MODELS AND ANALYSES OF GENOMIC EVOLUTION

David Sankoff*

*Centre de recherches mathématiques, Université de Montréal
C.P. 6128 Succursale RAS, Montréal, Québec H3C 3J7 Canada

15.1 Introduction

Individual genes evolve through the substitution, insertion and deletion of nucleotides. Genomes, containing all the genes of the organism, are by definition evolving whenever their component genes evolve. Additional evolutionary mechanisms operate at the genomic level without affecting the composition of individual genes, and these are the focus of this paper. Entire genes, or segments of chromosomes made up of a series of genes, are inserted or removed as a single evolutionary event. Other segments migrate, are "transposed" from one region of the genome to another. A segment of a chromosome can be inverted. In multi-chromosomal organisms, reciprocal translocation can exchange segments between two chromosomes. Genomic comparison, as an approach to inferring evolutionary divergence, must take account of all of these processes.

15.2 Mechanisms of Genomic Evolution

Mathematical models of evolution at the genomic level, and the inferential apparatus associated with them, are qualitatively different from the traditional theory of macromolecular sequence comparison. Even if the well-known processes of insertion and especially of deletion of nucleotides have their counterparts at the genomic level, this is not the case for the predominant process, that of the substitution of one nucleotide for another. At the genomic level, other processes take on importance. These mechanisms can involve two or more remote regions of the genome, in contrast to processes like insertion, deletion and substitution of nucleotides, which are all local operations. We will discuss, in a common framework, analyses of the movement of segments of genomes within a single chromosome (transpositions), of the reciprocal translocation of segments between two chromosomes (*e.g.*, Ref. 1), and of the inversion of segments (*e.g.*, Refs. 2,3).

At the outset, we may ask whether the frequency and the regularity of these processes justify their use as a statistical basis for the evaluation of the similarity, the distance, or the evolutionary divergence between species, in analogy with nucleic acid sequence comparison. We will describe a number of studies that try to answer this question.

We point out that at this level, the problem is not to find alignments, nor to identify homologous genes in two or more organisms. We assume that these homologies have already been established by other means and that our data consists, for some set of genes occurring in a number of organisms, of the relative order of these genes on chromosomes in the organisms. The problem then becomes one of accounting for differences in gene order in terms of genomic level processes. Whereas in traditional sequence comparison, obtaining the alignment is almost the same thing as inferring the evolutionary history, in genomic comparison, because of the non-local nature of the processes, the relationship between the given correspondences

of gene positions in different organisms and the series of genomic changes responsible for them can be quite complex.

15.3 Properties of Random Permutations

The mechanisms of genome "shuffling" during the evolution of species inspired the construction of a stochastic model.[4] Consider a random permutation of n genes numbered from 1 to n. An "intersection" is defined as any pair of genes in the permutation where the number on the first is greater than the number on the second. Y_n represents the total number of intersections in the random permutation. It can be shown that the mean and variance of this quantity are $E(Y_n) = n(n-1)/4$ and $Var(Y_n) = n(2n^2 + 3n - 5)/72 \approx n^3/36$.

In extending these results to circular genomes, we encounter certain problems, such as how to define an intersection, since the choice of starting point for numbering the genes in both of the genomes is arbitrary. Nevertheless, with the help of relatively natural conventions about how to number the genes in circular alignments and how to decide whether a correspondence between two genes should be considered in the clockwise or counterclockwise direction, it is possible to show that $E(Y_n) \approx n^2/6$.

Making use of the distribution of Y_n, we can carry out a statistical test to verify whether we have arrived, in the shuffling process, at an equilibrium. We can model a simple process of transposition by choosing a term in the sequence at random, and moving it to another position, also chosen at random. This process fits naturally into the framework of the "card-shuffling" theory of Aldous and Diaconis in Refs. 5,6.

In Ref. 4, it is shown how to simulate standardized curves of $E[Y_n(t)]$, where t represents the time measured in terms of the number of elementary transpositions carried out (Fig. 1). The results show that the waiting time to arrive at the equilibrium situation (or "almost": $Y_n(t) > (1-\varepsilon)Y_n$) is of the order of $n \log n$.

15.4 The Evolution of Gene Order in Certain Bacteria

The results of Ref. 4 provide a theoretical framework for the comparison of gene order in five bacteria — *Escherichia coli* (EC), *Salmonella typhimurium* (ST), *Bacillus subtilis* (BS), *Caulobacter crescentus* (CC) and *Pseudomonas aeruginosa* (PA).[7,8] This comparison necessitated as a first step the construction of a normalized data base of gene names and their descriptors[9] adapting that of Ref. 10, since each species in the study has had its own terminological traditions.

An index of evolutionary divergence was defined by normalizing the number of intersections in the alignment of each pair of species, *i.e.*, dividing this number by the expected number for random circular permutations having the same n. Constructed on the basis of this index, a phylogenetic tree (Fig. 2) groups the five species in a way which is also biologically the most plausible. This result shows that gene order contains a great deal of information about the phylogenetic relationships among genomes.

15.5 Testing the Hypothesis of Reciprocal Translocations

Nadeau and Taylor[1] analyzed the divergence between mouse and man in terms of the number of translocations inferred to have occurred during the separate evolution of two species. Their analysis is based on the identification of "conserved" chromosomal segments, segments containing the same genes in the two species. Under the hypothesis that these segments

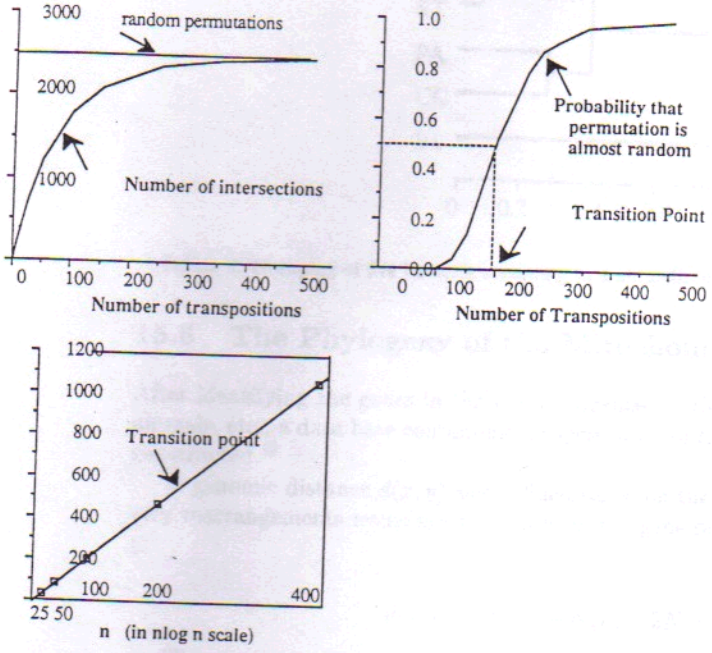

Figure 1: Upper left: average number of intersections as a function of number of transpositions $n = 100$. Upper right: Number of transpositions until 50% of permutations are almost random. Lower left: Number of transpositions to equilibrium, as a function of n.

represent the set of genes not rearranged by reciprocal translocation events, they were able to estimate the number of these events. Fig. 3 contain genes in common, based on the recent compilation in Ref. 11.

Under a model of random reciprocal translocations between chromosomes, we tested the hypothesis that these events could account for the configuration observed in the matrix. One prediction of the model is that there should be a large proportion of pairs (A,B) of human chromosomes where each member of the pair shares segments with two mouse chromosomes (C,D). These can be identified in the matrix of Fig. 3 as four filled cells at the corners of a rectangle (as in the inset). For example. human chromosomes 2 and 7 and mouse chromosomes 7 and 12 contain evidence of a reciprocal translocation.

If the data were produced by reciprocal translocation only there should be more such rectangles in Fig. 3 than there would be if 91 filled cells were scattered at random throughout the matrix. The expected number of rectangles under the latter null hypothesis can be shown to be about 84. But there are only 60 rectangles, significantly less than expected, not the direction predicted by the reciprocal translocation model. Indeed, though there are 91 pairs of human and mouse chromosomes that contain genes in common, the authors identify only 64 conserved segments using criteria more stringent than just one gene in common, since other mechanisms may well have moved some genes from chromosome to chromosome.

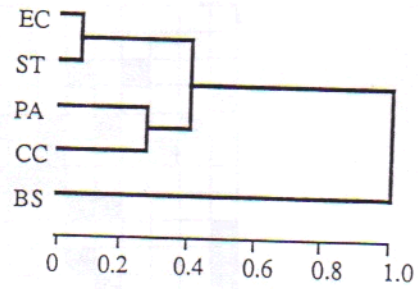

Figure 2: Grouping of five bacteria according to a normalized measure of the number of intersections.

15.6 The Phylogeny of the Mitochondrial Genome

After identifying the genes in the DNA sequences of the mitochondrial genome in 16 fungi, animals, etc., a data base containing the gene order of these mitochondrial chromosomes was constructed.[12]

A genomic distance $d(x, y)$ was defined based on the minimum number $e(x, y)$ of elementary rearrangements necessary to transform the gene order of genome x to that of another, y:

$$d(x, y) = N(x) + N(y) - 2N(xy) + e(x, y).$$

The rearrangements consist of inversions or transpositions of segments of the chromosome. N measures the number of genes in x, y or common to both. The formulation of a "branch-and-bound" algorithm for calculating the genomic distance d enabled the development of the "DERANGE" software.[13] The distances estimated by this program applied to the mitochondrial data were input to a phylogenetic analysis whose results (Fig. 4) correspond in large measure to the evolutionary relationships generally accepted among these 16 organisms.

15.7 Inversion Distance

The problem of inferring the number of rearrangements necessary to transform one permutation into another has no rapid solution. The DERANGE program can only provide an upper bound for the answer for n larger than 10 or 12, since computing time becomes excessive unless only the most promising regions of the space of possible solutions are searched.[14] Research into more powerful algorithms has concentrated on the simpler problem of computing the shortest series of inversions only that transform one permutation to another.[15–16] In Ref. 16, an exact branch-and-bound algorithm is developed that finds an optimal solution rapidly as long as n does not exceed 30. This algorithm makes use of maximum weight matchings, shortest paths, and linear programming. For large n, we can use the rapidly calculated upper and lower bounds as estimates, since for $n = 50$, say, they differ only by 3 on the average. Moreover, if we restrict our analysis to permutations likely to be generated by a reasonable number of inversions, these bounds are even tighter. In a series of experiments on permutations generated by k random inversions, we find that the average upper bound estimate of k only differs from k by at most 1, for $k < n/2$ and n up to 100.

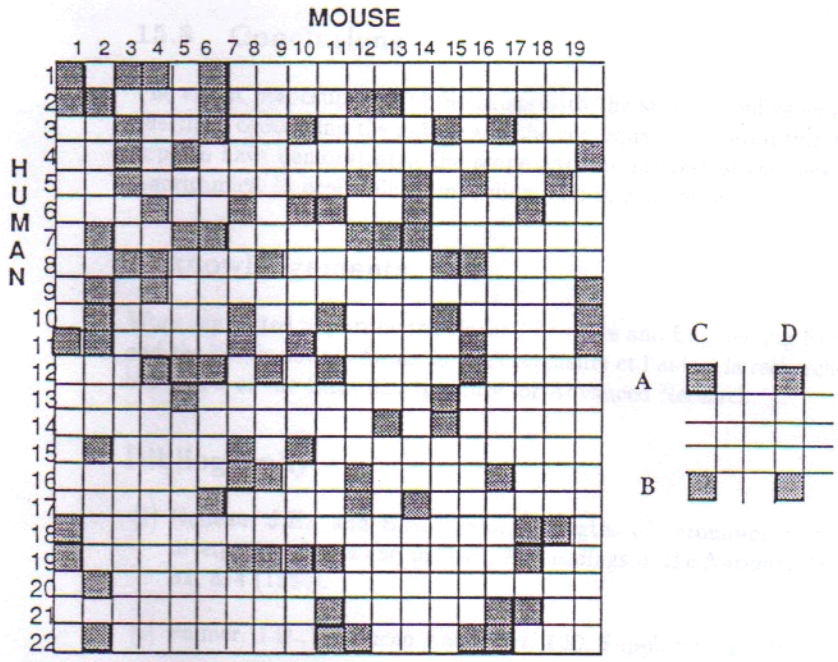

Figure 3: Tabulation of human and mouse autosomes with genes in common (filled cells). Inset: Pattern resulting from reciprocal translocation.

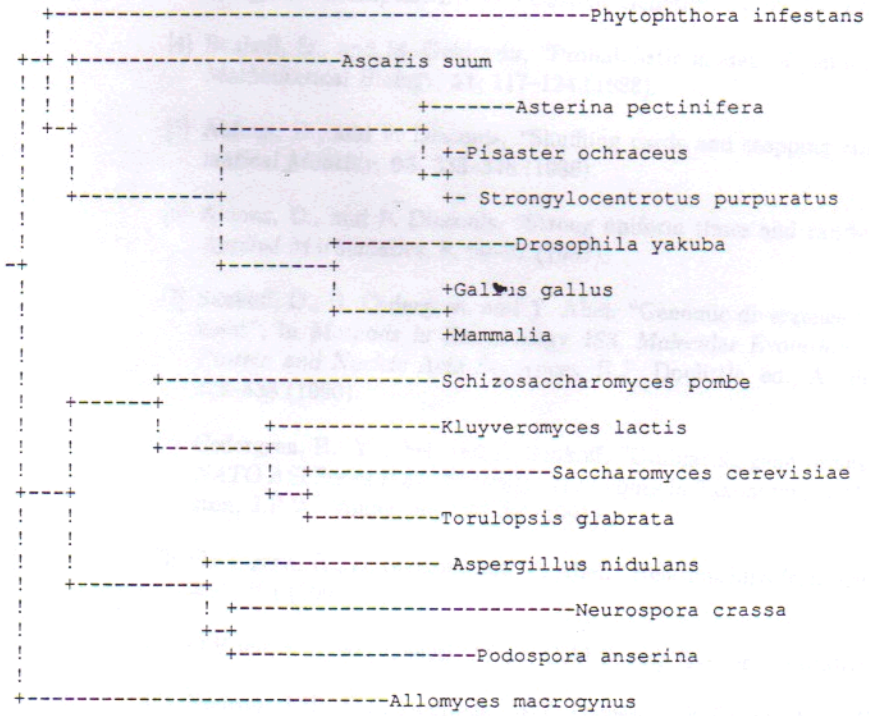

Figure 4: Eukaryote phylogeny based on the order of genes in the mitochondrial genome.

15.8 Conclusions

The recent preoccupation of biologists with the study of entire genomes leads inevitably to questions concerning the nature and the consequences of evolution at the genomic level. We hope to have demonstrated the scope and the interest of the new problems thus raised in algorithmics, in probabilistic modeling, and in simulation.

Acknowledgements

Work supported in part by the Natural Sciences and Engineering Research Council (Canada) and the Fonds pour la formation de chercheurs et l'aide a la recherche (Quebec). The author is a fellow of the Canadian Institute for Advanced Research.

Bibliography

[1] Nadeau, J.H., 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 (1984).

[2] Palmer, J.D., *American Naturalist*, **130**, Suppl. S6–S29 (1987).

[3] Palmer, J.D., B. Osorio, and W.F. Thompson, "Evolutionary significance of inversions in legume chloroplast DNAs", *Current Genetics*, **14**, 65–74 (1988).

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

[5] Aldous, D., and P. Diaconis, "Shuffling cards and stopping times", *American Mathematical Monthly*, **93**, 333–348 (1986).

[6] Aldous, D., and P. Diaconis, "Strong uniform times and random walks", *Advances in Applied Mathematics*, **8**, 69–97 (1987).

[7] Sankoff, D., R. Cedergren, and Y. Abel, "Genomic divergence through gene rearrangement", in *Methods in Enzymology 183, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, R.F. Doolittle, ed., Academic Press, San Diego, 428–438 (1990).

[8] Cedergren, R., Y. Abel, and D. Sankoff, "Evaluating gene versus genome evolution", in *NATO ASI Series H 57, Molecular Techniques in Taxonomy*, G.M. Hewitt, A.W.B. Johnston, J.P.W. Young, eds., 87–99 (1991).

[9] Cedergren, R., D. Sankoff, and Y. Abel, "Relationships from gene sequences", *Nature*, **345**, 484 (1990).

[10] O'Brien, S.J., ed., *Genetic Maps*, Cold Spring Harbor Laboratory, (1987).

[11] Nadeau, J.H., 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*, in press (1992).

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

[13] Sankoff, D., G. Leduc, and D. Rand, DERANGE — Minimum weight generation of oriented permutation by block inversions and block movements, Macintosh Application, Centre of recherches mathématiques, Université de Montréal, (1991).

[14] Sankoff, D., "Edit distance for genome comparison based on non-local operations", *Combinatorial Pattern Matching 92*, A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, eds., Lecture Notes in Computer Science, Springer-Verlag, Berlin, 121–135 (1992).

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

[16] Kececioglu, J., and D. Sankoff, "Exact and approximation algorithms for the reversal distance between two permutations", *Algorithmica*, in press (1993).