ELSEVIER

Discrete Applied Mathematics 71 (1996) 247–257

DISCRETE APPLIED MATHEMATICS

Conserved synteny as a measure of genomic distance

David Sankoff[a,*], Joseph H. Nadeau[b]

[a] *Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Qué., Canada H3C 3J7*

[b] *Department of Human Genetics, Montreal General Hospital, McGill University, 1650 Cedar Avenue, Montreal, Que., Canada H3G 1A4*

Received 29 May 1995; revised 11 May 1996; accepted 30 May 1996

Abstract

The number of chromosomal segments conserved during the evolution of two species can be used to measure their genomic distance. The number of conserved segments containing homologous genes can be estimated by comparing synteny relations within and between the two genomes. There are three sources of underestimation, however. The first stems from conserved segments in which genes are yet to be identified in one or both species. The second results from repeated translocations or transpositions resulting in not just one, but several conserved segments from a chromosome in one species being located on a single chromosome in the other. We characterize the bias due to both effects and propose correct measures of syntenic distance. We also discuss underestimation due to intrachromosomal rearrangements such as inversion.

1. Introduction

During evolution, inter- and intra-chromosomal exchanges such as reciprocal translocation, transposition and inversion disrupt the order of genes along the chromosome, as sketched in Fig. 1.

As depicted in the illustration, however, gene order remains fixed between any two neighboring breakpoints resulting from these rearrangements. These conserved segments tend to become shorter with time as they are disrupted by new events, and by the same token, the number of segments increases with each disruption. The number of chromosomal segments conserved during the divergence of two species can be used to measure their genomic distance, insofar as this number reflects the number of chromosomal rearrangement events that have occurred since the divergence of the lineages leading to the two species [5]. In most genomes, however, few genes have been precisely mapped, though for many genes the chromosomal assignment has been determined. Two or

* Corresponding author. 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. Fellow of the Canadian Institute for Advanced Research.

PII S0166-218X(96)00067-4

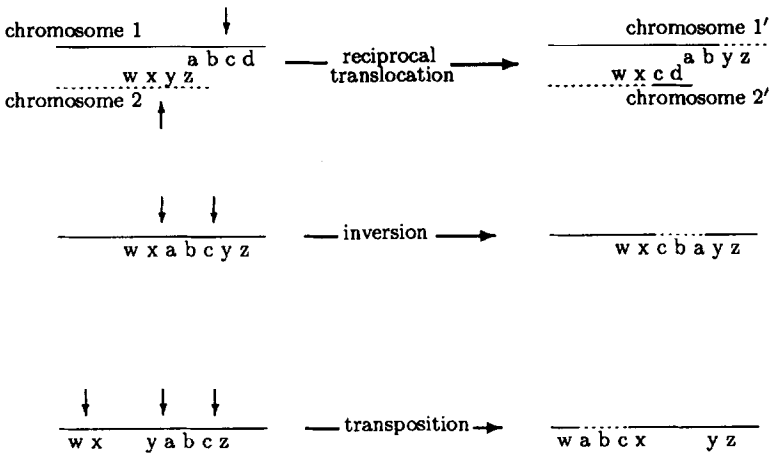

Fig. 1. Schematic view of genome rearrangement processes. Letters represent positions of genes. Vertical arrows at left indicate breakpoints introduced into original genome. Reciprocal translocation (top) exchanges end segments of two chromosomes. Inversion (center) reverses the order of genes between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

more genes are said to be *syntenic* if they are on the same chromosome, as demonstrated by recombination experiments, by cytological methods, or otherwise. While a lack of recombination maps means we cannot directly characterize conserved segments in two genomes, we can make indirect comparisons through the study of conserved synteny. We can hypothesize that, as a first approximation, a subset of genes that are all syntenic in both of the two species, i.e. a conserved synteny, is evidence of a single conserved segment somewhere on the respective chromosomes in the two species. An interesting aspect of using synteny data this way is that they are not perturbed by intrachromosomal rearrangements and thus lead to measures of genomic distance that reflect interchromosomal events only.

Trying to infer the number of conserved segments, and hence genomic distance, through conserved synteny, however, leads to the problem of underestimation. There are at least three sources for this bias. The first stems from conserved segments in which genes have not yet been identified in one or both species. The same problem was solved in [5] for the case where a recombination map is available for at least one of the two species being compared, but this solution is of no help when a recombination map is not available for either of the two species, i.e. when we must rely on synteny data by themselves. This bias 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 syntenies, and hence the conserved segments, are not represented in the comparison, and genomic distance may be severely underestimated.

The second source of bias is multiple interchromosomal exchanges resulting in not just one, but several conserved segments from a chromosome in one species being

located on a single chromosome in the other, so that a conserved synteny set consists of not just one, but several independently conserved segments. Here syntenies disrupted by previous translocations or transpositions are partially *reconstituted* by later exchanges between chromosomes. The resulting underestimate is greatly aggravated as genomic distance increases. A third bias in identifying the number of conserved segments with the number of conserved syntenies is due to inversions and intrachromosomal transpositions. This, however, is of no consequence for measures of genomic distance based on the number of interchromosomal events, as mentioned above and as discussed further in Section 6.

Previous work on conserved synteny measures [1, 7] has relied on normalizing the number of pairs of genes in conserved syntenies by the number of syntenic pairs in the individual genomes, to try to circumvent these biases to some extent. These scores, though based on generally valid statistical principles, have inherent problems due to the non-independent contributions of the pairs in any one synteny set (cf. [3]), and are not based on any model of the biological mechanisms: breakage, translocation and transposition. Nor do they take acount of the experimental background: what is the spatial relation among the genes experimentally identified in a given chromosome?

To model genomic divergence processes in this paper, we assume a spatially homogeneous model of random chromosome breakage, and independence of one interchromosomal exchange event from the preceding ones. It is further assumed that the experimental identification of genes asociated with a given chromosome does not depend on their position on the chromosome, nor on their proximity to each other. We can then derive the probability distribution for the number of identified genes per conserved synteny, in the first instance by assuming that each conserved synteny set is equivalent to a conserved segment, and then by relaxing this assumption, taking into account that a conserved synteny set may originate from several interchromosomal exchanges and may thus contain several conserved segments. In each case we show how to estimate the number of conserved segments, including unobserved ones, i.e. those that contain only genes as yet unidentified. We then apply our method to conserved synteny data for humans and mice in order to compare our results to map-based knowledge of conserved segments.

2. Conserved syntenies and conserved segments

Aside from the human genome and a limited number of "model" genomes where considerable mapping has been done, the primary data in comparative genomics are synteny sets, namely the sets of genes or markers known to be on the same chromosome in a given species. From the synteny sets for two species, by the simple operation of set intersection, we can construct conserved synteny sets, namely sets of markers that are syntenic in both species.

As a first approximation, it is assumed that these conserved synteny sets represent conserved segments, though of course this is not always the case. Repeated inter- and

intra-chromosomal rearrangement eventually results in conserved synteny sets that contain genes from not one, but several conserved segments. Indeed, the number of conserved segments can theoretically increase indefinitely (limited only by the number of genes in the genome), while the number of conserved syntenies cannot exceed the product of the numbers of chromosomes in the two species. As long as the number of rearrangement events is not too large with respect to the number of chromosomes, however, the assumption that one conserved synteny equals one conserved segment should not be too far off the mark, and this will be the basis of the first model we will investigate.

For two species with the same number $c_1 = c_2 = c$ of chromosomes, and the number of conserved synteny sets or conserved segments equal to the number of chromosomes, the obvious inference is that there have been no translocations or interchromosomal transpositions. After one translocation or transposition, the number of conserved segments increases by two, and after t such exchanges, by $2t$, disregarding for the moment the possibility that two or more disruptions have at least one breakpoint in common. If the numbers of chromosomes in the two species are not the same, but differ by ϕ, this may be accounted for by ϕ chromosome fissions (or fusions). Then the observed number s of conserved synteny sets (and hence conserved segments) may be accounted for by

$$s = \min(c_1, c_2) + 2t + \phi,$$

so that

$$T = \frac{s - \min(c_1, c_2) - \phi}{2}$$

is an estimate of the total number of interchromosomal exchange events that have transpired in the divergent lineages leading to the two species.

In Section 3 we discuss how s, and hence t, is underestimated by this approach as long as the number and chromosomal distribution of homologous markers identified to date is not yet large enough to saturate all the conserved segments. In Section 4, we extend our model to take into account that as t increases, the genes from two or more conserved segments may be on a single chromosome in both of the two species, and thus form a single, albeit reconstituted, conserved synteny.

3. Unobserved segments

In this section, we model the genome as a single long unit broken at random into c chromosomes, with each chromosome further broken into a number of segments, each segment representing a length of chromosome within which the gene order has been conserved with reference to some other genome. Our only concern will be to estimate the number of segments that contain no markers, and during the analysis we will not distinguish between breakpoints separating two successive segments and concatenation

boundaries separating two successive chromosomes. It has been demonstrated that a uniform random breakage model accounts well for the distribution of segment lengths [5], and chromosome size distributions are reasonably well modeled by uniformly random breakage of the genome [6].

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

$$P(r) = \frac{nm!\,(n+m-r-1)!}{(n+m)!\,(m-r)!}.$$

Proof. It is a property of the uniform distribution that the segment length x between any two adjacent breakpoints has probability density $f(x) = n(1-x)^{n-1}$, for $0 \leqslant x \leqslant 1$. The same holds for the segment between 0 and the first breakpoint, as well as that between the nth breakpoint and 1.

For a segment of length x, what is the probability that it contains r genes? This is just the binomial probability

$$B(m,x;r) = \binom{m}{r} x^r (1-x)^{m-r},$$

for $0 \leqslant r \leqslant m$. For an arbitrary segment, the probability that it contains r genes is then

$$\begin{aligned} P(r) &= \int_{x=0}^{1} f(x)B(m,x;r)\,dx \\ &= n\binom{m}{r}\int_{x=0}^{1} x^r(1-x)^{n+m-r-1}\,dx \\ &= \frac{nm!\,(n+m-r-1)!}{(n+m)!\,(m-r)!}. \end{aligned}$$

□

The estimation of n is complicated by the non-independence of the number of genes in different segments. When n is small, say $n = 2$ or $n = 3$, the lengths of any two segments will tend to be inversely related in a rather strong way and hence the number of identified genes falling within them will also tend to be inversely related. This relationship will be stronger than if the number of genes in each segment were independently decided by sampling from the distribution P, even if this were conditioned by the total sample size m. For practical purposes, however, we can hypothesize that for large, and even moderate, values of n, the effects of this non-independence are negligible, in probability. We can then make use of the following theorem, which is a direct consequence of the definition of likelihood.

Theorem 2. *Let* $M(v, \{p(r)\}_{r=0}^{m} ; \{v_r\}_{r=0}^{m})$ *be a multinomial law, where* $m > 0$ *and where* v_0 *is unobservable and hence* v *is an unobservable parameter of* p. *Let* $N_1, ..., N_m$ *be observed experimental values for* $v_1, ..., v_m$. *Then the likelihood is given by*

$$L(v) = \frac{(v)!}{N_1!...N_m!(v - \sum_{i=1}^{m} N_i)!} p(0)^{v - \sum_{i=1}^{m} N_i} \prod_{r=1}^{m} p(r)^{N_r}.$$

To estimate n we set $v = n + 1$, and identify $p = P$. We define $N_1, ..., N_V$ to be the number of segments observed to contain $1, ..., V$ genes, respectively, where V is the largest number of genes in any segment. The parameter m is $\sum_{i=1}^{V} iN_i$. The value of n that maximizes L in Theorem 2 can be found by explicitly calculating $\log L$ for a sufficiently large range of values of n, where $n + 1 > \sum_{i=1}^{V} N_i$.

4. Synteny sets that are not conserved segments

In Section 3, we purposely neglected the distinction between conserved synteny sets and conserved segments. It is obvious that as long as breakpoints are truly randomly distributed and hence coincide with negligible frequency, the number of segments continues to increase as twice the number of reciprocal translocation events since each such event breaks one segment in both chromosomes [1] But the number s of conserved synteny sets on all chromosomes can never exceed c_1c_2. It follows that using the lower bound

$$T = \frac{s - \min(c_1, c_2) - \phi}{2},$$

to estimate the number of interchromosomal events becomes increasingly inadequate as evolution progresses. We can, however, correct this underestimate, and more accurately infer the number of conserved segments and hence the evolutionary divergence of the two species.

Suppose that the true number of breakpoints on some chromosome in species 1 is some unknown number n, so that the number of conserved segments is $n + 1$. This is the quantity that we wish to estimate since, summed together with the corresponding figures from the other chromosomes, it is our basic measure of evolutionary divergence. In Theorem 3, given n we will first calculate the probability that a conserved synteny set consists of $h \geqslant 0$ conserved segments, based on the assumption that each segment of a chromosome in species 1 is equally likely to occur on any one of the chromosomes

[1] The situation is somewhat more complicated with interchromosomal transpositions. One, two or three additional segments are created depending on whether a chromosome end or an internal chromosomal segment is transposed, and whether it is transposed to the end of another chromosome or to an internal site. This does not alter our point that the number of breakpoints theoretically increases without limit as rearrangement events accumulate. In any case there is remarkably little direct evidence for the existence of transpositions, for example in cytogenetic assays, and we may safely ignore this process in the ensuing discussion.

in species 2, and vice-versa.[2] This enables us to calculate the probability density f^* of the total proportion of chromosome length implicated in a single conserved synteny set, and hence the distribution $Q(r)$ of the number of genes in the set, given that m genes are located at random along the chromosome. Note that $Q(r)$ pertains to genes per conserved synteny set, but its unknown parameter is n, the number of segments (less 1).

Theorem 3. *Consider any chromosome of species* 1 *as linear interval of length* 1, *with* $n > 0$ *uniformly distributed breakpoints that partition the interval into* $n+1$ *segments. Each segment is randomly and independently identified as also occurring on exactly one of the* c_2 *chromsomes of species* 2, *and the union of those segments that occur on a single chromosome in species* 2 *constitutes a conserved synteny set. Suppose there are m genes also distributed uniformly on the interval between* 0 *and* 1, *and independently of the breakpoints. For each conserved synteny set, the probability that it contains r genes is then* $Q(r)$ *for* $0 < r < m$; $Q(0) + (1 - c_2^{-1})^{n+1}$ *for* $r = 0$, *and* $Q(m) + (c_2^{-1})^{n+1}$ *for* $r = m$, *where*

$$Q(r) = \binom{m}{r}\left(\frac{c_2 - 1}{c_2}\right)^{n+1} \sum_{h=1}^{n} \frac{h(n+1-h)}{(h+r)(n+1)}(c_2 - 1)^{-h}\binom{n+1}{h}^2 / \binom{n+m}{h+r}.$$

Proof. Consider any chromosome A in species 1 and any chromosome B in species 2. What is the probability that h of the $n+1$ conserved segments on A also occur on B?[3] Under the random model, this is just

$$\pi_h = \binom{n+1}{h}\left(\frac{1}{c_2}\right)^h \left(\frac{c_2 - 1}{c_2}\right)^{n+1-h}.$$

For $0 < h < n+1$, the probability density of the combined length of the h conserved segments from B is, without loss of generality, the density of the position of the hth breakpoint, namely,[4]

$$f_h(x) = h\binom{n}{h} x^{h-1}(1-x)^{n-h}.$$

For $h = 0$, we may assume that the "combined length" of segments from B is zero, with probability 1. For $h = n+1$, the combined length of segments from B is 1, with probability 1. Then the probability density of the total proportion x of the length of

[2] There are not yet enough data to systematically verify this assumption. The analysis could equally well have been carried out under the hypothesis that the probability that a segment occurs on a chromosome in species 2 is proportional to the size of that chromosome.

[3] In other words, what is the probability that the conserved synteny between chromosomes A and B involves h conserved segments?

[4] In this notation, the density f in Theorem 1 becomes f_1.

chromosome A corresponding to segments from chromosome B is

$$f^*(x) = \sum_{h=1}^{n} \pi_h f_h(x) + F + G$$
$$= \sum_{h=1}^{n} \frac{h(n+1)}{n+1-h} \binom{n}{h}^2 \frac{(c_2-1)^{n+1-h}}{c_2^{n+1}} x^{h-1}(1-x)^{n-h} + F + G,$$

where F and G are atoms at $x = 0$ and 1 of probability π_0 and π_{n+1}, respectively.

Now the probability that a set of total length x contains r out of m genes identified on chromosome A is, as in Theorem 1, just $B(m,x;r)$, so that the probability that an arbitrary one of the sets contains r genes is

$$Q(r) = \int_{x=0}^{1} f^*(x) B(m,x;r)\,dx$$
$$= \binom{m}{r} \sum_{h=1}^{n} \frac{h(n+1-h)}{n+1} \binom{n+1}{h}^2 \frac{(c_2-1)^{n+1-h}}{c_2^{n+1}}$$
$$\times \int_{x=0}^{1} x^{h-1+r}(1-x)^{n-h+m-r}\,dx,$$

for $0 < r < m$; $Q(0) + \pi_0$ for $r = 0$, and $Q(m) + \pi_{n+1}$ for $r = m$. Then

$$Q(r) = \binom{m}{r} \left(\frac{c_2-1}{c_2}\right)^{n+1}$$
$$\times \sum_{h=1}^{n} \frac{h(n+1-h)}{(h+r)(n+1)} (c_2-1)^{-h} \binom{n+1}{h}^2 \Big/ \binom{n+m}{h+r}. \qquad \square$$

The problem of estimating n from synteny data generated by Q differs subtly from the problem in Section 3 of estimating n from segment data generated by P. Here we can observe not only $N_1,\ldots,N_m$, the number of conserved synteny sets on chromosome A containing $1,\ldots,m$ genes, respectively, but also N_0, the number of possible conserved synteny sets involving chromosome A which are not identified as containing genes, since $N_0 = c_2 - \sum_{r=1}^{m} N_r$. Note, however, that N_0 includes both genuinely unobservable segments, corresponding to the $Q(0)$ term, and a count of the number of chromosomes in species 2 that contain *no* conserved segments in common with chromosome A, corresponding to the π_0 term. Both terms, as well as the $Q(r), r > 0$, depend on n. Thus instead of using Theorem 2, we can estimate n by maximizing[5]

$$L = \frac{c_2!}{N_1!N_2!\ldots N_m!N_0!} (Q(0)+\pi_0)^{N_0} \prod_{r=1}^{m-1} Q(r)^{N_r} (Q(0)+\pi_{n+1})^{N_m},$$

[5] As in Section 3, and for the same reason, L is not strictly speaking the likelihood under our model for generating segments, but this is of consequence only for very small values of n.

Calculation of $\log L$ for a range of values of n, where $n + 1 \geqslant \sum_{i=1}^{m} N_i$, is feasible, despite the more complicated expression for Q in comparison with P in Section 3.

Estimation of the n's can be carried out independently for each of the c_1 chromosomes of one species, and the values then summed. For small c, say $c = 2$ or 3, this may be an unwarranted assumption, since the n for one chromosome clearly is related to the n for another, but treating chromosomes independently is justified when c is of the order of 20, as with humans and most mammals.

5. An empirical test

To test our models and estimators, it would be desirable to apply them to data from many pairs of species where we have detailed knowledge of conserved segments. As stressed in the Introduction, however, detailed maps have not been made for many species. And maps from different species do not share a high proportion of homologous genes. The one comparison where suitable data exist, however, is between mouse and human (cf. [4]). Analyzing the maps of the two genomes by the methods of [5] indicates some 140 conserved segments of which 101 have been identified, leaving around 40 to be discovered [2]. Some of the disjoint segments are on the same chromosome on both species. Most of these instances are undoubtedly due to intrachromosomal disruptions of single conserved segments, but a few likely arise through multiple interchromosomal exchanges. Nevertheless, these figures provide the best available knowledge against which to compare our methods. At the time of writing, the Mouse Genome Database documented 87 conserved syntenies containing at least one pair of homologous genes assigned to specific chromosomes in both mouse and humans. [6]

Application of the estimator in Section 3 results in an estimate of 94 breakpoints (including the 21 "concatenation boundaries" between successive chromosomes that are set up in the mathematical model), or 95 inferred conserved segments. Thus this method suggests that only $95 - 87 = 8$ conserved segments (less than 10%), remain to be identified.

Applying the method in Section 4 to each of the 22 human chromosomes individually gives a total estimate of 90 breakpoints or 112 conserved segments, since the number of segments on each chromosome is one more than the number of breakpoints. This not only takes into account the unobserved conserved syntenies, those containing yet to be identified homologous genes, but also syntenies that consist of more than one conserved segment due to multiple interchromosomal events. Because this analysis involves asymmetric roles for the two genomes, there is no guarantee that the results from considering mouse as species 1 will be the same as when the human genome plays this role. When the method is applied to the 19 mouse chromosomes, however, the total estimate is 88 breakpoints, or 107 conserved segments, not very different

[6] We thank J.S. Ehrlich for extracting these data from the MGD database.

from the original estimate of 112. According to this model, the number (around 30) of observed conserved segments in excess of this estimate is due to intrachromosomal rearrangements.

6. Discussion and conclusion

In Section 4 we showed how a conserved synteny approach to genomic distance could take into account the fact that with increasing distance, an apparently conserved synteny set is more likely to consist of several conserved segments, each resulting from a separate interchromosomal event. On the other hand, those conserved synteny sets that are the consequence of a unique interchromosomal event can also consist of two or more distinct conserved segments scattered along the chromosome as a result of *intrachromosomal* events such as inversion or transposition. To our methodology based only on conserved synteny data, however, such events are invisible, and even the multiple conserved segments we infer to make up each conserved synteny in Section 4 result from interchromosomal events only. Granted that systematic gene map data can furnish more information about genomic evolution, including intrachromosomal developments, the approach based on conserved synteny only is *not biased* by the lack of map data; it is only somewhat less comprehensive as an indicator of genomic divergence. And lacking complete genetic maps for many species for which considerable synteny data is available, the conserved synteny approach greatly widens the scope of phylogenetic studies based on genomic comparison.

A number of sources of error are apparent in this approach. One is its sensitivity to mistaken preliminary identification of homologous genes. This will tend to have the effect of increasing the number of apparently conserved segments containing only one (incorrectly) identified gene and hence the total number of conserved segments. The same effect will occur if a gene in species 1 is identified as homologous with one of a set of duplicated genes in species 2, but not the one in the original conserved segment.

Another error that can affect preliminary data is the failure to recognize that two or more groups of linked genes are actually syntenic. The groups may simply be too far apart for recombination experiments to detect linkage. Once again, the effect is to increase the apparent genomic divergence by increasing the number of (short) conserved syntenies.

In addition, chromosome rearrangments including deletion in somatic cell hybrids can complicate synteny assignments.

The most important focus for further research in this area is the random process postulated to model both the occurrences of genome rearrangement along the length of the chromosome and the distribution of identified genes on the chromosome. That we used a uniform distribution in both cases is not important; the key is that we used the same model for both phenomena. It would not matter that genes are more likely to be identified in euchromatic, non-telomeric regions, as long as breakages were also more frequent in these regions. This may not be the case, however. Gene-rich regions

could be slightly less suscepible to breakage, for example. This would result in more segments with zero genes, more with very many genes, and fewer around the average number, than predicted by our current model. How to refine these models is a question both for mathematical modeling and statistical analyses on existing data.

References

[1] B.O. Bengsston, K.K. Levan and G. Levan, Cytogenetics and Cell Genetics 64 (1993) 198–200.
[2] N.G. Copeland, N.A. Jenkins, D.J. Gilbert, J.T.Eppig, L.J. Maltais, J.C. Miller, W.F. Dietrich, A. Weaver, S.E. Lincoln, R.G. Steen, L.D. Stein, J.H. Nadeau and E.S. Lander, A genetic linkage map of the mouse: current applications and future prospects, Science 262 (1993) 57–66.
[3] J.S. Ehrlich, D. Sankoff and J.H. Nadeau, Synteny conservation and chromosome rearrangements during mammalian evolution, Centre de recherches mathématiques, Université de Montréal Technical Report, 1996.
[4] J.H. Nadeau, D.P. Doolittle, M.T. Davisson, P. Grant, A.L. Hillyard, M. Kosowsky and T.H. Roderick, Comparative map for mice and humans. Mammalian Genome 3 (1992) 480–536.
[5] 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, Vol. 81 (1984) 814–818.
[6] D. Sankoff and V. Ferretti, Karotype distributions in a stochastic model of reciprocal translocation, Genome Research 6, (1996) 1–9.
[7] I.A. Zakharov, V.I. Nikiforov and E.V. Stepanyuk, Genetika (translated from Soviet Genetics) 28 (1992) 77–81.