

# Edit Distance for Genome Comparison Based on Non-Local Operations \*

David Sankoff

Centre de recherches mathématiques, Université de Montréal  
C.P. 6128, succursale "A", Montréal, Québec H3C 3J7  
(sankoff@ere.umontreal.ca)

**Abstract.** Detailed knowledge of gene maps and even complete nucleotide sequences for small genomes has led 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. In this paper, we define a number of measures of gene order rearrangement, describe algorithm design and software development for the calculation of some of these quantities in single-chromosome genomes, and report on the the results of applying these tools to a database of mitochondrial gene orders inferred from genomic sequences.

## 1 Role of Rearrangements in Evolution

Genes evolve largely through the processes of nucleotide substitution, insertion and deletion. Genomes, containing the entire genetic complement of an organism, evolve because the genes in them evolve. But there are other mechanisms of evolution at the genomic level as well. For example, entire genes, or segments of chromosomes containing several genes, may be deleted or inserted as a single event. Genes or chromosomal segments may migrate (i.e. be transposed) from one region of the genome to another. A segment of the chromosome may become inverted:  $\dots xabcdy\dots$  becomes  $\dots xdcbay\dots$ , where the inverted segment is transferred to the opposite strand of double stranded DNA so that it is still read in the original order.

## 2 Phylogenetic Inference

The mathematical study of evolution at the genomic level based on operations such as insertion, deletion, transposition and inversion of chromosomal segments containing one or more genes is necessarily different from traditional studies involving nucleotide

---

\* Work supported by operating and infrastructure grants from the Natural Science and Engineering Research Council (Canada) and a team grant from the *Fonds pour la formation de chercheurs et l'aide à la recherche (Quebec)*. The author is a Fellow of the Canadian Institute for Advanced Research. This paper was begun while a guest of Professor Claude Weber at the University of Geneva in 1991. I thank Natalie Antoine, Robert Cedergren, Franz Lang and Bruno Paquin who constructed the database from which the mitochondrial gene orders used here were drawn, and especially Guillaume Leduc for his perseverance in implementing DERANGE.

sequence comparison because some of these operations are non-local; inversion and transposition in particular may involve arbitrarily distant (at least in theory) terms in the gene order.

Nevertheless, since these processes are widespread and accumulate over time, it is interesting to ask whether the comparison of genomes, i.e. of their gene orders, can provide us a statistical basis for assessing the similarity or distance or evolutionary divergence between organisms, in analogy with nucleotide sequence comparison [1, 4, 5, 6, 8]. In this paper, we motivate and describe an algorithm capable of inferring the minimal set of inversions and transpositions necessary to convert one genome order into another. This, together with a measure of gene deletion, not only provides an estimate of the degree of evolutionary divergence, but characterizes it in terms of the predominant processes separating related organisms from their ancestors.

### 3 The Mitochondrial Data

The mitochondrion constitutes an ideal model for studying eukaryotic evolution through genome rearrangement. The organellar genome is small enough ( $\approx 50,000$  bp) to be tractable by current sequencing technology, so that about 20 sequences are known completely or almost completely, with nearly all genes identified.

**Table 1.** Organisms providing mitochondrial genomes used in this study. References in [8]

Organism	Number of Genes
FUNGI	
Fission Yeast	
<i>Schizosaccharomyces pombe</i>	35
Budding Yeasts	
<i>Torulopsis glabrata</i>	34
<i>Saccharomyces cerevisiae</i>	39
<i>Kluyveromyces lactis</i>	31
Filamentous fungi	
<i>Neurospora crassa</i>	50
<i>Aspergillus nidulans</i>	44
<i>Podospora anserina</i>	45
MULTICELLULAR	
Human	37
<i>Strongylocentrotus purpuratus</i> (sea urchin)	37
<i>Drosophila yakuba</i> (fruitfly)	37
<i>Ascaris suum</i> (nematode)	36

These are widely dispersed among subgroups of the eukaryotes so that the phylogenetic range of our methodology may be evaluated. Most important, despite its minuscule size compared to nuclear genomes, the mitochondrion appears to undergo

the same processes of genome rearrangement which interest us here as do nuclear chromosomes: mainly deletion, inversion and transposition of chromosomal segments.

We have sampled 11 mitochondrial genomes from the database of [8]. As listed in Table 1, this contains four multicellular organisms, extending from nematodes to humans, and seven fungi, of which three are budding yeasts, one is a fission yeast, and three are filamentous fungi. In our analyses, whenever we compare all pairs of genomes, we will distinguish comparisons within the two groups of fungi (grouping budding and fission yeasts), within the multicellular organisms, and among the three groups.

## 4 Genes in Common

Let us first consider only those differences between genomes due to gene deletion and insertion. The simplest distance measure between two genomes based on these processes is simply  $D(x, y)$ , the number of genes present in either one of the genomes but not the other. This is calculated as

$$D(x, y) = N(x) + N(y) - 2N(xy),$$

where  $xy$  represents the genes in common between  $x$  and  $y$ , and  $N(s)$  is the number of genes in  $s$ . The intra- and intergroup averages of  $D$  are presented in Figure 1.

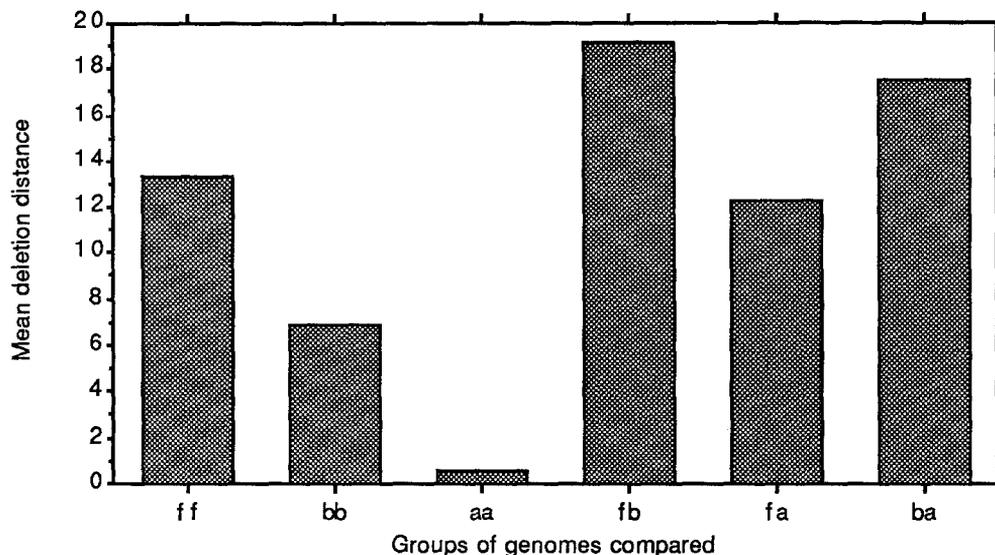


Fig. 1. Intra- and intergroup deletion distances  $D(x, y)$ , based on difference in gene content only. f=filamentous fungi, b=budding yeasts and *S. pombe*, and a=multicellular eukaryotes.

In some contexts  $D(x, y)$  might be a perfectly reasonable measure; but it has at least three weaknesses. If genomes differ markedly in the total number of genes each contains, then a given distance between two small genomes may represent a greater

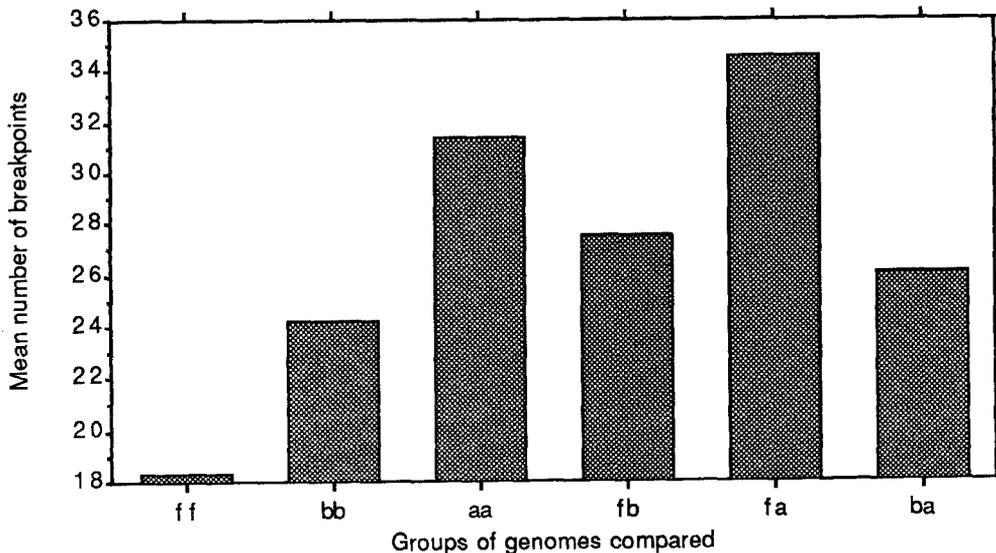
amount of evolution than the same distance between two large genomes. This could possibly be compensated for by normalisation, such as by dividing the distance by half the sum of the two genome lengths. This, however, would implicitly involve the questionable assumption that deletion from already small mitochondrial genomes, even those affecting less “superfluous” genes, happens as readily as deletions from large mitochondrial genomes.

Second, if the deletion or insertion processes may involve a block of several contiguous genes, it may not seem appropriate to count each gene in such a block separately. We will discuss this problem further below.

Finally, it is clear that within the multicellular group of genomes (aa), deletion is too rare an event to be used as an indicator of phylogenetic relationships, though it does distinguish between these and the fungi (fa,ba), and within the fungal group, the yeasts (bb) show lower values of  $D$  among themselves than between them and the non-yeasts (fb).

## 5 Breakpoints

Suppose  $ABCDEFGH$  and  $ABCEFGDH$  are two genomes. Except for the adjacent pairs  $CD, DE$  and  $GH$  in the first genome and  $CE, GD$  and  $DH$  in the second, the gene orders are the same. We say that each order has three “breakpoints” with respect to the other. The number of such breakpoints is often considered an index of how much genome rearrangement has occurred (e.g. [2, 10]). Figure 2 displays the quantity  $B(x, y)$  measuring the number of breakpoints between the genomes  $x$  and  $y$  when all genes absent from one or the other are excluded.



**Fig. 2.** Average number of breakpoints  $B(x, y)$  in the comparisons between groups of mitochondrial genomes. f=filamentous fungi, b=budding yeasts and *S. pombe*, and a=multicellular eukaryotes.

While Figure 2 suggests that the number of breakpoints is indicative of phylogenetic relationships, especially within the fungi, we should take into account that  $B(x, y)$  can be expected to be closely related to  $N(xy)$ , the number of genes in both genomes.

On the average, in fact, for two random genomes  $x$  and  $y$  having  $n = N(xy)$  genes in common, the number of breakpoints is  $n(n - 2)/(n - 1)$ . This can be seen by considering the probability that in a random circular genome the right-hand neighbour of gene  $i$  is gene  $i + 1$  - otherwise there is a breakpoint between the two<sup>2</sup>. For  $n \geq 2$ , this probability is obviously  $1/(n - 1)$ . Then the expected number of such genes  $i$  is  $n/(n - 1)$ , and the expected number of breakpoints is then  $n - n/(n - 1) = n(n - 2)/(n - 1)$ , which is very close to  $n - 1$  for the range of  $n$  that interests us.

Figure 3 presents the difference between  $N(xy)$ , the number of genes in common and  $B(x, y)$ , the number of breakpoints depicted in Figure 2. It can be seen that there is little if any non-randomness detectable in the comparisons between the fungal and the other mitochondria since the distribution is centered around the value 1, but within the fungi and within the multicellular organisms, there are clearly fewer breakpoints than between random genomes since the means of the distributions are much greater than 1.

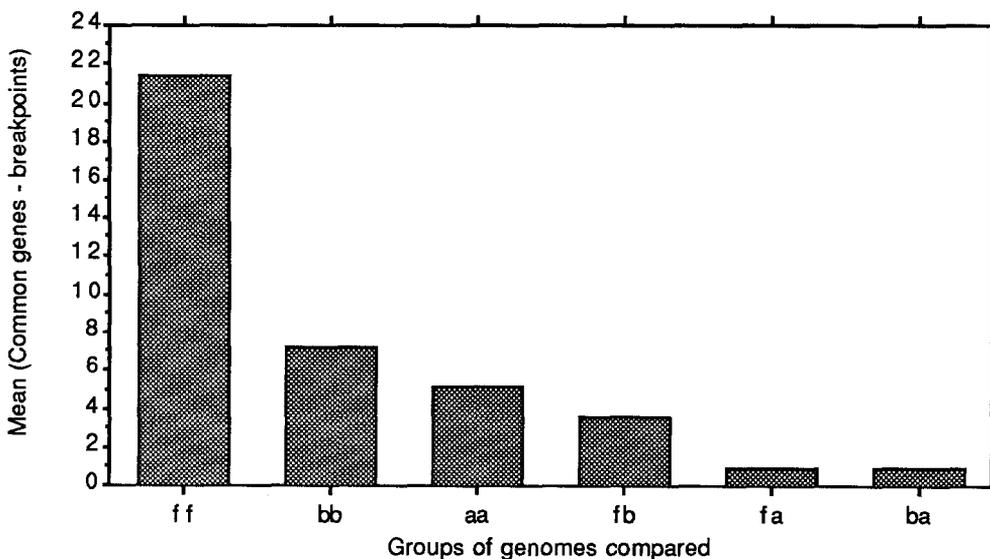


Fig. 3. Number of breakpoints compared to number of genes in common within and between groups. f=filamentous fungi, b=budding yeasts and *S. pombe*, and a=multicellular eukaryotes.

Figure 4 takes departures from randomness into account in presenting a nor-

<sup>2</sup> When, as a result of inversion, a genome contains genes of both left-to-right and right-to-left orientations, the latter is indicated by a minus sign. Thus if the  $j$ -th gene has right-to-left orientation, we write  $-j$ , and the condition for a breakpoint is the the right-hand neighbour not be  $-(j - 1)$

malised value, namely  $100 \times B(x, y)/N(xy)$ .

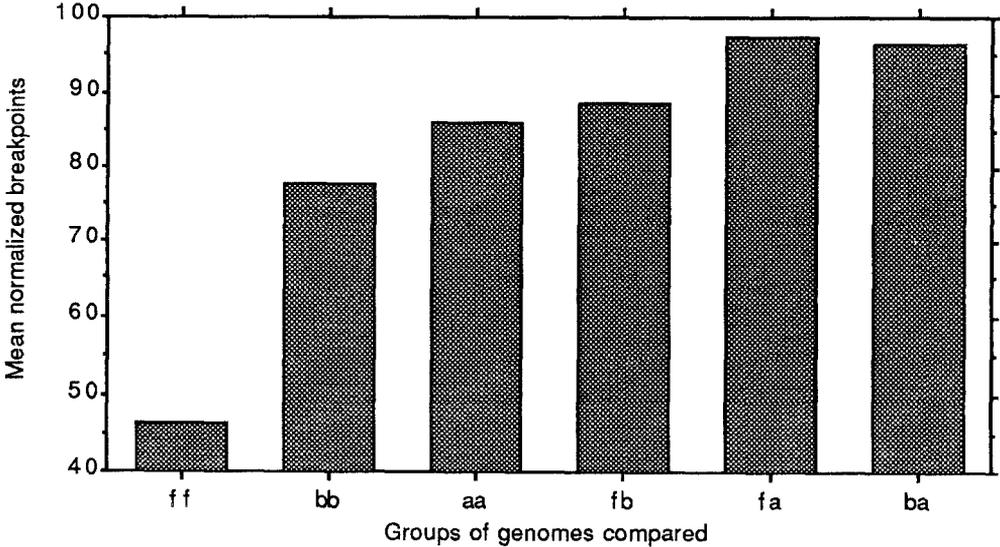


Fig. 4.  $B(x, y)$  normalized, i.e. multiplied by  $100/N(xy)$ . f=filamentous fungi, b=budding yeasts and *S. pombe*, and a=multicellular eukaryotes.

The normalised values in Figure 4 are consistent with phylogenetic relationships within the non-fungal group, within the yeast group and within the filamentous fungi. In addition the comparison between the two groups of fungi is closer than between either of them and the multicellular group.

## 6 Edit Distance

In the example at the beginning of the previous section, *ABCDEFGH* versus *ABCEFGDH*, it is clear that the second genome could have been produced from the first by the simple operation of excising the gene *D* from between *C* and *E*, and transposing it to a position between *G* and *H*. Such a transposition creates three breakpoints. If we now transposed *B* in the second genome to the position between *F* and *G*, this would produce the order *ACEFBGDH*, which has six breakpoints compared to the first order, each transposition having created three. But if instead of transposing *B*, we had transposed *C* in the second genome to the position between *F* and *G*, this would have produced *ABEFCGDH*, which has only five breakpoints with respect to *ABCDEFGH*. And had we transposed *C* to the position after *D*, this would have created *ABEFGDCH*, which has only four breakpoints with respect to the original order (or three if we don't count *DC* as a breakpoint). Thus the number of breakpoints, while roughly reflecting the degree of rearrangement in comparing two genomes, is not necessarily an accurate measure of the number of rearrangement events which have occurred to produce one genome from the other. This motivates the definition of a rearrangement edit distance between two genomes,

namely the minimal number of rearrangement events necessary to convert one to the other. In our analysis of the mitochondrial genome, we count each inversion of a number of contiguous genes, each transposition of a number of contiguous genes, and each deletion of a single gene<sup>3</sup>, as one event, though these events could well have been weighted differently.

## 7 The DERANGE Program

The inference of a minimal series of rearrangements is carried out by a program called DERANGE [9]. The key technique is that of alignment reduction. Two or more gene symbols that are adjacent in both genomes and are either of the same orientation and order, such as the four pairs linked by dotted lines in Figure 5, or of opposing orientation and in reverse order, such as the two groups of three pairs linked by solid lines in Figure 5, may be combined and replaced by a single symbol, since an optimal solution for the reduced (after this combination operation) problem is also optimal for the original (before combination) problem.

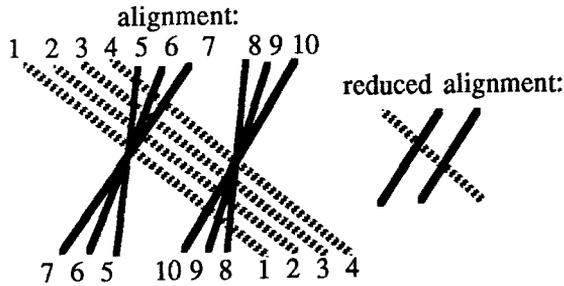


Fig. 5. Example of alignment reduction. Dotted lines represent homologous genes with same orientation in both genomes, solid lines indicate genes with opposite orientation.

Alignments may also be reduced by first applying inversion and transpositions so as to produce such combinable pairs of terms. This is illustrated with three inversions in Figure 6.

Inversions may reduce an alignment by up to two linkages, and transpositions by up to three. The weight associated with each rearrangement operation (weight = 1 in the present study) contributes to the total distance between the genomes. Reducing an alignment through the combining operation of course costs nothing, since this is just a formalism for indicating that we have obtained a partial solution to the problem. The algorithm stops when the alignment has been reduced to exactly one linkage, representing the fact that one genome has been transformed so that all its terms are in the same order as in the other genome, with the same orientation. This algorithm is a generalization of a sub-optimal method of [10] for inversions only,

<sup>3</sup> This was calculated in Section 4 and, in the present formulation, is independent of the analysis of transposition and inversion.

in the single circular chromosome case without taking account of orientation; its current implementation [9] is based on a branch-and bound search to completely solve<sup>4</sup> the general case of an unknown number of overlapping inversions (e.g. data such as that of [3], requiring five superimposed inversions), and transpositions, with user-imposed weights on the different types of events.

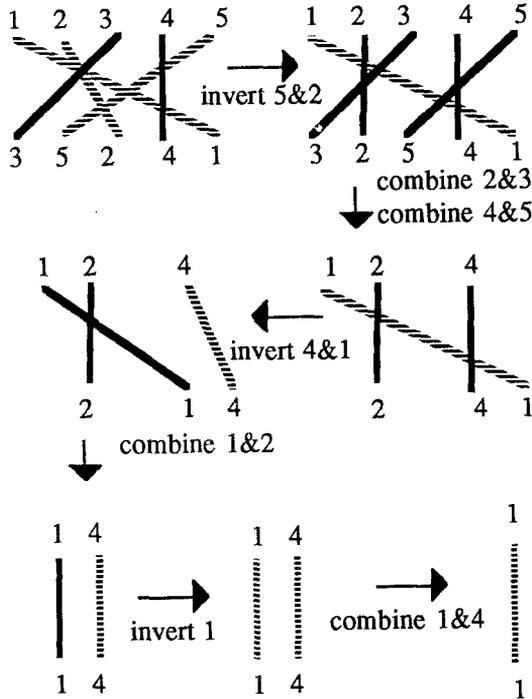


Fig. 6. Reducing an alignment using three inversions.

## 8 DERANGE Results

Figure 7<sup>5</sup> presents the  $e(x, y)$ , the output of DERANGE, once the genes absent from either member of the pair being compared are excluded, and combining adjacent genes not separated by a breakpoint.

In Figure 7, it appears as if  $e(x, y)$  contains much phylogenetic information, in that for pairs  $(x, y)$  within the two fungal groups, or between them, the distance  $e$  tends to be small, while it is much greater between fungi and multicellular organisms. The multicellular organisms are relatively diverse as measured by  $e$ , but this was

<sup>4</sup> For even the small genomes discussed here, time and space requirements are prohibitive. Thus the search is limited in a number of ways, as discussed below.

<sup>5</sup> Figure 7 represents the bulk of the computation in this research. About 150 hours of computing time were distributed among a battery of various Macintosh models.

also visible in the analysis of  $B$  in Figure 4. Indeed, it must be remembered that for a pair where  $B$  is small, that the corresponding permutation is easily converted into another may simply be due to the fact that the reduced alignment is short. Thus, similar to the way we normalised  $B$  by dividing by  $N$ , we also normalise  $e$  by dividing by  $B$ . The results appear in Figure 8.

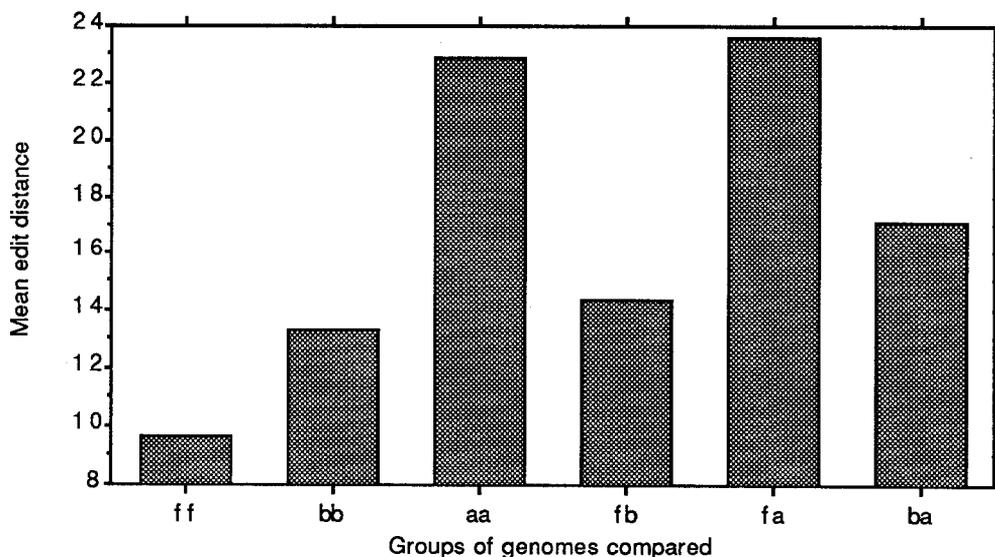


Fig. 7. Average intra- and intergroup values of  $e(x, y)$ , the number of inversions and transpositions necessary to convert genome  $x$  into  $y$  (discounting deletions). f=filamentous fungi, b=budding yeasts and *S. pombe*, and a=multicellular eukaryotes.

Normalizing the edit distances subtly changes the configuration seen in Figure 7. Within the fungi, the grouping of filamentous fungi versus the yeasts is lost. Nevertheless the fungal genomes appear consistently more closely related among themselves than they are to the non-fungal organisms.

We may conclude that  $e$  contains not only the phylogenetic information in  $B$ , but somewhat more.

## 9 Weights

As mentioned above, in our analysis of the mitochondrial genome, the two types of rearrangement event were assigned the same weight, and so was the deletion of each gene. This decision is of course somewhat arbitrary. Let us examine first the relative weights of transpositions versus inversions. Figure 9 shows some typical results of varying the relative weights on four comparisons. It can be seen that be when the effect of changing the weights is to increase the number of transpositions, the number of inversions decreases almost twice as fast.

We find that this relationship is quite general. Despite the fact that inversions can reduce an alignment by two lines, in practice few of these are found, and the

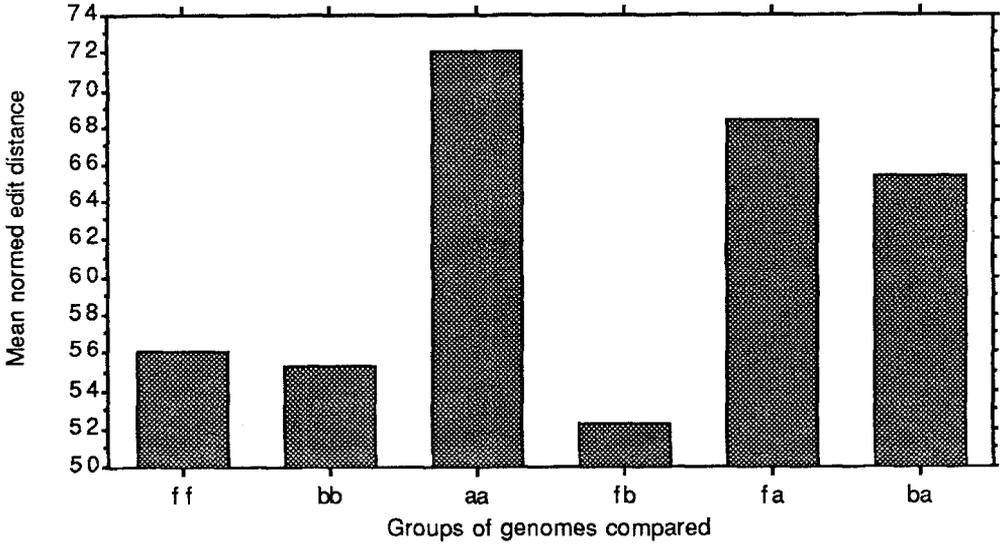


Fig. 8. Edit distances  $e(x, y)$  normalized by number of breakpoints  $B(x, y)$ .

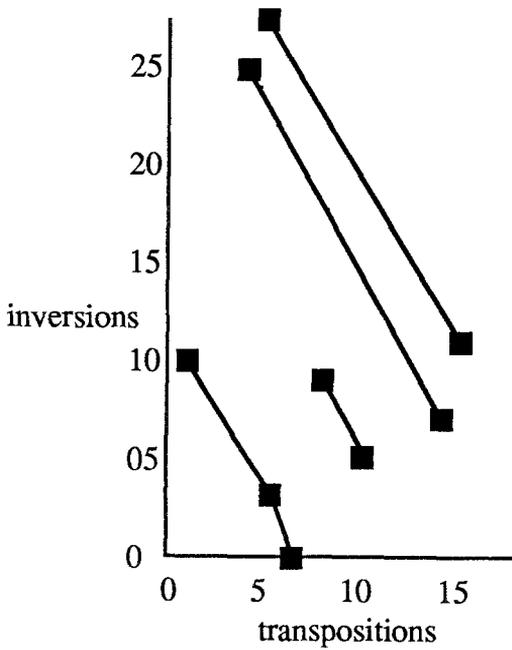


Fig. 9. Effect of changing inversion vs. transposition weights. Examples involve comparison of *Neurospora* with, from upper right to lower left, human, *Drosophila*, *S.pombe* and *K.lactis*.

remainder reduce by only one line. Similarly, transpositions reducing alignments by three lines are also rare. But transpositions saving two lines are rather more frequent, and transpositions saving only one linkage are avoided by the program for the sake of biological realism and computational efficiency. Thus the effect of  $x$  transpositions typically requires  $2x$  inversions to achieve (indeed the effect of any transposition can be achieved by two inversions). The fact that this holds widely (though not exclusively) throughout the comparisons we make means that there is no mechanism internal to the analysis which enables us to choose one weighting scheme over another.

Nevertheless, there are at least two reasons to believe that equal weights are appropriate, in addition to the simple lack of justification for favouring one over the other. First, since the fungal mitochondria show little evidence of inversion, all genes in all genomes having a common orientation, with only one exceptional gene in the data we used, we could expect transpositions to occur exclusively, or at least to predominate, in the intra-fungal comparisons. This is indeed the case, several of the comparisons indicating no inversions, and the others a relatively small number. Second, the results we have are fairly stable over a wide range of weightings. By increasing the weight on inversions rather substantially, the number of inversions remains the same or decreases relatively slowly. It is somewhat easier to increase the number of of inversions, but only by weighting transpositions very heavily, and this leads to unlikely numbers of inversions in the intra-fungal comparisons.

As for gene deletion, its analysis could be integrated with that of inversion and transposition, but this does not seem biologically justifiable at the present time. If it were to be found that deletion often involves several contiguous genes at the same time, the cost of deletion and insertion of  $k$  contiguous genes should not be simply  $k$  times the cost for one gene, but a more slowly growing (concave) function of  $k$ . In addition, the insertion or deletion of a number of contiguous genes should be allowed to occur at any time during the transformation of one genome into another, rather than calculated at the outset as we do presently, since a number of genes deleted<sup>6</sup> together at one point in 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.

## 10 Search depth

The computing requirements for DERANGE, both time and memory, grow rapidly as the number of breakpoints increases. When there are more than 12-15 breakpoints (there are up to 36 in the present study), it is no longer practical, in the current version of the program, to carry out the entire search for a minimal series of rearrangement events. Instead, the program incorporates an user-imposed limit on

<sup>6</sup> It would be important to require of such events that they involve only the genes that are present in one genome and not the other; otherwise the minimizing sequence of genome changes might artificially turn out to involve the deletion of a large segment of one genome followed by the insertion of a long segment of the other, where both segments contain many of the same genes, biologically an implausible event.

the number of possible search paths to be explored, and eliminates those that seem less likely. Figure 10 indicates the effect of varying this “search depth” parameter on five typical genome comparisons where the “true” number of events ranges from 10 to 26. It can be seen that the number of events is badly overestimated when the search depth is 50 or 100, but that it approaches the true value when the depth is set at 500 and only diminishes slowly if at all with depths over 1000. In this study, a depth of 4000 was used throughout. Though undoubtedly some of the distances are overestimated, we can be sure that these are not large errors, and all the distances are subject to proportionately the same risk of error, so that the phylogenetic results should not be biased in any particular direction.

There are a number of ways we use to check whether a comparison gives artificially high results due to an insufficient search depth. One way is to reverse the order in which the algorithm receives the two genomes to compare. If converting one genome to the other is more costly than the conversion in the opposite direction, the first of these two analyses is not optimal. Second, for selected comparisons only we can afford to increase the search depth to 10 000 or more to see if a more economical solution is found. Third, our results should obey the triangular inequality: if converting  $x$  to  $z$  costs more than the cost of converting  $x$  to  $y$  plus the cost of converting  $y$  to  $z$ , then the  $x$  to  $z$  result is not optimal.

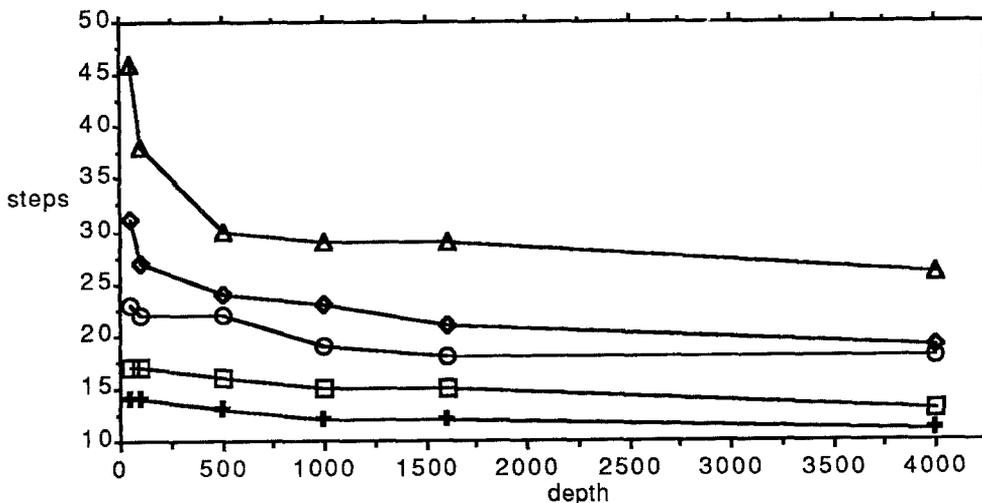


Fig. 10. The effect of limiting search depth on the number of events inferred. Comparison of humans with, from top to bottom, *Ascaris*, *S. pombe*, *S. purpuratus*, *Drosophila* and *Allomyces macrogynus*.

## 11 Nonrandomness

In Section 8, we found that normalizing  $e$  removes much of the apparent phylogenetic information it contains. We may go further and ask whether the edit distance

$e(x, y)$  between the various pairs  $(x, y)$  of mitochondrial genomes is significantly below the random noise level, in analogy to the discussion above of the departure from randomness of the breakpoint number  $B(x, y)$  for closely related genomes. Because no analytical results are available for  $e$ , we generated a number of pairs of random permutations of various lengths, combined adjacent genes not separated by a breakpoint, and submitted them to the same analysis as the mitochondrial genomes. The results will be reported elsewhere, but it is clear that comparisons within and between the two fungal groups are clearly non-random and that the comparisons involving multicellular organisms are generally indistinguishable from random.

## 12 Genomic Distance

We have seen that there are three independent aspects of the differences among genomes, measured by  $D$ ,  $B$  normalized by  $N$ , and  $e$  normalized by  $B$ . Each of these is pertinent to different subgroupings of the genomes.  $D$  clearly distinguishes the fungal mitochondria from the others and within the fungi, sets the yeasts apart from the others.  $B$ , when normalized by  $N$ , also distinguishes the yeasts from the other fungi, but shows much less striking homogeneity within the multicellular organisms. The only clear distinction made by  $e$ , when the effects of  $B$  are normalized away, distinguishes the fungi as a whole from the other organisms, but fails to capture the division within fungal group.

It is not our purpose, however, to see which of these three measures contains the most phylogenetic information. Rather we set out to construct a global measure of genome rearrangement and, now that we have characterised the three components of the distance between genomes, we may combine them in a principled way.

Figure 11 shows the results of applying the formula

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

to the 55 pairs of genomes in our data set. The distance  $d = D + e$  takes into account deletion through  $D$ , transposition and inversion through  $e$ . Of course  $B$  also measures genomic differentiation, but as we have seen,  $e$  subsumes the information in  $B$ , and contains additional subgrouping information. The problem of the sensitivity of  $e$  (and  $B$ ) to the length of the permutation is taken care of by the combination with  $D$ , which is also sensitive to this, but in a compensating direction. This is not to say that there would be no better way to differentially weight  $D$  and  $e$ , but there is as yet no justification for any particular weighting.

Figure 11 depicts more clearly than any we have seen in previous sections, the relatively low within-group distances for all three groups (ff, bb, and aa), and the somewhat closer relationship among the two fungal groups (fb) than between them and the multicellular group (fa, ba).

## 13 Conclusions

While gene sequences may in many cases serve as good indicators of evolutionary relationships, they do not in themselves bear on many phylogenetically important

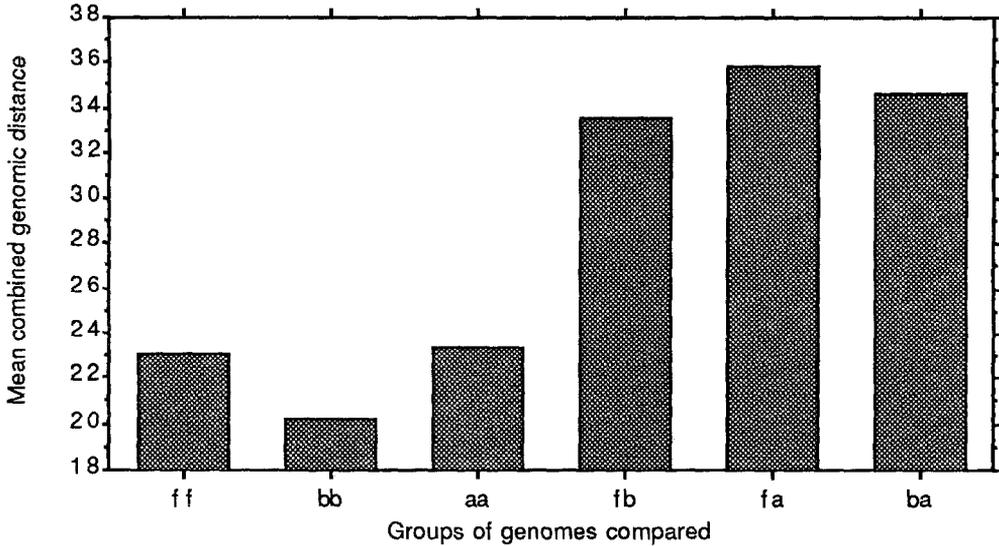


Fig. 11. Genomic rearrangement distance between pairs of genomes.

events. For example, gene-based phylogenies are constructed from the analysis of point mutations, with only minor insertions and/or deletions also being tabulated. In recent years, however, we have seen that evolutionarily relevant processes such as inversion, duplication, transposition, reciprocal translocation and deletion of entire segments of chromosomes, as well as recombination resulting in gene conversion generate much greater change, which is ignored (and in fact is effectively invisible) in gene sequence comparisons. In addition, in phylogenies near the radiative origin of many orders, phyla and kingdoms, current reconstructions of the branching order are too imprecise to provide clear distinctions among alternate tree topologies. Changes in gene sequence during the relatively short periods of time involved are often not extensive enough, and too difficult to infer precisely, to distinguish between opposing scenarios of evolution. On the other hand, major events at the level of genome arrangement may be able to discriminate decisively among competing phylogenetic accounts. In a broader sense, an understanding of the role of genome-level processes in evolutionary divergence and their consequences for the organization of life processes is of crucial importance to biology. Organellar DNAs are sufficiently small that one can realistically hope to have available the complete sequence of a great many of them in the near future, drawn from phylogenetically key organisms, thereby obtaining the type and amount of data required to evaluate these issues.

On the methodological level, progress is being made on the general problem of minimally generating elements of the group of permutations using inversions only, which is very closely related to the task of DERANGE. Better bounds for branch-and-bound algorithms, approximation algorithms and the statistics of solutions to this problem have been obtained by J. Kececioğlu [7].

## References

1. Cedergren, R., Abel, Y. and Sankoff, D. (1991) in *Molecular Techniques in Taxonomy* (G. M. Hewitt, A.W.B. Johnston and J.P.W. Young, eds.) Springer-Verlag. 87-99.
2. Nadeau, J.H. and Taylor, B.A. (1984) *Proceedings of the National Academy of Sciences USA* 81, 814-8.
3. Palmer, J.D., Osorio, B. and Thompson, W.F. (1988) *Current Genetics* 14, 65-74.
4. Sankoff, D. (1989) *Bulletin of the International Statistical Institute* 47.3, 461-475
5. Sankoff, D., Cedergren, R. and Abel, Y. (1990) in *Methods in Enzymology* 183. *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, (R.F. Doolittle, ed.) Academic Press. 428-438.
6. Sankoff D. and Goldstein, M. (1988) *Bulletin of Mathematical Biology* 51, 117-124.
7. Sankoff, D., Kececioglu, J. and Leduc, G. (1992) manuscript in preparation.
8. Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., and Cedergren, R. (1992) *Proceedings of the National Academy of Sciences USA* to appear.
9. Sankoff, D., Leduc, G., and Rand, D. (1991) DERANGE. Minimum weight generation of oriented permutation by block inversions and block movements. Macintosh application, Centre de recherches mathématiques, Université de Montréal.
10. Watterson, G.A., Ewens, W.J., Hall, T.E. and Morgan, A. (1982) *Journal of Theoretical Biology* 99, 1-7.