# Mathematics of Evolution and Phylogeny

Edited by Olivier Gascuel

CLARENDON PRESS · OXFORD
2004

# CONTENTS

# 1

# CONSERVED SEGMENT STATISTICS AND REARRANGEMENT INFERENCES IN COMPARATIVE GENOMICS

David Sankoff

**Abstract**

The statistical treatment of chromosomal rearrangement has evolved along with the biological methods for producing pertinent data. We trace the development of conserved segment statistics through the mouse linkage/human chromosome assignment data analyzed by Nadeau and Taylor in 1984, through the comparative gene order information on organelles (late 1980's) and prokaryotes (mid-1990's), to higher eukaryote genome sequences, whose rearrangements have been studied without prior gene identification. Each new type of data suggested new questions and led to new analyses. We focus on the problems introduced when small sequence fragments are treated as noise in the inference of rearrangement history.

## 1.1 Introduction

The history of modeling and quantitative analysis for comparative genomics has been largely determined by the kinds of experimental data available at various periods (see the timeline in Table 1.1). For over eighty years, recombination-based linkage maps have been used for studying genome rearrangements . Through studies of giant salivary gland chromosomes in *Drosophila*, band structure became a valuable tool seventy years ago, allowing the visualization of inverted segments and the localization of their breakpoints with the microscope, and enabling the first rearrangement-based phylogeny.

Cytogenetics blossomed in the intervening years but modern banding techniques for human and other eukaryotic chromosomes are little more than thirty years old. These soon led to phylogenies for primates and and a number of other groups. The last thirty years also saw the development of radiation hybrid methodology as well as well as a number of sequence-level molecular biological techniques, first for gene assignment to chromosomes, then for constructing chromosomal maps of genes and other features at increasing levels of resolution. Complete genome sequencing resulted in the first complete virus map in 1975 and the first complete organelle map in 1981, and increasing number of these became available for comparative work in the mid-eighties. It is less than ten

| | | |
|---|---|---|
| 1921 | recombination maps | Sturtevant [48] |
| 1933 | chromosome bands (*Drosophila*) | Painter [30] |
| 1970 | chromosome bands (*human*) | Caspersson *et al.* [41] |
| 1975 | radiation hybrid | Goss and Harris [16] |
| 1975 | virus genome sequences | Sanger *et al.* [35] |
| 1981 | organelle genome sequences | Anderson *et al.* [1]) |
| 1995 | prokaryotic genome sequence | Fleischmann *et al.* [13] |
| 1996 | eukaryotic genome sequences | Goffeau *et al.* [15] |
| 2001 | human genome sequence | [53,21] |

**Table 1.1** *Availability of comparative genomic data.*

years since whole genome sequences of prokaryotes could be compared, and it is within the last five years that comparative genomics of eukaryotes could be based on whole genome sequences.

There are at least two mathematically-oriented literatures in comparative genomics that go beyond traditional quantitative reports of intensive variables such as base composition or codon usage or summary variables such as genome size or gene content. One is the statistical analysis of genetic maps to quantitatively characterize the chromosomal segments conserved in both of two genomes being compared, as well as the breakpoints between these segments, dating to the fundamental paper of Nadeau and Taylor [28]. The other is the algorithmic inference of rearrangement processes, highlighted by the remarkable work of Hannenhalli and Pevzner [18,17,19], based on the comparison of complete gene orders. In this chapter, we review statistical analyses based on recombination distance and on gene order, as well as, very briefly, algorithms based on complete gene order, before focusing on the convergence of statistics and algorithmics in the comparison of whole genome sequences.

## 1.2 Genetic (recombinational) distance

At the time Nadeau and Taylor developed their approach to conserved segment statistics, distance along chromosomes was quantified in terms of linkage disequilibrium in recombination experiments, measured in centimorgans. They observed that some genes known to be located on the same human chromosome had homologous genes clustered on the same mouse chromosome linkage map, generally in the same order or in exactly the inverse order. Their insight was that the position of the mouse genes in these clusters could be used to determine the average size $\mu$ of conserved segments and hence the total number $n$ of conserved segments, where the known total genome length is $|G| = n\mu$. Since the different clusters generally did not overlap, they made the assumption that each cluster represented a sample of the genes in a single conserved segment. The data to be considered was then of the form represented in Fig. 1.1.

Then the simplest form of the inference, though not exactly in Nadeau and Taylor's terms, is as follows: Let $x_1 < \cdots < x_h$ be order statistics based on $h$ independent samples from a uniform distribution on $[a, b]$. There are a number

$$a \quad x_1 \quad x_2 \quad \cdots \quad x_h \quad b$$

FIG. 1.1. Genes of known position $x_1 < \cdots < x_h$ in a conserved segment with unknown endpoints (breakpoints) $a$ and $b$.

of ways of estimating $b - a$: the maximum likelihood estimate is $x_h - x_1$ but for small $h$ this is obviously very biased towards underestimation. An unbiased estimate of $b - a$, but one which is only defined for $h \geq 2$, is

$$(\widehat{b-a}) = \frac{h+1}{h-1}(x_h - x_1). \tag{1.1}$$

Nadeau and Taylor could not calculate $\mu$ by simply averaging the estimates for the different segments, for two reasons. First, they could not observe those segments containing no mapped genes, and for data quality reasons, they did not consider segments containing only one gene. (In any case, the length estimators do not give meaningful estimates for segments containing one gene.) Second, the expected number of genes observed in a segment is proportional to the length of that segment, which itself approximately follows an exponential distribution, assuming a uniform distribution of breakpoints. Thus, the set of observed segment estimates must be fit to an exponential length distribution, weighted by the probability that a segment of a specific length contains at least two mapped genes. The parameter of this distribution is an estimate $\hat{\mu}$ of $\mu$, the average segment length. An estimate of the number of segments is then $\hat{n} = |G|/\hat{\mu}$.

### 1.3 Gene counts

We can use the uniformly distributed breakpoints component of the Nadeau-Taylor procedure to estimate $n$ without first estimating the size in centimorgans of the observed segments, simply by counting the number of genes in each observed segment.

We model the genome as a single long unit broken at $n - 1$ random breakpoints into $n$ 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 [38]. If the total number of genes is $m$, the marginal probability that a segment contains $r$ genes, $0 < r < m$, has been shown [42] to be :

$$P(r) = \left(1 + \frac{m}{n+1}\right)^{-1} \frac{\binom{m}{r}}{\binom{n+m}{r}}. \tag{1.2}$$

We cannot directly compare the theoretical distribution $P(r)$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, and hence $n$ is unknown .

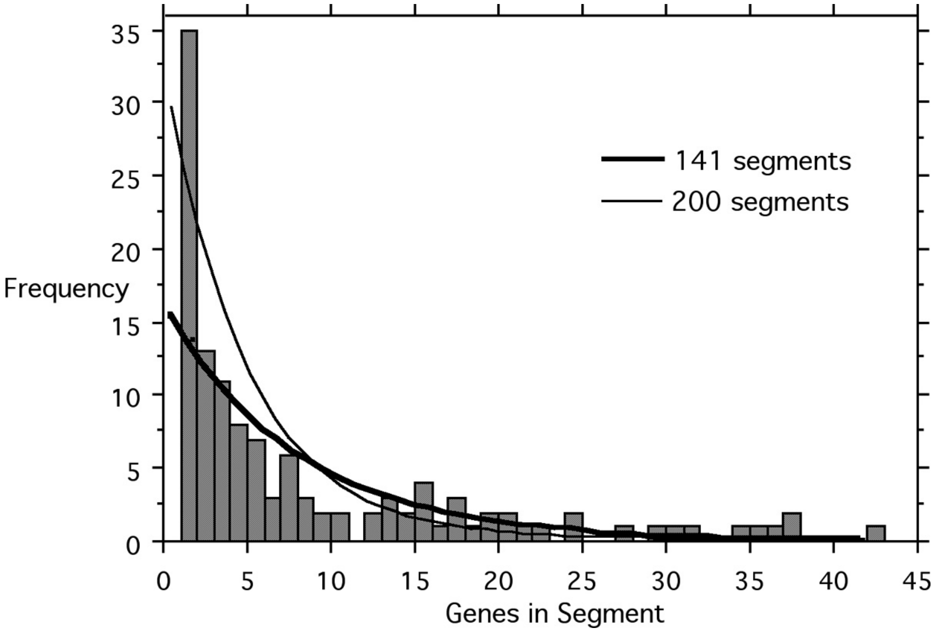

FIG. 1.2. Comparison of relative frequencies $n_r, r > 0$ of segments containing $r$ genes, with predictions of the Nadeau-Taylor model, for MLE $\hat{n} = 141$ and Kolmogorov-Smirnov-based estimator $\hat{n} = 200$ [41].

We can, however, compare the frequencies $n_r$ with the predicted frequencies $\hat{n}P(r), r > 0$, for various estimators $\hat{n}$, as illustrated in Fig. 1.2, our first analysis based on the $m$ =1423 human-mouse orthologies documented in 1996.

The largest discrepancy is the comparison between $n_1$ and $\hat{n}P(1)$, due at least in part to error in the identification of orthologous genes or other experimental error in chromosome assignment, but also possibly to a genuine shortfall in the model when predicting the number of short segments. We will return to this question in Section 1.8 on genome sequences.

## 1.4 The Inference Problem

It might seem undeniable that the number of segments $n_r$ observed to contain $r$ genes, for $r = 1, 2, \cdots, m$ would be useful data for inference about the Nadeau-Taylor model, in particular about $n$, the unknown number of segments . It is remarkable, then, that to estimate $n$ from $m$ and $n_r$, only the number of non-empty segments $a = \sum_{r>0} n_r$ is important, since for practical purposes it behaves like a sufficient statistic for the estimation of $n$ [41], although sufficiency is not strictly satisfied [20].

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

$$P(a,m,n) = \frac{\binom{m-1}{a-1}\binom{n}{a}}{\binom{n+m-1}{m}}, \tag{1.3}$$

which is a constrained hypergeometric distribution with mean and variance:

$$\mu_a = \frac{mn}{m+n-1}, \qquad \sigma_a^2 = \frac{n(n-1)m(m-1)}{(n+m-2)(n+m-1)^2}. \tag{1.4}$$

Note that this model reduces to a classical occupancy problem of statistical mechanics ( [12], p62).

The maximum likelihood estimate $\hat{n}$, given $m$ and $a$, is the value of $n$ which maximizes $P$. For given $m$ and $n$, the expectation and the variance of $\hat{n}$ can be calculated making use of the probability distribution in eqn (1.3), except in the special case of few data ($m \leq n$) and every gene in a separate segment ($a = m$), where the estimates are undefined.

Substituting $a$ for $\mu_a$ in eqn (1.4) gives Parent's estimator [31]

$$\hat{n} = \frac{a(m-1)}{m-a}, \tag{1.5}$$

which, when rounded to the nearest integer, coincides with the maximum likelihood estimator over the range of $a, m$ and $n$ likely to be experimentally interesting, as long as some segments contain at least two genes .

Alternatives, extensions and generalizations of the Nadeau-Taylor and the gene count approaches have been investigated by a number of researchers. Schoen has shown that high marker (e.g. gene) density and high translocation/inversion ratios greatly improve the accuracy of estimation [45,46]. Waddington *et al.* have developed the theory in the direction of allowing different densities of breakpoints for each chromosome [54] and have compared various approaches for their performance in avian genomes, with their distinctly bimodal distribution of chromosome sizes [55]. The evolution of chromosome sizes have been studied analytically, through simulation and empirically [38, 4, 10]. Marchand [27] initiated the statistical study of inhomogeneities in breakpoint densities and gene densities on the chromosome. Housworth and Postlethwait [20] showed how the number of observed conserved syntenies, i.e., pairs of chromosomes – one in each genome – that share at least one ortholog, has some better statistical properties than the number of observed segments.

## 1.5 What can we infer from conserved segments?

The comparative study of whole-genome maps makes no formal reference to the processes that create the breakpoints while progressively fragmenting the conserved segments, except for an implicit assumption that the number of breakpoints and segments increases roughly in parallel with the number of rearrangement events affecting either of the two genomes being compared. In observing the order of segments along the chromosomes in one genome while noting to which chromosomes they correspond in the other genome, however, we can extract additional information about the relative proportion of intra-chromosomal and inter-chromosomal events that gave rise to this pattern. Considering only autosomes, i.e., setting aside the sex chromosomes, which are essentially excluded from inter-chromosomal exchanges, let the total number of segments on a human chromosome $i$ be

$$n^{(i)} = t + u + 1, \tag{1.6}$$

where $t$ is the number due to inter-chromosomal transfers, and $u$ the number due to local rearrangements. Under a random exchange model we can try to predict how often two or more segments from the same mouse chromosome will co-occur on the same human chromosome through inter-chromosomal events. By then compiling co-occurrence frequencies from the empirical comparison of the two genomes, we can estimate the the relative proportion of intra-chromosomal and inter-chromosomal events.

We label the ancestral chromosomes $1, \cdots, c$, ignoring for the moment that there may have been changes in the number of chromosomes due to fusions and/or fissions in the human or mouse lineages or both. We model each chromosome as a linear segment with identified left-hand and right-hand endpoints. A reciprocal translocation between two chromosomes $h$ and $k$ consists of breaking each one, at some interior point, into two segments, and rejoining the four resulting segments such that two new chromosomes are produced, each containing a left-hand part of one of the original chromosomes and the right-hand part of the other. We label each new chromosome according to which left-hand it contains, but for each of its constituent segments, we retain the information of which ancestral chromosome it derived from.

At the outset, assume the first translocation on the human lineage involves ancestral chromosome $i$. We assume that its partner can be any of the $c-1$ other ancestral autosomes with equal probability $\frac{1}{c-1}$, so that the probability that the new chromosome labelled $i$ contains no fragment of ancestral chromosome $h$, where $h \neq i$, is exactly $1 - \frac{1}{c-1}$. For small $t$, after chromosome $i$ has undergone $t$ translocations, the probability that it contains no fragment of the ancestral chromosome $h$ is approximately $(1 - \frac{1}{c-1})^t$, with some correspondingly small corrections, for example to take into account the event that $h$ previously translocated with one or more of the $t$ chromosomes that then translocated with $i$, and that a secondary transfer to $i$ of material originally from $h$ thereby occurred.

Then the probability that the new (i.e., human) chromosome $i$ now contains at least one fragment from $h$ is approximately $1 - (1 - \frac{1}{c-1})^t$ and the expected number of ancestral chromosomes with at least one fragment showing up on human chromosome $i$ is

$$E(c^i) \approx 1 + (c-1)[1 - (1 - \frac{1}{c-1})^t], \tag{1.7}$$

where the leading 1 counts the fragment containing the left-hand endpoint of the ancestral chromosome $i$ itself.

We assume that our random translocation process is stochastically reversible. This assumption should not introduce much error as long as chromosome sizes do not deviate too much from their stationary distribution. Then we can treat the mouse genome as ancestral and the human derived (or vice versa), instead of considering them as diverging independently from a common ancestor. Now $E(c^i)$ represents the expected number of mouse chromosomes with at least one fragment showing up on human chromosome $i$.

As $t$ increases for all the chromosomes, so that each human chromosome contains segments from several mouse chromosomes, eqn (1.7) will start to overpredict $c^i$, since each translocation will be more likely to transfer fragments of the same origin as those already contained in the chromosome $i$. Nevertheless, substituting $c^i$ for $E(c^i)$ in eqn (1.7) gives us

$$\hat{t} = \frac{\log(c-1) - \log(c-c^i)}{\log(c-1) - \log(c-2)}, \tag{1.8}$$

a good first estimate of $t$, where $c = 19$, the number of mouse autosomes. To illustrate, for the 22 human autosomes, a 100 Kb resolution construction [52] indicates 350 autosomal segments, while the sum of the $c^i$ is 109. Applying eqn (1.8) to each chromosome and summing the 22 values of $\hat{t}$ gives a total of 130 segments. In other words, for $130 - 109 = 21$ segments, two (or more) segments from the same mouse chromosome are found on the same human chromosome *because of independent translocational events*. By eqn (1.6), this leaves unaccounted for

$$\begin{aligned} \sum u &= \sum n^{(i)} - \sum t - 22 \\ &= 350 - 130 - 22 \\ &= 198 \end{aligned}$$

segments, which must be attributed to local rearrangements such as inversion.

Table 1.2 shows the results of these calculations for this and a number of other maps of various levels of resolution, based on genomic sequence or gene maps. Of interest in the genome sequence-based results is the relative stability of the estimates of the number of reciprocal translocations or other inter-chromosomal events versus the great increase in local rearrangements over the analyses based on gene maps. This reflects the discovery of high numbers of smaller-scale local arrangements recognizable from genomic sequence [25,8] compared to gene maps. As resolution increases, a greater proportion of these local rearrangements have no effect on gene order and more of the conserved segments identified will contain no genes. At the same time, many of the conserved segments identified in the recent gene maps contain a number of genes in a relatively small stretch of sequence, too short to even show up as a conserved segment in the sequence-based analyses (cf [5,51]). Thus the congruence apparent between large conserved segments in the genome sequence and the gene map data breaks down as we zoom down to smaller segments, with small segments of conserved sequence containing

| resolution of<br>comparative map | autosomal<br>segments<br>$\sum n^{(i)}$ | segment-sharing<br>chromosome pairs<br>$\sum c^i$ | inter-<br>chromosomal<br>$\sum t$ | intra-<br><br>$\sum u$ |
|---|---|---|---|---|
| 100 Kb [52] | 350 | 109 | 130 | 198 |
| 300 Kb [8] | 370 | 107 | 128 | 220 |
| 1 Mb [14] | 270 | 100 | 117 | 131 |
| 2K genes [47] | 192 | 99 | 114 | 59 |
| 12K genes [7],NCBI | 213 | 113/120 | 137/149 | 64/41 |

**Table 1.2** *Inference of inter- and intra-chromosomal rearrangements based on number of conserved segments and number of segment-sharing autosome pairs in the two genomes. Sources: [52] based on UCSC Genome Browser, [8, 14] on anchor-sequence constructions, [47] on outdated human mouse homology data cited in [40], [7] on MGI 2004 Oxford grid cells containing at least three/two genes for the $c^i$, and on NCBI Human Mouse Homology Map for the $n^{(i)}$.*

no genes and the small segments containing genes passing beneath the radar of sequence-based analyses.

Many of the single-gene orthologies on comparative maps are undoubtedly due to paralogy and other errors in assignment, but a significant proportion will certainly prove to be valid, opening questions about the nature of the processes creating them. The repertoire of inversions, reciprocal translocations and Robertsonian translocations popular with modelers may have to be expanded to include such processes as transpositions or jump translocations, within and between chromosomes and non-tandem duplication processes, with or without loss of functionality.

## 1.6 Rearrangement algorithms

We can derive much more detailed inferences about the processes responsible for a particular comparative map if we are willing to work within the framework of a sufficiently restrictive model, though we must then be vigilant that our results are really consequences of the data rather than simply artifacts of the model restrictions. The types of chromosomal rearrangement most often modeled are inversions, reciprocal translocations and fissions and fusions, including Robertsonian translocations. The basic aim is to efficiently transform a given genome, represented as a set of idealized disjoint chromosomes made up of an ordered subset of genes, into another given genome made up of the same genes but differently partitioned among chromosomes, in a minimum number of steps $d$. The algorithm outputs $d$ and a sequence of $d$ rearrangements that carry out the desired transformation. The literature on this problem area (highlighted by the Hannenhalli-Pevzner discoveries [17–19] and reviewed in [37]) is extensive and has seen much recent progress (cf [50,58,29,49] and chapters in this volume), and we will not go into details here. Some points will be important in the ensuing sections:

- Each reciprocal translocation or inversion increases or decreases the number of segments by at most two; i.e., it adds or removes at most two breakpoints between adjacent segments. (Other rearrangements, such as transpositions or “jump translocations” can change the number of segments by three, but these are thought to be rare.)
- In the Hannenhalli-Pevzner algorithms and their improvements, virtually all moves decrease the number of segments, i.e., decrease the number of breakpoints, by two or one.
- In general, there are a large number of optimal solutions.
- The algorithms consider all operations to have the same cost, independent of whether they are inversions or translocations, and independent of how many genes are in their scope. This is essential to the algorithms. If we wish to modify the problems or to change the objective function, the mathematical basis of the algorithm is lost.

We will return to other aspects of these algorithms in Section 1.8.

## 1.7 Loss of signal

To what extent does the sequence of rearrangements reconstructed by rearrangement algorithms actually reflect the true evolutionary history? It is well-known that past a threshold of $\theta n$, where $n$ is the number of genes and $\theta$ is in the range of 1/3 to 2/3, the inferred value of $d$ tends to underestimate the number of events that actually occurred [22–24].

Whether any signal is conserved as to the actual individual events themselves, and which ones, is even more problematic.

Lefebvre *et al.* [26] carried out the following test: For a genome of size $n = 1000$, they generated $u$ inversions of size $l$ at random (for $l =$5, 10, 15, 20, 50, 100, 200), and then reconstructed the optimal inversion history, for a range of values of $u$. Typically, for small enough values of $u$, the algorithm reconstructed the true inversion history, although inversions that do not overlap may be reconstructed in any order. Above a certain value of $u$, however, depending on $l$, the reconstructed inversions manifest a range of sizes, as illustrated in Fig. 1.3, reflecting the ability of the algorithm to find alternative solutions, and eventually solutions where $d < u$, with the concomitant decay of the evolutionary signal.

For each $l$, they then calculated
$\overline{s}_l = \max(u|\text{reconstruction has at most 95\% error})$
$\underline{s}_l = \min(u|\text{reconstruction has at least 5\% error})$
where any inversion having length different from $l$ is considered to be an error. Fig. 1.4 plots $\overline{s}$ and $\underline{s}$ as a function of $l$ and shows how quickly the detailed evolutionary signal decays for large inversions. Only for very small inversions is a clear signal preserved long after longer ones have been completely obscured.

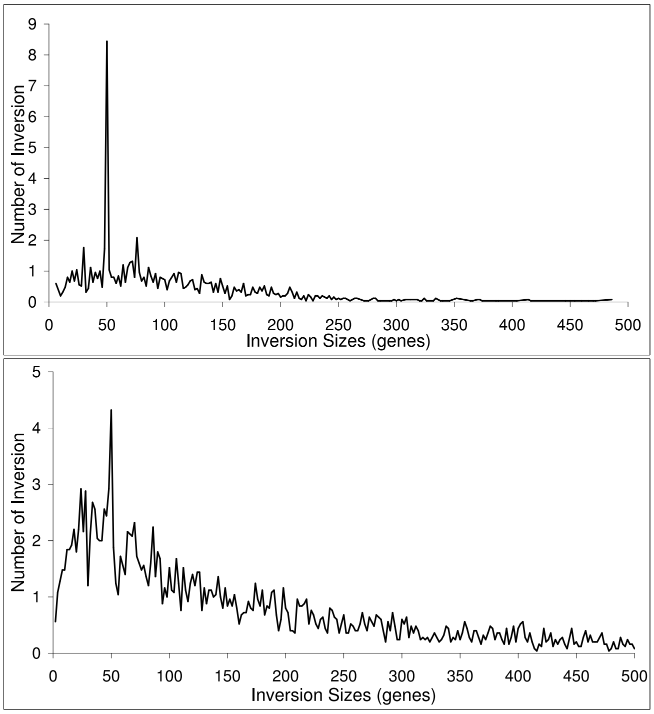

Fig. 1.3. Frequency of inversion sizes inferred by the algorithm for random genomes obtained by performing $u$ inversions of size $l = 50$. Top: $u = 80$. Bottom: $u = 200$.

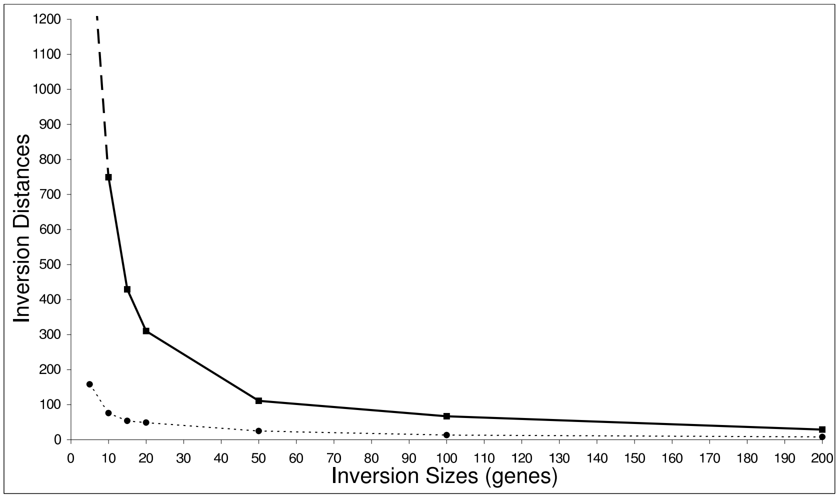

Fig. 1.4. Solid line: values of $\overline{s}$. Dotted line: values of $\underline{s}$.

## 1.8 From gene order to genomic sequence

Gene order rearrangement algorithms can handle many thousands of genes in reasonable computing time. Faced with large nuclear genome sequences, particularly from the higher eukaryotes, however, uncertainties in global alignments, lack of complete consensus inventories of genes and the difficulties of distinguishing among paralogs widely distributed across the genome, constitute apparently insurmountable impediments to the direct application of the algorithms.

### 1.8.1 *The Pevzner-Tesler approach*

In comparing drafts of the human and mouse genomes , Pevzner and colleagues [32–34, 8] adopt an ingenious stratagem to leap-frog the global alignment, gene finding and ortholog identification steps. In their first study, on the human-mouse comparison, they analyzed almost 600,000 relatively short (average length 340 bp) anchors of highly aligned sequence fragments as a starting point for building blocks of conserved synteny, and then amalgamated neighboring sub-blocks using a variety of criteria to avoid disruptions due to "microrearrangements" less than 1Mb. This procedure inferred a set of 281 blocks larger than 1 Mb, which is basically what is reported in [14] and [8]. (The latter also improve the resolution down to 300 Kb – see Table 1.2.) They then used the order of these blocks on the 23 chromosomes as input to a gene order rearrangement algorithms in order to reconstruct optimal sequences of $d$ inversions and translocations to account for the divergent arrangements of the two genomes.

### 1.8.2 *The re-use statistic $r$.*

One of the key results reported by Pevzner and Tesler pertains to the "re-use" of the breakpoints between the $b'$ syntenic blocks on the $c$ chromosomes used as input to their rearrangement algorithms. Basic to the combinatorial optimization approach to inferring genome rearrangements are the bounds $\frac{b}{2} \leq d \leq b$, where $b = b' - c$. (This type of bound was first found in 1982 [57].) We define breakpoint re-use as $r = \frac{2d}{b}$. Then

$$1 \leq r \leq 2.$$

The lower value $r = 1$ is characteristic of an evolutionary trajectory where each inversion or translocation breaks the genome at two sites specific to that particular rearrangement; no other inversion or translocation breaks the genome at either of these sites. High values of $r$, near $r = 2$, are characteristic of evolutionary histories where each rearrangement after the first one breaks the genome at one new site and at one previously broken site. In their comparison of the human and mouse genomes, Pevzner's group find that $r$ is somewhere between 1.6 and 1.9, depending on the resolution of their syntenic block construction and on whether they discard telomeric blocks or not, and argue that this is evidence that evolutionary breakpoints are concentrated in fragile regions covering a relatively small proportion of the genome.

Now it is easily shown that for random permutations of size $n$, the expected value of $b$ is very close to $n$ [39], and it is an observed property [23] of such permutations that the number of inversions needed to sort them ($d$) is also very close to $n$, and thus breakpoint re-use is close to 2. Without getting into the substantive claim about fragile regions persisting across the entire mammalian class, for which the evidence is controversial [25, 11, 43, 36], we may ask what breakpoint re-use in empirical genome comparison really measures: a bonafide tendency for repeated use of breakpoints or simply the degree of randomness of one genome with respect to the other at the level of synteny blocks.

### 1.8.3 *Simulating rearrangement inference with a block-size threshold*

To see whether a high inferred rate of breakpoint re-use necessarily reflects a real phenomenon or is an artifact of methodology, Sankoff and Trinh [44] generated model genomes with NO breakpoint re-use ($r = 1$), then mimicked the Pevzner-Tesler imposition of a block-size threshold by discarding random parts of the genome before applying the Hannenhalli-Pevzner algorithm to the remainder of the genome to infer $d$ and hence $r$.

Each genome consisted of a permutation of length $n = 1000$ or $n = 100$ terms generated by applying $d$ “two-breakpoint” inversions to the identity permutation $(12 \cdots n)$. A two-breakpoint inversion is one that disrupts two hitherto intact adjacencies in the starting (i.e. identity) permutation. At each step, the two breakpoints were chosen at random among the remaining original adjacencies. This represents the extreme hypothesis of no breakpoint re-use at all during evolution, which is not unreasonable given the 3 x $10^9$ distinct dinucleotide sites available in a mammalian genome.

Of course, the terms are just abstract elements in the permutation and have no associated size, and indeed the Hannenhalli-Pevzner procedures do not involve any concept of block size. Thus, one way to imitate the effect of imposing a block-size threshold involves simply deleting a fixed proportion of the terms at random, the same terms from both the starting and derived genomes, relabeling the remaining terms according to their order in the starting (identity) genome, and applying the Hannenhalli-Pevzner algorithm.

It can be shown that before any deletions, the Hannenhalli-Pevzner algorithm will recover exactly $d$ inversions. At each step it will find a configuration of form $\cdots gh \mid -(i-1) \cdots -(h+1) \mid ij \cdots$ and will “undo” the inversion between $h$ and $i$, removing two breakpoints. There being $b = 2d$ breakpoints, breakpoint re-use is 1.0.

What happens as terms are deleted? Suppose $j \neq i+1$ in the above example, and $i$ is deleted. Then the two-breakpoint inversion from $-(i-1)$ to $-(h+1)$ is no longer available to undo. An inversion that erases the breakpoint between $h$ and $-(i-1)$ will not eliminate a second breakpoint. So while the distance $d$ drops by 1, the number of breakpoints $b$ also only drops by 1, and $r$ increases.

The probability that one, two, or more two-breakpoint inversions are “spoiled’ in this way depends on the number of terms deleted.

Figure 1.5 shows how $r$ increases with the proportion of terms deleted, for different values of $d$, for $n = 100$ and $n = 1000$.

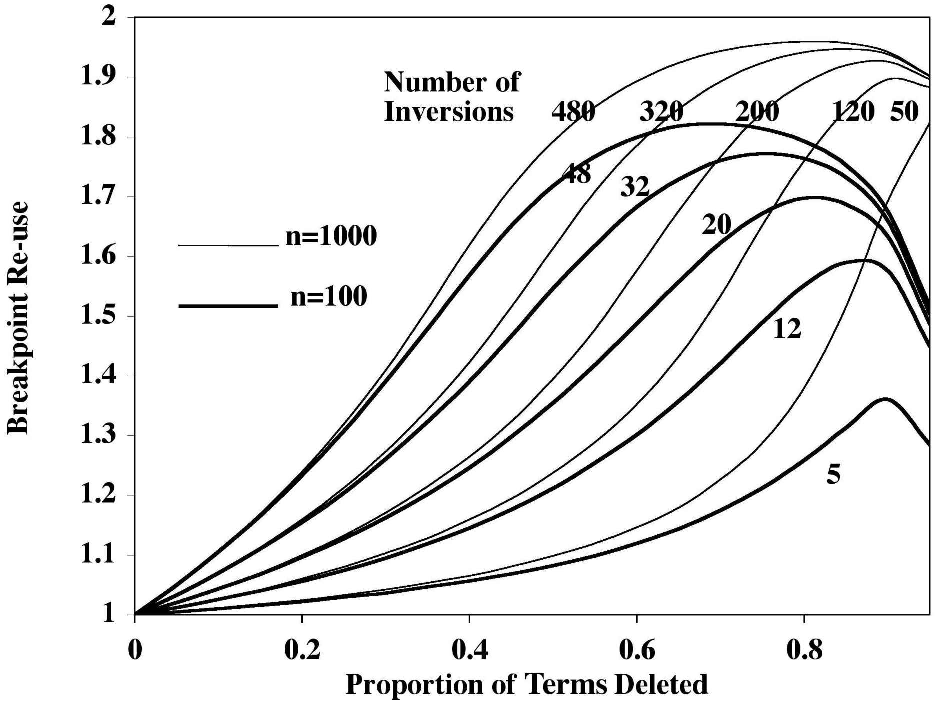

Fig. 1.5. Effect of deleting random terms on breakpoint re-use, as a function of proportion of terms deleted, for various levels of rearrangement of the genome.

Note

- $r$ increases more rapidly for more highly rearranged genomes.
- the initial rate of increase of $r$ depends only on $d/n$
- the increase in $r$ levels off well below $r = 2$ and then descends sharply. The maximum level attained increases with $n$.

The first of these is readily explained. In more rearranged permutations, the deletion of term $i$ is more likely to cause the configuration change described above, i.e. $\cdots gh \mid -(i-1) \cdots -(h+1) \mid ij \cdots$, simply because it is more likely that $j \neq i+1$.

The third observation is also easily understood. For large $n$, the re-use rate $r$ approaches 2 for random permutations. As $n$ decreases, however, expected re-use drops as indicated in Table 1.3. As more and more terms are dropped from a permutation, it loses its "structure", i.e., the pairs of breakpoints involved in the original inversions are wholly or partially deleted, and the remaining permutation becomes essentially random. We may consider that after a curve in Fig. 1.5 attains its maximum, it is entering into the "noisy" region where the historical

| $n$ | $r$ |
|---:|---|
| 5 | 1.53 |
| 25 | 1.83 |
| 50 | 1.90 |
| 100 | 1.94 |
| 250 | 1.97 |

**Table 1.3** *Expected re-use for random permutations as a function of $n$. Estimated from samples of size 500.*

signal becomes thoroughly hidden.

### 1.8.4 *A model for breakpoint re-use*

This section explains the second observation above about the pertinence of $d/n$ for the initial shape of the curves. Suppose a genome $G$ has $b$ breakpoints with respect to $12\cdots n$ and the inversion distance is $d = d_2 + d_1$, where $d_1$ and $d_2$ represent the number of one-breakpoint inversions and two-breakpoint inversions required to sort $G$ optimally. Then $2d_2 + d_1 = b$.

Suppose now that we delete one gene $i$ at random and relabel genes $j = i+1, \cdots, n$ as $j = i, \cdots, n-1$, respectively. The number of breakpoints changes, and quantities $b, d_1, d_2$ and $d$ can change only if the original gene $i$ was flanked by two breakpoints. The probability of this event is $b(b-1)/n(n-1)$. The various configurations in which the two breakpoints may be involved, their probabilities and the effects of deleting $i$ on $d_1$ and $d_2$ ($b$ always decreases by 1, except in case 21 where it decreases by 2) are summarized in Table 1.4 and discussed in some detail in [44].

| | | | effect | on |
|---|---|---|---|---|
| Case | configuration | probability | $d_1$ | $d_2$ |
| 11 | $-(i+1) \mid i \mid j$ | $\frac{d_1}{b(b-1)}$ | -1 | 0 |
| 12 | $g \mid -(i-1)\cdots\cdots h \mid i \mid j \cdots -(i+1) \mid k$ | $\frac{d_1(d_1-2)}{4b(b-1)}$ | -1 | 0 |
| | $g \mid -(i-1)\cdots h \mid i \mid -(k-1)\cdots j \mid k$ | $\frac{d_1(d_1-2)}{2b(b-1)}$ | -1 | 0 |
| | $g \mid h \cdots -(g+1) \mid i \mid -(k-1)\cdots j \mid k$ | $\frac{d_1(d_1-2)}{4b(b-1)}$ | -1 | 0 |
| 21 | $-(i+1) \mid i \mid\ -(i-1)$ | $\frac{1}{d_2-1}\ \frac{2d_2(2d_2-1)}{b(b-1)}$ | 0 | -1 |
| 22 | $g \mid -(i-1)\cdots -(g+1) \mid i \mid -(k-1)\cdots -(i+1) \mid k$ | $\frac{d_2-2}{d_2-1}\ \frac{2d_2(2d_2-1)}{b(b-1)}$ | +1 | -1 |
| 3 | $g \mid -(i-1)\cdots -(g+1) \mid i \mid -(k-1)\cdots j \mid k$ | $\frac{d_1 d_2}{b(b-1)}$ | +1 | -1 |
| | $g \mid -(i-1)\cdots -(g+1) \mid i \mid j \cdots -(i+1) \mid k$ | $\frac{d_1 d_2}{b(b-1)}$ | +1 | -1 |

**Table 1.4** *Probabilities and usual effects of discarding gene $i$ in various configurations, given it is flanked by two breakpoints. Probabilities include those of inverted or nested versions (not listed) of configurations shown. Special cases of configurations with order $O(1/n)$ probabilities not distinguished, e.g. $g \mid -(i-1)\cdots h \mid i \mid h+1\cdots -(i+1) \mid k$.*

These considerations are, of course only valid insofar as the inversions associated with the endpoints are directly available in $G$ (in fact, some are set up by other inversions later, during the sorting of $G$), but they give us an idea of the dynamics of the situation and motivate the deterministic model:

$$d_2(t+1) = d_2(t) + \frac{b(t)(b(t)-1)}{(n-t)(n-t-1)}(-p_{21}(t) - p_{22}(t) - p_3(t))$$

$$= d_2(t) - \frac{2d_2(2d_2-1) + 4d_1d_2}{(n-t)(n-t-1)},$$

$$d_1(t+1) = d_1(t) + \frac{b(t)(b(t)-1)}{(n-t)(n-t-1)}(p_{22}(t) + p_3(t) - \max[0, p_1(t)])$$

$$= d_1(t) + \frac{\frac{d_2-2}{d_2-1}2d_2(2d_2-1) + 4d_1d_2 - \max[0, d_1(d_1-1)]}{(n-t)(n-t-1)}.$$

$$b(t+1) = 2d_2(t+1) + d_1(t+1),$$

where $t$ ranges from 0 to $n$ and with initial conditions $b(0) = 2d_2(0) = 2d(0)$ and $d_1(0) = 0$. (N.B. All the $d$ terms on the RHS of the recurrence should be understood as indexed by $t$.)

Figure 1.6 shows how the recurrence models closely the average evolution of $r$ as the number of terms randomly deleted increases, particularly at the outset, before there are large numbers of one-breakpoint inversions in the Hannenhalli-Pevzner reconstruction. As $d_1$ increases, the model renders less well the changing structure of optimal reconstructions. Finally, the loss of historical signal in the noisy zone for the reconstructions is not built into the model, which thus attains $r = 2$ as the last terms of the permutation are deleted rather than the values in Table 1.3.

Let $\theta = t/n$ represent the proportion of terms deleted. Formally, since $r = 2d/b$, and $d$ is constant in a neighbourhood of $t = 0$, while $db/dt \approx -(b/n)^2$, we can write that $dr/d\theta|_{\theta=0} = 2d/n$. This explains the coincidence between the curves for $n = 100$ and $n = 1000$ in Fig. 1.5.

### 1.8.5 *A measure of noise?*

After investigating the effect of threshold size on $r$, albeit indirectly by varying the rate of random deletion of blocks, Sankoff and Trinh [44] carried out simulations that showed how amalgamations exacerbate the re-use artifact caused by deleting small blocks.

Though Pevzner and Tesler used $r$ to infer relative susceptibility of genomic regions to rearrangement, the simulations described in this section show that it serves rather to measure the loss of signal of evolutionary history, due to the

FIG. 1.6. Plot of $r$ predicted by the recurrence compared to true value estimated by simulation.

imposition of thresholds for retaining syntenic blocks and for repairing microrearrangements. Indeed, breakpoint re-use of the same magnitude as found by Pevzner's group may very well be artifacts of the use of thresholds in a context where NO re-use actually occurred. Indeed, while this may not have been their goal, Pevzner and Tesler have invented a statistic that is a measure of the noise affecting a genomic rearrangement process at the sequence level. Given some information about the parameters of rearrangement, the number of blocks and the size of the thresholds, the re-use rate tells us whether we can have confidence in evolutionary signal reconstructed, whether it must be considered largely random, or whether we are in the “twilight” zone.

### 1.9 Between the blocks

The syntenic blocks are reconstructed by algorithms that bridge the gaps between neighbouring well-aligned regions on both genomes, such as the Pevzner-Tesler method described in Section 1.8.1 above or the approach used in the UCSC Genome Browser [25].

Generally the two syntenic blocks on either side of a breakpoint on, say, a human chromosome do not abut directly, but are rather separated by a short region where there is little similarity with the mouse genome. The obverse of analyzing the order of the reconstructed syntenic blocks as in Section 1.8 is the investigation of these regions, the largely unaligned stretches of genomic DNA left over once the blocks are identified.

Pevzner and Tesler interpret the lack of sustained human-mouse similarity in the regions containing breakpoints as suggestive of the “fragility” of these regions, their susceptibility to frequent rearrangement, in line with their claimed inference of breakpoint re-use. Previous documentation of evolutionary subtelomeric translocational hotspots and pericentromeric duplication and/or transpositional hotspots [11] can be adduced to support the strong hypothesis that potential breakpoints are largely restricted to a limited number (e.g. $< 500$) of very small regions in the genome, and that this regional susceptibility is conserved over considerable evolutionary time scales. Further lines of evidence for this viewpoint include the high rates of recurrence of certain breakpoints in the clinical study of tumor cell karyotypes, and the existence of certain physically fragile regions in human chromosomes under laboratory conditions.

But how can we reconcile the apparently contradictory notions of evolutionarily conserved fragility of breakpoint regions and the lack of human-mouse similarity in these regions? If conserved fragility is based on some substantial primary sequence signal, why is this not picked up by the alignment protocol and how is it conserved if the region is being churned by rearrangements? There are, of course, many possible answers: the signals may be too short, they may be removed by repeat-masking prior to the reconstruction of the syntenic blocks, they may involve conserved secondary but not primary structures, they may involve GC-poorness or other gross sequence characteristics, or they may even be determined by unknown epigenetic considerations. There is no evidence, how-

FIG. 1.7. Effect of meiotic non-alignment of regions surrounding breakpoints in heterokaryotypes.

FIG. 1.8. Hypothetical human chromosome with breakpoint region (space) containing three types of small fragment. Shading of syntenic blocks B1-B5 and fragments keyed to aligned portions of mouse chromosomes. a = archipelago, c = compatriot, f = foreigner.

ever, for any of these nor, as we argued in Section 1.8, for the contention that the breakpoint regions contain multiple breakpoints.

This notion of "fragile regions" or *a priori* proclivity for breakage as interpretation of the evidence is rejected in [52], where a combination of the following three factors is suggested to explain the limited amounts of similarity in the neighborhood of breakpoints.

- The algorithms [32, 25] that reconstruct the syntenic blocks bridge gaps as long as appropriate similarity exists at both ends of the gap. A rearrangement event with one breakpoint within a gap destroys the match between the homologies at each end. This effect would show up only *after* the breakage event.
- For a rearrangement to become established in a population, the process of meiosis has to tolerate the coexistence of different rearrangement haplotypes through many generations of heterokaryotypy. The mechanism of this tolerance may be seen in quadrivalent meiotic figures (in the case of reciprocal translocations), as depicted in Fig. 1.7, and in looped figures (in the case of inversions). Though there does not appear to be any direct molecular cytogenetic evidence, it is hypothesized that there is an increase of aberrant processes, such as recombination errors, deletion, duplication or retroposition in the necessarily unapposed chromosomal regions in the immediate vicinity of breakpoints in such figures, during the heterokaryotypy period before the rearrangement becomes fixed. Note that this process is operative only *after* the rearrangement event, and is consistent with breakpoints occurring randomly over virtually the entire genome and not confined to a small number of regions.
- To the extent that breakage occurs disproportionately in intergenic regions, these tend to undergo more rapid sequence evolution than regions containing exons and introns. This is not the same as the fragile regions hypothesis: the number of intergenic regions is almost two order of magnitudes greater than the supposed number of fragile regions, and the intergenic regions cover most of the genome! Rather, accelerated intergenic sequence evolution would compound, *after* breakage, the effects of the preceding two paragraphs. Note that in general the breakpoint regions contain many genes [25] and, depending on the criteria used to delimit the regions, parts of genes.

Fig. 1.9. Top: Length distribution for fragment categories. Bottom: Number of chromosomes for which the null hypotheses of identical size fragments is rejected or accepted.

### 1.9.1 *Fragments*

The largely unaligned region between two syntenic blocks on a human chromosome usually contains a number of smaller regions (or fragments) that are aligned with regions on various mouse chromosomes . As depicted in Fig. 1.8, these fragments fall into three categories:

- If a fragment is aligned with a region on the same mouse chromosome as one of the two adjacent syntenic blocks on the left or right of the space, it is said to be in the *archipelago*.
- Fragments aligned with regions on other mouse autosomes sharing syntenic blocks with the same human chromosome are called *compatriots*. (Recall that the X chromosome generally does not participate in inter-chromosomal exchanges.)
- Fragments aligned with regions on mouse chromosomes, including X, sharing no syntenic blocks with the same human chromosome, are *foreigners*.

Trinh *et al.* [52] undertook a statistical assessment of the three types in the hopes of revealing the formative processes of the breakpoint regions.

Based on the construction in the UCSC Genome Browser comparison of the mouse and human genomes, and using a 100Kb threshold for the minimum size of a syntenic block, they extracted 320 inter-block spaces on the human genome for analysis, excluding pericentromeric spaces subject to repetitive segmental duplication and/or transposition [3]. Their median length was 120 Kb, about the same as the shortest blocks. For about half the spaces, the two adjacent syntenic blocks were from different mouse chromosomes. The spaces contained 12930 smaller aligned fragments as identified by the browser, and these were labeled as archipelago (N=4139), compatriot (N=2706) or foreigner (N=6085) .

The archipelago fragments are considerably longer than the compatriot and foreigner fragments as can be seen from the distributions of fragment length in Fig. 1.9. The median length of the archipelago fragments is twice as large as either of the other two in most chromosomes . The disparity prevails throughout the genome as can be seen in the plot of the number of chromosomes for which a one-tailed Kolmogorov-Smirnov test rejects the null hypothesis that the different types of fragment have the same distribution of lengths.

Figure 1.9 also shows that the compatriot fragments are systematically longer than the foreigner ones, though the difference is less marked than that between either of these categories and the archipelago. Fourteen of the 18 chromosomes for which there are sufficient data have longer mean fragment size for compatriots than foreigners, and eight of these are significantly so at the 5 % level.

Trinh *et al.* [52] also showed that :

- Archipelago fragments tended to be much more frequent in an inter-block space than compatriot fragments, in proportion to the number of different mouse chromosomes in which the two types could originate. In turn, compatriots tended to be much more frequent than foreigners, again relative to the number of different mouse chromosomes in which the two types could originate.
- The proportion of the inter-block space covered by archipelago fragments is much greater than that of the compatriots, which in turn is greater than that of the foreigners.
- The archipelago fragments in spaces defined by blocks from two different mouse chromosomes, though somewhat interspersed, tended to segregate towards the corresponding block.
- The archipelago fragments tend to correspond to regions in the mouse chromosome close to the homolog of the adjacent block. The compatriot fragments tend to correspond to regions in the mouse chromosome close to the homolog of one of the blocks on the same human chromosome.

These observations about the different kinds of fragments suggest that they derive from at least three separate types of process. All or most of the foreigners but a smaller proportion of the compatriots and a much smaller proportion of the archipelago, probably come from some common processes such as retroposition of mRNA, or small jumping translocation or transposition events originating randomly across the genome and correlating roughly with chromosome size. Compatriots represent either a greater propensity for retroposition to the same chromosome originating, due to geometrical considerations (mRNA is more concentrated around the chromosome from which it is transcribed) or, in some lesser proportion, from some intrachromosomal shuffling process, such as inversion or transposition. Finally, the larger archipelago blocks seem to be hived off the large syntenic blocks on either side, and are the results, in some proportion, of two types of process. One is the residual similarity exceeding whatever thresholds are required by the alignment algorithms. These islands of similarity "peeking through" the noise may be either a natural consequence of the variable degree of similarity across all regions of the genome, or indicate the sporadic way the algorithms fail near breakpoints, or both. Second, these fragments may be chunks of the two surrounding syntenic blocks that have been thrown from near the ends of these blocks into the space by the same processes of local rearrangement that affect the interior of the blocks. That the archipelago fragments corresponding to two syntenic blocks are partially interspersed is evidence that such rearrangement continues to occur post-rearrangement, and that they are not solely the residues of decaying measures of similarity .

One process that is not invoked in explaining these statistics is the repeated use of the same breakpoints by several large-scale genomic rearrangements. The archipelago fragments only attest to local rearrangements, and the numerous small compatriot and foreigner fragments, including many fragments of X chromosome origin, do not seem like the residue of repeated large-scale rearrangements.

## 1.10 Conclusions

Genome rearrangement analysis has not scaled up directly to genomic sequences, not because of any computational difficulty, but because this new information is not as neat as the gene order data of organelles. Whatever the loss of evolutionary signal from divergent organellar or prokaryotic genomes, this problem is compounded in nuclear genomes by the difficulties of gene finding and ortholog identification at the gene level, and the lack of congruence of genomic sequence rearrangement and gene order rearrangement. Whereas the former involves movement of material that may not involve any genes, the latter may sometimes operate on gene-containing fragments too short to be picked up by syntenic block construction algorithms .

Genome sequence data has thus proved to be more of a problem for comparative genomics than a solution to old problems.

### Acknowledgements

Research supported by grants from the Natural Sciences and Engineering Research Council (NSERC). The author holds the Canada Research Chair in Mathematical Genomics and is a Fellow in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

### References

[1] Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R. and Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, **290,** 457–65.

[2] Bader, D. A., Moret, B. M., and Yan, M. (2001). A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of Computational Biology*, **8,** 483–91.

[3] Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D,, Myers, E.W,, Li, P.W. and Eichler, E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297,** 1003-7.

[4] Bed'hom, B. (2000). Evolution of karyotype organization in *Accipitridae*: A translocation model. In Sankoff, D. and Nadeau, J. H. (eds) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families.* Dordrecht, NL, Kluwer, 347–56.

[5] Bennetzen, J.L. and Ramakrishna, W. (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Molecular Biology*, **48,** 821–7.

[6] Bergeron, A. (2001). A very elementary presentation of the Hannenhalli-Pevzner theory. In Amihood A. and Landau, G. M. (eds), *Proceedings of the*

*12th Annual Symposium on Combinatorial Pattern Matching (CPM 2001).* Lecture Notes in Computer Science **2089**, Springer-Verlag, Berlin, 106–17.

[7] Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A. , Eppig, J. T. and the members of the Mouse Genome Database Group (2003). MGD: The Mouse Genome Database. *Nucleic Acids Research*, **31,** 193–5.

[8] Bourque, G., Pevzner, P. A. and Tesler, G. (2004). Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes. *Genome Research*, **14**, 507–16.

[9] Caspersson, T., Zech, L., Johansson, C. and Modest, E. J. (1970). Identification of human chromosomes by DNA-binding fluorescent agents. *Chromosoma*, **30**, 215–27.

[10] De, A., Ferguson, M., Sindi, S. and Durrett, R. (2001). The equilibrium distribution for a generalized Sankoff-Ferretti model accurately predicts chromosome size distributions in a wide variety of species. *Journal of Applied Probability*, **38,** 324–34.

[11] Eichler, E. and Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science*, **301,** 793–7.

[12] Feller, W. (1965). *Introduction to Probability Theory and its Applications, Volume 1* (2nd edn). John Wiley and Son, New York.

[13] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.F., Dougherty, B. A., Merrick, J. M. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd. Science*, **269,** 496–512.

[14] Gibbs, R. A., Weinstock, G. M., Metzker, M.L. *et al.* ( 2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428,** 493–521.

[15] Goffeau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Louis, E., Mewes, H., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. (1996). Life with 6000 genes. *Science*, **274,** 546,563–567.

[16] Goss, S. J. and Harris, H. (1975). New method for mapping genes in human chromosomes. *Nature*, **255,** 680.

[17] Hannenhalli, S. (1996). Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Applied Mathematics*, **71,** 137–51.

[18] Hannenhalli, S. and Pevzner, P. A. (1995). Transforming men into mice (polynomial algorithm for genomic distance problem). *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science*. 581–92.

[19] 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.

[20] Housworth, E. A. and Postlethwait, J. (2002).Measures of synteny conservation between species pairs. *Genetics*, **162,** 441–8.

[21] IHGC (International Human Genome Sequencing Consortium). (2001). Initial sequencing and analysis of the human genome. *Nature*, **409,** 860–921.

[22] Kececioglu. J. and Sankoff, D. (1993). Exact and approximation algorithms for the inversion distance between two chromosomes. In Apostolico, A., Crochemore, M., Galil, Z. and Manber, U. (eds), *Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching (CPM 1993).* Lecture Notes in Computer Science **684**, Springer Verlag, Berlin 87–105.

[23] Kececioglu. J. and Sankoff, D. (1994). Efficient bounds for oriented chromosome inversion distance. In Crochemore, M. and Gusfield, D. (eds), *Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching (CPM 1994).* Lecture Notes in Computer Science **807**, Springer-Verlag, Berlin, 307–25.

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

[25] Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003). Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences, USA*, **100,** 11484–9.

[26] Lefebvre, J.-F., El-Mabrouk, N., Tillier, E. and Sankoff, D. (2003). Detection and validation of single-gene inversions. *Bioinformatics*, **19,** Suppl. 1, i190–i196.

[27] Marchand, I. (1997). *Généralisations du modèle de Nadeau et Taylor sur les segments chromosomiques conservés.* MSc thesis, Département de mathématiques et de statistique, Université de Montréal.

[28] Nadeau, J. H. and Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences, USA*,**81,** 814-8.

[29] Ozery-Flato, M. and Shamir, R. (2003). Two notes on genome rearrangements. *Journal of Bioinformatics and Computational Biology*, **1,** 71–94.

[30] Painter, T. S. (1933). A new method for the study of chromosome rearrangements and the plotting of chromosome maps. *Science*, **78,** 585-6.

[31] Parent, M.-N. (1997). *Estimation du nombre de segments vides dans le modèle de Nadeau et Taylor sur les segments chromosomiques conservés.* MSc thesis, Département de mathématiques et de statistique, Université de Montréal.

[32] Pevzner, P. A. and Tesler, G. (2003). Genome rearrangements in mammalian genomes: Lessons from human and mouse genomic sequences. *Genome Research*, **13,** 37-45

[33] Pevzner, P. A. and Tesler, G. (2003). Transforming men into mice: The Nadeau-Taylor chromosomal breakage model revisited. *Proceedings of RECOMB 03, Seventh International Conference on Computational Molecular Biology.* ACM Press, 247–56.

[34] Pevzner, P. A. and Tesler, G. (2003). Human and mouse genomic sequences

reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences, USA*, **100,** 7672-7

[35] Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C A., Hutchison, C. A., Slocombe, P. M. and Smith, M. (1977) Nucleotide sequence of bacteriophage ΦX174 DNA. *Nature*, **265,** 687–95

[36] Sankoff, D. (2003). Rearrangements and chromosomal evolution. *Current Opinion in Genetics and Development* , **13,** 583–7.

[37] Sankoff, D. and El-Mabrouk, N. (2002). Genome rearrangement. In Jiang, T., Smith, T., Xu, Y. and Zhang, M. (eds) *Current Topics in Computational Biology.* Cambridge, MA: MIT Press; 135–55.

[38] Sankoff, D. and Ferretti, V. (1996). Karotype distributions in a stochastic model of reciprocal translocation. *Genome Research*, **6,** 1-9.

[39] Sankoff, D. and Goldstein, M. (1988). Probabilistic models for genome shuffling. *Bulletin of Mathematical Biology*, **51,** 117-124.

[40] Sankoff, D., Parent, M.-N. and Bryant, D. (2000). Accuracy and robustness of analyses based on numbers of genes in observed segments. In Sankoff, D. and Nadeau, J. H. (eds) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families.* Dordrecht, NL, Kluwer, 299-306.

[41] Sankoff, D., Parent, M.-N., Marchand, I. and Ferretti, V. (1997). On the Nadeau-Taylor theory of conserved chromosome segments. In Apostolico, A. and Hein, J. (eds) *Combinatorial Pattern Matching. Eighth Annual Symposium.* Lecture Notes in Computer Science **1264,** Springer Verlag, 262–74.

[42] Sankoff, D. and Nadeau, J.H. (1996). Conserved synteny as a measure of genomic distance. *Discrete Applied Mathematics*, **71,** 247–57.

[43] Sankoff, D. and Nadeau, J.H. (2003). Chromosome rearrangements in evolution: From gene order to genome sequence and back. *Proceedings of the National Academy of Sciences, USA*, **100,** 11188–9.

[44] Sankoff, D. and Trinh, P. (2004). Chromosomal breakpoint re-use in the inference of genome sequence rearrangement. *Proceedings of RECOMB 04, Eighth International Conference on Computational Molecular Biology.* New York: ACM Press, 30–5.

[45] Schoen, D. J. (2000a). Comparative genomics, marker density and statistical analysis of chromosome rearrangements. *Genetics*, **154,** 943–52.

[46] Schoen, D. J. (2000b). Marker density and estimates of chromosome rearrangement. In Sankoff, D. and Nadeau, J. H. (eds) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families.* Dordrecht, NL, Kluwer, 307–319.

[47] Seldin, M. F. (1999). The Davis human/mouse homology map. `http://www.ncbi.nlm.nih.gov/Homology/`".

[48] Sturtevant, A. H. (1965). *A history of genetics.* New York: Harper and Row.

[49] Tannier, E. and Sagot, M. F. (2004). Sorting by reversals in subquadratic time. *INRIA Research Report*, **RR-5097**.

[50] Tesler, G. (2002). GRIMM: Genome rearrangements web server. *Bioinformatics*, **18,** 492-3.

[51] Thomas, J. W, and Green, E. D. (2003). Comparative sequence analysis of a single-gene conserved segment in mouse and human. *Mammalian Genome*, **14,** 673–8.

[52] Trinh, P., McLysaght, A. and Sankoff, D. (2004). Genomic features in the breakpoint regions between syntenic blocks. *Bioinformatics* (in press).

[53] Venter, J. C., Adams, M. D., Myers, E. W. *et al.* (2001). The sequence of the human genome. *Science*, **291,** 1304–51.

[54] Waddington, D., Springbett, A. J. and Burt, D. W. (2000). A chromosome-based model for estimating the number of conserved segments between pairs of species from comparative genetic maps. *Genetics*, **154,** 323–32.

[55] Waddington, D. (2000). Estimating the number of conserved segments between species using a chromosome-based model. In Sankoff, D. and Nadeau, J. H. (eds) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families.* Dordrecht, NL, Kluwer, 321–332.

[56] Waterston, R. *et al.* (2002) Initial sequencing and analysis of the mouse genome. *Nature*, **420,** 520–62.

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

[58] Zhu, D. M. and Ma, S. H. (2002). Improved polynomial-time algorithm for computing translocation distance between genomes. *The Chinese Journal of Computers*, **25,** 189–196.

# CONTRIBUTORS

**David Sankoff**
Department of Mathematics and Statistics
University of Ottawa
585 King Edward Avenue
Ottawa, Canada K1N 6N5