On the Nadeau-Taylor Theory of Conserved Chromosome Segments

David Sankoff * , Marie-Noelle Parent, Isabelle Marchand, Vincent Ferretti

Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale
Centre-ville, Montréal, Québec, H3C 3J7

Abstract. The quantification of comparative genomics dates from 1984 with the work of Nadeau and Taylor on estimating interchromosomal exchange rates based on the rearrangement of chromosomal segments in human versus mouse genomes. We reformulate their analysis in terms of a probabilistic model based on spatial homogeneity and independence of breakpoints and gene distribution. We study the marginal distribution of the number of genes per segment and the distribution of the number of non- empty segments as a function of the number of genes and segments. We propose a rapid algorithm for identifying a given number of conserved segments in noisy comparative map data. Finally, we propose a model which incorporates a degree of inhomogeneity in the distribution of genes and/or breakpoints. Comparative maps of human and mouse genomes serve as test data throughout.

1 Introduction

During evolution, inter- and intrachromosomal exchanges such as reciprocal translocation, transposition and inversion disrupt the order of genes along the chromosome (Figure 1).

In comparing two divergent genomes, a contiguous stretch of chromosome in which the number and order of homologous genes is the same in both species, i.e. has not been interrupted by any of the rearrangement processes that have occurred in either lineage, is called a *conserved segment.* The number of conserved segments increases as they are disrupted by new events, so that they tend to become shorter over time. The number of chromosomal segments conserved during the divergence of two species can be used to measure their genomic distance.

An early and influential contribution to the quantitative methodology of comparative genomics was made by Nadeau and Taylor in 1984 [3], focusing on interchromosomal exchange as the major mechanism in the rearrangement of mammalian

* E-mail: sankoff@ere.umontreal.ca. Research supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. DS is a fellow of the Canadian Institute for Advanced Research.

genomes. Our formulation of the Nadeau-Taylor model of genomic divergence 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).

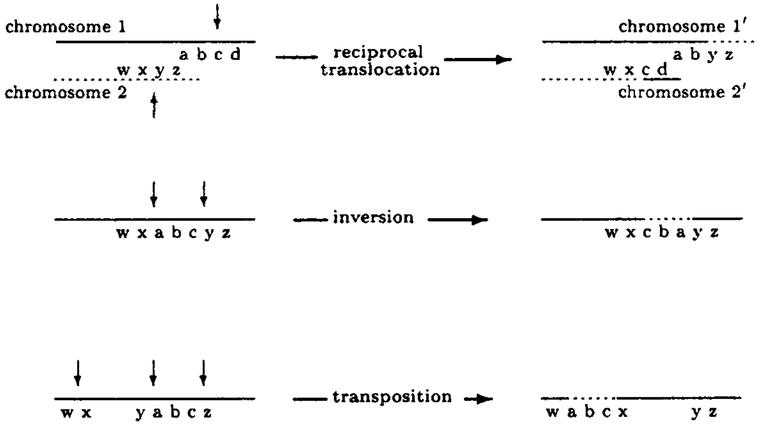

Fig. 1. Schematic view of genome rearrangement processes. Letters represent positions of genes. Vertical arrows at left indicate breakpoints introduced into original genome. Reciprocal translocation (top) exchanges end segments of two chromosomes. Inversion (center) reverses the order of genes between 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 conserved (possibly inverted) within segments.

2 The Marginal Probability of r-gene Segments

In trying to count the number of conserved segments for the quantification of evolution, we must deal with underestimation due to conserved segments in which genes have not yet been identified in one or both species. There are two in Figure 2: 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 [3] in 1984 could only treat 13 segments out of the 130 or so now known to exist (see Section 4.3 below).

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

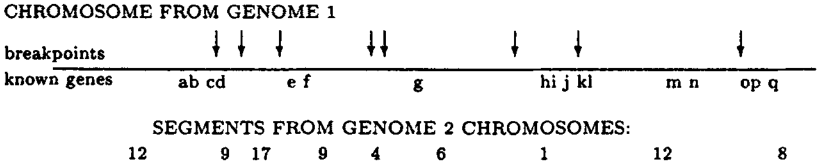

Fig. 2. Fictitious example of conserved segments indicated on a chromosome from Genome 1, with each segment labeled between its endpoints (adjacent arrows) as to which chromosome it is found on in Genome 2. Homologous genes that have been discovered to date are indicated with letters.

concatenation boundaries separating two successive chromosomes [5].) The marginal probability that a segment contain r genes is given by the following theorem [7].

Theorem 1. *Consider a linear interval of length* 1, *with* $n > 0$ *uniformly distributed breakpoints that partition the interval into* $n + 1$ *segments. Suppose there are* m *genes also distributed uniformly on the interval between* 0 *and* 1, *and independently*

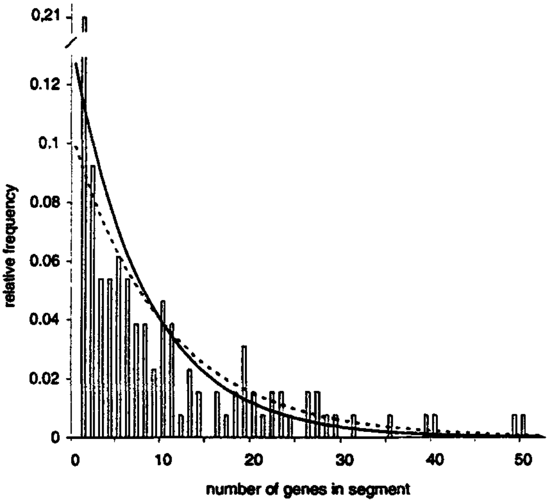

Fig. 3. Comparison of relative frequencies $n_r / \sum_{r>0} n_r$ of segments containing r genes with predictions of Nadeau-Taylor model. Value of n in formula for Q is taken to be 141 (dotted curve) or 181 (uninterrupted curve), as estimated by the maximum likelihood method of Section 3.2 or the Kolmogorov-Smirnov method of Section 5, respectively. Both curves show values for $Q(0)$, though zero is not in the range of the conditional distribution, to permit a comparison of the estimated number $K(m,n)Q(0)$ of unobserved (empty) segments with the predictions $K(m,n)Q(r)$ for positive r, where $K(m,n) = (n+1)m/(n+m)$. Three data points are off-scale, with $r = 54, 65$ and 83 and the vertical axis is interrupted to allow an expanded scale, facilitating more detailed visualization of $f(r)$ and $Q(r)$, $r > 1$.

of the breakpoints. For an arbitrary segment, the probability that it contains r genes, $0 \leq r \leq m$, is then

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

We can only partially compare the theoretical distribution $\Pi(r)$ 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 at least compare the relative frequencies $f(r) = \frac{n_r}{\sum_{r>0} n_r}$ with the conditional probabilities $Q(r) = \Pi(r \,|r > 0)$. This is seen in Figure 3, where the largest discrepancy is the comparison between $f(1)$ and $Q(1)$. We will discuss this discrepancy, how to interpret it, and the consequences of ignoring it, in Section 5.

3 The Inference Problem

It might seem that the number of segments n_r observed to contain r genes, for $r = 1, 2...$, would be useful data for inference about the Nadeau-Taylor model, in particular about n, the unknown number of breakpoints. Though we will see in Section 5 that these data are indeed useful for generalizing the model, they are not necessary for the basic distribution given in Theorem 1.

3.1 The Sufficiency of the Number of Observed Segments

It is remarkable that to estimate n from m and the n_r, for $r = 1, 2...$, only the number of non-empty segments $a = \sum_{r>0} n_r$ is important [4].

Theorem 2. *The variable a is a sufficient statistic for the estimation of n.*

3.2 Estimating n from a

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:

Theorem 3.

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

After observing m and a it is an easy matter to find the value of n which maximizes P, i.e. the maximum likelihood estimate.

Another approach, for extremely large values of the parameters, is to use the mean and variance of $P(a,m,n)$:

$$E(a,m,n) = \frac{(n+1)m}{(n+m)}, \quad \mathrm{Var}(a,m,n) = \frac{(n+1)nm(m-1)}{(n+m-1)(n+m)^2}$$

A gaussian approximation allows accurate calculation for high values of m and n. To do maximum likelihood estimation, the log of the gaussian density with $\mu = E(a,m,n), \sigma^2 = \mathrm{Var}(a,m,n)$ is differentiated with respect to n and set equal to zero. The solution is the only positive root of the following degree 6 polynomial:

$$\begin{gathered}
m^3 - 2am^3 + a^2m^3 - 2m^4 + 4am^4 - 2a^2m^4 + m^5 - 2am^5 + a^2m^5 - 2am^2n \\
+a^2m^2n + m^3n + 2am^3n - a^2m^3n - 3m^4n + 4am^4n - 2a^2m^4n + 2m^5n \\
-4am^5n + 2a^2m^5n + mn^2 - a^2mn^2 - 4m^2n^2 - 4am^2n^2 + 7a^2m^2n^2 \\
+4m^3n^2 + 10am^3n^2 - 13a^2m^3n^2 - 2m^4n^2 - 4am^4n^2 + 7a^2m^4n^2 + m^5n^2 \\
-2am^5n^2 - a^2n^3 - mn^3 - 2amn^3 + 9a^2mn^3 - m^2n^3 + 2am^2n^3 - 17a^2m^2n^3 \\
+3m^3n^3 + 6am^3n^3 + 8a^2m^3n^3 - 2m^4n^3 - 4am^4n^3 + 3a^2n^4 - 3mn^4 - 2amn^4 \\
-8a^2mn^4 + 4m^2n^4 + 8am^2n^4 + 2a^2m^2n^4 - 3m^3n^4 - m^4n^4 - a^2n^5 - mn^5 \\
+2amn^5 - 2a^2mn^5 + 4am^2n^5 - 2m^3n^5 - a^2n^6 + 2amn^6 - m^2n^6
\end{gathered}$$

The approximation is not necessary for current data levels, but both methods give, for $m = 1423, a = 130$, valid values for the man-mouse comparison in the summer of 1996 (cf. Section 4.3), an estimate of 141 for n, suggesting that less than 10% of the segments have not yet been observed.

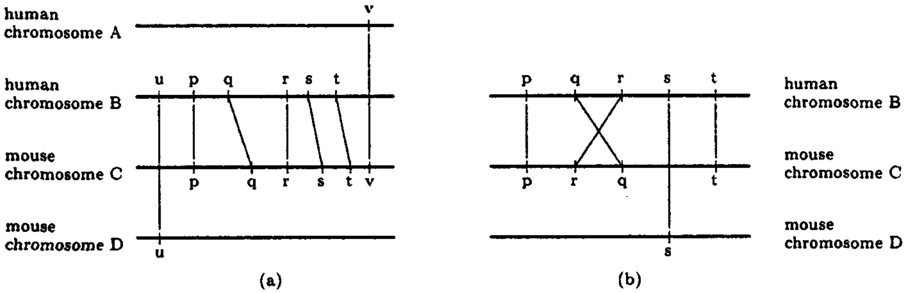

Fig. 4. (a). Schematic example of conserved segment in a human chromosome B and a mouse chromosome C. Genes u and v have homologues elsewhere in the mouse and human genomes, respectively, and thus limit the leftward and rightward extension of the segment. (b). Experimental mistake in the chromosomal 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.

4 The Identification of Conserved Segments

Conserved segments were defined in the Introduction to be regions of chromosomes in two related species in which both gene content and gene order are parallel (Figure 4(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. In addition, in the comparison of multichromosomal 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 and transpositions of small regions of chromosomes (Figure 4(b)).

Our hypothesis is that we can recover 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.

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 [6].

4.1 The Objective Function

Let $c \leq c_1c_2$ be the total number of conserved synteny sets, where c_1 and c_2 are the number of chromosomes in species 1 and species 2, respectively. c is also the smallest number of segments that can be produced by any analysis, grouping all genes belonging to a conserved synteny, no matter how dispersed they are along the chromosome, into a single conserved segment, not allowing for a single conserved synteny to be the result of two or more translocation events. At the other extreme, if we assume that each gene defines a different conserved segment and that genes are adjacent in two genomes only by coincidence, we obtain m segments, the total number of homologous genes identified in the two genomes. All solutions lie somewhere between these two extremes. For an appropriate choice of weighting parameters, α, β, γ, and for all a, $c \leq a \leq m$, we wish to find the subgroupings of conserved syntenic genes into a segments so as to minimize

$$D = \sum_{i=1}^{a} D_i,$$

where D_i is a weighted measure of the compactness, density and integrity of segment i. Formally,

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

where $x \epsilon i(j)$ 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.

4.2 The Algorithm

Direct minimization of $D = \sum D_i$ is generally not feasible, because what is included in segment i impacts the quality of other segments and vice-versa. Instead we propose a rapid stepwise upper-bound algorithm and show sufficient conditions for it to calculate D exactly. An advantage of this method is that it constructs solutions for all a in one pass.

Our procedure starts with the extreme solution where $a = m$, then combines step by step genes syntenic in both genomes into conserved segments.

Basic to the algorithm is the notion of a rooted binary branching tree T_i with the leaves, or terminal nodes, associated with the m_i genes in conserved synteny i. This is illustrated in Figure 5.

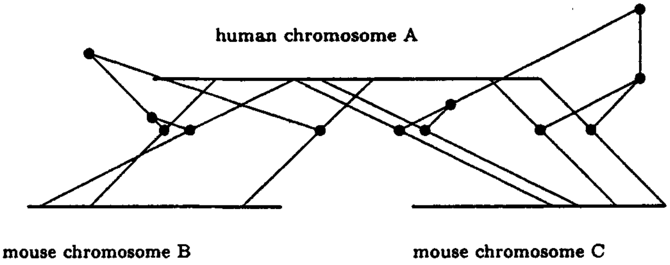

Fig. 5. Two rooted binary trees each representing successive solutions to the problem of identifying conserved segments within two conserved syntenies. Thin lines connect homologous genes in the two genomes. Note that the conserved syntenies overlap on the human chromosome and that the number of segments from the synteny on the right intervening between genes on the left changes as the trees are constructed from bottom up.

Each nonterminal node v denotes the formation of a segment from two smaller segments v_1, v_2 of distance $d(v_1, v_2) = D(v)$ apart. Note that d is a not a metric, and it is defined only for two segments v_1 and v_2 containing genes in the same synteny sets.

After precalculating all the distances d among the terminal nodes (segments consisting of single genes), we apply the following:

Algorithm conseg

Let m_k be the number of genes in the k-th conserved synteny. Set $a = m = \sum m_k$, the total number of homologous pairs of genes, and let seg to be the set of all these genes. For all k, set $S_k = -\beta m_k$. Initial construction step for T_k: Identify the terminal nodes with the m_k genes in the conserved synteny.

```
while there remains a conserved synteny with $\geq 2$ segments in $seg$,
    Find the two segments $v_1$ and $v_2$ that minimize $d(v_1, v_2)$.
    Combine $v_1, v_2$ to form $v$. Add $v$ to $seg$. Remove $v_1$ and $v_2$.
    if $v$ contains genes in the $k$-th synteny
        Update $T_k$ to indicate branching of $v$ to $v_1, v_2$
        Set $S_k = S_k + D(v) - D(v_1) - D(v_2)$.
    endif
    Set $a = a - 1$, and output configuration of the $a$ segments in $seg$.
    Recalculate all distances $d$ given the decrease in number of segments in $seg$.
    Set $D^* = \sum S_k$.
endwhile
```

A relatively literal implementation of this algorithm has worst-case performance in time cubic in m, the number of genes. Within the **while** loop, the distance update can take quadratic time (without any sophisticated data structures), though with small proportionality factor, and the loop itself must be executed $m - 1$ times. The search step is carried out at the same time as the update step. Improvement, possibly to quadratic performance, could be achieved by tracking which segments intervene in which other segments. With available data, however, there is little need for improved code.

The clustering procedure may seem a roundabout way of approaching the objective function, but to the extent that segments are disjoint, or overlap to a very limited extent, the following theorem [6] becomes pertinent:

Theorem 4. *For any a, the upper bound D^* achieved by the algorithm is equal to the objective D if no segment intervenes in any other segment by virtue of more than one gene.*

4.3 How Many Segments?

What value of a is the most reasonable? To answer this, we compare the number U_i of different human chromosomes represented among the a_i segments on a single mouse chromosome i, with the number u_i expected under a random hypothesis:

$$u_i = 22[1 - (\frac{21}{22})^{a_i}].$$

We chose the parameter values and a so that

$$\sum_{i=1}^{19} u_i = \sum_{i=1}^{19} U_i.$$

In our data set, these values are $a = 130$, $\alpha = 30$, $\gamma = 1$ and $\beta = 0.3$. There are 113 conserved syntenies in the data. Since we infer 130 segments, this means that about one conserved synteny per chromosome consists of more than one conserved segment, or that almost all the observed fragmentation of conserved syntenies is due to intrachromosomal movement and not interchromosomal events.

5 Gene Clumping and Non-uniform Densities

In Section 3.2, we used the value of $a = 130$ satisfying the criterion of Section 4.3 and 130 segments produced by the identification procedures in Section 4.2 as data for the maximum likelihood estimation of the total number, observed and unobserved, of segments. This was calculated making use of the exact values of (or, equivalently, the gaussian approximation to) $P(a, m, n)$, a valid procedure insofar as the basic Nadeau-Taylor model represents reality, with uniformly distributed breakpoints and uniformly and independently distributed genes. One check on this is the comparison in Figure 3 of the distribution predicted by the model $Q(r) = \Pi(r \mid r > 0)$ (dotted curve in the figure) with the $f(r), r = 1, \ldots,$ the relative frequency of segments containing r genes, $r = 1, \ldots.$

Based on data for 1423 genes and an analysis giving $a = 130$ segments, we find two major discrepancies. First, $f(1)$ is far greater than $Q(1)$, and second, $f(r)$ is systematically less than $Q(r)$ for r in the range [3,18]. To the extent the basic Nadeau-Taylor model needs refinement, we must rely less on Theorem 2 and maximum likelihood estimation based on it. Instead we use in this section a method which is most sensitive to a systematic discrepancy between $f(r)$ and $Q(r)$ over a range of values of r, namely a Kolmogorov-Smirnov approach. To estimate n, we simply choose the value which minimizes $\sup_r |F(r) - G(r)|$, where F and G are the cumulative distributions of f and Q, respectively. As is reflected in $Q(0)$ particularly and in the first few other inflated values of $Q(r)$ in Figure 3 (uninterrupted curve), compared to the maximum likelihood estimate of 141, the Kolmogorov-Smirnov-based estimate for n is 181, due to its sensitivity to the large $|F(1) - G(1)|$ discrepancy. (Indeed, $\sup_r |F(r) - G(r)| = |F(1) - G(1)| = f(1) - Q(1) = 0.095$.)

The excess observations accounting for the value of $f(1)$ may include a good proportion of experimental error, as we previously [6] noticed from changes in the data set over time for many of the chromosomal assignments involved. By removing the case $r = 1$ from the analysis (involving 27 of 130 observed segments), and conditioning both f and Q by $r \geq 2$, we obtain a better fit as seen in Figure 6. With the effect of f(1) removed, n is estimated at 129, greatly diminished from the exaggerated value of 181. The statistic $\sup_r |F(r) - G(r)|$ is dramatically reduced from 0.095 to 0.043. The range for which $f(r)$ is systematically less than $Q(r)$ is contracted to [12,18].

We undertook two approaches to modifying our basic model, relaxing the hypotheses of independence of gene distribution and uniformity of gene and breakpoint distributions [2].

Instead of distributing the genes one at a time according to the uniform distribution, we constructed a model where z genes, where z was fixed to be 2,3, or more, were positioned at the same point. (Thus, only $\frac{m}{z}$ points were sampled from the uniform.) This non-independence of gene distribution turned out to have little effect on the general shape of the predicted frequency curve, despite its effect on the first few values of r.

A second type of modified model divided the genes into two fractions and the breakpoints into two fractions and distributed the first fraction of genes among the first fraction of breakpoints and the rest of the genes among the remaining breakpoints.

The inhomogeneities of distribution rectify to some extent the discrepancies between the predictions and the observed results, both when data on $r = 1$ are retained and when they are excluded. For example, when the genes are divided into two equal groups, and the breakpoints are divided unevenly, proportion α in one part of the genome and $1 - \alpha$ in the other, the best fit, as obtained by minimizing $\sup_r |F(r) - G(r)|$ with respect to α is illustrated in Figures. 7 and 8. In the case where $r = 1$ data are included, half the genes are distributed within a portion of the genome containing 20% of the 157 breakpoints and the other half among the other 80%. Note that 157 is a distinct reduction from the 181 needed in the homogeneous model, and the statistic of goodness-of-fit is reduced from 0.095 to 0.079. The fit of the model to the data is improved both for $r = 1$ and in the range [12,18]. In the case where the $r = 1$ segments are excluded, the best fit is with $n = 118$ and the split of the breakpoints is 29% *vs.* 71%. Here the improvement in $\sup_r |F(r) - G(r)|$ is from 0.043 to 0.036 as the fit is improved for $r = 2$ and in the range [12,18].

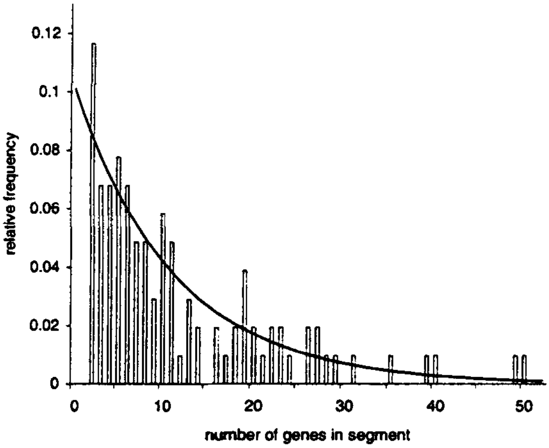

Fig. 6. Comparison of relative frequencies $f(r)/(1 - f(1))$ of segments containing $r \geq 2$ genes with predictions of Nadeau-Taylor model. Value of n in formula for Q (curve shown also conditioned for $r \geq 2$) is taken to be 129, as estimated by minimizing a Kolmogorov-Smirnov-type statistic. Values shown for $Q(0)$ and $Q(1)$, though [0,1] is outside the range of the conditional distribution, to permit a comparison of the estimated number of empty or single-gene segments with the predictions for $r \geq 2$. Three data points are off-scale, with $r = 54, 65$ and 83.

6 Discussion

The analytic insights of Nadeau and Taylor [3] and the prophetic accuracy of their estimation of the number of segments conserved between the mouse and human genomes have become increasingly relevant with the recent massive increases in the available genomic data, whether genetic maps, physical maps or complete sequences. Their work serves as a starting point for a variety of algorithmic, probabilistic, statistical and other applications of mathematical science.

6.1 The Original Approach of Nadeau and Taylor

In the intellectual climate of the early 80's, Nadeau and Taylor used $r \geq 2$ as a criterion for the existence of a conserved segment, in contradistinction to a model of random gene scrambling throughout the genome. Their analysis was based on the estimation of average segment length, in centimorgans, prior to the estimation of of the number of segments. This work involved a good number of mathematical assumptions and approximations that, while justifiable, turn out to be unnecessary within our formulation of the key assumptions of spatial homogeneity and independence of breakpoint and gene distributions in Sections 2 and 3.

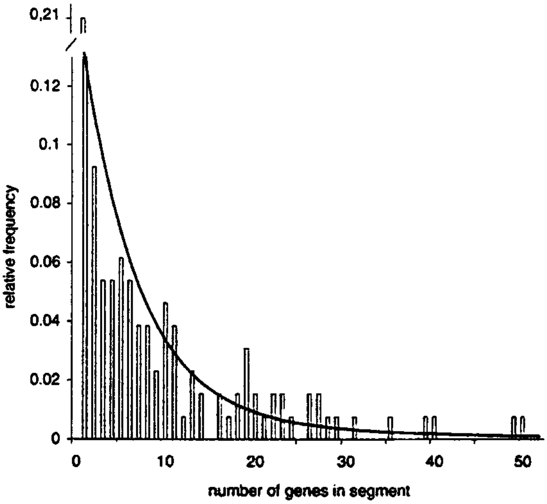

Fig. 7. Comparison of relative frequencies $f(r)$ of segments containing r genes with predictions of inhomogeneous genome model. Values of n and α are taken to be 157 and 0.2, respectively, as estimated by minimizing Kolmogorov-Smirnov-type statistic.

6.2 The Distribution $P(a, m, n)$

When appropriately formulated, the probabilistic model fundamental to the Nadeau-Taylor theory derives from a classical occupancy problem related to statistical mechanics ([1], p. 62). As such, it makes no reference to the linear nature of chromosomes, though considerations of order are central to the identification of segments in Section 4, prior to statistical analysis.

6.3 Why So Few Segments?

The applications of our method in this paper were all based on the estimate of a in Section 4.3. This estimate of 130, contrasting with the 140-185 segments seen elsewhere in the literature may be considered low for reasons definitional, methodological, or biological.

The criterion in Section 4.3 is designed to estimate the number of reciprocal translocations based on the total number of conserved syntenies detected on each chromosome, and is not influenced by how fragmented each of these syntenies may be. This choice follows from our goal specified in Section 4 of recovering the history of translocation and ignoring the effects of intrachromosomal rearrangement. It is not, however, a fundamental aspect of our methodology; we could have chosen a somewhat larger value of a in the hope that the **conseg** algorithm would identify segments created by inversions and intrachromosomal transposition as well as translocation, for example, while excluding multiple counts of single segments due simply to small mapping errors. This new value of a and the corresponding n_r could have equally well served to draw Figures 3 and 6-8, and to do the calculations in Sections 3.2 and 5.

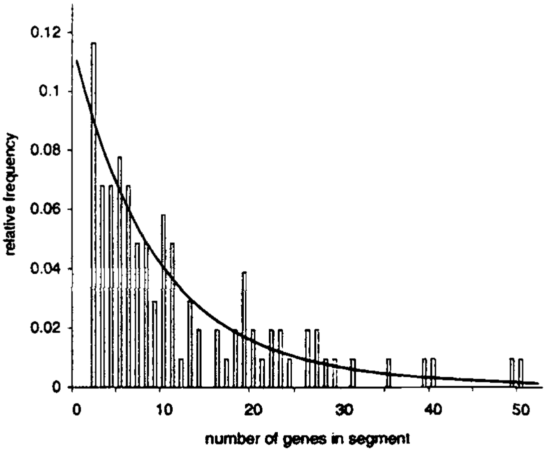

Fig. 8. Comparison of relative frequencies $f(r)/(1 - f(1))$ of segments containing $r \geq 2$ genes with predictions of inhomogeneous genome model. Values of n and α are taken to be 118 and 0.29, respectively, as estimated by minimizing Kolmogorov-Smirnov-type statistic.

Another explanation of the small estimate of a is the rather simple formula used in Section 4.3. A more detailed analysis of how segments are distributed, symmetric with respect to the two organisms, and using likelihood techniques, could have resulted in a larger value of a, though not very much so. This is a direction for future research.

A final type of explanation would depend on the cellular mechanisms, as yet unassessed, resulting in the fixation of a chromosomal aberration such as reciprocal translocation. These explanations might invoke differences in chromosome size or differential tendencies among chromosomes for synteny preservation, fusion, fission and translocation. For the time being these considerations remain purely speculative, but they have the greatest potential for revising and deepening our analysis of conserved segments.

6.4 The Study of Inhomogeneities

In our study of the fit of the distribution Π, or its version conditioned on $r \geq 1$, to the relative frequency f of segment sizes, the greatest discrepancy would seem to be for $r = 1$, which is most likely a reflection of error in the identification of homologous genes or other experimental error in chromosome assignment. Nevertheless, when this source of error is removed, there is clear evidence that allowing inhomogeneity in breakpoint and gene distributions offers a closer fit to the data. A refinement of our model of inhomogeneity, and associated statistical tests, are potential directions for combined empirical and theoretical research.

References

1. W. Feller. *An Introduction to Probability Theory and its Applications, Vol.1. 3d ed.* New York: John Wiley and Son, 1968.
2. I. Marchand. *Généralisations du modèle de Nadeau et Taylor sur les segments chromosomiques conservés.* MSc thesis, Département de mathématiques et de statistique, Université de Montréal. 1997.
3. J.H. Nadeau and B.A. Taylor Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences USA*, 81: 814-818, 1984.
4. M.-N. Parent. *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. 1997.
5. D. Sankoff and V. Ferretti. Karotype distributions in a stochastic model of reciprocal translocation. *Genome Research* 6, 1-9, 1996.
6. D. Sankoff, V. Ferretti and J.H. Nadeau. Conserved segment identification. *RECOMB 97. Proceedings of the First Annual International Conference on Computational Molecular Biology.* New York: ACM Press, 1997, pp. 252-256.
7. D. Sankoff and J.H. Nadeau. Conserved synteny as a measure of genomic distance. *Discrete Applied Mathematics* 71, 247-257, 1996.