Canadian Mathematical Society
Conference Proceedings

Comparative Genomics Via Phylogenetic Invariants For Jukes-Cantor Semigroups

David Sankoff and Mathieu Blanchette

Dedicated to Professor Donald A. Dawson

Abstract. We review the theory of invariants as it has been developed for comparing the DNA sequences of homologous genes from phylogenetically related species, with particular attention to the semigroups used to model sequence evolution. We also outline the computational theory of genome rearrangements, including the optimization problems in calculating edit distances between genomes and the simpler notion of breakpoint distance. The combinatorics of rearrangements, involving non-local changes in the relative order of genes in the genome, are more complex than the base substitutions responsible for gene sequence evolution. Nevertheless we can construct models of gene order evolution through symmetry assumptions about disruptions of gene adjacencies. Based on the extended Jukes-Cantor semigroup that emerges from this modeling, we derive a complete set of linear phylogenetic invariants. We use these invariants to relate mitochondrial genomes from a number of animal phyla and compare the results to parsimony trees also based on gene adjacencies and to minimal breakpoint trees.

1. Invariants for models of sequence evolution

Consider the aligned DNA sequences: $X_1^{(1)} \cdots X_n^{(1)}, \cdots, X_1^{(N)} \cdots X_n^{(N)}$, all of length 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 Figure 1. The terminal vertices represent observed, or present-day, species. The non-terminal 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 $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. 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 [**46**], Kimura [**51, 52**], Tajima and Nei [**75**],

1991 *Mathematics Subject Classification.* Primary 92D15, 60J27; Secondary 68Q25, 13P10.
Key words and phrases. phylogenetic trees, linear invariants; Jukes-Cantor semigroup; genomics; sorting by reversals; breakpoints; Metazoa; mitochondria.

Hasegawa *et al.* [**43**], Cavender [**17**], Jin and Nei [**45**], Tamura [**76**], Nguyen and Speed [**57**], Tamura and Nei [**77**], Steel [**70**] and Ferretti and Sankoff [**32**].

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 $(X_i^{(1)}, \cdots, X_i^{(N)})$, where $X_i^{(J)}$ is the i-th base in the J-th DNA sequence.

In DNA evolution, rates of change between any two elements of $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ tend to be symmetric. With this type of data, 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 non-terminal vertices of degree 3), hence $|V| = 2N - 2$, $|E| = 2N - 3$, and will confine ourselves to symmetric transition matrices.

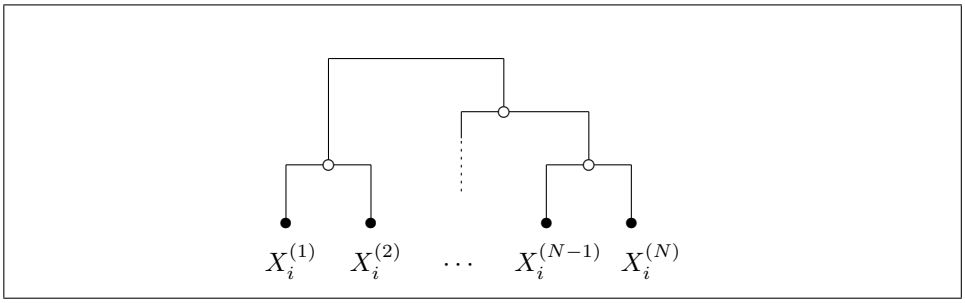

FIGURE 1. Sample trajectory $X_i^{(\cdot)}$. Indexing tree $\mathbf{T}$ is unknown, but the same for all $i = 1, \cdots, n$. Filled dots at terminal vertices indicate N present-day species at which values of the process can be observed, unfilled dots represent unobservable ancestral species.

Phylogenetic invariants are predetermined functions of the probabilities of the observable N-tuples. These functions are identically zero only for $\mathbf{T}$ (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 $\mathbf{T}$ 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 which do not take account of the model used to generate the data may be susceptible to an artifact which tends to group long lineages together and short lineages together.

Lake [**53**] introduced linear invariants in 1987, studying the case $N = 4$ for a 2-parameter semigroup originally suggested by Kimura [**51**]. In the same year, Cavender and Felsenstein [**18**] published quadratic invariants for a 1-parameter semigroup of 2×2 matrices. Subsequently a great deal of research has been carried out into both linear invariants, by Cavender [**17**], Fu [**34**], Nguyen and Speed [**57**],

Steel and Fu [**71**], Hendy and Penny [**44**], and polynomial invariants, by Drolet and Sankoff [**23**], Sankoff [**61**], Felsenstein [**28**], Ferretti *et al.* [**31, 29, 32, 33**], Evans and Speed [**25**], Steel *et al.* [**72**], Szekeley *et al.* [**74**], Steel [**70**], Evans and Zhou [**26**], Hagedorn [**37**] and Hagedorn and Landweber [**38**].

2. The study of gene order evolution.

The aligned DNA sequences discussed in Section 1 represent the internal structure of the same gene in N related species or, more correctly, homologous genes (i.e., diverging from the same ancestral gene) in these species. For the rest of this paper, we turn to a higher level of analysis, the *genome*, comparing the linear ordering of all the different genes in the chromosomes of these species. For these purposes, we do not take into account variation in the DNA sequences among a set of N homologous genes, but simply consider them identical.

As individual genes evolve through the Markovian base substitution discussed in Section 1, as well as other *local* processes such as base deletion or insertion, several additional, *non-local*, evolutionary mechanisms also operate, at the genomic level.

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 certain *rearrangement* processes. For single-chromosome genomes, this has been formulated as the problem of calculating an edit distance between two linear orders on the same set of objects, representing the ordering of homologous genes in two genomes. In the most realistic version of the problem, a sign (plus or minus) is associated with each object in the linear order, representing the direction of transcription of the corresponding gene, which depends on which of the two complementary DNA strands the gene is located. The elementary edit operations (Figure 2) may include one or more of the processes:

1) **inversion**, or **reversal**, of any number of consecutive terms in the ordered set, which, in the case of signed orders, also reverses the sign of each term within the scope of the inversion. Kececioglu and Sankoff [**50**] re-introduced the problem—earlier posed by Waterson *et al.*[**80**], and even earlier in the genetics literature, e.g. [**73**]—of computing the minimum reversal distance between two given permutations in the unsigned case, and gave approximation algorithms and an exact algorithm feasible for moderately long permutations. Bafna and Pevzner [**4**] gave improved approximation algorithms and Caprara [**14**] showed this problem to be NP-complete. Kececioglu and Sankoff [**49**] also found tight lower and upper bounds for the signed case and implemented an exact algorithm which worked rapidly for long permutations. Indeed, Hannenhalli and Pevzner [**41**] showed that the signed problem is only of polynomial complexity, improvements to their algorithm were given by Berman and Hannenhalli [**6**] and by Kaplan *et al.* [**47**].

2) **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. Computation of the transposition distance between two permutations was considered by Bafna and Pevzner [**5**], but its NP-completeness has not yet been confirmed. An

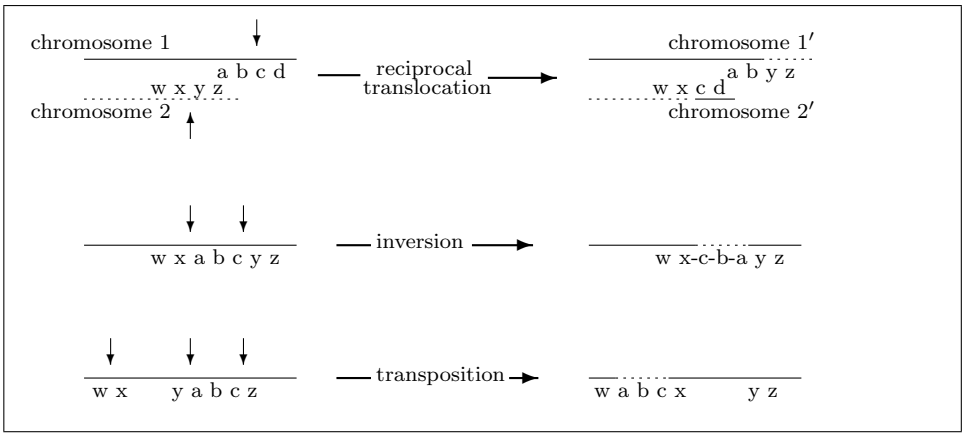

FIGURE 2. Schematic view of genome rearrangement processes. Letters represent positions of genes. Vertical arrows at left indicate breakpoints introduced into original genome. Reciprocal translocation exchanges end segments of two chromosomes. Inversion reverses the order and sign of genes between two breakpoints (dotted segment). Transposition removes a segment defined by two breakpoints and inserts it at another breakpoint (dotted segment), in the same chromosome or another.

edit distance which is a weighted combination of inversions, transpositions and deletions has been studied by Sankoff [**62**] and Blanchette *et al.* [**8**].

There is no difficulty in reformulating these notions to apply to circular rather than linear genomes as is often the case of unichromosomal genomes (e.g. bacteria or mitochondria). For multichromosomal genomes, the most important role is played by:

3) **reciprocal translocation.** Kececioglu and Ravi [**48**] began the investigation of translocation distances, and Hannenhalli [**39, 42**] has shown that two formulations of the problem are of polynomial complexity. Ferretti *et al.* [**30**] proposed a relaxed form of translocation distance applicable when chromosomal assignment of genes, but not their order, is known. The complexity of its calculation was shown to be NP-complete by DasGupta *et al.* [**22**] and its structure was further investigated by Liben-Nowell [**54**].

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 all N species, conserved chromosomal segments in comparative genetic maps (cf. Nadeau and Sankoff [**56**]) or, indeed, the results of any decomposition of the chromosome into disjoint ordered fragments, each identifiable in all N genomes.

3. 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 2 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 [**15**]. An example is the "median" problem: find the "ancestor" genome which is closest to three given genomes. Heuristics are available [**40, 69**], but they are only feasible for small genomes.

The breakpoint distance between two genomes containing the same genes, is the number of pairs of adjacent genes in one genome which are not adjacent in the other [**80**]. (Figure 2 illustrates how breakpoints are created.) 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. Nevertheless its extension to three or more genomes is also NP-hard [**58, 13**]. It does have a simple reduction to the Traveling Salesman Problem [**64**] 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 [**7, 65**], and ultimately to the search for optimal topologies [**9**].

In this kind of phylogenetic inference, breakpoint distance is used as a *parsimony* criterion. And parsimony methods are among those which, under the simplest probabilistic models of mutation, may sometimes reconstruct trees incorrectly when there are some very short and some very long branches [**27**]. This problem, together with the computational complexity of all versions of the multiple genome rearrangement problem, leads us to investigate the potential of phylogenetic invariants for studying gene order evolution. Not only do they avoid branch length problems, but they require negligible calculation time.

But can we find invariants for gene order data? After all, the various sets of breakpoints in a multi-genome 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 inversions 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 fifteen binary unrooted trees where $N = 5$.

4. Probability models for breakpoint distances

We will propose models for inversion on unsigned and signed circular genomes, as well as for transpositions on unsigned genomes. 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 [**9, 63**]. Recall that, as in Figure 2, two different pairs must be disrupted for each inversion, and three for each transposition.

First we will provide detailed notation for the operations described in Section 2 in the case of circular genomes. Consider such a 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 $g_i =$ -γ_i.) By convention, we "view" the circle from the side which 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 $d = a$). An example of an inversion is the operation which 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 Figure 3).

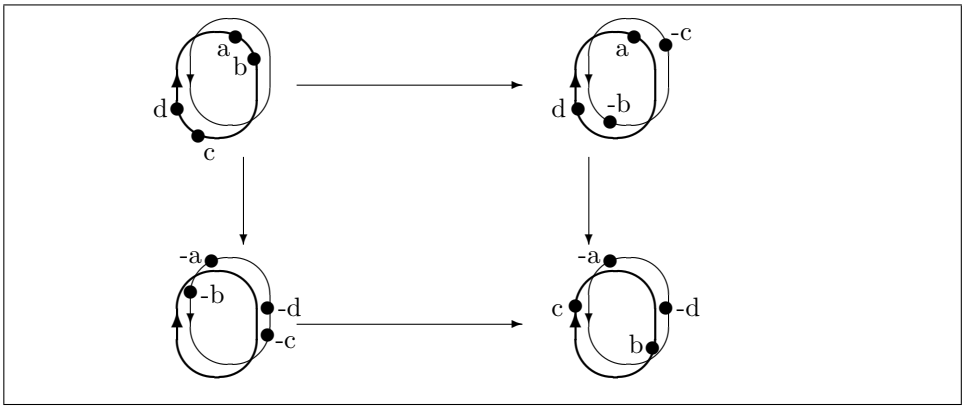

FIGURE 3. Reading direction, sign assignment to genes, and inversion. Reading direction is indicated by arrowheads on each DNA strand. The two genomes on the left are biologically identical; one view can be derived from the other by flipping the genome over and assigning signs to each gene according 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 inverting the segment from gene b to gene c, inclusive. The commutivity of flipping and inversion accords with the fact that it does not matter biologically from which side we view the genome.

We may also consider *unsigned* genomes where the reading direction (or strand) of each gene is unknown. In Figure 3, we may imagine the two strands superimposed and ignore the signs on the genes. In this case, the inversion 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 Figure 3, genes $b \cdots c$ were in the *scope* of the inversion; in the latter, reading clockwise at the bottom right, these genes were not in the scope of the inversion. Though flipping the genome does not change its identity, considering the two representations separately will be important for probabilistic modeling in the next section.

Consider any three pairs of adjacent genes ab, cd and fg, where fg occurs in the interval $d \cdots a$. The operation which 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 an inversion as well.

4.1. Inversions, unsigned case. For an unsigned circular genome, consider a continuous time process with rate $\lambda = 1$. Each change of state involves an inversion, where any two pairs of adjacent genes fg and hk are is equally likely to be disrupted. We focus on a particular gene f and, after each inversion, choose the representation of the resultant genome where f has *not* been in the scope of the inversion. 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 inversion 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 1's, and $1-(n-1)\alpha = e^{-t/(n-2)}$. This is a generalization to $n-1 \times n-1$ matrices of the Jukes-Cantor [**46**] semigroup of 4×4 matrices.

4.2. Inversions, signed case. A model consisting only of random inversions on signed genomes, however, is quite different. Suppose all genes are on the same strand and have positive sign. Then if fg is disrupted by an inversion, 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 definitely does not hold. Moreover, after the first inversion, the strandedness of some genes will have changed, so that for the next inversion, 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 inversion we flip the genome if necessary so as to ensure gene 1 always has 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 i-th position. After k inversions, the probability that the i-th position will be occupied by gene h is

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

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

TABLE 1. Approach of $P_k(x_2 = h)$ to equiprobability.

k \ h	2	3	4	-2	-3	-4
1	0.500	0	0	0.167	0.167	0.167
2	0.333	0.111	0.083	.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

It can be seen that it takes a relatively large number of inversions 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 i-th 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 inversion on signed genomes model for $n = 4$.

TABLE 2. Approach of $P_t(x_2 = h)$ to Jukes-Cantor probabilities.

	random inversions						Jukes-Cantor	
t \ h	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

This table shows that the transition probabilities remain rather inhomogeneous for a considerable time, even for n as small as 4. For $t = 4$, there have been about 8 opportunities on the average for each of the four adjacencies to be disrupted (two per inversion); nonetheless the probabilities are decidedly non-uniform, even among the genes where $h \neq 2$. 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 inversions 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 neighbour of f. But the fact that g cannot replace itself as it could in the unsigned inversion model, leads to the same sort of difficulty as with signed inversion. 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 inversion, or any other particular process, is the only mechanism of genome rearrangement. Inversion, transposition or single-gene movement could also 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 inversions 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 from g 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 which 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 which 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:

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

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

6. The invariants

Using the algorithm of Fu [**34**], 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 Figure 4.

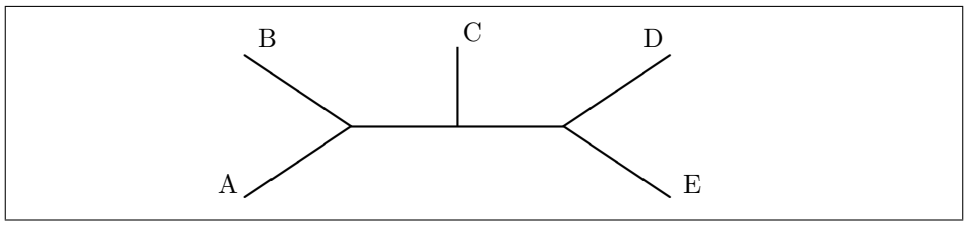

FIGURE 4. Unrooted binary tree ((AB)C(DE)). The other 14 trees are obtained by permuting the 5 labels

The term "complete" is used in the sense that these eleven 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).

$$xyzyx - xyzyw - xyzzx + xyzzw$$

$$xyzyz - xyzyx - xyzwz + xyzwx$$

$$xyzxy - xyzxw - xyzzy + xyzzw$$

$$xyzxz - xyzxy - xyzwz + xyzwy$$

$$xyzzx - xyzzy - xyzwx + xyzwy$$

$$xxyxy - xxyyx + xxyyz - xxyxz - xxyzy + xxyzx$$

$$xyyxy - xyyyx + xyyyz - xyyzy - xyyxz + xyyzx$$

$$xyxxy - xyxyx + xyxyz - xyxxz - xyxzy + xyxzx$$

$$xyxzy - xyxyz + xyyxz - xyyzx + xyzyw - xyzxw - xyzwy + xyzwx$$

$$xyxxy - xyxxz - xyxyy + xyxzz - xyyyx + xyyyz$$
$$+xyyxx - xyyzz + xyzyy - xyzxx - xyzwy + xyzwx$$

$$xyxyy - xyxzz - xyyxx + xyyzz - xyzyy + xyzxx$$
$$+k(xyxzy - xyxzw - xyyzx + xyyzw - xyzwy + xyzwx)$$

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 inversion model in Section 4.1, $k = n - 1$.

6.1. Remarks on the invariants. In examining the eleven 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 which characterize one (top three configurations), the other (second three configurations), or both (bottom 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 non-informative configurations above and other configurations requiring no "extra steps" such as $xyyzw$ or $xyyyz$). Were all the data concentrated on these fourteen configurations, then all the eleven 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 which 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 non-zero 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 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 [**9**]. 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.

TABLE 3. Coelomate mitochondrial genomes compared in this investigation, with higher taxonomic levels. Citations: HU [**2**], SS [**3**], BA [**16**], DR [**20**], KT [**10**], LU [**11**].

	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)

We will evaluate the eleven 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 which 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, e.g. [**1, 19**]; among the groupings in Table 3, only the split between deuterostomes and protostomes seems undisputed. Eernisse *et al.* [**24**], Giribet and Ribera [**36**] 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. [**59**], of a traditional grouping (*Articulata*) of annelids and arthropods as sister taxa. Hemichordates have been grouped with the chordates as in Brusca and Brusca [**12**] or in the "Tree of Life" [**55**], but evidence by Wada and Satoh [**79**] has led many to group them closer to the echinoderms, e.g. [**60, 78**].

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. [**68**] and [**9**], Figures 4a and 4b, an artifact of the mitochondrial genome of the latter being highly divergent, the former two relatively conservative.

Figure 5 contrasts three phylogenies, one representing the "Tree of Life" [**55**], another the summary phylogeny by Valentine [**78**] on the University of California Museum of Paleontology website, and the third the *Drosophila*-human artifact.

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 inversions, 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

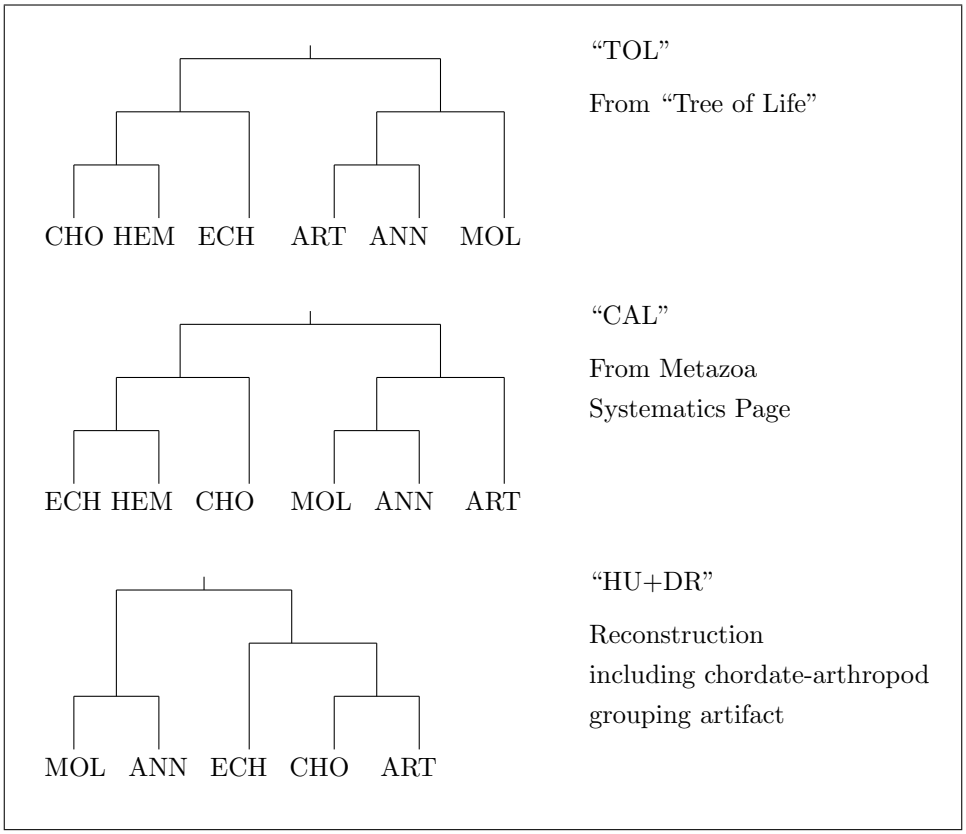

FIGURE 5. Three alternative views of coelomate evolution.

[**65**] and dividing by $2\theta(2N-3)$, each inversion 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.

A more clear-cut result of our method would see the tree $\mathbf{T}$ 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. Simulations show, however, that for genomes with only 37 genes and the degree of divergence present on the coelomate data, this criterion will select the true tree at most 40% of the time, using the most favorable choice of Ψ [**66**]. Thus we rely here on the summary score based on statistical significance.

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 [**35**] 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 3, 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,

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

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 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 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 [**68**] and [**9**]. In this case $R = 10$.

The best three trees manifested scores of $2 \times 10^{-12}, 6 \times 10^{-15}, 7 \times 10^{-17}$. The first of these was consistent with the CAL tree in Figure 5, 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 which 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 $10^{-7}, 10^{-7}, 6 \times 10^{-8}$, 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 Figure 5, 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 [**9**]. It can be seen that the two methods operating on gene-

TABLE 5. Comparison of three methods on three data sets (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 which are tied with at least one other. Parentheses indicate 6-species supertree based on best tree in 5-species analysis without BA, combined with three best trees in 5-species analysis without LU. Note that for 5-species analyses, ranks are out of 15, while for 6 species, they are out of 105.

	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*

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 a ECH+CHO grouping (which is of little credibility) as to the ECH+HEM analysis, whereas the other methods favour a more traditional CHO+HEM grouping.

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,[**21**]).

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. [**72**]) and others, and biological interpretation in the breakpoint context is possible. Even though an *exact* representation of models such as random inversions 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, 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 towards 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 inversions, transpositions and reciprocal translocations. Increasing n only linearly increases the time to compute configuration frequencies, which is negligible. Our simulations [**66**] 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, nor 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 [**9**], 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 book-keeping 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 from NSERC and the Canadian Genome Analysis and Technology program to DS and an NSERC graduate scholarship to MB. DS is a Fellow of the Canadian Institute for Advanced Research. We are grateful to the referees for their comments, and 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. Much of this material also appears in [**67**].

References

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

[2] 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., *Sequence and organization of the human mitochondrial genome*, Nature **290** (1981), 457-465.

[3] Asakawa, S., Himeno, H., Miura, K. and Watanabe, K., *Nucleotide sequence and gene organization of the starfish Asterna pectinifera mitochondrial genome* (1993), unpublished.

[4] Bafna, V. and Pevzner, P.A., *Genome rearrangements and sorting by reversals*, SIAM Journal of Computing **25** (1996), 272-289.

[5] Bafna, V. and Pevzner, P.A., *Sorting by transpositions*, In: Combinatorial Pattern Matching. 6th Annual Symposium, Z.Galil and E. Ukkonen, eds., Lecture Notes in Computer Science 937, Springer Verlag, New York (1995), 614-623.

[6] Berman, P. and Hannenhalli, S., *Fast sorting by reversal*, In: Combinatorial Pattern Matching. 7th Annual Symposium, D. Hirschberg and G. Myers, eds., Lecture Notes in Computer Science 1075, Springer Verlag, New York (1996), 168-185.

[7] Blanchette, M., Bourque, G. and Sankoff, D. *Breakpoint phylogenies*. In: Genome Informatics 1997, S. Miyano and T. Takagi, eds., Universal Academy Press, Tokyo (1997), 25-34.

[8] Blanchette, M., Kunisawa, T. and Sankoff, D., *Parametric genome rearrangement*, Gene **172** (1996), GC 11-17.

[9] Blanchette, M., Kunisawa, T., Sankoff, D., *Gene order breakpoint evidence in animal mitochondrial phylogeny*, Journal of Molecular Evolution **49** (1999), 193-203.

[10] Boore, J.L. and Brown, W.M., *Complete DNA sequence of the mitochondrial genome of the black chiton, Katharina tunicata*, Genetics **138** (1994), 423-443.

[11] Boore, J.L. and Brown, W.M., *Complete sequence of the mitochondrial DNA of the annelid worm Lumbricus terrestris*, Genetics **141** (1995), 305-319.

[12] Brusca, R.C. and Brusca, G.J., *Invertebrates*, Sinauer, Sunderland, MA, (1990).

[13] Bryant, D., *Complexity of the breakpoint median problem*, manuscript, Centre de recherches mathématiques (1998).

[14] Caprara, A., *Sorting by reversals is difficult*, In: Proceeedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97), ACM, New York (1997), 75-83.

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

[16] Castresana, J., Feldmaier-Fuchs, G. and Paabo, S., *Codon reassignment and amino acid composition in hemichordate mitochondria*, Proceedings of the National Academy of Sciences (U.S.A.) **95** (1998), 3703-3707.

[17] Cavender, J.A., *Mechanized derivation of linear invariants*, Molecular Biology and Evolution **6** (1989), 301-316.
[18] Cavender, J.A. and Felsenstein, J., *Invariants of phylogenies: Simple case with discrete states*, Journal of Classification **4** (1987), 57-71.
[19] Christofferson, M.L. and Araújo-de-Almeida, E.A., *A phylogenetic framework of the enterocoela (Metameria: Coelomata)*, Revista Norestina de Biologia (Brazil) **9** (1994), 172-208.
[20] Clary, D.O. and Wolstenholme, D.R., *The mitochondrial DNA molecular of Drosophila yakuba: nucleotide sequence, gene organization, and genetic code*, Journal of Molecular Evolution **22** (1985), 252-271.
[21] Dalkie, K.-S., *Analysis of breakpoints for genome rearrangement*, Honours essay, Department of Mathematics, University of Canterbury, New Zealand (1998).
[22] DasGupta, B., Jiang, T., Kannan, S., Li, M. and Sweedyk, Z., *On the complexity and approximation of syntenic distance*, In: Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97), ACM, New York, (1997), 99-108.
[23] Drolet, S. and Sankoff, D., *Quadratic invariants for multivalued characters*, Journal of Theoretical Biology **144** (1990), 117-129.
[24] Eernisse, D.J., Albert, J.S. and Anderson, F.E., *Annelida and Arthropoda are not sister taxa. A phylogenetic analysis of spiralian metazoan morphology*, Systematic Biology **41** (1992), 305-330.
[25] Evans, S.N. and Speed, T.P., *Invariants of some probability models used in phylogenetic inference*, Annals of Statistics **21** (1993), 355-377.
[26] Evans, S.N. and Zhou, X., *Constructing and counting phylogenetic invariants*, Journal of Computational Biology **5** (1998), 713-724.
[27] Felsenstein, J., *Cases in which parsimony or compatibility methods will be positively misleading*, Systematic Zoology **27** (1978), 401-410.
[28] Felsenstein, J., *Counting phylogenetic invariants in some simple cases*, Journal of Theoretical Biology **152** (1991), 357-376.
[29] Ferretti, V., Lang, B.F. and Sankoff, D., *Skewed base compositions, asymmetric transition matrices and phylogenetic invariants*, Journal of Computational Biology **1** (1994), 77-92.
[30] Ferretti, V., Nadeau, J.H. and Sankoff, D., *Original synteny*, In: Combinatorial Pattern Matching. 7th Annual Symposium, D. Hirschberg and G. Myers, eds., Lecture Notes in Computer Science 1075, Springer Verlag, New York (1996), 159-167.
[31] Ferretti, V. and Sankoff, D., *The empirical discovery of phylogenetic invariants*, Advances in Applied Probability **25** (1993), 290-302.
[32] Ferretti V. and Sankoff D., *Phylogenetic invariants for more general evolutionary models*, Journal of Theoretical Biology **173** (1995), 147-162.
[33] Ferretti, V. and Sankoff, D., *A remarkable nonlinear invariant for evolution with heterogeneous rates*, Mathematical Biosciences **134** (1996), 71-83.
[34] Fu Y. X., *Linear invariants under Jukes' and Cantor's one-parameter model*, Journal of Theoretical Biology **173** (1995), 339-352.
[35] Gallut, C., *Codage de l'ordre des génes du génome mitochondrial animal en vue d'une analyse phylogénétique*, Mémoire de DEA, Université Paris VI Pierre et Marie Curie (1998).
[36] Giribet, G. and Ribera, C., *The position of Arthropods in the animal kingdom: A search for as reliable outgroup for internal arthropod phylogeny*, Molecular Phylogenetics and Evolution **9** (1998), 481-488.
[37] Hagedorn, T.R., *On the number and structure of phylogenetic invariants*, Advances in Applied Mathematics (in press).
[38] Hagedorn, T.R. and Landweber, L.F., *Phylogenetic invariants and geometry*, manuscript, College of New Jersey and Princeton University (1999).
[39] Hannenhalli, S., *Polynomial algorithm for computing translocation distance between genomes.* In: Combinatorial Pattern Matching. 6th Annual Symposium, Z.Galil and E. Ukkonen, eds., Lecture Notes in Computer Science 937, Springer Verlag, New York (1995), 162-176.
[40] Hannenhalli, S., Chappey, C., Koonin, E.V. and Pevzner, P.A., *Genome sequence comparison and scenarios for gene rearrangements: a test case*, Genomics **30** (1995), 299-311.
[41] Hannenhalli, S. and Pevzner, P.A., *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 (1995), 178-189.

[42] Hannenhalli, S. and Pevzner, P.A., *Transforming men into mice (polynomial algorithm for genomic distance problem)*, In: Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science (1995), 581-592

[43] Hasegawa, M., Kishino, H. and Yano, T., *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA*, Journal of Molecular Evolution **22** (1985), 160-174.

[44] Hendy, M.D. and Penny, D., *Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption*, Journal of Computational Biology **3** (1996), 19-31.

[45] Jin, L. and Nei, M., *Limitations of the evolutionary parsimony method of phylogenetic analysis*, Molecular Biology and Evolution **7** (1990), 82-102.

[46] Jukes, T.H. and Cantor C.R., *Evolution of protein molecules*, In: Mammalian Protein Metabolism, H.N. Munro, ed., Academic Press, New York (1969), 21-132.

[47] Kaplan, H., Shamir, R. and Tarjan, R.E., *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 (1997), 344-351.

[48] Kececioglu, J. and Ravi, R., *Of mice and men. Evolutionary distances between genomes under translocation*, In: Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms, (1995), 604-613.

[49] Kececioglu, J. and Sankoff, D., *Efficient bounds for oriented chromosome inversion distance*, In: Combinatorial Pattern Matching. 5th Annual Symposium, M. Crochemore and D. Gusfield, eds., Lecture Notes in Computer Science, 807, Springer Verlag, New York (1994), 307-325.

[50] Kececioglu, J. and Sankoff, D., *Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement*, Algorithmica **13** (1995),180-210.

[51] Kimura, M., *A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences*, Journal of Molecular Evolution **16** (1980), 111-120.

[52] Kimura, M., *Estimation of evolutionary sequences between homologous nucleotide sequences*, Proceedings of the National Academy of Sciences (U.S.A.) **78** (1981), 454-458.

[53] Lake, J.A., *A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony*, Molecular Biology and Evolution **4** (1987), 167-191.

[54] Liben-Nowell, D., *On the structure of syntenic distance*, In: Combinatorial Pattern Matching. 10th Annual Symposium, M. Crochemore and M. Paterson, eds., Lecture Notes in Computer Science 1645, Springer Verlag, New York (1999), 50-65.

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

[56] Nadeau, J.H. and Sankoff, D., *Counting on comparative maps*, Trends in Genetics **14** (1998), 495-501.

[57] Nguyen, T and Speed, T.P., *A derivation of all linear invariants for a nonbalanced transversion model*, Journal of Molecular Evolution **35** (1992), 60-76.

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

[59] Rouse, G.W. and Fauchald, K., *The articulation of annelids*, Zoologica Scripta **24** (1995), 269-301

[60] Ruppert, E.E. and Barnes, B.D., *Invertebrate Zoology*, Saunders, Philadelphia (1994).

[61] Sankoff, D., *Designer invariants for large phylogenies*, Molecular Biology and Evolution **7** (1990), 255-269.

[62] Sankoff, D., *Edit distance for genome comparison based on non-local operations*, In: Combinatorial Pattern Matching. 3rd Annual Symposium, A. Apostolico, M. Crochemore, Z. Galil and U. Manber, eds., Lecture Notes in Computer Science 644, Springer Verlag, New York (1992), 121-135.

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

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

[65] Sankoff, D. and Blanchette, M., *Multiple genome rearrangement and breakpoint phylogeny*, Journal of Computational Biology **5** (1998), 555-570.

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

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

[68] Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F and Cedergren, R.J., *Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome*, Proceedings of the National Academy of Sciences (U.S.A.) **89** (1992), 6575-6579.

[69] Sankoff, D., Sundaram, G. and Kececioglu, J., *Steiner points in the space of genome rearrangements*, International Journal of the Foundations of Computer Science **7** (1996), 1-9.

[70] Steel, M.A., *Recovering a tree from the leaf colorations it generates under a Markov model*, Applied Mathematics Letters **7** (1994), 19-23.

[71] Steel, M. A. and Fu, Y.X., *Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model*, Journal of Computational Biology **2** (1995), 39-47.

[72] Steel, M. A., Szekeley, L.A., Erdos, P.L. and Waddell, P., *A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model*, New Zealand Journal of Botany **31** (1993), 289-296.

[73] Sturtevant, A.H. and Novitski, E., *The homologies of chromosome elements in the genus Drosophila*, Genetics **26** (1941), 517-541.

[74] Szekeley, L.A., Steel, M. A. and Erdos, P.L., *Fourier calculus on evolutionary trees*, Advances in Applied Mathematics **14** (1993), 200-216.

[75] Tajima, F. and Nei, M., *Estimation of evolutionary distance between nucleotide sequences*, Molecular Biology and Evolution **1** (1984), 269-285.

[76] Tamura, K., *Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases*, Molecular Biology and Evolution **9** (1992), 678-687.

[77] Tamura, K. and Nei, M., *Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees*, Molecular Biology and Evolution **10** (1993), 512-526.

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

[79] Wada, H., and Satoh, N., *Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18S rDNA*, Proceedings of the National Academy of Sciences (U.S.A.) **91** (1994), 1801-1804.

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

Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec H3C 3J7

E-mail address: sankoff@ere.umontreal.ca

Department of Computer Science & Engineering, University of Washington, Seattle, Washington 98195-2350

E-mail address: blanchem@cs.washington.edu