i is a weighted measure of the compactness, density and integrity of segment *i*. Formally,

$$D_i = \gamma \max_{x,y \in i(1)} |x-y| + \alpha\, s[i(1)] + \gamma \max_{x,y \in i(2)} |x-y| + \alpha\, s[i(2)] - \beta\, r(i),$$

where $x \in i(j)$ refers to a gene (or its map coordinate) in segment i in species j, $r(i)$ indicates the number of homologous gene pairs in segment i and $s[i(j)]$ denotes the number of *other* segments with elements within the range of segment i in species j.

We do this with a variant of single link stepwise cluster analysis performed simultaneously on all conserved synteny sets (sets of genes occurring in common on one human chromosome and one mouse chromosome), with the interim results from each cluster analysis affecting the current state of all other cluster analyses (Sankoff *et al.*, 1997b). The algorithm provides, in one pass, solutions ranging from $a=m$, the number of genes, down to $a=c$, the number of conserved syntenies. The appropriate value for a is estimated as the value associated with the solution where the number of times conserved syntenies are realized as multiple segments is closest to what would be expected given the total number of segments on the chromosome.

The Dynamics of Average Segment Size

Several recent gene mapping efforts have discovered conserved segments that seem to be smaller than expected for the Nadeau-Taylor model of random translocation. To address the significance of these short segments, we calculated the distribution of the lengths of conserved segments that remain to be discovered (Nadeau and Sankoff, 1998b), assuming the model of random translocation. The mean of this distribution is $1/(m+n+1)$. With 2097 genes in the comparative map for humans and mice in mid-1997, the predicted segment length is approximately 0.6 cM, a result that is remarkably consistent with the lengths of many new segments that are being discovered.

Genomic Distances

The algorithmic study of comparative genomics (Sankoff, 1989) has focused on inferring the most economical explanation for observed differences in gene orders in two or more genomes in terms of a limited number of rearrangement processes. For single-chromosome genomes, this has been formulated as the problem of calculating an edit distance between two linear orders on the same set of objects, representing the ordering of homologous genes in two genomes. Note that the mathematical objects in these orders are called “genes” in the literature, though may actually represent conserved segments. In the most realistic version of the problem, a sign (plus or minus) is associated with each object in the linear order, representing the direction of transcription, or strandedness, of the corresponding gene. As illustrated in Figure 4, the elementary edit operations may include one or more of:

1) 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. Caprara (1997) showed the *unsigned* problem to be NP-hard. On the other hand, Hannenhalli and Pevzner (1995a) showed that the *signed* problem is only of polynomial complexity, and an improved polynomial algorithm was given by Kaplan *et al.* (1997).

2) Transposition of any number of consecutive terms from their position in the order to a new position between any other pair of consecutive terms. This may or may not also involve an inversion. The computational complexity of this problem is not yet known, nor is its biological significance. Sankoff *et al.* (1992) and Blanchette *et al.* (1996) implemented and applied heuristics to compute an edit distance which is a weighted combination of inversions and transpositions.

In addition, for multi-chromosome genomes, a major role is played by:

3) Reciprocal translocation. Hannenhalli and Pevzner (1995b) have also shown that a formulation of this problem is of polynomial complexity.

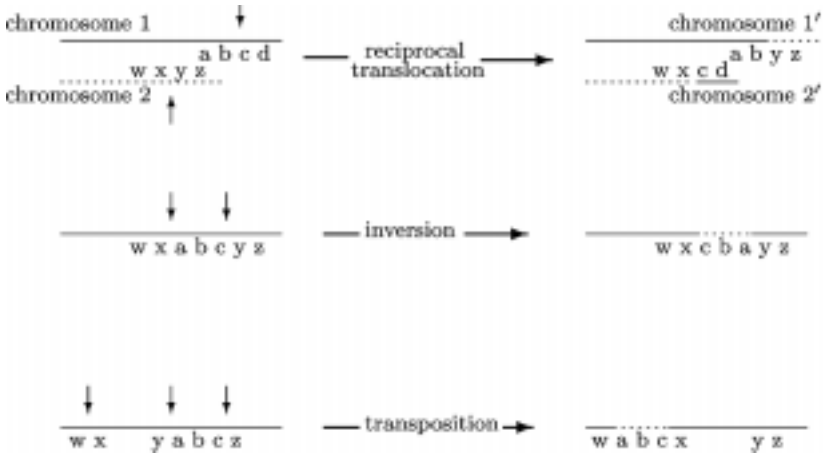

FIGURE 4. Schematic view of genome rearrangement processes. Letters represent positions of genes. Vertical arrows at the left indicate breakpoints introduced into the original genome. Reciprocal translocation (top) exchanges end segments of two chromosomes. Inversion (center) reverses the order of genes between two breakpoints (dotted segment at right). Transposition (bottom) removes a segment defined by two breakpoints and inserts it at another breakpoint (dotted segment at right), in the same chromosome or another. Gene order is conserved (possibly inverted) within segments.

Genome Rearrangement with Gene Families

The theory of genome rearrangements discussed above takes for input two different orders of the same set of genes. Implicit is the assumption that homologies between all pairs of corresponding genes in the two genomes have previously been established. While this may be appropriate for some small genomes — viruses, mitochondria, plastids — or for closely related species where long chromosomal segments, not just individual genes, correspond in the two genomes, it is clearly unwarranted for divergent species where several copies of the same gene, or several highly homologous (paralogous) genes may be scattered across the genome.

We have formulated a generalized version of the genome rearrangement problem where each gene may be present in a number of copies (Sankoff, 1999). Two versions of this problem have been investigated: exemplar breakpoints distance (EBD) and exemplar reversals distance (ERD). For each genome, an *exemplar* string 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.

As an example, suppose the two genomes are -b -a b a -c d c and a -a c a -c b d. Note that in this context, the adjacency a b in one genome and -b -a in the other does not constitute a breakpoint, but a b versus b a does. Also non-identical initial or terminal genes count as breakpoints. Based on the exemplar strings -b -a -c d and c a b d, the EBD equals 2 and the ERD equals 1.

We have developed a branch and bound approach for solving both versions of the problem, which is efficient for moderate sized instances, e.g., ten gene families with two to ten genes per family in each genome, plus hundreds (EBD) or dozens (ERD) of additional genes with only one copy per genome.

Hybridization

Several types of biological processes, perhaps most widespread in the plant kingdom, give rise to hybrids. We have explored several mathematical problems that arise in trying to model these processes. These problems differ according to the process responsible for hybridization, the kinds of data available, and the aim of the evolutionary reconstruction.

Resolution of Tetraploidy; Ancestral Synteny Unknown. 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 stabilized as a diploid (e.g., Gaut and Doebley, 1997). The two homologous versions of each gene, one from each parent species, may diverge functionally to create a gene family. From the moment of hybridization till the present, the two parent species may also undergo chromosomal rearrangement. Thus we have direct access to neither the ancestral hybrid genome nor the two contributing strains. The goal is to infer an ancestral hybrid genome, i.e., made up of chromosomes, each of which contains genes from one parent species only, such that the number of translocations necessary to derive the modern hybrid is minimized. We have found a linear-time algorithm for reconstructing the ancestral hybrid, given the order of the genes on the chromosomes of the modern hybrid, as well as data (obtained, for example, from sequence analysis) on which of these genes originated from each of the parent species (El-Mabrouk and Sankoff, 1999). As might be expected from the limited type of data input to it, there are generally many equally good solutions to this problem.

Resolution of Tetraploidy; Ancestral Synteny and Gene Order Inferred. A second version of the hybridization problem uses additional data, namely the modern configurations (synteny and gene order) of the two parent genomes *A* and *B*, as well as of the hybrid *G*, to infer the three ancestral genomes *A'*, *B'* and *G'* at the moment of hybridization, as on the left of Figure 5. Note that *G'* consists of the chromosomes in *A'* plus the chromosomes in *B'*.

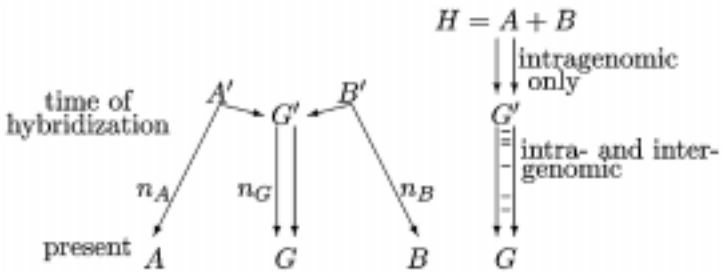

FIGURE 5. Localization of the ancestral hybrid immediately before intergenomic translocations.

As a first step, we infer the total number *n* of evolutionary steps required to produce *G* from a construct *H* consisting of the chromosomes of *A* and the chromosomes of *B*, as on the right of Figure 5. We assume that *G'* is one of the intermediate steps in this evolution, so that $n=n_A+n_B+n_G$, where n_X is the number of steps from genome *X'* to genome *X*, for $X \in \{A,B,G\}$.

Under the assumption that one of the first translocations to occur in the stabilization of the hybrid will be an *intergenomic* one, involving chromosomes from both *A'* and *B'*, we could locate *G'* at the last step on the path from *H* to *G* before the first intergenomic translocation, as on the right of Figure 5. Among the many

paths characterized by the same cost *n*, we argue that the most likely is the one where n_G is as small as possible, to allow for evolution in the parent species as well as the hybrid.

To minimize n_G among all paths, we adapted the techniques of Hannenhalli and Pevzner (1995a, 1995b) in a heuristic for separating the intragenomic and intergenomic stages and gave upper and lower bounds for the optimal transition point between them.

Hybridization through Interspecific Fertility. In hybrids formed through the exceptional fertilization of two distinct though related species, the parent species A and B may differ from each other by numerous genome rearrangements (e.g., Rieseberg *et al.,* 1995; Ungerer *et al.,* 1998). The hybrid G' is able to survive and propagate despite the difference between the two haploid components of its diploid genome. Genome rearrangement of the hybrid rapidly ensues, however, first until a normal symmetric diploid configuration G* is attained, and then while further stabilization of the new genome occurs. This scenario is illustrated in Figure 6. The rapid evolution of the hybrid means that we may assume the relative stability of the parent genomes A and B if the evolutionary scale is not too lengthy.

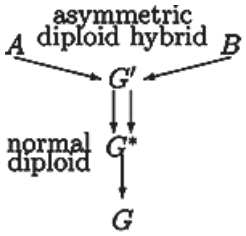

FIGURE 6. Rearrangement before and after development of a symmetric diploid.

We hypothesized that the key stage in the stabilization of the hybrid genome *G**, is to be found by calculating the "median" of the three diploid genomes, the two parents *A* and *B*, and the hybrid *G*. The median is defined to be that genome for which the sum of breakpoint distances to *A*, *B* and *G* is minimum. Though computationally complex, this problem can be solved for moderate size genomes through a reduction to the traveling salesman problem (Sankoff and Blanchette, 1998), for which much software is available.

Genome duplication

Perhaps the most spectacular cause of gene duplication is tetraploidization of the genome. If this doubling of the genome can be resolved in the organism and eventually fixed as a normalized diploid state in a population, it represents a simultaneous duplication of the entire genetic complement. It transcends other mechanisms for gene duplication, in that not only is one copy of each gene free to evolve its own function (or become a pseudogene), but it can evolve in concert with any subset of the thousands of other extra gene copies, with the consequent emergence of whole new physiological pathways. For gene loss versus divergence rates, see Nadeau and Sankoff (1997).

Evidence for the effects of genome doubling has shown up across the eukaryote spectrum. It is particularly well-documented in yeast (Wolfe and Shields, 1997) and plants (e.g., Gaut and Doebley, 1997). Four hundred million years ago, the vertebrate genome may have undergone two duplications (Ohno *et al.*, 1968), though at least one of these remains controversial (Skrabanek and Wolfe, 1998).

Originally, a duplicated genome contains two identical copies of each chromosome, but through inversion or other intrachromosomal movement, the gene orders in each pair of chromosomes change independently, and through reciprocal translocation, parallel linkage patterns in the two copies are disrupted. Eventually,

all that can be detected are chromosome segments of greater or lesser length, each of which appears twice in the genome, containing many paralogous genes in parallel orders. These are analogous to conserved segments in inter-genome comparisons.

We have proposed a suite of "genome halving" problems (El-Mabrouk *et al.*, 1998) 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 minimized. This paper also offered a heuristic algorithm for one version of the problem when only synteny (i.e., chromosomal assignment of each gene) is known, and the ancestral syntenies are to be constructed. In a later paper, El-Mabrouk et al. (1999) found an exact, polynomial-time algorithm for the case where gene order is known and is to be reconstructed. The problem of assuring appropriate centromere distribution during the reconstructed translocational history has been investigated, but not yet solved.

Towards a Synthesis of the Algorithmic and Statistical Approaches

Though they deal with essentially identical phenomena, each approach has fundamental shortcomings when viewed from the perspective of the other. For example, the algorithmic inference of rearrangement history takes for granted that all the breakpoints are given, and its results depend crucially on how they are ordered in the genome, while from the comparative viewpoint we know that many or even most of the breakpoints may not yet have been identified; indeed we can only estimate how many there are. On the other hand, the statistical approach assumes that the number of conserved segments, appropriately corrected for as yet unobserved segments, is proportional to the degree of evolutionary divergence, while we know that the length of algorithmically reconstructed histories may differ by a factor of two (if they contain only inversions and reciprocal translocations) or three (transpositions). Neither approach can be convincingly realistic as it stands. At least two aspects of the research in this field need to be pursued.

Comprehensive Modeling

The idea of a random translocation model makes many geneticists uneasy. It is clear that at the population level, some chromosomes are more susceptible to translocation than others and some chromosomal regions are more prone than others. Indeed, some translocations recur often at exactly the same sites, which would hardly occur in a purely random context. Nevertheless, even if these considerations could be extrapolated to the evolutionary time scale, which is not at all assured, they reflect a simplified notion of randomness. True, the original Nadeau-Taylor model is consistent with a pattern of uniform distribution of breakpoints along the chromosome. But the alternate hypothesis that should be explored retains an element of randomness. Some chromosomes, regions, or even sites, are more translocation-prone, but the "next" translocation to occur cannot be predicted with certainty. Thus a realistic probability model must discard the uniform distribution in favour of a random breakage process parameterized by regional chromosomal characteristics.

We know that several regional characteristics are correlated: G-C rich isochores, R-banding, telomeric proximity, early S-phase replication, recombination rates, oncogenic rearrangements and, most important for our considerations, gene richness and evolutionary rearrangements (Holmquist, 1992; Saccone *et al.*, 1996). Translocation rates have been correlated in some detail for many of these features in human populations (Cohen *et al.*, 1996). All this information must be incorporated into the modeling process in order to characterize the distribution of the number of genes per segment and other predictions of the theoretical model.

Model-based Algorithms

The many new and fascinating combinatorial optimization problems motivated by questions of genomic rearrangement have given rise to a rich literature. We have cited but a small sample biased towards our own interests. In this work, the optimality criterion is basically one of parsimony, and hence suffers some of the shortcomings of this approach, chiefly underestimation, sometimes severe, of the total number of events, the tendency to weight all events equally, occasional lack of robustness to small changes in the data, and multiple solutions rated equally.

The algorithms that have been produced are sometimes extremely intricate and it would not be feasible to try to generalize them for other criteria or to include weightings. Nevertheless the problems should be reformulated in a more realistic manner, making use of the quantitative aspects of the biological context and parameterized probability models. The focus on exact global optima may sometimes have to be sacrificed in favour of heuristic algorithms, local optimization, and simulation. Some steps in these directions are represented in the simulations of Seoighe and Wolfe (1997) of the formation of duplicate segments post-genome-doubling, the algorithm of Blanchette *et al.* (1996) for parameterized genome rearrangement, and the probabilistic modeling of the production of breakpoints in the phylogenetic context by Sankoff and Blanchette (1998).

Acknowledgments

This research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. The author is a Fellow of the Canadian Institute for Advanced Research.

References

Blanchette, M., T. Kunisawa, and D. Sankoff, 1996. Parametric genome rearrangement. Gene 172:GC11-17.

Caprara, A., 1997. Sorting by reversals is difficult. In: Proceeedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97), ACM, New York, NY. 75-83.

Cohen, O., C. Cans, M. Cuillel, J.L. Gilardi, H. Roth, M. A. Mermet, P. Jalbert, and J. Demongeot, 1996. Cartographic study: breakpoints in 1574 families carrying human reciprocal translocations. Human Genetics 97:659–667.

El-Mabrouk, N., J.H. Nadeau, and D. Sankoff, 1998. Genome halving. Pages 235-250 *in:* Combinatorial Pattern Matching. 9th Annual Symposium. M. Farach-Colton, ed. Lecture Notes in Computer Science 1448. Springer Verlag, New York, NY.

El-Mabrouk, N., and D. Sankoff, 1999. Hybridization and genome rearrangement. *In:* Combinatorial Pattern Matching. 10th Annual Symposium. Lecture Notes in Computer Science. Springer Verlag, New York, NY. (In press.)

El-Mabrouk, N., D. Bryant, and D. Sankoff, 1999. Reconstructing the Pre-Doubling Genome. Pages 154–163 *in:* Proceeedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99), ACM, New York, NY. (In press.)

Gaut, B.S., and J.F. Doebley, 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Science (U.S.A.) 94:6809–6814.

Hannenhalli S., and P.A. Pevzner, 1995a. Transforming cabbage into turnip. (polynomial algorithm for sorting signed permutations by reversals). Pages 178–189 *in:* Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing.

Hannenhalli S., and P.A. Pevzner, 1995b. Transforming men into mice (polynomial algorithm for genomic distance problem). Pages 581-592 *in:* Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science.

Holmquist, G.P., 1992. Chromosome bands, their chromatin flavors, and their functional features. American Journal of Human Genetics 51:17–37.
Kaplan H., R. Shamir, and R.E. Tarjan, 1997. Faster and simpler algorithm for sorting signed permutations by reversals. Pages 344-351 *in:* Proceeedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms. ACM, New York, NY.
Nadeau, J.H., and D. Sankoff, 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147:1259–1266.
Nadeau J.H., and D. Sankoff, 1998a. Counting on comparative maps. Trends in Genetics 14:495–501.
Nadeau, J.H., and D. Sankoff, 1998b. The lengths of undiscovered segments in comparative maps. Mammalian Genome 9:491–495.
Nadeau, J.H., and B.A. Taylor, 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. Proceedings of the National Academy of Sciences (USA) 81: 814–818.
Ohno, S., U. Wolf, and N.B. Atkin, 1968. Evolution from fish to mammals by gene duplication. Hereditas 59:169–187.
Parent, M.-N., 1999. Estimation du nombre de segments vides dans le modèle de Nadeau et Taylor sur les segments chromosomiques conservées. Master's thesis, Université de Montréal.
Rieseberg, L.H, C. Van Fossen, and S.M Desrochers, 1995. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. Nature 375:313–316.
Saccone S., S. Caccio, J. Kusuda, L. Andreozzi, and G. Bernardi, 1996. Identification of the gene-richest bands in human chromosomes. Gene 174:85–94.
Sankoff, D., 1989. Mechanisms of genome evolution: models and inference. Bulletin of the International Statistical Institute 47.3:461–475.
Sankoff, D., 1999. Genome rearrangement with gene families. Technical Report, Centre de recherches mathématiques, Université de Montréal.
Sankoff, D., and M. Blanchette, 1998. Multiple genome rearrangement and breakpoint phylogeny. Journal of Computational Biology 5:555–570.
Sankoff, D., and M. Blanchette, 1999. Probability models for genome rearrangement and linear invariants for phylogenetic inference. Pages 302-309 *in:* Proceeedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99), ACM, New York, NY, NY. (In press.)
Sankoff, D., and V. Ferretti, 1996. Karyotype distributions in a stochastic model of reciprocal translocation. Genome Research 6:1–9.
Sankoff, D., V. Ferretti, and J.H. Nadeau. 1997b. Conserved segment identification. Journal of Computational Biology 4:559–565.
Sankoff, D., G. Leduc, N. Antoine, B. Paquin, B.F. Lang and R. Cedergren, 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proceedings of the National Academy of Sciences (USA) 89:6575–6579.
Sankoff, D., and J.H. Nadeau, 1996. Conserved synteny as a measure of genomic distance. Discrete Applied Mathematics 71:247–257.
Sankoff, D., M.-N. Parent, I. Marchand, and V. Ferretti, 1997a. On the Nadeau-Taylor theory of conserved chromosome segments. Page 262-274 *in:* Combinatorial Pattern Matching. A. Apostolico and J. Hein, eds. 8th Annual Symposium Lecture Notes in Computer Science 1264, Springer Verlag.
Seoighe, C., and K.H. Wolfe, 1998. Extent of genomic rearrangement after genome duplication in yeast. Proceedings of the National Academy of Sciences (USA) 95:4447–4452.
Skrabanek, L., and K.H. Wolfe, 1998. Eukaryote genome duplication — where's the evidence? Current Opinion in Genetics and Development 8:694–700.
Ungerer, M.C., S.J.E. Baird, J. Pan, and L.H. Rieseberg, 1998. Rapid hybrid speciation in wild sunflowers. Proceedings of the National Academy of Sciences (U.S.A.) 95:11757–11762.
Wolfe, K.H., and D.C. Shields, 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713.