Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 6575–6579, July 1992
Evolution

Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome

(genomics/algorithm/inversions/edit distance/conserved segments)

DAVID SANKOFF*, GUILLAUME LEDUC*, NATALIE ANTOINE†, BRUNO PAQUIN†, B. FRANZ LANG†, AND ROBERT CEDERGREN†

*Centre de recherches mathématiques, Université de Montréal, CP 6128, Montréal H3C 3J7, Canada; and †Département de Biochimie, Université de Montréal, CP 6128, Montréal H3C 3J7, Canada

Communicated by Russell F. Doolittle, April 20, 1992

ABSTRACT **Detailed knowledge of gene maps or even complete nucleotide sequences for small genomes leads to the feasibility of evolutionary inference based on the macrostructure of entire genomes, rather than on the traditional comparison of homologous versions of a single gene in different organisms. The mathematical modeling of evolution at the genomic level, however, and the associated inferential apparatus are qualitatively different from the usual sequence comparison theory developed to study evolution at the level of individual gene sequences. We describe the construction of a database of 16 mitochondrial gene orders from fungi and other eukaryotes by using complete or nearly complete genomic sequences; propose a measure of gene order rearrangement based on the minimal set of chromosomal inversions, transpositions, insertions, and deletions necessary to convert the order in one genome to that of the other; report on algorithm design and the development of the DERANGE software for the calculation of this measure; and present the results of analyzing the mitochondrial data with the aid of this tool.**

Evolutionary inference based on DNA sequences traditionally compares homologous versions of a single gene in different organisms. These comparisons are generally reliable indicators of phylogenetic relationships, even for very divergent organisms, but are limited in being based on point mutations only. In particular, homology between related mitochondrial genes may become difficult to distinguish from noise levels due to rapid nucleotide substitution (1), and this is not the only context in which the degree of sequence homology between genes having common origin is not a useful measure. Availability of complete nucleotide sequence for organellar genomes suggests the possibility of inferring phylogenetic distances from their gene orders instead of from sequences of individual genes (2).

Analyses of evolution at the genome level necessarily differ from sequence comparisons of individual genes. Though the processes of insertion and deletion of sequence elements have direct counterparts at the genomic level, the predominant process, nucleotide substitution, does not, whereas other processes assume major importance, such as the transposition of a segment from one region of a chromosome to another or the inversion of a chromosomal segment. Here we propose a quantitative analysis of transposition, inversion, and insertion/deletion, leading to the reconstruction of a mitochondrial phylogeny.

Though the inference of evolutionary history through genomic rearrangements is well-established (3–5), it has been the goal of our work to define a general edit distance that combines a variety of order-disrupting events, to devise and implement a combinatorial algorithm capable of estimating this distance, and to apply these tools in a uniform way across a wide spectrum of eukaryotic organisms to generate input suitable for phylogenetic tree construction methods. Our results generally agree with evolutionary relationships inferred from gene sequences.

Mutation at the gene level may be neutral or it may be directly linked to specific changes in function. Analogously, genomic level changes such as inversion, transposition, and duplication may have no apparent functional consequence or they may affect levels of expression of unchanged functional molecules or, more dramatically, permit functional differentiation through gene duplication and divergence or cause interruption in important relationships of coregulation. At both levels, it is the tendency over time to accumulate more and more changes, neutral or not, that permit the statistical analysis of differences among organisms with a view to phylogenetic inference.

DATA

The mitochondrion constitutes an ideal model for studying evolution due to genome rearrangement. Where high rates of nucleotide substitution may reduce gene homology to the level of noise, gene order may still retain traces of phylogenetic relationship. In addition, the organellar genome is small enough to be tractable by current sequencing technology, so that nearly 20 genomes have been extensively sequenced. This provides gene order data from a widely dispersed set of eukaryotes (Table 1) in which a convenient number of genes (i.e., not too many genes to be handled in reasonable computational time by our program and not too few to give statistically meaningful numbers of inferable rearrangement events when comparing genomes) has been conserved across major evolutionary distances.

METHODS

An Edit Distance. Our analysis is based on the notion of an evolutionary edit distance, $E(a, b)$, the number of elementary events—inversions, transpositions, and deletions/insertions—necessary to change the gene order of one circular genome a into that of another, b. Note that $E(b, a) = E(a, b)$, since each inversion or transposition may be reversed by a corresponding inversion or transposition and each deletion may be reversed by an insertion and vice versa. For the mitochondrial genome, we know that there have generally been no gene insertions, so an apparent insertion of a gene as part of a transformation of a into b really reflects a deletion from the common ancestor of a and b during the evolution of a.

To evaluate E, we first consider separately only those differences between genomes due to gene deletions and insertions. The deletion/insertion distance is defined as $D(a, b)$, the total number of genes present in either one of the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. Mitochondrial genomes compared

Genome	Genes, no.
Fungi	
Fission yeast	
Schizosaccharomyces pombe (6)*	35
Budding yeast	
Torulopsis glabrata (7)	34
Saccharomyces cerevisiae (8)	39
Kluyveromyces lactis (9)	31
Filamentous Ascomycetes	
Neurospora crassa (10)	50
Aspergillus nidulans (11)	44
Podospora anserina (12)	45
Chytridiomycetes	
Allomyces macrogynus (B.P. and B.F.L., unpublished data)	35
Protist	
Phytophthora infestans (B.F.L., unpublished data)	31
Animals	
Vertebrates	
Mammalia (13, 14)	37
Gallus gallus (chicken) (15)	37
Echinoderms	
Strongylocentrotus purpuratus (sea urchin) (16)	37
Asterina pectinifera (star fish) (17)	37
Pisaster ochraceus (sea star) (18)	36
Insect	
Drosophila yakuba (19)	37
Nematode	
Ascaris suum (20)	36

Excluded from our analysis are mitochondria whose sequences were not sufficiently known at the time of the analysis (July 1991). In particular, the *Paramecium* sequence has not been used because it contains a high proportion of unidentified open reading frames. The *Chlamydomonas reinhardtii* mitochondrion was excluded, because it contains too few genes in common with the other genomes to allow for reliable quantitative analysis. We have not taken account of intron open reading frames. The gene maps used in our analysis are available from D.S.

*European Molecular Biology Laboratory data library, accession no. MISPCG.

genomes but not the other. We then define a rearrangement distance $R(a, b)$ between the two genomes—namely, the minimal number of inversion and transposition events necessary to convert one to the other, ignoring those genes that are absent from either one. Thus, $E = D + R$.

The Rearrangement Distance. The distance R is roughly related to the easily calculated number of "conserved chromosomal segments" C counted by Nadeau and Taylor (4). When an inversion affects a chromosome in one organism and not in another that previously had the same gene order, it generally results in three segments in which the order in each segment is the same in both genomes (ignoring the "directionality" of the inverted segment). Increasing R by one inversion should then correspond to increasing C by two segments. When an endpoint of one inversion coincides with an endpoint of a previous inversion, however, C only increases by one segment, so that all that can be said is that C is no greater than $2R$ for a circular genome or that R is no less than $C/2$. The same lack of a precise relationship applies to the effect of transpositions on the value of C. Thus there is no way of calculating R from observing C when there have been a number of possibly overlapping rearrangement events (3). Indeed, it is generally thought that finding R is a computationally hard problem, requiring computing time that can increase exponentially with the number of genes in the genome.

Our calculation of R is carried out by a branch-and-bound search implemented in a program called DERANGE (D.S., G.L., and D. Rand, unpublished program). The key technique is that of alignment reduction, as illustrated in Fig. 1*a*. The genes combined in this operation may be considered to constitute a conserved segment of the chromosome (4); that is, they participate as a unit in any recombinational event. By applying a rearrangement operation such as transposition or inversion of a segment (which is equivalent to "undoing" a rearrangement event that has occurred during evolution) to a reduced alignment, we may produce a situation where the alignment may be further reduced by combining a number of linked pairs. The new configuration may be considered a hypothetical most-recent ancestral genome of b, with fewer and larger conserved segments in common with a, before the last transposition or inversion took place. Up to three pairs of links can be combined in a transpositional operation or up to two such pairs can be combined in an inversion. In the example illustrated in Fig. 1*b*, inverting the order of genes 5 and 2 allows the combination of gene 2 with gene 3 and gene 4 with gene 5. At each stage all the possible ancestral genomes produced at the previous step may be tested to see whether there are transpositions or inversions that will lead to further reductions (i.e., to more remote ancestors) and this continues until the smallest number $R(a, b)$ of operations is

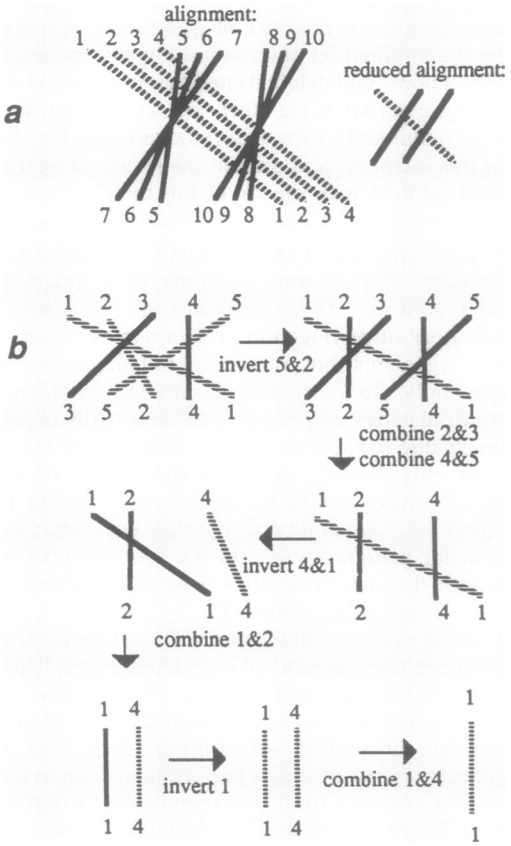

FIG. 1. Examples of alignment reduction. (*a*) Two or more homologous pairs of genes that are adjacent in both genomes and are either of the same orientation and order, such as the four pairs linked by dotted lines, or of opposing orientation and in reverse order, such as the two groups of three pairs linked by solid lines, may be combined and replaced by a single symbol representing a "conserved segment," since the minimum number of recombinatory events for the reduced (after this combination operation) problem is the same as the original (before combination) problem. (*b*) Reducing an alignment while finding a three-inversion solution for the minimal events distance.

found that leads to a completely reduced alignment—exactly one link.

For pairs of genomes containing 25–35 or more genes in common, it is not computationally feasible to examine all possible ancestral genomes at each stage of the algorithm. Mathematical results to be presented elsewhere enable us to abandon some search paths that cannot possibly lead to a minimal value of R. In addition, probabilistic considerations allow us to discard the most unlikely potential intermediate genomes, so that at any one time only a restricted number are under active consideration—we have used 5000 as an upper limit, since increasing the capacity to 10,000 rarely results in a lower value of R, and even when it does, it is only by one or, very rarely, two rearrangement events (29). Finally, we have excluded from consideration whenever possible inversions and transpositions that do not result in any reduction of the alignment and transpositions that reduce the number of links by only 1, though it is theoretically possible (though very unlikely) that these can occur in a series of arrangements that minimize R.

Even with these algorithmic simplifications, the applications of DERANGE reported below required considerably more than 300 h of computing time on a battery of Macintosh computers of various types.

RESULTS

It is clear in Table 2 that within the group of animal genomes, deletion/insertion is too rare an event to be used by itself as an indicator of phylogenetic relationships, though it does distinguish between these and the fungi. Moreover, the yeasts show lower D values among themselves than when compared to the filamentous fungi, cleanly separating the two groups within the fungal branch. Nevertheless, $D(a, b)$ seems too crude a measure to use by itself for phylogenetic purposes, particularly as it is not known how it may depend statistically on the total number of genes in a and in b.

The values of $R(a, b)$ calculated by DERANGE for the organisms in Table 1, once the genes absent from either member of the pair being compared are excluded, are given in Table 2 along with the values of $D(a, b)$.

Validation. We may ask whether the rearrangement distance $R(a,b)$ between the various pairs (a, b) of mitochondrial genomes is significantly different from the random noise level. Because no analytical results are available for the probabilistic behavior of R, we generated a number of pairs of random circular permutations of various lengths and submitted them to the same analysis as the circular mitochondrial genomes. Note that, for pairs of random permutations of length n, the average number of genes that can be combined, as in Fig. 1a, is only two (21) (i.e., the number of conserved segments C is $n - 1$), but the average R is about $0.8n - 3.0$. Compared to this, most of the values of R within the fungal group and within the animal group are clearly nonrandom, indicating that the gene orders have not evolved to the random noise level and justifying our use of R to assess the phylogenetic relationships within these two groups. The comparisons between the two groups, however, are indistinguishable from random.

We note that the number of conserved segments C is also much lower than random within the fungi and within the animal groups and is quite highly correlated with R ($r^2 = 0.81$). When we normalize R by dividing by C, however, the quotient still shows within-group values that are lower than random, indicating that R contains phylogenetic information in addition to that contained in C (29). Nevertheless, given the relative ease of calculating C and the cost of calculating R, it would not be unjustified to use C (or better, $0.75C - 2.8$) as a rough estimate of R when calculating E, at least in preliminary studies.

Phylogeny. Fitting $E(a, b) = D(a, b) + R(a, b)$ to an additive tree model using a weighted least-squares criterion (22) produces the tree in Fig. 2. The branching order within the nonfungal group corresponds almost perfectly to accepted evolutionary knowledge, with successively deeper branch-

Table 2. Distances between genome pairs

	Distance																		
	Mam	Gal	Str	Ast	Pis	Dro	Asc	Phy	All	Sch	Tor	Klu	Sac	Asp	Neu	Pod
Mam		1	18	16	19	13	25	12	— 18	— 21	— 17	16	19	— 23	26	27
Gal	0		19	17	17	12	26	13	— 22	— 21	— 17	17	19	— 23	24	26
Str	0	0		2	1	26	27	13	— 21	— 19	— 19	16	20	— 24	27	25
Ast	4	4	4		1	22	25	13	— 20	— 16	— 14	18	18	— 24	25	25
Pis	1	1	1	5		23	24	12	— 17	— 20	— 17	16	19	— 24	24	22
Dro	0	0	0	4	1		28	11	— 19	— 21	— 17	19	17	— 26	26	27
Asc	1	1	1	5	2	1		11	— 15	— 14	— 13	13	12	— 16	20	16
Phy	26	26	26	28	25	26	25		— 10	— 10	— 10	8	10	— 12	11	15
	—	—	—	—	—	—	—	—	— —	—	—	—	—	—	—	—
All	12	12	12	14	13	12	13	30	—	15	— 14	13	14	— 17	17	16
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Sch	14	14	14	18	13	14	15	28	— 18		15	15	18	— 18	19	18
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Tor	17	17	17	19	16	17	18	29	— 19	— 7		11	10	— 15	12	15
Klu	18	18	18	20	17	18	19	28	— 20	— 8	— 3		11	— 11	11	13
Sac	20	20	20	24	19	20	21	32	— 24	— 10	— 5	8		— 13	13	15
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Asp	11	11	11	13	12	11	12	31	— 13	— 13	— 16	17	21	—	10	10
Neu	15	15	15	19	16	15	16	35	— 19	— 17	— 22	23	25	— 14		9
Pod	10	10	10	14	11	10	11	30	— 16	— 14	— 19	20	22	— 11	15	

Deletion distance D is to the lower left and the rearrangement distance R is to the upper right. Double dashes separate fungal and nonfungal mitochondria; single dashes distinguish among *Schizosaccharomyces pombe*, *Allomyces*, the budding yeasts, and the filamentous Ascomycetes. Calculation of R was performed by DERANGE, a Pascal program implemented in a Macintosh application. In this analysis, the two types of rearrangement event (inversion and transposition) were assigned the same weight, although the program allows differential weighting. Mam, *Mammalia*; Gal, *Gallus*; Str, *Strongylocentrotus purpuratus*; Ast, *Asterina*; Pis, *Pisaster*; Dro, *Drosophila yakuba*; Asc, *Ascaris*; Phy, *Phytophthora*; All, *Allomyces*; Sch, *Schizosaccharomyces pombe*; Tor, *Torulopsis*; Klu, *Kluyveromyces lactis*; Sac, *Saccharomyces cerevisiae*; Asp, *Aspergillus*; Neu, *Neurospora*; Pod, *Podospora*.

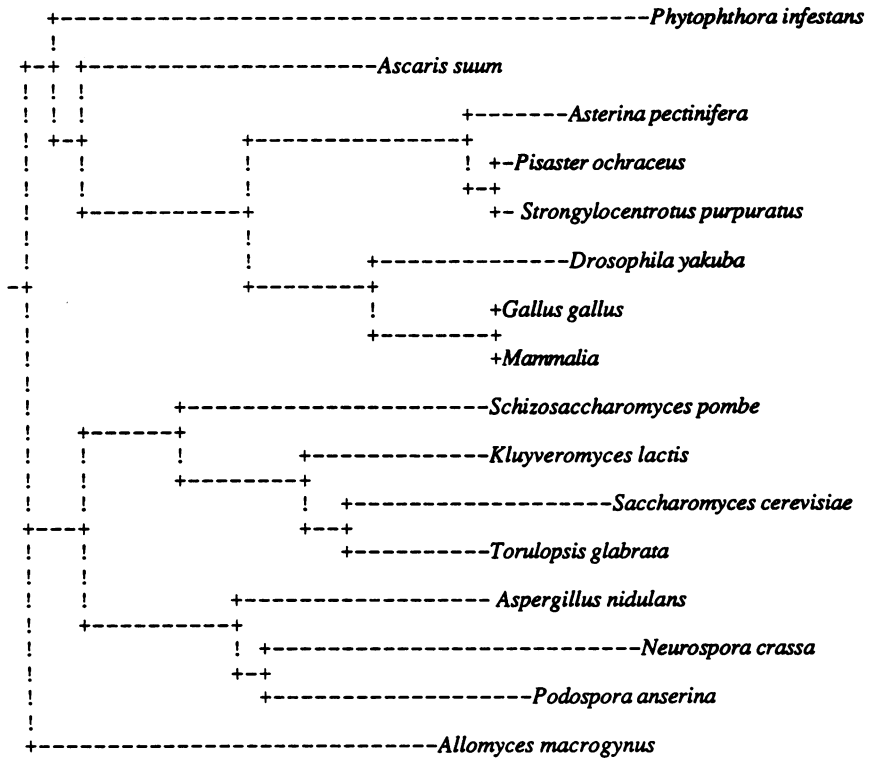

FIG. 2. Tree derived from the evolutionary edit distance E (= D + R) between mitochondrial genomes in Table 2. Root (not found by tree algorithm) was placed between fungi and other eukaryotes for ease of interpretation.

ings for the vertebrates, the insect, the nematode, and the protist. Only the echinoderms, which should share a more recent branching with the chordates, appear to branch before *Drosophila*. This may be an artifact of the relative mobility of tRNA genes within the echinoderms, which perhaps should not have been weighted as heavily as other rearrangement events (see *Discussion*). The branching of *Gallus* and *Mammalia* represents a small divergence of both from a common ancestor, though we know from comparative *Xenopus* data that only the *Gallus* genome has changed.

Within the fungi, the budding yeast group is well-defined, including the close pair *Saccharomyces* and *Torulopsis*, and the close branching of *Schizosaccharomyces pombe* accords with most mitochondrial single-gene phylogenies, though nuclear genes place it before the yeast–Ascomycetes divergence. The three filamentous Ascomycetes are together, with *Podospora* and *Neurospora* forming a subgroup. *Allomyces*, a true lower fungus, branches close to the divergence point of the fungi, consistent with sequence-level analyses on small ribosomal subunit RNA (B.F.L. and B.P., unpublished results). *Phytophthora*, which is not a fungus but is rather close to the chrysophyte algae such as *Ochromonas*, branches as expected outside the fungal subtree.

DISCUSSION

Overall Assessment. The coherence of our mitochondrial phylogeny, based entirely on the gene composition and gene order of mitochondrial genomes, offers strong validation of the hypothesis that the macrostructures of genomes contain quantitatively meaningful information for phylogenetic reconstruction, analogous to gene-level measures of sequence similarity in traditional molecular evolution studies.

Weighting. In this study, all inferred rearrangement events contributed the same amount to the evolutionary edit distance E: each insertion, deletion, inversion, and transposition was accorded a weight of 1. Our program includes an option for weighting inversions relative to transpositions, but until there is some empirical justification for unequal weights, the best we can hope to do is to undertake the (computationally costly) investigation of the stability of the reconstructed phylogeny with respect to different weightings. Preliminary trials indicate that weighting to favor (disfavor) inversions increases (decreases) the number of inversions twice as fast as the increase in the number of transpositions (29). This is understandable in terms of the mathematical fact that the effect of any transposition can also be achieved by at most two inversions and in terms of a bias built into DERANGE discussed above—to avoid a prohibitive increase in computing complexity—against transpositions that result in fewer than two links being combined. Thus in any area of the tree, such as within the fungi or within the nonfungal mitochondria, in which the proportions of inversions and transpositions do not vary too widely, the branching order will be relatively stable with respect to the choice of weighting. As for weighting D versus R, it is clear from Table 2 that changing this somewhat will not have any systematic effect on the branching order of the nonfungal organisms by themselves, nor on that of the fungi by themselves, though a heavy weighting on R may perturb the positioning of *Allomyces* and/or *Phytophthora* in the tree.

Our approach depends more on the combination of a variety of order-disrupting events—inversion, transposition, and deletion—to produce a statistically meaningful measure of evolutionary divergence than on any hypothesis that one or more of these processes occur at a steady rate across all phylogenetic lines. Indeed, although inversion is certainly of great importance in the nonfungal region of the tree, perhaps the most important process, there is little evidence that it plays a major role in fungal evolution since in almost all fungi the mitochondrial genes are all read in the same direction. On the other hand, differential deletion of genes among fungal mitochondrial genomes is quite striking, whereas from nematodes to mammals, the gene complement is very stable.

Assuming that each deletion of a gene is a separate event has allowed us to calculate D independently of R. Though it

seems biologically unlikely at present, it is possible that deletion could involve several contiguous genes at the same time. Were this so, the cost of deletion and insertion of k contiguous genes should not necessarily be simply k times the cost for one gene but, perhaps, a more slowly growing (convex) function of k. In addition, the insertion or deletion of a number of contiguous genes would have to be allowed to occur at any time during the transformation of one genome into another, rather than calculated at the outset as we do now, since a number of genes deleted together at one time might previously have been dispersed throughout the genome and then brought together through inversion and transposition. These changes would require integrating deletion into the DERANGE search algorithm, which would risk making computing requirements excessive.

More problematic than the uniform weights on the different classes of rearrangement events is their uniform application to all genes. It seems likely, for example, that tRNA genes are relatively mobile (23–25) and that their movement or deletion is less informative for relatively remote phylogenetic relationships than, say, that of large inversions. There are no technical difficulties in incorporating differential gene-specific weightings into the program, but reasonable estimates of these weightings must await more data.

Hidden Events. With increasing divergence times, more back mutations occur, reversing the effects of previous mutations, so that minimal edit-distance assessments of homology tend to underestimate longer distances. This is as true at the genome level as at the gene-sequence level. This may occasionally introduce errors in tree-construction algorithms by increasing the uncertainty attached to the early branching nodes and through the tree algorithm's fitting long, but systematically understated, distances at the expense of accurate representation of closer relationships. Simulation studies will eventually allow us to correct distance estimates in analogy with the exponential corrections often applied to gene-sequence distances (26).

Statistical Validation. In phylogenetic trees inferred from the comparison of a number of sequences with r aligned positions, the details of the branching order may be validated by means of a large number of "bootstrap" resamplings, with replacement, of r sequence positions from the pool of r different positions, each resampling followed by reapplication of the distance calculation and the phylogeny algorithm. The tendency of any branch in the tree to show up across all or most of the resamplings is a measure of its validity in the original tree (27, 28). This approach would not be appropriate to gene orders since sampling with replacement would generally yield several copies of some genes, which would not be compatible with the mathematical notion of a simple (circular) order, basic to our analysis. Nonetheless, other resampling schemes may be suitable for validating branching orders based on our edit distance, such as repeating the analysis n times, each time omitting one gene.

Computational Developments. The design of the DERANGE program is focused on finding the true value of R. The running time and storage requirements for even moderate-size genomes, however, are prohibitive without introducing some limits on the set of possible solutions to be examined. Thus, though by expanding this set as much as possible we can be fairly sure that all the values of R are within 1 or 2 events of their true value, we have nonetheless settled for a suboptimal algorithm. It may be worthwhile, then, to sacrifice the optimality-oriented design at the outset in favor of exploring various types of approximations known to be rapid and memory-efficient, such as greedy algorithms or iterative local improvement methods.

We thank David Rand and Yvon Abel for computing assistance and Professor Claude Weber, University of Geneva, for encouragement and facilities. This research was supported in part by operating and infrastructure grants from the National Science and Engineering Research Council of Canada, a team grant from the Fonds pour la formation de chercheurs et l'aide à la recherche of the government of Québec, and grants from the Medical Research Council of Canada. D.S., B.F.L., and R.C. are fellows of the Canadian Institute for Advanced Research.

1. Brown, W. M., George, M., Jr., & Wilson, A. C. (1984) *Proc. Natl. Acad. Sci. USA* **76**, 1967–1971.
2. Sankoff, D., Cedergren, R. & Abel, Y. (1990) *Methods Enzymol.* **183**, 428–438.
3. Watterson, G. A., Ewens, W. J., Hall, T. E. & Morgan, A. (1982) *J. Theor. Biol.* **99**, 1–7.
4. Nadeau, J. H. & Taylor, B. A. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 814–818.
5. Palmer, J. (1987) *Am. Nat.* **130**, Suppl., S6–S29.
6. Lang, F. B. (1984) *EMBO J.* **3**, 2129–2136.
7. Clark-Walker, G. D., McArthur, C. R. & Sriprakash, K. S. (1985) *EMBO J.* **4**, 465–473.
8. Grivell, L. A. (1990) in *Genetic Maps*, ed. O'Brien, S. J. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 3.50–3.57.
9. Wilson, C., Ragnini, A. & Fukuhara, H. (1989) *Nucleic Acids Res.* **17**, 4485–4491.
10. Collins, R. A. (1990) in *Genetic Maps*, ed. O'Brien, S. J. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 3.19–3.21.
11. Brown, T. A. (1990) in *Genetic Maps*, ed. O'Brien, S. J. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 3.109–3.110.
12. Cummings, D. J., McNally, K. L., Domenico, J. M. & Matsuura, E. T. (1990) *Curr. Genet.* **17**, 375–402.
13. Anderson, S., Bankier, A. T., Barrell, B. G., deBruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1981) *Nature (London)* **290**, 457–465.
14. Bibb, J. J., Van Etten, R. A., Wright, C. T., Walberg, M. W. & Clayton, D. A. (1981) *Cell* **26**, 167–180.
15. Desjardin, P. & Morais, R. (1990) *J. Mol. Biol.* **212**, 599–634.
16. Jacobs, H. T., Elliott, D. J., Math, V. B. & Farquharson, A. (1988) *J. Mol. Biol.* **202**, 185–217.
17. Asakawa, S., Kumazawa, Y., Araki, T., Himeno, H., Miura, K.-I. & Watanabe, K. (1991) *J. Mol. Evol.* **32**, 511–520.
18. Smith, M. J., Banfield, D. K., Doteval, K., Gorski, S. & Kowbel, D. J. (1990) *J. Mol. Evol.* **31**, 195–204.
19. Clary, D. O. & Wolstenholme, D. R. (1985) *J. Mol. Evol.* **22**, 252–271.
20. Wolstenholme, D. R., Macfarlane, J. L., Okimoto, R., Clary, D. O. & Wahleithner, J. A. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1324–1328.
21. Sankoff, D. & Goldstein, M. (1988) *Bull. Math. Biol.* **51**, 117–124.
22. Felsenstein, J. (1990) *PHYLIP Version 3.3* (Univ. of Washington, Seattle).
23. Cantatore, P., Gadaleta, M. N., Roberti, M., Saccone, C. & Wilson, A. C. (1987) *Nature (London)* **32**, 853–855.
24. Jacobs, H., Asakawa, S., Araki, T., Miura, K., Smith, M. J. & Watanabe, K. (1989) *Curr. Genet.* **15**, 193–206.
25. Pääbo, S., Thomas, W. K., Whitfield, K. M., Kumazawa, Y. & Wilson, A. C. (1991) *J. Mol. Evol.* **33**, 426–430.
26. Jukes, T. H. & Cantor, C. H. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. M. (Academic, New York), pp. 21–123.
27. Felsenstein, J. (1985) *Evolution* **39**, 783–791.
28. Sankoff, D., Abel, Y., Cedergren, R. J. & Gray, M. W. (1988) in *Classification and Related Methods of Data Analysis*, ed. Bock, H. H. (North Holland, Amsterdam), pp. 385–394.
29. Sankoff, D. (1992) in *Combinatorial Pattern Matching '92*, eds. Apostolico, A., Crochemore, M., Galil, Z. & Manber, U. Lecture Notes in Computer Science (Springer, Berlin), pp. 121–135.