JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 6, Numbers 3/4, 1999
Mary Ann Liebert, Inc.
Pp. 431–445

Phylogenetic Invariants for Genome Rearrangements

DAVID SANKOFF[1] and MATHIEU BLANCHETTE[2]

ABSTRACT

We review the combinatorial optimization problems in calculating edit distances between genomes and phylogenetic inference based on minimizing gene order changes. With a view to avoiding the computational cost and the "long branches attract" artifact of some tree-building methods, we explore the probabilization of genome rearrangement models prior to developing a methodology based on branch-length invariants. We characterize probabilistically the evolution of the structure of the gene adjacency set for reversals on unsigned circular genomes and, using a nontrivial recurrence relation, reversals on signed genomes. Concepts from the theory of invariants developed for the phylogenetics of homologous gene sequences can be used to derive a complete set of linear invariants for unsigned reversals, as well as for a mixed rearrangement model for signed genomes, though not for pure transposition or pure signed reversal models. The invariants are based on an extended Jukes–Cantor semigroup. We illustrate the use of these invariants to relate mitochondrial genomes from a number of invertebrate animals.

Key words: sorting by reversals, breakpoints, Jukes–Cantor semigroup, metazoan phylogeny, mitochondrial genome.

1. GENOMIC DISTANCES: HARD, MEDIUM, AND EASY

As individual genes evolve through the *local* processes of base substitution, deletion, or insertion, several additional, *nonlocal*, evolutionary mechanisms also operate, at the genomic level.

Consider a circular *signed* genome with gene order $\gamma_1 \cdots \gamma_n$. The origin is arbitrary so that the genome could also be written $\gamma_{i+1} \cdots \gamma_n \gamma_1 \cdots \gamma_i$. Label the genes found on one of the two complementary strands of the genome with a plus sign and those on the other with a minus, resulting in $g_1 \cdots g_n$. ($g_i = \gamma_i$ or $g_i = -\gamma_i$.) By convention, we "view" the circle from the side that ensures that the positively labeled strand is the one read in a clockwise manner, the other counterclockwise. Changing the sign on all genes is equivalent to viewing the circle from the "flip" side, and does not change the identity of the genome.

Consider any two pairs of adjacent genes ab and cd (possibly $b = c$ or $d = a$). The operation that takes $g_1 \cdots ab \cdots cd \cdots g_n$ to $g_1 \cdots a - c \cdots -bd \cdots g_n$ (or, equivalently, to $-g_n \cdots -db \cdots c - a \cdots -g_1$), as illustrated in Fig. 1, is an example of a reversal (or inversion).

We may also consider *unsigned* genomes where the reading direction (i.e., strand) of each gene is unknown. In Fig. 1, we may imagine the two strands superimposed and ignore the signs on the genes. In this case, the reversal transforms $g_1 \cdots ab \cdots cd \cdots g_n$ to $g_1 \cdots ac \cdots bd \cdots g_n$ or, equivalently, to $g_n \cdots db \cdots ca \cdots g_1$; in the former representation, reading clockwise at the top right of Fig. 1, genes $b \cdots c$ were in the *scope* of the

[1] Centre de recherches mathématiques, Université de Montréal, Montréal, Québec, Canada.

[2] Department of Computer Science, University of Washington, Seattle, Washington.

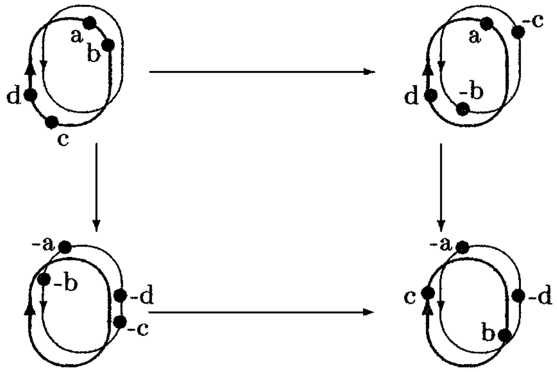

FIG. 1. Reading direction, sign assignment to genes, and reversal. Reading direction is indicated by arrowheads on each DNA strand. The two genomes on the left are biologically identical; one view can be derived from the other by flipping the genome over and assigning signs to each gene according to whether it is on the "front" (i.e., read clockwise) or the "back" (read counterclockwise) strand. The two views of the genome on the right result from reversing the segment from gene *b* to gene *c*, inclusive. The commutativity of flipping and reversal accords with the fact that it does not matter biologically from which side we view the genome.

reversal; in the latter, reading clockwise at the bottom right, these genes were not in the scope of the reversal. Though flipping the genome does not change its identity, considering the two representations separately will be important for probabilistic modeling in Section 4.

Consider any three pairs of adjacent genes ab, cd, and fg, where fg occurs in the interval $d \cdots a$. The operation that takes $g_1 \cdots ab \cdots cd \cdots fg \cdots g_n$ to $g_1 \cdots ad \cdots fb \cdots cg \cdots g_n$ is a transposition. In some models, $g_1 \cdots ad \cdots f - c \cdots -bg \cdots g_n$ can also be produced by a single transposition; in other models, it requires a reversal as well.

The study of comparative genomics has focused on inferring the most economical explanation for observed differences in gene orders in two or more genomes in terms of one or more of these *rearrangement* processes. This has been formulated as the problem of calculating an edit distance between two linear or circular permutations of the same set of objects, representing the ordering of homologous genes in two genomes. Kececioglu and Sankoff (1994) introduced the problem of computing the minimum reversal distance between two given permutations in the signed case, found tight lower and upper bounds, and implemented an exact algorithm that worked rapidly for long permutations. Indeed, Hannenhalli and Pevzner (1995) showed that this problem is only of polynomial complexity, and improved algorithms were given by Berman and Hannenhalli (1996) and by Kaplan *et al.* (1997). Watterson *et al.* (1982) originally posed the problem for the unsigned case. Kececioglu and Sankoff (1995) gave approximation algorithms and an exact algorithm feasible for moderately long permutations. Bafna and Pevzner (1996) gave improved approximation algorithms and Caprara (1997) showed it to be NP-complete.

Computation of the transposition distance between two permutations was considered by Bafna and Pevzner (1995), but its NP-completeness has not yet been confirmed. Edit distances that are a combination of reversals, transpositions, and deletions have been studied by Sankoff (1992), Blanchette *et al.* (1996), and Gu *et al.* (1997).

The breakpoint distance between two genomes containing the same genes (Watterson *et al.*, 1982) is the number of pairs of adjacent genes in one genome that are not adjacent in the other. (In signed genomes, if a and b are adjacent, so are $-b-a$, but not $-a-b$, $-ab$, $a-b$, ba, $-ba$, or $b-a$.) This is not an edit distance, but tends to be highly correlated with such distances and has the advantage of being computable in linear time.

Note that although our discussion in this paper is phrased in terms of the order of genes along a chromosome, the key aspect for mathematical purposes is the order and not the fact that the entities in the order are genes. They could as well be blocks of genes contiguous in a number of species, conserved chromosomal segments in comparative genetic maps (cf. Nadeau and Sankoff, 1998), or, indeed, the results of any decomposition of the chromosome into disjoint ordered fragments, each identifiable in two or more genomes.

2. PHYLOGENY BASED ON GENE ORDER

The extension of edit distances for gene order data to finding globally optimal phylogenetic trees is inherently difficult. Not only are some of the measures of genomic edit distance in Section 1 computationally complex, but the extension of any of them, even the reversals distance for signed genomes (itself only of quadratic complexity), to three or more genomes—multiple genome rearrangement—is NP-hard (Caprara, 1999). An example is the "median" problem: find the "ancestor" genome that is closest to three given genomes. Heuristics for multiple genome rearrangement are available (Hannenhalli *et al.*, 1995; Sankoff *et al.*, 1996), but they are feasible only for small genomes.

The breakpoint distance has the advantage of being computable in linear time. Nevertheless its extension to three or more genomes is also NP-hard (Pe'er and Shamir, 1998; Bryant *et al.*, 1999). It does have a simple reduction to the Traveling Salesman Problem (Sankoff and Blanchette, 1997) and can thus benefit from relatively efficient software available for the latter to solve examples on three genomes with moderate-sized n. This can then be extended to the optimization of fixed-topology phylogenies (Blanchette *et al.*, 1997; Sankoff and Blanchette, 1998a), and ultimately to the search for optimal topologies (Blanchette *et al.*, 1999).

In this kind of phylogenetic inference, breakpoint distance is used as a *parsimony* criterion. And parsimony methods are among those that, under the simplest probabilistic models of mutation, may sometimes reconstruct trees incorrectly when there are some very short and some very long branches (Felsenstein, 1978). This problem, together with the computational complexity of all versions of the multiple genome rearrangement problem, leads us to investigate the potential of *branch-length invariants* for inferring phylogeny based on gene order comparisons. Phylogenetic invariants are based on probabilistic models of evolution, and in the next section we will review how they were developed in the context of sequence evolution. Following this, in Section 4 we will develop probabilistic models for gene order evolution preparatory to deriving invariants for genome-level evolution.

3. INVARIANTS FOR MODELS OF SEQUENCE EVOLUTION

Consider the aligned DNA sequences of length n: $X_1^{(1)} \cdots X_n^{(1)}, \ldots, X_1^{(N)} \cdots X_n^{(N)}$ representing N species whose history of evolutionary divergence, or *phylogeny*, is represented by a tree $\mathbf{T}$ with vertex set V and edge set E, as in Fig. 2. The terminal vertices represent observed, or present-day, species. The nonterminal vertices represent hypothetical ancestral species. For each i, the $X_i^{(J)}$ are the terminal points of a trajectory indexed by $\mathbf{T}$, taking on values in the alphabet of bases $\{A, C, G, T\}$. This trajectory is a sample from a process described by $|E|$ 4×4 Markov matrices with positive determinant all belonging to the same semigroup, one matrix associated to each of the edges in E. Such semigroups have been proposed by Jukes and Cantor (1969), Kimura (1980, 1981), Tajima and Nei (1984), Hasegawa *et al.* (1985), Cavender (1989), Jin and Nei (1990), Tamura (1992), Nguyen and Speed (1992), Tamura and Nei (1993), Steel (1994), and Ferretti and Sankoff (1995).

Aside from the fact that it has N terminal vertices, the tree $\mathbf{T}$ is unknown. In particular, the $|E|$ matrices associated with the edges are unknown, though the common semigroup from which they are drawn is given. The central problem of phylogenetic inference is to estimate $\mathbf{T} = (V, E)$, given only n data vectors each consisting of the values at the N terminal vertices of the trajectory, of form $(X_i^{(1)}, \ldots, X_i^{(N)})$, where $X_i^{(J)}$ is the ith base in the Jth DNA sequence.

In DNA evolution, it is simplest to consider rates of change between any two elements of $\{A, C, G, T\}$ to be symmetric. With this assumption, it is usually preferable not to try to locate a root, or earliest ancestor node, in the tree. Thus in this paper we will make the simplifying assumption that $\mathbf{T}$ is an unrooted binary branching tree (all nonterminal vertices of degree 3), hence $|V| = 2N - 2$, $|E| = 2N - 3$, and will confine ourselves to symmetric transition matrices.

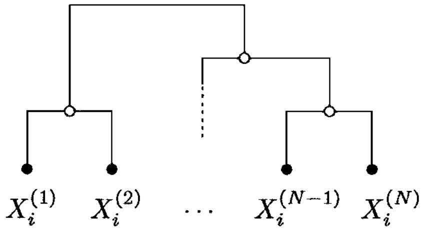

FIG. 2. Sample trajectory $X_i^{(\cdot)}$. Indexing tree $\mathbf{T}$ is unknown, but the same for all $i = 1, \ldots, n$. Filled dots at terminal vertices indicate N present-day species at which values of the process can be observed; open dots represent unobservable ancestral species.

Phylogenetic invariants are predetermined functions of the probabilities of the observable N-tuples. These functions are identically zero only for **T** (and possibly a limited number of other trees), no matter which $|E|$ matrices are chosen from the semigroup. Evaluating the invariants associated with all possible trees, using observed N-tuple frequencies as estimates of the probabilities, enables the rapid inference of the (presumably unique) tree **T** for which all the invariants are zero or vanishingly small.

The chief virtue of the method of invariants is that it is not sensitive to "branch length," i.e., to which $|E|$ matrices are chosen from the semigroup; for a matrix M, this length may be taken to be $-\log \det M$. Methods of phylogenetic reconstruction that do not take account of the model used to generate the data may be susceptible to an artifact that tends to group long lineages together and short lineages together.

Lake (1987) introduced linear invariants, studying the case $N = 4$ for a two-parameter (representing transversion versus transition probabilities) semigroup originally suggested by Kimura (1980). At the same time, Cavender and Felsenstein (1987) published quadratic invariants for a one-parameter semigroup of 2×2 matrices. Subsequently a great deal of research has been carried out into both linear invariants, by Cavender (1989), Fu (1995), Nguyen and Speed (1992), Steel and Fu (1995), Hendy and Penny (1996), and polynomial invariants, by Drolet and Sankoff (1990), Sankoff (1990), Felsenstein (1991), Ferretti *et al.* (1993, 1994, 1995, 1996), Evans and Speed (1993), Steel *et al.* (1993), Szekeley *et al.* (1993), Steel (1994), Evans and Zhou (1998), Hagedorn (1999), and Hagedorn and Landweber (1999).

Can we apply this theory to comparative genomics? After all, the various sets of breakpoints in a multigenome comparison do not resemble a multiple alignment of sequences in any way, so that the phylogenetic invariants developed in the context of DNA base sequence data are not applicable. In Section 4 we will present models for genome rearrangement processes analogous to the base substitution models for gene sequence evolution, and examine the evolution of the adjacencies of pairs of genes over time. In one case, that of reversals on unsigned genomes, we obtain a matrix semigroup of transition probabilities among these adjacencies. In the other cases, the time-indexed matrices of transition probabilities do not form a semigroup. Nevertheless, in Section 5, we propose a simpler model for the evolution of breakpoints, not based on any assumptions about the rearrangement processes responsible for them, and use this to calculate a complete set of linear invariants for the 15 binary unrooted trees where $N = 5$.

4. PROBABILITY MODELS FOR BREAKPOINT DISTANCES

We will propose models for reversals on unsigned and signed circular genomes, as well as for transpositions (where it suffices to consider unsigned genomes only). We will assume in all three models that all pairs of adjacent genes fg are equally likely to be disrupted, though this is a simplification of biological reality (Blanchette *et al.*, 1999; Sankoff, 1999). Recall that, as in Fig. 1, two different pairs must be disrupted for each reversal, and three for each transposition.

4.1. Reversals, unsigned case

For an unsigned circular genome, consider a continuous time process with rate $\lambda = 1$. Each change of state involves a reversal, where any two pairs of adjacent genes fg and hk are equally likely to be disrupted. We focus on a particular gene f and, after each reversal, choose the representation of the resultant genome where f has *not* been in the scope of the reversal. If fg is one of the two pairs chosen (probability $1/n$), any of the $n-1$ genes other than f is equally likely to replace g. In the case $g = h$, gene g replaces itself and the reversal is "invisible." The matrix of transition probabilities for the occupant, at a specific time t, of the slot in the genome originally occupied by g, whose columns and rows are labeled by the $n-1$ candidate genes, is of form $[1-(n-1)\alpha]I + \alpha J$, where I is the identity and J the matrix of 1s, and $1-(n-1)\alpha = e^{-t/(n-2)}$. This is a generalization to $n-1 \times n-1$ matrices of the Jukes–Cantor (1969) semigroup of 4×4 matrices.

4.2. Reversals, signed case

A model consisting only of random reversals on signed genomes, however, is quite different. Suppose all genes are on the same strand and have a positive sign. Then if fg is disrupted by a reversal, the new successor to f will necessarily have negative sign. All negatively signed genes (other than $-f$) will have probability $1/(n-1)$ of replacing the successor to f. All positively signed genes will have probability zero. So the Jukes–Cantor equiprobability among the $2n-3$ possible new successors does not hold. Moreover, after the first reversal, the standedness of some genes will have changed, so that for the next reversal, some of the transition probabilities for successors will change. In other words, the process cannot be modeled by a semigroup of matrices as in the unsigned case.

Without loss of generality, we label the genes from 1 to n, and after each reversal we flip the genome if necessary so as to ensure gene 1 always has a positive sign. In addition, we designate the position occupied by gene 1 to be position 1, the position occupied by its successor to be position 2, and so on. Let x_i be the occupant of the ith position. After k reversals, the probability that the ith position will be occupied by gene h is

$$\begin{aligned} P_k(x_i = h) &= P_{k-1}(x_i = h)\mathrm{Pr}[h \text{ not in scope of } k\text{th reversal}] \\ &\qquad + \sum_{j=2}^{n} P_{k-1}(x_j = -h)Pr[k\text{th reversal moves } h \text{ from } j \text{ to } i] \\ &= P_{k-1}(x_i = h)\left[1 - \binom{n}{2}^{-1}(i-1)(n+1-i)\right] \\ &\qquad + \binom{n}{2}^{-1}\sum_{j=2}^{n} P_{k-1}(x_j = -h)\min\left\{\begin{matrix} i-1, & n+1-j, \\ j-1, & n+1-i \end{matrix}\right\} \end{aligned}$$

For $n = 4$, this recurrence produces the pattern in Table 1.

It can be seen that it takes a relatively large number of reversals to “scramble” the genome enough so that the successor to gene 1 is equally likely to be any other gene, with either sign.

To compare this rearrangement process to the one generated by the Jukes–Cantor semigroup, we define $P_t(x_i = h)$ as the probability that the ith position will be occupied by gene h at time t. Then

$$P_t(x_2 = h) = \sum_{k=0}^{\infty} \frac{e^{-t}t^k}{k!} P_k(x_2 = h)$$

Table 2 illustrates the approach to Jukes–Cantor probabilities of the reversal on signed genomes model for $n = 4$.

Table 2 shows that the transition probabilities remain rather inhomogeneous for a considerable time. For $t = 4$, there have been about eight opportunities on the average for each of the four adjacencies to be disrupted (two per reversal); nonetheless the probabilities are decidedly nonuniform, even among the genes where $h \neq 2$.

TABLE 1. APPROACH OF $P_k(x_2 = h)$ TO EQUIPROBABILITY

	h					
k	2	3	4	−2	−3	−4
1	0.500	0	0	0.167	0.167	0.167
2	0.333	0.111	0.083	0.167	0.139	0.167
4	0.205	0.154	0.143	0.170	0.158	0.170
8	0.169	0.166	0.165	0.167	0.166	0.167

TABLE 2. APPROACH OF $P_t(x_2 = h)$ TO JUKES–CANTOR PROBABILITIES

	h							
	Random reversals						Jukes–Cantor	
t	2	3	4	−2	−3	−4	2	*Others*
1	0.632	0.032	0.025	0.106	0.099	0.106	0.672	0.066
2	0.433	0.078	0.065	0.145	0.133	0.145	0.473	0.105
4	0.258	0.134	0.123	0.166	0.154	0.166	0.279	0.144
8	0.178	0.163	0.160	0.168	0.164	0.168	0.182	0.164

For larger n, such as the case $n = 37$ of interest in Section 7 below, the situation is analogous. Even after all the original adjacencies have had ample opportunity to be disrupted, the transition probabilities remain quite different from Jukes–Cantor, especially for low or high values of h, e.g., $\pm h = 2, 3, 36$, or 37. But the values of t of biological interest will be those during which a fair proportion of the original adjacencies will be conserved. In other words, for those lengths of time for which we wish to apply these methods, the Jukes–Cantor semigroup is not a good approximation for the random reversals model.

4.3. *Transpositions*

Finally, consider transpositions on unsigned circular genomes. Again, we assume a uniform probability rate $\lambda = 1$ of such events occurring. At each event, any choice of three different pairs of adjacent genes ab, cd, and fg is equally likely to be disrupted. Any of the $n - 2$ genes other than f or g is equally likely to play the role of b in replacing g as the neighbor of f. But the fact that g cannot replace itself as it could in the unsigned reversal model leads to the same sort of difficulty as with signed reversal. A Jukes–Cantor model cannot be formulated.

5. EXTENDED JUKES–CANTOR MODEL FOR BREAKPOINTS

In this section, we construct a model for signed genomes. We will not assume that reversal, or any other particular process, is the only mechanism of genome rearrangement. Reversal, transposition, or single-gene movement could all play a role, in unknown proportions. Thus, we will not assume that only $-h$ can replace g, where h and not $-h$ appears in the original genome, as in the pure reversals case. Indeed, inspired by Jukes–Cantor, we assume that for any gene f, whose successor is g, the probability α that, over a given time interval, the successor to f will have changed to h, is the same for all pairs of genes f and g, and for all $h \neq g$. Note that $h = -g$ is not excluded. There are $2n - 3$ such changes possible. The probability that g will remain the successor is then $1 - (2n - 3)\alpha$. Note that $1 - (2n - 3)\alpha > \alpha$ since, for consistency's sake, this event, including both no change and reversed changes, is at least as likely as any other particular change.

We have in effect defined a $2n - 2 \times 2n - 2$ Jukes–Cantor matrix $M(\alpha)$, where the rows and columns are indexed by the $2n - 2$ possible signed genes different from f and $-f$. The entries are all α except for $1 - (2n - 3)\alpha$ on the diagonal. The model defines a semigroup that determines (stochastically) the trajectory of the occupant of the "successor to f" slot across a phylogeny. From it, if we were given the branch lengths, we could calculate the probabilities of all possible N-tuples at the terminal vertices.

We are not, however, given the branch lengths, nor are we directly interested in these lengths, since our goal is to find the correct tree topology in a way that is *insensitive* to them.

For a given f, and there are $2n$ of them, since we analyze f and $-f$ separately, the $(2n - 2)^N$ different N-tuples in the successor slot may be summarized by far fewer patterns. The 5-tuple *gghhh* has the same probability as *gg–h–h–h* or *hhkkk*, because of the symmetries in the model. We identify these configurations as follows: The first component of the N-tuple is labeled x, the second—if it is not also labeled x by virtue of being identical to the first—is labeled y. The label z is reserved for the third different gene name in the N-tuple, if there is one, and so on. If g and $-g$ occur in the same N-tuple, they require two distinct labels.

In the case of 37 genes (74 distinct gene names), instead of more than a billion 5-tuples there are only 52 distinct configurations. In effect, this is the fifth term in the Bell series:

$$a(N) = 1 + \sum_{i=1}^{N-1} a(i) \binom{N}{i} = 1, 2, 5, 15, 52, 203, \ldots,$$

which is the number of ways of distributing five indistinguishable objects into five labeled boxes.

6. THE INVARIANTS

Using the algorithm of Fu (1995), we find the following complete set of phylogenetic linear invariants for the $k \times k$ Jukes–Cantor semigroup on the unrooted binary tree [(AB)C(DE)], as in Fig. 3.

The term "complete" is used in the sense that these 11 invariants below form a basis for the ideal of invariants. We use the configuration label, e.g., $xyzxw$, as a shorthand for the configuration probability normalized by the number of N-tuples it represents, or for simply the probability of any one of these N-tuples, e.g., Prob(hg-ghk).

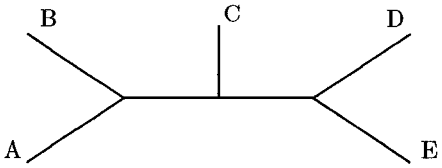

FIG. 3. Unrooted binary tree [(AB)C(DE)]. The other 14 trees are obtained by permutating the five labels.

$$
\begin{gathered}
xyzyx - xyzyw - xyzzx + xyzzw \\
xyzyz - xyzyx - xyzwz + xyzwx \\
xyzxy - xyzxw - xyzzy + xyzzw \\
xyzxz - xyzxy - xyzwz + xyzwy \\
xyzzx - xyzzy - xyzwx + xyzwy \\
xxyxy - xxyyx + xxyyz - xxyxz - xxyzy + xxyzx \\
xyyxy - xyyyx + xyyyz - xyyzy - xyyxz + xyyzx \\
xyxxy - xyxyx + xyxyz - xyxxz - xyxzy + xyxzx \\
xyxzy - xyxyz + xyyxz - xyyzx + xyzyw - xyzxw - xyzwy + xyzwx \\
xyxxy - xyxxz - xyxyy + xyxzz - xyyyx + xyyyz \\
+ xyyxx - xyyzz + xyzyy - xyzxx - xyzwy + xyzwx \\
xyxyy - xyxzz - xyyxx + xyyzz - xyzyy + xyzxx \\
+ k(xyxzy - xyxzw - xyyzx + xyyzw - xyzwy + xyzwx)
\end{gathered}
$$

In our context, $k = 2n - 2 = 72$. There are other invariants, but they are not *phylogenetic*, i.e., they are zero for all trees. For the unsigned reversal model in Section 4.1, $k = n - 1$.

6.1. Remarks on the invariants

In examining the 11 invariants, we observe that 14 of the 52 possible configurations enter into no invariant. Seven of these contain no information on the branching structure of the tree:

$$xxxxx \qquad xyyyy, xyxxx, xxyxx, xxxyx, xxxxy \qquad xyzwu$$

and so it is not surprising that they play no role here. The other seven are

$$xxyyy, xxyzw, xyzzz \qquad xxxyy, xxxyz, xyzww \qquad xxyzz$$

These are precisely the configurations that characterize one (first three configurations), the other (second three configurations), or both (last configuration) of the two internal edges of the tree [(AB)C(DE)]. They could be expected to be among the most frequent configurations (along with the seven noninformative configurations above and other configurations requiring no "extra steps" such as $xyyzw$ or $xyyyz$). Were all the data concentrated on these 14 configurations, then all the 11 invariant functions would be exactly zero.

6.2. Evaluating the invariants

To estimate the configuration probabilities, we analyze the successor slot for each of the $2n$ gene names, treating f and $-f$ separately, and calculating the relative frequency of each configuration, normalized by the number of different N-tuples that it contains. Though the configurations for different genes are not statistically independent, the expected value of a relative frequency is nonetheless the probability that generated it. By the linearity of the invariant functions, the expected value of each of the invariants evaluated using the relative frequencies is zero for [(AB)C(DE)] and nonzero for some other trees.

Note that with 37 genes as in the application of Section 7 below, or 74 data points, the 52 configurations will not all be estimated with any degree of accuracy. Neither will the invariant functions, especially since

TABLE 3. COELOMATE MITOCHONDRIAL GENOMES COMPARED IN THIS INVESTIGATION, WITH HIGHER TAXONOMIC LEVELS

	Organism		*Phylum*
HU	Human	CHO	Chordate (deuterostome)
SS	*Asterina pectinifera* (sea star)	ECH	Echinoderm (deuterostome)
BA	*Balanoglossus carnosus* (acorn worm)	HEM	Hemichordate (deuterostome)
DR	*Drosophila yakuba* (insect)	ART	Arthropod (protostome)
KT	*Katharina tunicata* (chiton)	MOL	Mollusc (protostome)
LU	*Lumbricus terrestris* (earthworm)	ANN	Annelid (protostome)

Citations: HU, Anderson *et al.* (1981); SS, Asakawa *et al.* (1993); BA, Castresana *et al.* (1998); DR, Clary and Wolstenholme (1985); KT, Boore and Brown (1994); LU, Boore and Brown (1995).

much of the data will be concentrated on the configurations that do not even appear in the invariant formulae. The situation would be much worse for $N = 6$ with 203 configurations, one of the reasons for not proceeding beyond $N = 5$ here.

7. AN APPLICATION TO METAZOAN PHYLOGENY

The mitochondrion is an "organelle" occurring in profusion in animal, plant, fungal, and most other eukaryotic cells. It has its own genome with a small number (<100) of genes, usually organized as a single circular chromosome. The mitochondrial genome of many metazoan animals has been completely sequenced and the genes they contain identified. The breakpoints in comparisons among the gene orders of these genomes have proven to contain much information pertinent to the inference of metazoan phylogeny (Blanchette *et al.*, 1999). The conservatism of certain genomes, such as human, *Drosophila*, and *Katharina tunicata* (a chiton), versus the extreme divergence of related lineages, such as echinoderms or snails, i.e., the presence of both short and long branches, is the chief difficulty in the reconstruction of this phylogeny. In the next sections we apply our theory of breakpoint invariants to explore three problems in the phylogeny of higher metazoans, the true coelomates, based on the species in Table 3. These problems pertain to the protostome–deuterostome split, the internal structure of the protostomes, and the internal branching order of the deuterostomes.

We will evaluate the 11 invariant relations, substituting the observed N-tuple frequencies for their probabilities; with larger genomes these frequencies should satisfy the invariant relations more closely, but with just 37 genes in the mitochondrial genome, we can only hope that the invariants associated with the true tree **T** are better satisfied than are those that are not associated with it. We carry out extensive simulations to assess to what extent the trees we infer are likely to be the correct ones.

8. METAZOAN PHYLOGENY

Aspects of coelomate metazoan phylogeny are controversial (cf. Aguinaldo *et al.*, 1997; Christofferson and Araújo-de-Almeida, 1998); among the groupings in Table 3, only the split between deuterostomes and protostomes seems undisputed. Eernisse *et al.* (1992), Giribet and Ribera (1998), and most others would group annelids and molluscs as sister groups, with arthropods related to these at a deeper level. But there are still proponents (e.g., Rouse and Fauchald, 1995) of a traditional grouping (*Articulata*) of annelids and arthropods as sister taxa. Hemichordates have been grouped with the chordates as in Brusca and Brusca (1990) or in the "Tree of Life" (Maddison and Maddison, 1995), but recent evidence by Wada and Satoh (1994) has led many to group them closer to the echinoderms (cf. Ruppert and Barnes, 1994; Valentine, n.d.).

Aside from these unsettled questions, efforts to infer phylogeny based on distances between mitochondrial gene orders have tended to group *Drosophila* closer to human than the echinoderms are (e.g., Sankoff *et al.*, 1992; Blanchette *et al.*, 1999, Fig. 4a and 4b), an artifact of the mitochondrial genome of the latter being highly divergent, the former two relatively conservative.

Figure 4 contrasts three phylogenies, one representing the "Tree of Life" (Maddison and Maddison, 1995), another the summary phylogeny by Valentine (n.d.) on the University of California Museum of Paleontology website, and the third the *Drosophila*–human artifact.

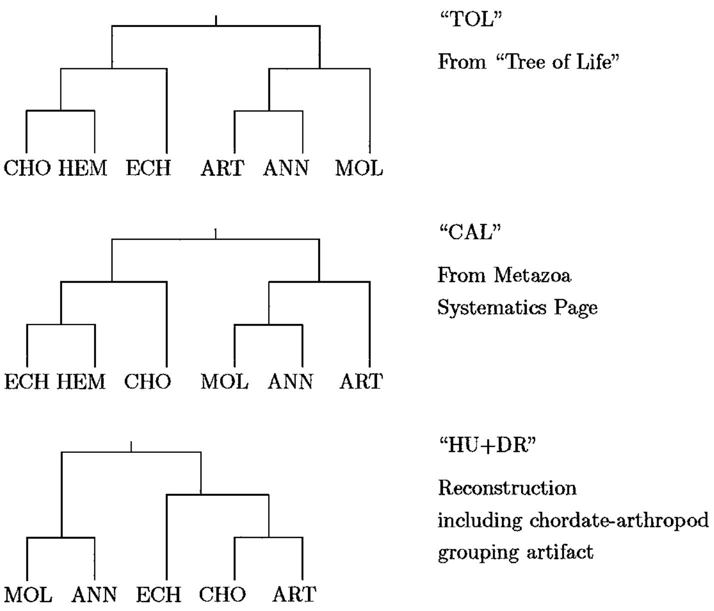

FIG. 4. Three alternative views of coelomate evolution.

9. TEST PROCEDURES

Different invariants contain different numbers of configurations and, when evaluated with frequency data on the correct and incorrect trees, have different ranges, so that it may be misleading to compare trees on the basis of how close they are to zero with respect to all the invariants. To standardize the comparisons, we simulated 10,000 trees of form [(AB)C(DE)] on 37-gene genomes, with all branches disrupted by R random reversals, and compiled the distribution of each the 11 invariants evaluated using the sample configuration frequencies. The value of R is determined by counting the number of breakpoints on a minimum breakpoint tree (Sankoff and Blanchette, 1998a) and dividing by $2\theta(2N-3)$, each reversal contributing up to two breakpoints, and there being $2N-3$ branches on an unrooted binary tree. The parameter θ corrects for "multiple hits"—we used $\theta = 0.75$. This only approximates the situation with the mitochondrial data (some lineages are clearly much longer than others), nonetheless the 11 test distributions constructed this way can serve as comparable scales to judge the fit of each of the 15 possible trees.

The score for each combination of tree and invariant can thus be transformed into a significance level. (Highly significant implies a poor fit.) A summary score for each tree can then be produced by taking the product of the 11 significance levels.

10. RESULTS

In this section, we will first present and comment on the trees selected by the use of invariants. We then compare these to the most parsimonious trees based on the same data, but simply minimizing the total number of changes (in terms of presence versus absence) over the tree for each possible adjacency of two genes. (See Gallut, 1998, for an approach based on three-gene adjacencies, also applied to mitochondrial gene orders of invertebrates.) These will both be compared to the minimum breakpoint trees in the sense of Section 2, i.e., minimizing the sum over all branches of the number of breakpoints between the genomes at each endpoint.

Note that the latter two methods, while both relying on a parsimony criterion, are quite distinct, and may have different results. It is not hard to show that if A is the cost of the most parsimonious tree, using presence versus absence of all possible gene adjacencies as characters, then $A \leq 2B$, where B is the minimum sum of the breakpoint costs over all branches of the same tree. Consider the three (unsigned) circular genomes in the top of Table 4.

In calculating the breakpoint distance, an optimal median genome is found to be 1 2 3 4 5 6, with breakpoint distance 3 from each of the first two data genomes and zero from the third, for a total tree distance of 6. The

TABLE 4. DATA SETS CONTRASTING ADJACENCY PARSIMONY AND BREAKPOINT DISTANCE

1	2	5	6	4	3
1	3	2	5	4	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	3	5	2	6	4

18 gene adjacencies in the data, however, consist of 9 pairs, each occurring twice and absent once, for a total cost of 9, so that $A < 2B$.

In contrast, for the three genomes in the bottom of Table 4, we again have $B = 6$ based on median genome 1 2 3 4 5 6, but $A = 12 = 2B$. The variability in the relation between A and B implies that the optimal breakpoint tree does not need to coincide with the most parsimonious tree in terms of adjacencies, and indeed it generally does not.

10.1. *Deuterostomes and protostomes*

The first subset of the data to be examined includes HU, SS, DR, KT, and LU, in order to compare the results with those of Sankoff *et al.* (1992) and Blanchette *et al.* (1999). In this case $R = 10$.

The best three trees manifested scores of $2 \times 10^{-12}, 6 \times 10^{-15}$, and 7×10^{-17}. The first of these was consistent with the CAL tree in Fig. 4, and the third was the artifactual tree in that figure. The second also contained the HU+DR artifact.

Nevertheless, according to the best tree, our method succeeded in correctly grouping CHO and ECH, despite the discordance of branch lengths that defeat distance-matrix-based attempts. And it also confirmed the ANN+MOL grouping in CAL versus the TOL grouping of ANN+ART.

10.2. *The Balanoglossus data*

The recently sequenced mitochondrial genome of *Balanoglossus carnosa* allows a more detailed investigation of deuterostome–protostome branching. Here we focus on the deuterostome–arthropod relationship, retaining *Katharina* as a second protostome, but dropping *Lumbricus* from the analysis. The simulations for constructing the statistical tests were redone with $R = 6$. The results in this analysis clearly confirm the deuterostome grouping. The three best trees, with summary scores 10^{-7}, 10^{-7}, and 6×10^{-8}, all group the deuterostomes together and no other tree scores better than 3×10^{-15} (which is the score when DR groups more closely with HU and BA than SS does). In this analysis the best tree is consistent with the TOL tree in Fig. 4, while the CAL tree is third best.

10.3. *A comparison of methods*

Table 5 shows how the candidate phylogenies fare under the method of invariants compared to two parsimony approaches. Both the methods of invariants and adjacency parsimony operate on the configuration frequencies in the gene-successor data described in Section 6.2, in the former case as detailed in that section, and in the latter by counting total "extra steps" required by all data configurations on each tree. The third method minimizes the sum of the breakpoint distances over all branches of the tree, involving the optimization of ancestral genomes (Blanchette *et al.*, 1999).

It can be seen that the two methods operating on gene-successor configuration frequencies tend to agree as to the best tree, although the method of invariants seems slightly less susceptible to the HU+DR artifact. The breakpoint distance method does not resolve among the best trees for two of the data sets, but whether this is a virtue or a shortcoming remains to be seen.

All of the analyses support the protostome–deuterostome split, and they all support the annelid–mollusc grouping as a sister group to the arthropods. On the other hand, they do not agree on the internal grouping of the deuterostomes. The breakpoint distance gives equal support to an ECH+CHO grouping (which is of little credibility) as to the ECH+HEM analysis, whereas the other methods favor a more traditional CHO+HEM grouping.

TABLE 5. COMPARISON OF THREE METHODS ON THREE DATA SETS

	Tree	Invariants	Adjacency parsimony	Breakpoint distance
TOL	(HU SS)[(DR LU) KT]	5*	5	3*
CAL	(HU SS)[DR (LU KT)]	1	1*	1
HU	[(HU DR) SS](LU KT)	3	1*	5*
+	[(HU DR) LU](SS KT)	4	6*	3*
DR	[(HU DR) KT] (LU SS)	2	3	2
TOL	[(HU BA) SS](DR KT)	1	1	1*
CAL	[HU (BA SS)](DR KT)	3	3	1*
	[(HU SS) BA](DR KT)	2	2	1*
HU+DR	[(HU BA) DR](SS KT)	4	4	4*
CAL	[HU (BA SS)][DR (LU KT)]	(3)	5	1*
	[(HU BA) SS][DR (LU KT)]	(1)	1	3
	[(HU SS) BA][DR (LU KT)]	(2)	4	1*
	[(HU BA) SS] [LU (DR KT)]		2*	13*
HU+DR	[(HU BA) DR][SS (LU KT)]		2*	13*

Top, without BA; middle, without LU; bottom, all six species. Figures indicate rank of trees built using the same method on the same data sets. Asterisks indicate ranks that are tied with at least one other. Parentheses indicate six-species supertree based on best tree in five-species analysis without BA, combined with three best trees in five-species analysis without LU. Note that for five-species analyses, ranks are out of 15, while for six species, they are out of 105.

10.4. *What is the sensitivity of our method with small genomes?*

A more clear-cut result of our method would see the tree **T** emerge with no invariant scoring less than Ψ and all other trees scoring less than Ψ (i.e., "significant") on at least one invariant, for some threshold Ψ. We simulated $N = 5$ data for a range of genome sizes, from $n = 8$ to $n = 140$, with $n/4$ random reversals disrupting gene order on each branch of the tree, with 2000 repetitions of the experiment for each n. (Recall from Section 9 that for $n = 37$, the minimum breakpoint tree warrants $R = 10$, approximately $n/4$.)

The form of the invariants ensures that each converges to a limit, zero for **T** and nonzero for each of several other trees. If Ψ is small enough, and n is large enough, only the invariants for **T** will be below the significance level. Results of our simulations in Fig. 5 indicate that for Ψ small enough to exclude all trees except **T**—a "true" threshold, we require $n = 140$ at least. For smaller n, the tree **T** is also likely to be "rejected" by at least one invariant. For $n = 37$, **T** is the sole tree to pass the tests of all 11 invariants with $\Psi = 0.01$ only 15% of the time. If Ψ is relaxed to a value that will maximize acceptance of **T** only, say $\Psi = 0.1$, only

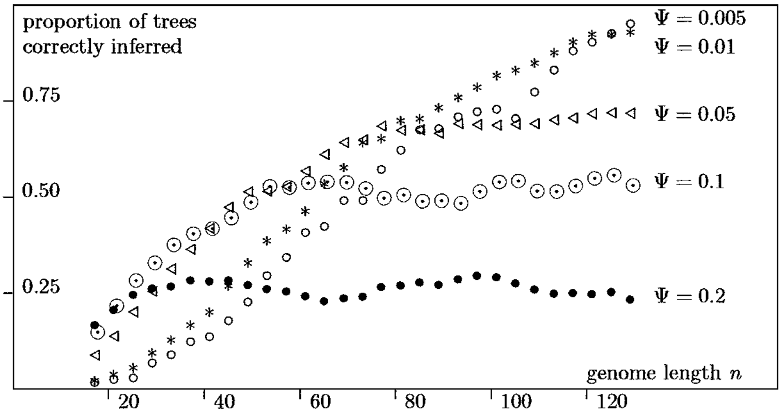

FIG. 5. Proportion of trees correctly inferred as a function of genome length and Ψ. Curves smoothed using a window size of three data points.

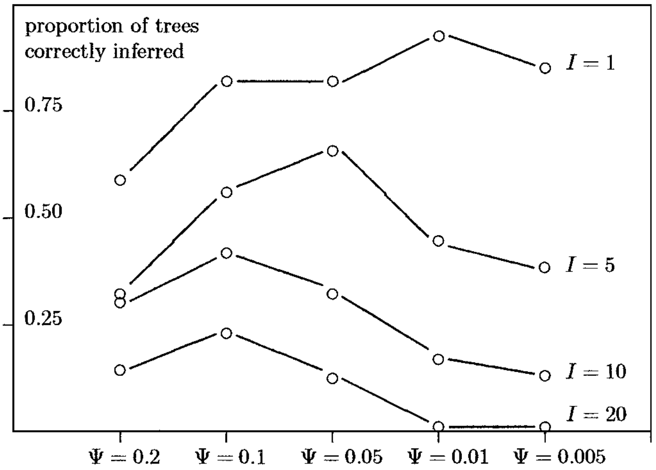

FIG. 6. Proportion of trees correctly inferred as a function of $\boldsymbol{\Psi}$ and number of rearrangements I.

40% can be attained for $n = 37$. This explains our recourse to more equivocal, statistical criteria described in Section 9. Another set of simulations tested the performance of our method on $N = 5$, $n = 37$ data for a range of branch lengths, the same on each branch. We used 10,000 simulations per branch length. As can be seen in Fig. 6, the rate of success drops off rapidly with branch length and decreasing $\boldsymbol{\Psi}$ so that with 10 random reversals per branch, successful discrimination in favor of **T** is 40% for a threshold value of $\boldsymbol{\Psi} = 0.1$, and for 20 reversals it is only 25%; with 10 reversals and $\boldsymbol{\Psi} = 0.01$, the success rate is only 17%.

11. FURTHER WORK

Though much probabilistic modeling of gene sequence changes has been incorporated into phylogenetic analysis, very little research has gone into mathematical approaches to phylogenetics based on gene order, and even less, previous to the present undertaking, into probability models for the evolution of gene order (see, however, Dalkie, 1998).

Of both mathematical and biological interest is whether this theory can be developed in the direction of other semigroups. Linear invariant theory is well developed, for the Kimura models (e.g., Steel *et al.*, 1993) and others, and biological interpretation in the breakpoint context is possible. Even though an *exact* representation of models such as random reversals only on signed data (Section 4) in terms of semigroups of matrices is not possible, significantly better approximations than Jukes–Cantor may well be feasible.

In comparing the invariant method to the two parsimony methods, we cannot do more based on these small applications than note that in one example, in Section 10.1 and Table 5, adjacency parsimony did not discriminate against a branch-length artifact, while the invariants, based on the same data, did.

Perhaps the most promising direction for the method of invariants lies toward larger genome size—plastids, prokaryotes, and, when more eukaryotes are completely sequenced, nuclear genomes. Multichromosomal genomes are handled as easily as single-chromosome ones, since the model pertains to single breakpoints and not to whole fragments, which behave differently in reversals, transpositions, and reciprocal translocations. Increasing n only linearly increases the time to compute configuration frequencies, which is negligible. Our simulations in Section 10.4 indicate that the method should be able to identify the true tree with a high degree of accuracy for large genomes. Note that heterogeneity of rates is not a problem with this approach, either from lineage to lineage or from gene to gene in their quantitative susceptibility to be adjacent to breakpoints; this stems from the linearity of the invariants. Thus the fact that tRNA genes may be more mobile (Blanchette *et al.*, 1999), either because they tend to be at the end of rearranged fragments or because they may be individually transposed in the genome, does not affect the results.

Enlarging the method to handle six species and perhaps more is quite feasible, though the bookkeeping involved with hundreds of invariants is considerable. Beyond this, some way of handling decomposition of the problem, such as we used in Sections 10.1 and 10.2, might be systematized.

The biological results obtained here include the relatively early branching of arthropods within the protostomes, and the grouping of the hemichordates with the chordates, though the latter is equivocal. Our method clearly distinguishes between deuterostomes and protostomes, which is not always the case with other approaches using rearrangement data.

ACKNOWLEDGMENTS

Research supported by grants to David Sankoff from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Genome Analysis and Technology program, and an NSERC fellowship for graduate studies to MB. David Sankoff is a Fellow of the Canadian Institute for Advanced Research. We are grateful for advice and discussion to Jeffrey Boore, Kenneth Halanych, Franz Lang, Mitchell Sogin, Veronique Barriel, Gonzalo Giribet, Martin Christofferson, Mary Mickevich, and Takashi Kunisawa, though we take full responsibility for all shortcomings remaining in this article. Parts of this paper were presented at a June 1998 International Conference on Stochastic Models in Ottawa, at Hennig XVII in São Paulo, and the Genome Informatics Workshop 98 (Sankoff and Blanchette, 1998b).

REFERENCES

Aguinaldo, A.M.A., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., and Lake, J.A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* (*London*) 387, 489–493.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, 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., and Young, I.G. 1981. Sequence and organization of the human mitochondrial genome. *Nature* (*London*) 290, 457–465.

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

Bafna., V., and Pevzner, P.A. 1995. Sorting by transpositions, 614–623. In Galil, Z., and Ukkonen, E., eds., *Proceedings of the Sixth Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science* 937. Springer-Verlag, New York.

Bafna, V., and Pevzner, P.A. 1996. Genome rearrangements and sorting by reversals. *SIAM J. Comput.* 25, 272–289.

Berman, P., and Hannenhalli, S. 1996. Fast sorting by reversal, 168–185. In Hirschberg, D., and Myers, G., eds., *Proceedings of the Seventh Symposium on Combinatorial Pattern Matching. Lecture Notes in Computer Science* 1075. Springer-Verlag, New York.

Blanchette, M., Kunisawa, T., and Sankoff, D. 1996. Parametric genome rearrangement. *Gene* 172, GC 11–17.

Blanchette, M., Bourque, G., and Sankoff, D. 1997. Breakpoint phylogenies, 25–34. In Miyano, S., and Takagi, T., eds., *Genome Informatics 1997*. Universal Academy Press, Tokyo.

Blanchette, M., Kunisawa, T., and Sankoff, D. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.* 49, 193–203.

Boore, J.L., and Brown, W.M. 1994. Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata. Genetics* 138, 423–443.

Boore, J.L., and Brown, W.M. 1995. Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris. Genetics* 141, 305–319.

Brusca, R.C., and Brusca, G.J. 1990. *Invertebrates*. Sinauer, Sunderland, MA.

Bryant, D., Deneault, M., and Sankoff, D. 1999. The breakpoint median problem. Manuscript, Centre de recherches mathématiques, Université de Montréal.

Caprara, A. 1997. Sorting by reversals is difficult, 75–83. In *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97).* ACM, New York.

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

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

Cavender, J.A. 1989. Mechanized derivation of linear invariants. *Mol. Biol. Evol.* 6, 301–316.

Cavender, J.A., and Felsenstein, J. 1987. Invariants of phylogenies: Simple case with discrete states. *J. Class.* 4, 57–71.

Christofferson, M.L., and Araújo-de-Almeida, E.A. 1994. A phylogenetic framework of the enterocoela (Metameria: Coelomata). *Rev. Norest. Biol. (Brazil)* 9, 172–208.

Clary, D.O., and Wolstenholme, D.R. 1985. The mitochondrial DNA molecular of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* 22, 252–271.

Dalkie, K.-S. 1998. Analysis of breakpoints for genome rearrangement. Honours essay, Department of Mathematics, University of Canterbury, New Zealand.

Drolet, S., and Sankoff, D. 1990. Quadratic invariants for multivalued characters. *J. Theoret. Biol.* 144, 117–129.

Eernisse, D.J., Albert, J.S., and Anderson, F.E. 1992. Annelida and Arthropoda are not sister taxa. A phylogenetic analysis of spiralian metazoan morphology. *Syst. Biol.* 41, 305–330.

Evans, S.N., and Speed, T.P. 1993. Invariants of some probability models used in phylogenetic inference. *Ann. Stat.* 21 355–377.

Evans, S.N., and Zhou, X. 1998. Constructing and counting phylogenetic invariants. *J. Comp. Biol.* 5, 713–724.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410.

Felsenstein, J. 1991. Counting phylogenetic invariants in some simple cases. *J. Theoret. Biol.* 152, 357–376.

Ferretti, V., and Sankoff, D. 1993. The empirical discovery of phylogenetic invariants. *Adv. Appl. Prob.* 25, 290–302.

Ferretti, V., and Sankoff, D. 1995. Phylogenetic invariants for more general evolutionary models. *J. Theoret. Biol.* 173, 147–162.

Ferretti, V., and Sankoff, D. 1996. A remarkable nonlinear invariant for evolution with heterogeneous rates. *Math. Biosci.* 134, 71–83.

Ferretti, V., Lang, B.F., and Sankoff, D. 1994. Skewed base compositions, asymmetric transition matrices and phylogenetic invariants. *J. Comput. Biol.* 1, 77–92.

Fu, Y.X. 1995. Linear invariants under Jukes' and Cantor's one-parameter model. *J. Theoret. Biol.* 173, 339–352.

Gallut, C. 1998. Codage de l'ordre des gènes du génome mitochondrial animal en vue d'une analyse phylogénétique. Mémoire de DEA, Université Paris VI Pierre & Marie Curie.

Giribet, G., and Ribera, C. 1998. The position of Arthropods in the animal kingdom: A search for a reliable outgroup for internal arthropod phylogeny. *Mol. Phylogenet. Evol.* 9, 481–488.

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

Hagedorn, T.R. 1999. On the number and structure of phylogenetic invariants. *Advances in Applied Mathematics*. In press.

Hagedorn, T.R., and Landweber, L.F. 1999. Phylogenetic invariants and geometry. Manuscript, College of New Jersey and Princeton University.

Hannenhalli, S., and Pevzner, P.A. 1995. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals), 178–189. In *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing.* ACM, New York.

Hannenhalli, S., Chappey, C., Koonin, E.V., and Pevzner, P.A. 1995. Genome sequence comparison and scenarios for gene rearrangements: A test case. *Genomics* 30, 299–311.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174.

Hendy, M.D., and Penny, D. 1996. Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption. *J. Comp. Biol.* 3, 19–31.

Jin, L., and Nei, M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* 7, 82–102.

Jukes, T.H., and Cantor, C.R. 1969. Evolution of protein molecules 21–132. In Munro, H.N., ed., *Mammalian Protein Metabolism*. Academic Press, New York.

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

Kececioglu, J., and Sankoff, D. 1994. Efficient bounds for oriented chromosome reversal distance, 307–325. In Crochemore, M., and Gusfield, D., eds., *Proceedings of the Fifth Symposium on Combinatorial Pattern Matching. Lecture Notes in Computer Science* 807. Springer-Verlag, New York.

Kececioglu, J., and Sankoff, D. 1995. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica* 13, 180–210.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.

Kimura, M. 1981. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. U.S.A.* 78, 454–458.

Lake, J.A. 1987. A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Biol. Evol.* 4, 167–191.

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

Nadeau, J.H., and Sankoff, D. 1998. Counting on comparative maps. *Trends Genet.* 14, 495–501.

Nguyen, T., and Speed, T.P. 1992. A derivation of all linear invariants for a nonbalanced transversion model. *J. Mol. Evol.* 35, 60–76.

Pe'er, I., and Shamir, R. 1998. The median problems for breakpoints are NP-complete. *Electronic Colloquium on Computational Complexity Technical Report* 98-071, http://www.eccc.uni-trier.de/ eccc.

Rouse, G.W., and Fauchald, K. 1995. The articulation of annelids. *Zool. Scripta* 24, 269–301.

Ruppert, E.E., and Barnes, B.D. 1994. *Invertebrate Zoology*. Saunders, Philadelphia.

Sankoff, D. 1990. Designer invariants for large phylogenies. *Mol. Biol. Evol.* 7, 255–269.

Sankoff, D. 1992. Edit distance for genome comparison based on non-local operations, 121–135. In Apostolico, A., Crochemore, M., Galil, Z., and Manber, U., eds., *Proceedings of the Third Symposium on Combinatorial Pattern Matching. Lecture Notes in Computer Science* 644. Springer-Verlag, New York.

Sankoff, D. 1999. Comparative mapping and genome rearrangement, 124–134. In Dekkers, J.C.M., Lamont, S.J., and Rothschild, M.F., eds., *From Jay Lush to Genomics: Visions for Animal Breeding and Genetics.* Ames, Iowa, Iowa State University, and *AgBiotechNet* 1, http://agbio.cabweb.org/ conference/ index.html.

Sankoff, D., and Blanchette, M. 1997. The median problem for breakpoints in comparative genomics, 251–263. In Jiang, T., and Lee, D.T., eds., *Computing and Combinatorics, Proceedings of COCOON '97. Lecture Notes in Computer Science* 1276. Springer-Verlag, New York.

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

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

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F, and Cedergren, R.J. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. U.S.A.* 89, 6575–6579.

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

Steel, M.A. 1994. Recovering a tree from the leaf colorations it generates under a Markov model. *Appl. Math. Lett.* 7, 19–23.

Steel, M.A., and Fu, Y.X. 1995. Classifying and counting linear phylogenetic invariants for the Jukes–Cantor model. *J. Comp. Biol.* 2, 39–47.

Steel, M. A., Szekeley, L.A., Erdos, P.L., and Waddell, P. 1993. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *NZ J. Bot.* 31, 289–296.

Szekeley, L.A., Steel, M.A., and Erdos, P.L. 1993. Fourier calculus on evolutionary trees. *Adv. Appl. Math.* 14, 200–216.

Tajima, F., and Nei, M. 1984. Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 1, 269–285.

Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversio n and G + C-content biases. *Mol. Biol. Evol.* 9, 678–687.

Tamura, K., and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526.

Valentine, J.W. no date. University of California Museum of Paleontology Metazoa Systematics Page, http://www. ucmp.berkeley.edu/ phyla/metazoasy.html.

Wada, H., and Satoh, N. 1994. Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18S rDNA. *Proc. Natl. Acad. Sci. U.S.A.* 91, 1801–1804.

Watterson, G.A., Ewens, W.J., Hall, T.E., and Morgan, A. 1982. The chromosome inversion problem. *J. Theoret. Biol.* 99, 1–7.

Address reprint requests to:
David Sankoff
Centre de recherches mathématiques
Université de Montréal
CP 6128 Succursale Centre-Ville
Montréal, Québec H3C 3J7, Canada

Email: sankoff@ere.umontreal.ca