J Mol Evol (1999) 49:193–203

JOURNAL OF MOLECULAR EVOLUTION

Gene Order Breakpoint Evidence in Animal Mitochondrial Phylogeny

Mathieu Blanchette,[1] Takashi Kunisawa,[2] David Sankoff[3]

[1] Department of Computer Science & Engineering, University of Washington, Seattle, WA 98195-2350, USA
[2] Department of Applied Biological Sciences, Science University of Tokyo, Noda 278, Japan
[3] Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7, Canada

Received: 15 July 1998 / Accepted: 5 March 1999

Abstract. Multiple genome rearrangement methodology facilitates the inference of animal phylogeny from gene orders on the mitochondrial genome. The *breakpoint distance* is preferable to other, highly correlated but computationally more difficult, genomic distances when applied to these data. A number of theories of metazoan evolution are compared to phylogenies reconstructed by ancestral genome optimization, using a minimal total breakpoints criterion. The notion of *unambiguously reconstructed segments* is introduced as a way of extracting the invariant aspects of multiple solutions for a given ancestral genome; this enables a detailed reconstruction of the evolution of non-tRNA mitochondrial gene order.

Key words: Genomic distance — Genome rearrangement — Breakpoint analysis — Mitochondrial gene order — Conserved segment — Metazoan phylogeny

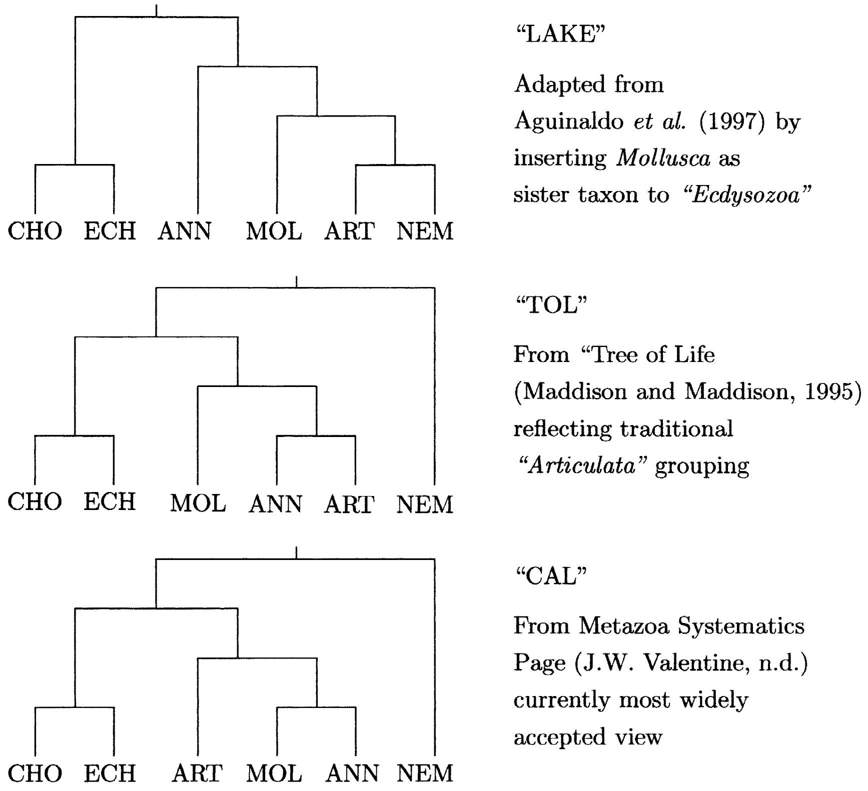
Introduction

In comparative genomics, the quantitative comparison of gene order differences can be used for phylogenetic inference about a set of organisms. This generally involves methods based on distance matrices (e.g., Sankoff et al. 1992), though it would be of more interest to employ a method which reconstructs some aspect of the genome at each ancestral node as an essential part of the optimization of a global objective function on the set of trees (e.g., Sankoff et al. 1996). Unfortunately, most of the distances of comparative genomics [e.g., the minimum edit distances calculated by Hannenhalli and Pevzner (1995a, b)] are conducive only to the distance-matrix approach, simply because the generalization to the comparison of more than two genomes has not proved computationally feasible for even moderately sized genomes (cf. Caprara 1999). An exception to this is the breakpoint distance (Watterson et al. 1982). Though an NP-hard problem (Pe'er and Shamir 1998), generalization of breakpoint distance to the simultaneous comparison of three or more genomes—multiple genome rearrangement—can be reduced to an instance of the Traveling Salesman Problem (TSP), which is quite tractable for moderate-size genomes (Sankoff and Blanchette 1997).

We have shown how to incorporate multiple genome arrangement into an iterative heuristic for the optimization of ancestral genome reconstruction on a fixed-topology phylogenetic tree, demonstrated through simulation the precision that can be achieved through careful initialization, and assessed the relative proximity of the different optimal solutions (Blanchette et al. 1997; Sankoff and Blanchette 1998a). The present paper represents the first application of this methodology to real data, an investigation of animal phylogeny through the gene order on mitochondrial genomes.

We first discuss metazoan phylogeny, including the ongoing debates about the relationships among the major metazoan branches, in the following section, as well as the available genomic data. We then situate the breakpoint distance in the context of genomic distances in general (under Genome Rearrangement Distances) and evaluate its significance in relationship to other measures

Correspondence to: D. Sankoff; *e-mail:* sankoff@ere.umontreal.ca



"LAKE"

Adapted from Aguinaldo *et al.* (1997) by inserting *Mollusca* as sister taxon to *"Ecdysozoa"*

"TOL"

From "Tree of Life (Maddison and Maddison, 1995) reflecting traditional *"Articulata"* grouping

"CAL"

From Metazoa Systematics Page (J.W. Valentine, n.d.) currently most widely accepted view

Fig. 1. Three alternative views of metazoan evolution.

when applied to the data. In the section on Tree Inference we compare a number of theories of metazoan evolution to phylogenies reconstructed from the breakpoint distance matrix and from ancestral genome optimization. Here we also discuss the effects of small gene (i.e., tRNA) mobility and the effect of "long branches," i.e., highly divergent genomes. Under Nonuniqueness, we introduce the notion of "unambiguously reconstructed segments" as a way of extracting the invariant aspects of multiple solutions for a given ancestral genome and apply this to characterize rather closely the evolution of non-tRNA mitochondrial gene order.

The Mitochondrial Genome and Problems in Animal Phylogeny

Our goal here is to investigate gene order evidence pertinent to the phylogenetic relationships among the following major metazoan groupings: chordates (CHO), echinoderms (ECH), arthropods (ART), mollusks (MOL), annelids (ANN), and nematodes (NEM). Aspects of metazoan phylogeny are controversial; among the groupings analyzed here, only the link of echinoderms and chordates seems undisputed. Many scholars would group annelids and mollusks as sister taxa, with arthropods related to these at a deeper level. Nematodes would represent the earliest branch on the metazoan phylogeny. But Rouse and Fauchald (1995) have proposed reviving a traditional grouping (*Articulata*) of annelids and arthropods as sister taxa, which had been discredited by Eernisse et al. (1992) and others. Lake has recently advocated a radical change linking arthropods and nematodes (Aguinaldo et al. 1997). Contending trees for metazoan phylogeny are compared in Fig. 1. One of our goals here is to investigate how mitochondrial gene order evidence discriminates among various theories of metazoan evolution. Another will be to see to what extent ancestral genomes can be reconstructed during phylogenetic inference.

As of April 1998, mitochondrial gene order was known for 55 metazoan species, including 36 chordates, 7 arthropods, 1 annelid, 5 echinoderms, 3 mollusks, and 3 nematodes. Thirty-seven genes are present in all species, except for atp8 (i.e., ATPase 8), absent from the nematode genomes, and one of the two tRNA-Ser and one of the two tRNA-Leu genes, absent from the snail *Cepaea nemoralis.*

To simplify the analysis and presentation, while retaining as much phylogenetic information as possible, we included in our analyses exemplars of the most diverse members of each group, excluding closely related species with identical or nearly identical mitochondrial gene orders. For example, the various chordate mitochondrial gene orders differ from the human order by one or two inversions or transpositions, so we retained only the human one for our analysis. Similarly we selected only one insect, two echinoderms, and two nematodes. The spe-

Table 1. Mitochondrial genomes compared in this investigation, with assumed monophyletic groupings[a]

	Organism		Group
HU	Human	CHO	Chordate
SS SU	*Asterina pectinifera* (sea star) *Strongylocentrotus purpuratus* (sea urchin)	ECH	Echinoderms
DR AF	*Drosophila yakuba* (insect) *Artemia franciscana* (crustacean)	ART	Arthropods
AC CN KT	*Albinaria coerulea* (snail) *Cepaea nemoralis* (snail) *Katharina tunicata* (chiton)	MOL	Mollusks
LU	*Lumbricus terrestris* (earthworm)	ANN	Annelid
AS OV	*Ascaris suum* *Onchocerca volvulus*	NEM	Nematodes

[a] Citations: HU, Anderson et al. (1981); SS, Asakawa et al. (1993); SU, Jacobs et al. (1988); DR, Clary and Wolstenholme (1985); AF, Perez et al. (1994); AC, Hatzoglou et al. (1995); CN, Terrett et al. (1996); KT, Boore and Brown (1994); LU, Boore and Brown (1995); AS, Okimoto et al. (1992); OV, Keddie and Unnasch (n.d.).

cies studied are listed in Table 1. The table also presents the major groups for which we assume monophyly in some of our analyses to reduce the size of phylogenetic computations.

Genome Rearrangement Distances

Edit Distances

The algorithmic study of comparative genomics has focused on inferring the most economical explanation for observed differences in gene orders in two genomes in terms of a limited number of rearrangement processes. For single-chromosome genomes such as in the mitochondrion, this has been formulated as the problem of calculating an edit distance between two circular permutations of the same set of genes. For these purposes, degradation of homology at the sequence level of individual genes is not pertinent; once homology is established, the two genes are considered to be identical. A sign (plus or minus) is associated with each gene in a genome, representing the direction, or orientation, of its transcription. The elementary edit operations may be as follows.

Inversion, or reversal, of any number of consecutive terms, which also reverses the polarity of each term within the scope of the inversion. Increasingly efficient exact algorithms for this problem have been given by Kececioglu and Sankoff (1994), Hannenhalli and Pevzner (1995a), Berman and Hannenhalli (1996), and Kaplan et al. (1997), whose version runs in quadratic time.

Transposition of any number of consecutive terms from their position in the order to a new position between any other pair of consecutive terms. This may or may not also involve an inversion. No efficient exact algorithm is available for this problem.

A combination (weighted) of the above two. Sankoff et al. (1992), Sankoff (1992), Blanchette et al. (1996), and Gu et al. (1997) implemented and applied heuristics to compute edit distances which combine inversions and transpositions, in some cases also allowing deletions, and differential weighting of all these operations.

genome A 1 2 3 | 4 | 5 | 6 7 |

genome B −2 − 1 | − 5 | 6 7 | 4 | − 3

Fig. 2. Defining breakpoints for circular genomes. Reading direction assumed to be left to right for + genes. Thus gene 7 precedes, and is read immediately before, gene 1 in genome A, gene −3 precedes −2 in genome B, but 2 is read immediately before 3. There are four breakpoints; their positions are indicated by *vertical lines.*

There are a number of problems associated with the use of these distances. The first is that there is no a priori reason for using one of them rather than another, e.g., transposition distance instead of inversion distance. Even for combined distances, the appropriate weights for the different operations may differ from context to context. Second, the reconstruction of evolutionary history implicit in calculating the distance is biased toward too few events and is highly nonunique. Third, no exact algorithm is known for extending the distances to three or more genomes.

Breakpoint Analysis

Consider two genomes $A = a_1 \ldots a_n$ and $B = b_1 \ldots b_n$ on the same set of genes $\{g_1, \ldots, g_n\}$, where each gene is signed (+ or −), though we may suppress the + without ambiguity. We say a_i precedes a_{i+1} in *A*, and a_n precedes a_1, as illustrated in Fig. 2. If gene *g* precedes *h* in *A* and neither *g* precedes *h* nor $-h$ precedes $-g$ in *B,* they determine a breakpoint in *A.* We are interested in the number of breakpoints in *A*, which is clearly equal to the number of breakpoints in *B.*

The number of breakpoints between two genomes not only is the most general measure of genomic distance, requiring no assumptions about the mechanisms of genomic evolution (inversion versus transposition) underlying the data, but also is the easiest to calculate (linear in *n*).

Table 2. Distance matrices for all gene (triangular matrices, top to bottom: breakpoints, minimal inversion, combined inversion/transposition)[a]

	AS-35.1	OV-35.3	HU-30.0	SS-33.4	SU-33.3	DR-30.9	AF-31.3	AC-35.1	CN-33.0	KT-29.1	LU-31.7
AS											
OV	25										
HU	36	36									
SS	36	36	27								
SU	36	36	26	6							
DR	36	36	21	33	33						
AF	36	36	23	33	33	6					
AC	35	36	36	35	35	35	35				
CN	33	34	34	33	33	33	33	7			
KT	34	33	29	34	34	22	23	34	32		
LU	34	35	32	34	34	29	30	34	31	24	
AS		21	35	33	36	36	35	35	32	33	33
OV	21.0		35	33	35	33	34	35	33	32	33
HU	34.3	35.3		28	24	20	23	35	33	28	31
SS	32.5	31.6	26.1		5	33	34	35	32	33	34
SU	34.3	34.2	24.2	3.0		32	33	34	31	32	32
DR	34.2	33.4	19.4	31.2	32.2		6	35	32	21	28
AF	35.2	33.3	21.2	32.0	31.3	4.0		34	31	21	28
AC	34.1	34.1	34.2	33.3	32.2	33.4	32.5		8	34	34
CN	31.3	31.3	33.2	30.5	29.5	30.6	31.2	5.1		31	29
KT	32.2	30.3	28.2	32.3	32.4	19.3	21.2	32.3	31.2		22
LU	32.5	32.4	30.4	32.1	32.2	26.3	26.5	33.2	29.4	22.2	

[a] Average breakpoint distance to nongroup members (e.g., excluding MOL comembers CN and KT for AC, excluding ECH comember SU for SS, and no exclusions for HU) on the top row. Note that AS and OV each have only 36 genes and CN has 35, while the other genomes have 37.

Empirical Comparison of the Distances

There are 55 comparisons among the genomes in the data. Table 2 contains the results of calculating the number of breakpoints, the minimal inversions distance, using Hannenhalli's (n.d.) software, and a combined inversion and transposition distance, the latter with a relative cost of 2.1 imposed on transpositions [see Blanchette et al. (1996) for the method and a justification of this parameter value]. It can be seen from the number of breakpoints that many of the gene orders seem to be random or near-random permutations of each other. (Random genomes with n genes would have $n - \frac{1}{2}$ breakpoints with each other, on the average.) Some of this is due simply to the high rates of rearrangement in some groups—nematodes, echinoderms, snails—as roughly indicated by their average breakpoint distance to nongroup members. Another contribution to randomness is the relatively high mobility of tRNA genes within the genome. When the calculations are repeated after deleting the tRNAs, the results are given in Table 3. Figure 3 is a scattergram of the data in both Table 2 and Table 3 of the breakpoint distance and the combined inversion/transposition distance. (An almost-identical pattern is revealed with simple inversion distance versus breakpoints.) Clearly the number of breakpoints is highly predictive of the other distances, though theoretically (Watterson et al. 1982), all that can be said is that (breakpoints/2) $\leq$ inversion distance.

Tree Inference

In this section we compare, in the light of theories of metazoan evolution, three criteria for optimum tree topology: neighbor joining, Fitch–Margoliash normalized sum of squared errors, and minimum breakpoint. The first two operate on the genome data as reduced to the breakpoint distance matrix in Table 2, slightly modified as described below; the third is based on the gene orders themselves. The first two produce ancestral nodes characterized only by their linear distance from neighboring (colinear) nodes; the third actually reconstructs hypothetical ancestral genomes which are jointly optimal with the tree topology.

A modified data set is used for these analyses so that each genome contains the same number of genes. This means adding two tRNA genes to the CN genome—it is relatively clear where these should go to minimize changes in the pattern of distances, through analogy with the related AC genome—and adding an atp8 gene to the nematode genomes or deleting this gene from all the other genomes. Since different choices of insertion location for atp8 in the nematodes have different consequences for the entire distance matrix, we repeated our calculations, first with atp8 deleted from all genomes and then with atp8 inserted in the nematode genomes directly adjacent to the atp6 gene, where it is most often located in the other genomes, including the conservative *Katharina tunicata* genome. There are two reasons for these

Table 3. Distance matrices omitting tRNA genes (triangular matrices, top to bottom: breakpoints, minimal inversion, combined inversion/transposition)[a]

	AS-13.3	OV-13.4	HU-9.1	SS-12.0	SU-11.8	DR-10.1	AF-10.1	AC-14.3	CN-13.6	KT-8.9	LU-11.4
AS											
OV	6										
HU	13	14									
SS	14	14	6								
SU	13	13	6	2							
DR	13	13	5	10	10						
AF	13	13	5	10	10	0					
AC	14	14	14	15	15	14	14				
CN	14	14	13	14	14	13	13	3			
KT	13	13	6	11	11	5	5	14	13		
LU	13	13	9	14	14	8	8	14	14	7	
AS		5	11	12	11	11	11	12	13	12	11
OV	4.2		13	12	11	11	11	12	12	12	11
HU	11.4	13.5		5	6	3	3	12	12	5	7
SS	12.3	12.2	4.1		1	7	7	14	12	9	11
SU	11.1	11.5	4.2	1.0		7	7	14	13	9	11
DR	11.3	11.3	3.0	7.1	7.1		0	12	12	4	7
AF	11.3	11.3	3.0	7.1	7.1	0.0		12	12	4	7
AC	12.4	12.1	12.2	14.2	14.2	12.2	12.2		3	11	12
CN	13.3	12.1	12.1	12.3	13.2	12.3	12.3	2.1		11	10
KT	12.2	12.2	5.1	9.1	9.2	3.1	3.1	11.2	11.2		5
LU	11.3	10.5	7.2	11.2	11.3	7.2	7.2	12.2	10.2	5.1	

[a] Average breakpoint distance to nongroup members (e.g., excluding MOL comembers CN and KT for AC, excluding ECH comember SU for SS, and no exclusions for HU) on the top row. Note that AS and OV each have only 14 genes, while the other genomes have 15.

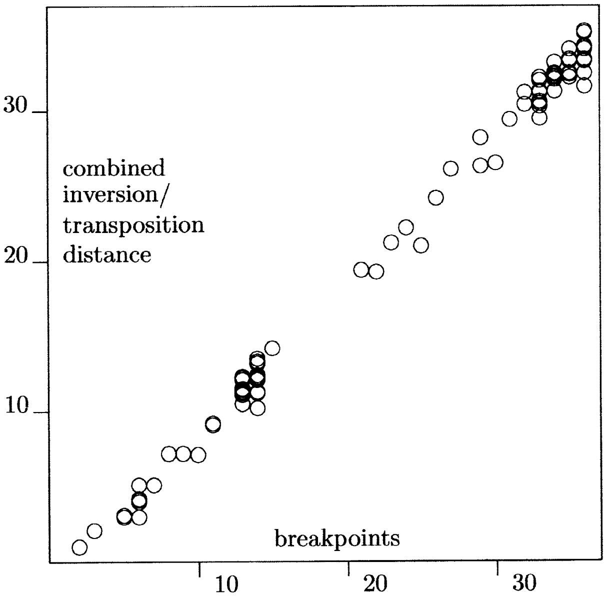


Fig. 3. Relationship of breakpoint distance to combined inversion/transposition distance. Data from Tables 2 and 3.

modifications: first, they remove biases in all the analyses due to unequal numbers of genes—genomes with fewer genes would otherwise have systematically lower breakpoint distances and could thus tend to gravitate to a more central part of the tree than where they should be located. The second reason is purely technical and is discussed under Minimal Breakpoint Phylogeny (below).

All the methods produce unrooted trees.

Distance-Matrix Methods

Neighbor-joining analysis, either with or without the atp8 data, produces the tree in Fig. 4a. This tree disrupts the deuterostomes by grouping the arthropods with the human genome and, less problematic, disrupts the molluscs by grouping *Katharina tunicata* with the annelid *Lumbricus terrestris.*

The Fitch–Margoliash routine, which minimizes the sum of squared differences between distance matrix entries and total path length in the tree between two species, divided by the square of the matrix entry, produces the tree in Fig. 4b. The same tree is produced whether or not the atp8 data are included.

This tree also disrupts the deuterostomes by grouping the arthropods with the human genome and disrupts the mollusks by branching the echinoderms between the snails and the annelid. Indeed, the Fitch–Margoliash tree must be considered considerably worse than the neighbor-joining one; its branching order reflects little more than the overall rate of evolution of the lineages as measured by the average breakpoint distance in Table 2. The rapidly evolving lineages, nematodes, snails, and echinoderms, are grouped together, and the more conservative lineages are grouped together, thus completely disrupting both the deuterostome (CHO+ECH) grouping and the MOL grouping.

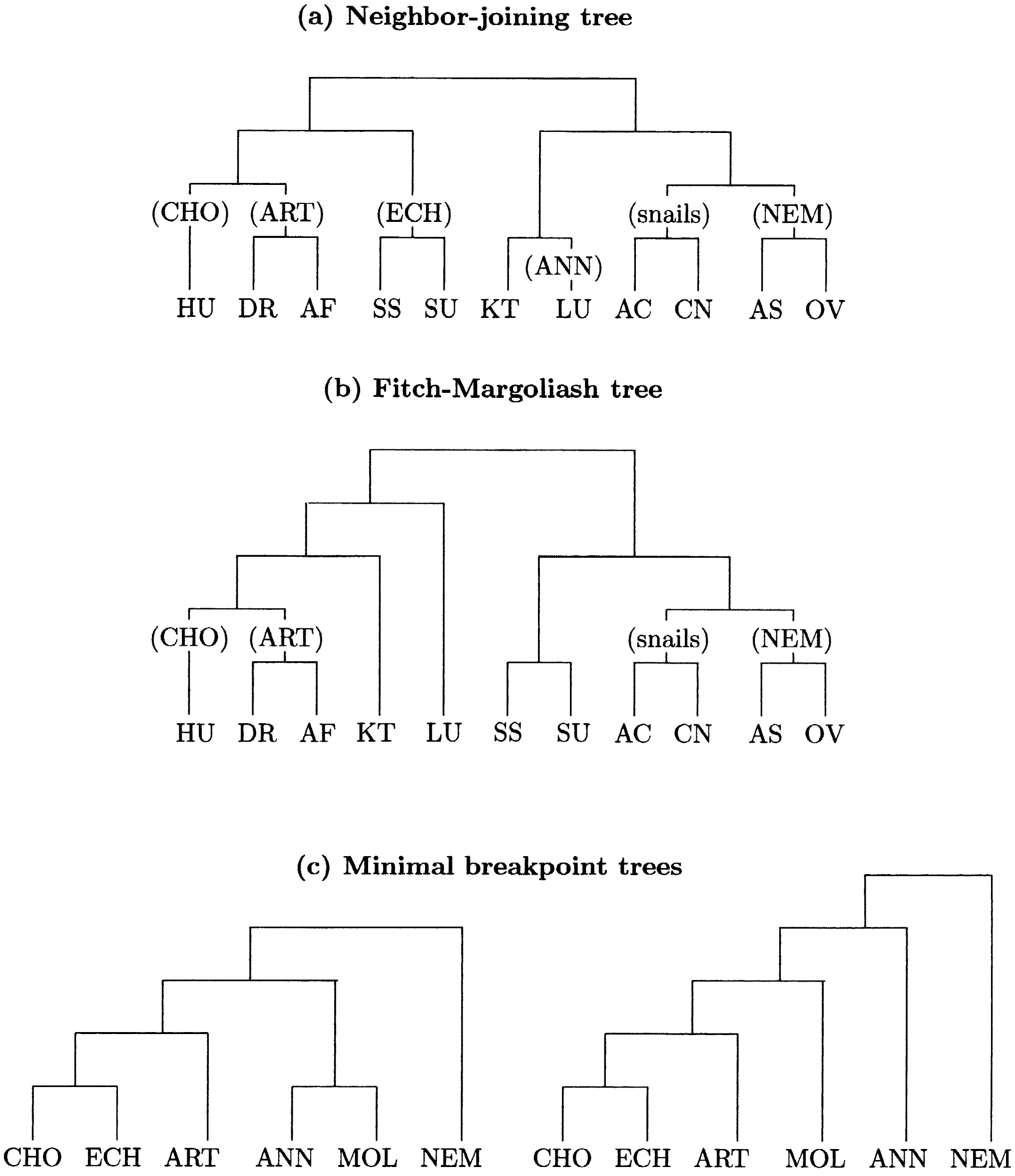

(a) Neighbor-joining tree
(CHO) (ART) (ECH) (snails) (NEM) (ANN)
HU DR AF SS SU KT LU AC CN AS OV
(b) Fitch-Margoliash tree
(CHO) (ART) (snails) (NEM)
HU DR AF KT LU SS SU AC CN AS OV
(c) Minimal breakpoint trees
CHO ECH ART ANN MOL NEM
CHO ECH ART MOL ANN NEM

Fig. 4. Comparison of trees produced by three methods from complete mitochondrial genome orders. Branches not drawn to scale. All the methods produce unrooted trees. The hierarchy is drawn for maximum comparability with the trees in Fig. 1, but the results are equally compatible with roots placed on any edge of the tree.

Without atp8 data, the Fitch–Margoliash normalized sum of squared differences is 0.401 for this tree, while for the trees in Fig. 1, it is 0.759 for CAL, 0.764 for TOL, and 0.765 for LAKE. With the atp8 data included, it is 0.359 for the tree in Fig. 4b, 0.694 for CAL, 0.702 for TOL, and 0.708 for LAKE. Thus, if we wished to restrain our choice of phylogeny to one of the three theories represented in Fig. 1 on the basis of Fitch–Margoliash, the CAL tree would be favored.

Minimal Breakpoint Phylogeny

A minimum breakpoint tree is one in which a genome is reconstructed for each ancestral node, the number of breakpoints is calculated for each pair of nodes, ancestral or given, directly connected by a branch of the tree, and the sum is taken over all branches, where this sum is minimal over all possible trees. This problem may be decomposed into an inner and outer component.

The inner problem starts with a given topology, or branching structure, for the tree and optimizes the ancestral genomes. In our method, the key to this solution is a technique for multiple genome rearrangement consensus: given three or more known genomes, find the median—the genome such that the sum of the number of breakpoints between it and each of the given genomes is minimal. Our solution to this, developed and tested over the past 2 years (Sankoff and Blanchette 1997, 1998a, 1999;

Blanchette et al. 1997) is based on a reduction to the Traveling Salesman Problem. As a first step toward a more widely applicable procedure, a program implementing this method runs relatively rapidly, as long as each of the three genomes contains the same set of genes, and this set is not too large (37 genes is well within its capabilities). While the algorithm also works for unequal gene sets, it is unwieldy for genomes containing more than about 20 genes. This is one of the reasons for equalizing the gene content across genomes prior to the analysis. New research promises a faster algorithm for the case of unequal gene sets in the near future.

The given phylogeny is then decomposed into a set of overlapping multiple genome rearrangement consensus problems, each one defined by one of the ancestral nodes as the median, to be found, with all the colinear nodes, ancestral or given, as the "known" genomes. With suitable initialization of the ancestral genomes (Blanchette et al. 1997), successive solution of all the overlapping consensus problems leads, after very few iterations, to convergence. In this study, we calculate the ancestral genomes in all trees using three different initializations and five passes of the successive optimization procedure. In the rare case where the three results are not identical, we retain the best one. Previous simulations (Blanchette et al. 1997) show that this virtually always succeeds in finding the global optimum for this size of problem.

To solve the outer problem, we simply evaluate every possible tree on the set of given data genomes. Since we are not questioning on the basis of our data the major groupings in Table 1, we discard all trees that disrupt them, leaving a total of 105 unrooted binary branching trees. Auxiliary tests (allowing, one at a time, ART, ECH, NEM, snails, or MOL to be ungrouped) indicate that all the assumed groupings are robust, with the exception of the conservative *Katharina tunicata* in MOL, which does not necessarily group with the highly diverged snails.

As with the neighbor-joining and Fitch–Margoliash methods, we first carried out our analysis on the genome data without the atp8 gene, then repeated it with this gene appropriately inserted in the nematode genomes and restored to its original position in the other genomes.

Minimal Breakpoint Phylogeny for Metazoans

The scores—minimal number of breakpoints—for the 105 trees evaluated, based on the data without atp8, are distributed from 199 to 218 as in Fig. 5. With atp8, the best two trees increase their scores to 201).

The two trees in Fig. 4c are clearly optimal (both with and without the atp8 data), but neither is biologically plausible, because they give the impression either that the deuterostomes (CHO+ECH) are a late-branching sister taxon of the arthropods or, rooted differently, that nematodes constitute a late-branching sister taxon to the annelids (or of ANN+MOL). Note, however, that in both of these trees, the deuterostome taxon is correctly found, while it did not emerge from either neighbor-joining or Fitch–Margoliash.

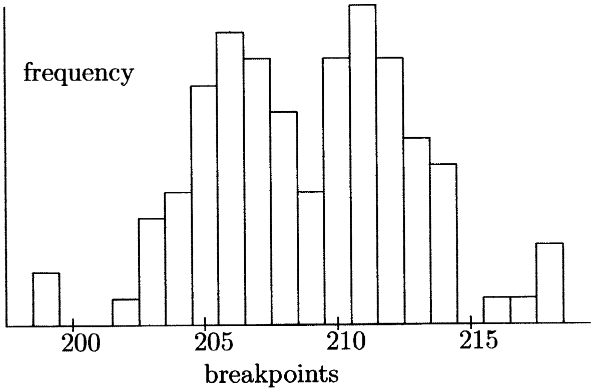
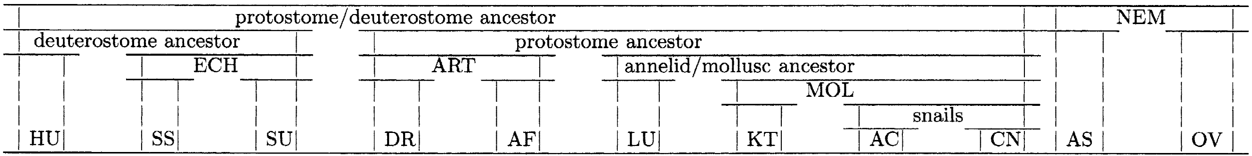


Fig. 5. Distribution of number of breakpoints in 105 trees.

What of the trees in Fig. 1? All are suboptimal; without atp8, there are no trees with scores between 199 and 202, one tree at 202, and score(CAL) = 203, score(TOL) = 204, and score(LAKE) = 206. Including the atp8 data, there are no trees with scores between 201 and 205, but score(CAL) = 205, score(TOL) = 206, and score(LAKE) = 209. Of the three theories, it is clear that these data favor CAL (tied for fourth best of 105 trees) and, somewhat less, TOL (tied for eight best of 105 trees), but the LAKE tree accounts for the configuration of gene orders about as well as a tree constructed randomly (tied for twenty-second best, along with 11 others).

To verify to what extent these results may be the result of tRNA gene mobility blurring out the conserved order of the protein coding and rRNA genes, the analysis was repeated using only the latter 15 genes. Here there was little to distinguish all the trees we have discussed, with the two in Fig. 4c and CAL scoring an optimal 59, while TOL and LAKE score 60. Including the atp8 data or not had no effect.

Returning to the full set of genes, what are we to make of the two trees in Fig. 4c? And what significance can a few points' difference in the total number of breakpoints have? To answer the first question, aspects of these trees seem to reflect the conservative versus rapidly evolving distinction among the groups, as with the neighbor-joining and, especially, the Fitch–Margoliash criteria. But surprisingly, since minimal breakpoints is a parsimony criterion, this method seems less affected than the other two by the "long branches attract" artifact, since it successfully identifies the CHO+ECH grouping. The echinoderm line has diverged almost to randomness, but the link with the human gene order is too strong for it to be detached. This result must be seen to be a strong point of the breakpoint phylogeny method.



HU		SS	ECH	SU		DR	ART	AF		LU		KT	MOL	AC	snails	CN	AS	NEM	OV
0	0	4	23	0	23	0	4	2	0	21	9	0	0	0	32	0	10	23	8
6	19	6	26	2	24	4	10	6	16	23	18	2	7	7	33	7	19	24	19
	11		4		21		1		11		12		11		7			25	
	1		2		4		1		4		2		2		1			1	
	18		6		19		4		13		8		8		3			19	

Fig. 6. Characteristics of optimal CAL trees. *First two rows:* Minimum and maximum length, among 20 optimal trees, of the branch leading upward from a specified node, i.e., toward the root. Note that the branch connecting the nematode and protostome/deuterostome ancestors is represented twice, since the precise position of the root on this branch is undetermined. *Third, fourth, and fifth rows:* The number of unambiguously reconstructed segments at ancestral nodes, considering all genes, non-tRNA genes, and only tRNA genes, respectively.

Our imposition of monophyly on the mollusks forces *Katharina tunicata* to group with the snails. However, when this constraint is relaxed, the trees in Fig. 4c remain optimal. Five other trees also score 199 when *Katharina tunicata* is unconstrained, but four of these involve only local restructuring of the tree configuration involving *Lumbricus terrestris* and *Katharina tunicata.* Just one of the seven optimal trees shows further susceptibility to the “long branches attract” artifact, as the snails break away from the MOL grouping and become attached to the echinoderms.

The second question can be answered through reference to the breakpoint distances in Table 2. All that distinguishes the comparison between *Ascaris suum,* on one hand, and *Katharina tunicata* and *Lumbricus terrestris,* on the other, from the complete randomness (36 breakpoints) of most other nematode comparisons are two instances of adjacent genes which occur in these three genomes. Boore et al. (1995) find that this fact alone is important evidence of arthropod monophyly. And it helps explain why the deep branching of both deuterostomes and nematodes cannot be detected by any method, based on these data.

Nonuniqueness

The study of genomic rearrangement inevitably encounters the problem of non-uniqueness. There are often many distinct solutions, all optimal, and many ways of arriving at these results.

For a given tree, we can construct a large number of optimal solutions for each node by running the algorithm on the same data but with different numbering of the genes, then assess which aspects of the solutions are constant and which are variable. Since it represents the biologically plausible tree most consistent with gene order data, we illustrate with the CAL model in Fig. 1.

We study two aspects of nonuniqueness, that of the branch lengths of reconstructed trees and that of the reconstructed genomes.

Branch Lengths

Twenty sets of optimal genomes were reconstructed. The CAL tree is schematized in Fig. 6. The first two rows of figures show the range of branch lengths we obtained over these 20 optimal reconstructions.

Much variability occurs with branches leading to sister terminal taxa: particularly the snails and the nematodes—a number of optimal genomes can be assigned to their immediate common ancestor, as one branch contracts and the other lengthens.

There are other sources of variability; the relatively short but variable branches in the lineage of *Katharina tunicata* are consistent with the fact that most of the variation among optimal trees involves different positioning of this species with respect to *Lumbricus terrestris.* Both types of local transformation (branch length, neighboring node interchange) in this region of the tree are compatible with optimality.

Finally, there is a great deal of variability associated with the protostome–deuterostome ancestor—it may be placed very close to the protostome ancestor or very close to the deuterostome ancestor without compromising optimality.

A consequence of this variability means that any single representation of the tree which attempts to portray branch lengths may be very misleading.

In the next section, we investigate the other aspect of nonuniqueness, the reconstructed genomes, and propose a solution to the representation of these genomes which escapes the problem of arbitrariness.

Reconstructing Ancestral Genomes

The genomes reconstructed at the ancestral nodes of a phylogeny are not generally unique; unless the genomes associated with the three adjacent nodes are all very similar, there will generally be many optimal solutions for their median. And if there are several connected ancestral nodes, the set of optimal genomes for each such node will differ depending on the specific optima chosen for the others.

Ancestral nematode: unique reconstruction

nad6 cob cox3 nad4L rns nad1 atp8 atp6 nad2 nad4 cox1 cox2 rnl nad3 nad5

Protostome/deuterostome ancestor: 4 segments

(nad2 cox1 cox2 atp8 atp6 cox3 nad3) (nad4L nad4 nad5) (nad6 cob) (rns rnl nad1)

Ancestral protostome: 4 segments

(nad2 cox1 cox2 atp8 atp6 cox3 nad3) (nad4L nad4 nad5) (nad6 cob) (rns rnl nad1)

Ancestral deuterostome: unique reconstruction

cox1 cox2 atp8 atp6 cox3 nad3 nad4L nad4 nad5 -nad6 cob rns rnl nad1 nad2

Ancestral echinoderm: 2 segments

(cox1 nad4L cox2 atp8 atp6 cox3 nad3 nad4 nad5 -nad6 cob rns) (nad1 nad2 rnl)

Ancestral arthropod: unique reconstruction

nad6 cob -nad1 -rnl -rns nad2 cox1 cox2 atp8 atp6 cox3 nad3 -nad5 -nad4 -nad4L

Annelid/mollusc ancestor: 2 segments

(nad6 cob) (-nad1 -rnl -rns cox3 nad3 nad2 cox1 cox2 atp8 atp6 -nad5 -nad4 -nad4L)

Ancestral mollusc: 2 segments

(nad6 cob) (-nad1 -rnl -rns cox3 nad3 nad2 cox1 cox2 atp8 atp6 -nad5 -nad4 -nad4L)

Ancestral snail: unique reconstruction

nad6 nad5 nad1 nad4L cob cox2 -atp8 -atp6 -rns -nad3 -cox3 nad4 nad2 cox1 rnl

Fig. 7. Unambiguously reconstructed segments (*in parentheses*). Note that the position and orientation of each segment with respect to the others are not reconstructed; only the order of the genes it contains.

It will generally be the case, however, that some aspects of the optimal genome for a node are invariant over all solutions, independent even of the particular solutions chosen for phylogenetically adjacent nodes. We can be quite certain that these aspects are correctly reconstructed.

We examined the 20 optimal solutions for each node on the CAL tree and extracted any strings of contiguous genes which recurred in all these solutions. This experiment was repeated for three sets of data as follows.

When the orderings of all mitochondrial genes are used as input, the number of invariant segments obtained for each node is given in the third row of figures in Fig. 6. It is clear that for several of the ancestral nodes, there is a great deal of uncertainty about the gene order, represented by a large number of predominantly short invariant segments.

When the 22 tRNA genes are excluded from the same analysis, however, a different picture emerges, as indicated in the fourth row in Fig. 6. There are few, much longer segments, so that the only aspects of the ancestral genomes not reconstructed are how these few segments are oriented and ordered among themselves. The segments are shown in Fig. 7.

The small numbers of unambiguously reconstructed segments are somewhat surprising, but it is confirmed by inspection of Fig. 7, which reveals patterns quite contrary to the impressions left by Tables 2 and 3 and the third row in Fig. 6. The non-tRNA mitochondrial gene order is seen to be relatively conservative.

Indeed, although we do not present the details for this limited data set, the reconstructed gene orders are relatively robust against small changes in the phylogeny, except of course in regions of the tree which are reconfigured and where there is no one-to-one correspondence between the ancestor nodes in the original and those in the modified trees.

This result, i.e., a reduced number of segments, is not due just to a reduced number of genes. When only the 22 tRNA genes are included, as in the fifth row in Fig. 6, the number of segments is proportionately still quite elevated.

It is the increased mobility of the tRNA genes which is responsible for the proliferation of alternative gene orders in the reconstructions and the consequent decrease in the length of invariant segments and increase in their number.

Conclusions

Breakpoint Distance

The relatively easy computability of multiple breakpoint distance makes it possible for the first time to do a systematic parsimony type of phylogenetic study based on genome rearrangements. The very high correlation between this distance and the edit distances hitherto used to compare genomes further validates the present approach.

Interpretation of Phylogenetic Results

How much phylogenetic history is contained in metazoan mitochondrial gene orders? What methods are most apt to infer this history correctly? Finally, what theory of metazoan evolution best accounts for the genomic data?

It is clear from the matrices in Tables 2 and 3 that a large proportion of the pairwise genomic comparisons reveals no trace of common ancestry, as far as gene order is concerned. Only the more conservative genomes retain deep phylogenetic parallels. Nevertheless, through the latter genomes and through the connections to one or another of them of the more rapidly evolving lines, all of the phylogenetic history can be inferred, with the exception of the earliest-branching nematode lineage. More genomic data from early-branching metazoan lineages, perhaps platyhelminthes, sponges, and other nematodes and more divergent deuterostomes, would resolve this difficulty.

Aside from the high rate of evolution in comparison to the metazoan time scale, the other major impediment to phylogenetic reconstruction is the difference in this rate from lineage to lineage, together with the lack of “tree additivity” in measures such as breakpoint distance, which attain a maximal upper value after a certain amount of evolution. Superficial linearizations of these measures could not compensate for the complete loss of resolution when divergences reach the level of random genomes. All the phylogenetic reconstructions we tested were susceptible to the “long branches attract” artifact resulting from these problems, especially the Fitch–Margoliash technique. The minimal breakpoint phylogeny was most resistant to this effect, and neighbor-joining was intermediate. Only the minimal breakpoint reconstruction produced results at all compatible with any major theory of metazoan evolution.

Among these theories, the currently most acceptable view represented by the CAL tree is the one most supported by the genomic data. Not only does it have near-optimal scores in terms of minimal breakpoints, with or without the atp8 data and with or without the tRNA genes, but also it scores better than TOL, and especially LAKE, when tested with the Fitch–Margoliash criterion, and among the three theories, it most resembles the neighbor-joining tree. While the LAKE hypothesis clearly has no support whatsoever from the gene order data, the TOL tree is only slightly disfavored in comparison with CAL. We emphasize again that the present study focuses entirely on mitochondrial gene orders: we do not suggest that these are superior to gene sequence evidence or other kinds of evidence; in any case, a number of additional genomes from diverse lineages are needed before confidence can be placed in this type of inference. In a related study (Sankoff and Blanchette 1998b, 1999a, 1999b) incorporating a recently sequenced hemichordate mitochondrial genome [*Balanoglossus carnosa* (Castresana et al. 1998)], the deuterostome–protostome split is even more unequivocally confirmed, although whether the hemichordate groups with the chordates or the echinoderms is not resolved.

Unambiguously Reconstructed Segments

The algorithmic reconstruction of ancestral genomes has been plagued by the problem of nonuniqueness. It is very difficult to assimilate, display and interpret the many solutions which may be simultaneously optimal. The technique we introduce here, focusing on segments which appear in all these solutions, and not trying to account for the sometimes combinatorially prohibitive number of ways they may be assembled, essentially solves this problem, at least for the kind of data studied here.

In addition, this approach provides an unexpectedly dramatic characterization of the differential evolutionary mobility of tRNA genes versus rRNA and protein-coding genes. Whereas the 60% of genes which code for tRNA account for about 70% of the evolutionary divergence, as measured by including and then excluding them from the calculation of breakpoint distance, the decrease in the number of unambiguously reconstructed segments when they are excluded is well over 80%. As a result, the relatively conservative pattern of gene order in metazoan mitochondrial genomes is highlighted.

Acknowledgments. We thank Franz Lang, Jeff Boore, and Ken Halanych for discussions of metazoan phylogeny, though we take full responsibility for the opinions expressed here. This research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. D.S. is a Fellow of the Canadian Institute for Advanced Research.

References

Aguinaldo AMA, Turbeville JM, Linford LS, et al. (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:29

Anderson S, Bankier AT, Barrell BG, et al. (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Asakawa S, Himeno H, Miura K, Watanabe K (1993) Nucleotide sequence and gene organization of the starfish Asterina pectinifera mitochondrial genome. Unpublished

Berman P, Hannenhalli S (1996) Fast sorting by reversal. In: Hirschberg D, Myers G (eds) Combinatorial pattern matching. 7th annual symposium. Lecture notes in computer science 1075. Springer, New York, pp 168–185

Blanchette M, Kunisawa T, Sankoff D (1996) Parametric genome rearrangement. Gene-Combis (on-line) and Gene 172:GC11–GC17

Blanchette M, Bourque G, Sankoff D (1997) Breakpoint phylogenies. In: Miyano S, Takagi T (eds) Genome informatics 1997. Universal Academy Press, Tokyo, pp 25–34

Boore JL, Brown WM (1994) Complete DNA sequence of the mitochondrial genome of the black chiton, Katharina tunicata. Genetics 138:423–443

Boore JL, Brown WM (1995) Complete sequence of the mitochondrial DNA of the annelid worm Lumbricus terrestris. Genetics 141:305–319

Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM (1995) Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376:163–165

Caprara A (1999) Formulations and hardness of multiple sorting by reversals. In: Istrail S, Pevzner P, Waterman M (eds) Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB 99). ACM, New York, pp 84–93

Castresana J, Feldmaier-Fuchs G, Paabo S (1998) Codon reassignment and amino acid composition in hemichordate mitochondria. Proc Natl Acad Sci USA 95:3703–3707

Clary DO, Wolstenholme DR (1985) The mitochondrial DNA molecule of Drosophila yakuba: Nucleotide sequence, gene organization, and genetic code. J Mol Evol 22:252–271

Eernisse DJ, Albert JS, Anderson, FE (1992) Annelida and Arthropoda are not sister taxa. A phylogenetic analysis of spiralian metazoan morphology. Syst Biol 41:305–330

Gu Q-P, Iwata K, Peng S, Chen Q-M (1997) A heuristic algorithm for genome rearrangements. In: Miyano S, Takagi T (eds) Genome informatics 1997. Universal Academy Press, Tokyo, pp 268–269

Hannenhalli S (no date) Software for reversal distance on signed permutations. http://www.hto-13.usc.edu/people/hannenhalli/signed_dist.c

Hannenhalli S, Pevzner PA (1995a) Transforming cabbage into turnip. (polynomial algorithm for sorting signed permutations by reversals). In: Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing, pp 178–189

Hannenhalli S, Pevzner PA (1995b) Transforming men into mice (polynomial algorithm for genomic distance problem). In: Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science, pp 581–592

Hatzoglou E, Rodakis, GC, Lecanidou R (1995) Complete sequence and gene organization of the mitochondrial genome of the land snail Albinaria coerulea. Genetics 140:1353–1366

Jacobs HT, Elliott DJ, Math VB, Farquharson, A (1988) Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. J Mol Biol 202:185–217

Kaplan H, Shamir R, Tarjan RE (1997) Faster and simpler algorithm for sorting signed permutations by reversals. In: Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms. ACM, New York, pp 344–351

Kececioglu J, Sankoff D (1994) Efficient bounds for oriented chromosome inversion distance. In: Crochemore M, Gusfield D (eds) Combinatorial pattern matching. 5th annual symposium. Lecture notes in computer science 807. Springer, New York, pp 307–325

Keddie EM, Unnasch TR (no date) Complete sequence of mitochondrial genome of Onchovcerca volvulus. Unpublished

Maddison D, Maddison W (1995) Tree of Life metazoa page. http://phylogeny.arizona.edu/tree/eukaryotes/animals/animals.html

Okimoto R, Macfarlane JL, Clary DO, Wolstenholme DR (1992) The mitochondrial genomes of two nematodes, Caenorhabditis elegans and Ascaris suum. Genetics 130:471–498

Pe'er I, Shamir R (1998) The median problems for breakpoints are NP-complete, Manuscript. University of Washington, Seattle

Perez ML, Valverde JR, Batuecas B, Amat F, Marco R, Garesse R (1994) Speciation in the Artemia genus: Mitochondrial DNA analysis of bisexual and parthenogenetic brine shrimps. J Mol Evol 38: 156–168

Rouse GW, Fauchald K, (1995) The Articulation of Annelids. Zool Scripta 24:269–301

Sankoff D (1992) Edit distance for genome comparison based on non-local operations. In: Apostolico A, Crochemore M, Galil Z, Manber U (eds) Combinatorial pattern matching. 3rd annual symposium. Lecture notes in computer science 644, Springer, New York, pp 121–135

Sankoff D, Blanchette M (1997) The median problem for breakpoints in comparative genomics. In: Jiang T, Lee DT (eds) Computing and combinatorics, Proceedings of COCOON '97. Lecture notes in computer science 1276. Springer, New York, pp 251–263

Sankoff D, Blanchette M (1998a) Multiple genome rearrangement and breakpoint phylogeny. J Comp Biol 5:555–570

Sankoff D, Blanchette M (1998b) Phylogenetic invariants for metazoan mitochondrial genome evolution. In: Miyano S, and Takagi T (eds) Genome Informatics 1998. Universal Academy Press, Tokyo, pp 22–31

Sankoff D, Blanchette M (1999a) Probability models for genome rearrangement and linear invariants for phylogenetic inference. In: Istrail S, Pevzner P, Waterman M (eds) Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB 99). ACM, New York, pp 302–309

Sankoff D, Blanchette M (1999b) Comparative genomics via phylogenetic invariants for Jukes-Cantor semigroups. In: Gorostiza L, Ivanoff G (eds) Proceedings of the International Conference on Stochastic Models. Conference proceedings series, Canadian Mathematical Society (in press)

Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren RJ (1992) Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. Proc Natl Acad Sci USA 89: 6575–6579

Sankoff D, Sundaram G, Kececioglu J (1996) Steiner points in the space of genome rearrangements. Int J Found Comput Sci 7:1–9

Terrett JA, Miles S, Thomas RH (1996) Complete DNA sequence of the mitochondrial genome of Cepaea nemoralis (Gastropoda: Pulmonata). J Mol Evol 42:160–168

Valentine JW (no date) Metazoa systematics page, http://www.ucmp.berkeley.edu/phyla/metazoasy.html. University of California Museum of Paleontology

Watterson WA, Ewens WJ, Hall TE, Morgan A (1982) The chromosome inversion problem. J Theor Biol 99:1–7