Probability Models for Genome Rearrangement and Linear Invariants for Phylogenetic Inference

David Sankoff * Mathieu Blanchette †

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 rearrangment 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 inversions on unsigned circular genomes and, using a non-trivial recurrence relation, inversions 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 inversions, as well as for a mixed rearrangement model for signed genomes, though not for pure transposition nor pure signed inversion 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.

1 Introduction.

1.1 Genomic distances: hard, medium and easy.

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

Consider a circular 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 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. The operation:
taking $g_1 \cdots ab \cdots cd \cdots g_n$ to $g_1 \cdots a\ \text{-}c \cdots \text{-}bd \cdots g_n$
(or, equivalently, to $\text{-}g_n \cdots \text{-}db \cdots c\ \text{-}a \cdots \text{-}g_1$, as illustrated in Figure 1) is called an inversion (or reversal). N.B. possibly $b = c$ or $d = a$.

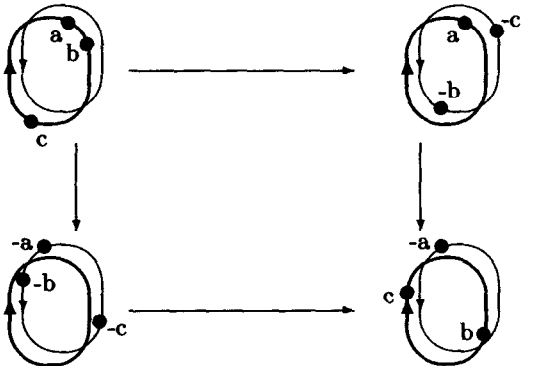

Figure 1: Reading direction, sign assignment to genes, and inversion. Reading direction is indicated by arrowheads on each DNA strand. The two genomes on the left are biologically identical; one view can be derived from the other by flipping the genome over and assigning signs to each gene according 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 commutativity 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 (i.e. strand) of each gene is unknown. In this case, an inversion transforms $\gamma_1 \cdots ab \cdots cd \cdots \gamma_n$ to $\gamma_1 \cdots ac \cdots bd \cdots \gamma_n$

Consider any three pairs of adjacent genes ab, cd and fg, where fg occurs in the interval $d \cdots a$. Transposition is an operation which takes $g_1 \cdots ab \cdots cd \cdots fg \cdots g_n$ to $g_1 \cdots ad \cdots fb \cdots cg \cdots g_n$

The study of comparative genomics, discussed more fully by Sankoff *et al.*(1997) and by Sankoff and Blanchette (1999),

*Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: sankoff@ere.umontreal.ca.

†Computer Science Department, University of Washington, Seattle, Washington 98195. E-mail: blanchem@cs.washington.edu.

RECOMB '99 Lyon France

has focused on inferring the most economical explanation for observed differences in gene orders in two or more genomes in terms of these kinds of operation. This reduces to the problem of calculating an edit distance between two circular orderings of the same set of objects (either signed or unsigned).

For the minimum inversion distance between two genomes in the unsigned case, Caprara (1997) showed the problem to be NP-complete. On the other hand, Hannenhalli and Pevzner (1995) showed that the signed problem is only of polynomial complexity. An improved algorithm was given by Kaplan *et al.* (1997).

Computation of the transposition distance between two permutations was considered by Bafna and Pevzner (1995), but its NP-completeness has not been confirmed. An edit distance which is a weighted combination of inversions and transpositions has been studied by Blanchette *et al.* (1996).

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 which are not adjacent in the other. 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.

1.2 Phylogeny based on gene orders

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 inversion distance for signed genomes (itself only of quadratic complexity), to three or more genomes — multiple genome rearrangement — is NP-hard (Caprara 1998).

Though breakpoint distance is easy to calculate, its extension to three or more genomes is also NP-hard (Pe'er and Shamir 1998; Bryant 1998). 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.* 1998).

In this kind of phylogenetic inference, breakpoint distance is essentially 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. This problem, together with the computational compexity 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.

1.3 Invariants for models of sequence evolution.

Consider the N aligned DNA sequences of length n: $X_1^{(1)} \cdots X_n^{(1)}, \cdots, X_1^{(N)} \cdots X_n^{(N)}$ representing N species related through an unknown phylogenetic tree $\mathbf{T} = (V, E)$, as in Figure 2. For each i, the $X_i^{(J)}$ are the terminal points of

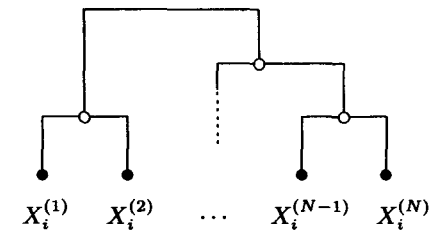

Figure 2: Sample trajectory $X_i^{(\cdot)}$. Indexing tree $\mathbf{T}$ is unknown, but the same for all $i = 1, \cdots, n$. Filled dots at terminal vertices indicate N present-day species at which values of the process can be observed, unfilled dots represent unobservable ancestral species.

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 $|E|$ (unknown) 4×4 Markov matrices with positive determinant all belonging to a (known) semigroup, e.g. any of the semigroups 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) or Ferretti and Sankoff (1995). One of the $|E|$ matrices is associated to each edge of $\mathbf{T}$. Only the values at the N terminal vertices of the trajectory are observed, giving a data vector of form $(X_i^{(1)}, \cdots, X_i^{(N)})$, where $X_i^{(J)}$ is the i-th base in the J-th DNA sequence.

The 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.

A 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 (1987) introduced linear invariants, studying the case $N = 4$ for a 2-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 1-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 (1998) and Hagedorn and Landweber (1998).

Can we apply this theory to comparative genomics? Gene orderings and breakpoint sets 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 the next section, we present simple models for genome rearrangement processes in analogy to the base substitution models for gene sequence evolution. Then, in Section 5, we use these models to calculate complete sets of linear invariants for the fifteen binary unrooted trees where $N = 5$.

2 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 adjacent pairs of genes fg are equally likely to be disrupted.

2.1 Inversions, unsigned case.

Consider first inversions on unsigned circular genomes. We will assume a uniform probability rate $\lambda = 1$ of inversion events occurring. At each event, any choice of two pairs of adjacent genes fg and hk is equally likely to be disrupted. If fg is chosen (with 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 $\alpha = \log(1-e^{-t})$. This is an extended form of the Jukes-Cantor semigroup of matrices (1969), characterized by $n-2$ equal transition probabilities per row.

2.2 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.

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

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, let the probability that the i-th position will be occupied by gene h be $P_k(x_i = h)$

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

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

It can be seen that it takes a relatively large number of 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 $n = 4$.

t \ h	random inversions 2	3	4	-2	-3	-4
1	0.485	0.001	0.001	0.044	0.042	0.044
2	0.499	0.041	0.034	0.100	0.092	0.100
4	0.330	0.108	0.096	0.153	0.141	0.153
8	0.190	0.159	0.154	0.167	0.163	0.167

t \ h	Jukes-Cantor 2	others
1	0.672	0.066
2	0.473	0.105
4	0.279	0.144
8	0.182	0.164

Table 2: Approach of $P_t(x_2 = h)$ to Jukes-Cantor model.

This table shows that the transition probabilities remain remarkably inhomogeneous for a considerable time, even for n as small as 4. Note that for $t = 4$, there have been 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.

2.3 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.

3 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 all play a role, in unknown proportions. Thus, where h appears in the original genome, we will not assume that only $-h$ can replace g, contrary to the pure inversions case. We assume instead 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 $h \neq g$, including $h = -g$. 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 undistinguishable objects into five labeled boxes.

4 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)). We use the configuration label as a shorthand for the configuration probability normalized by the number of N-tuples it represents.

$$xyzyx - xyzyw - xyzzx + xyzzw$$

$$xyzyz - xyzyx - xyzwz + xyzwx$$

$$xyzxy - xyzxw - xyzzy + xyzzw$$

$$xyzxz - xyzxy - xyzwz + xyzwy$$

$$xyzzx - xyzzy - xyzwx + xyzwy$$

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

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

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

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

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

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

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, $k = n-1$.

4.1 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, 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, as discussed by Sankoff and Blanchette (1999). The situation would be much worse for $N = 6$ with 203 configurations, one of the difficulties in proceeding beyond $N = 5$.

	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

Table 3: Coelomate mitochondrial genomes compared in this investigation, with higher taxonomic levels. 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).

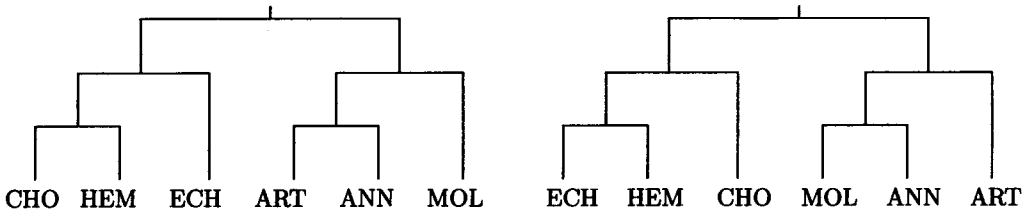

Figure 3: Two views of coelomate evolution. The phylogeny on the left, "TOL", represents the "Tree of Life" (Maddison and Maddison 1995). The other, "CAL", is from the University of California Museum of Paleontology (Valentine, nd).

5 An application to metazoan phylogeny.

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.* 1998). The conservatism of certain genomes, such as human, *Drosophila* and *Katharina tunicata* (a chiton), versus the extreme divergence of related taxa, 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 this section 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 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.

Aspects of coelomate metazoan phylogeny are controversial; among the taxa 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 taxa, 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.). See Figure 3.

Aside from these unsettled questions, efforts to infer phylogeny based on distances between mitochondrial gene orders have tended to group HU+DR rather than HU+SS (e.g. Sankoff *et al.* 1992), an artifact of the echinoderm genome being much more divergent than the other two.

6 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 (Sankoff and Blanchette 1998a) and dividing by $2\theta(2N-3)$, each inversion contributing up to 2 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 posssible 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 sig-

nificance levels.

7 Results.

7.1 Deuterostomes and protostomes.

The first subset of the data to be examined includes HU, SS, DR, KT and LU. 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 and by far the best of these was consistent with the CAL tree in Figure 3, while the the HU+DR artifact was relegated to the much lower-scoring second and third trees.

Thus our method succeeded in grouping CHO and ECH, despite discordancy of branch lengths which defeats distance-matrix-based attempts. It also confirmed the ANN+MOL grouping in CAL versus the TOL grouping of ANN+ART.

7.2 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}. In this analysis the best tree is consistent with the TOL tree in Figure 3, while the CAL tree is third best.

Discussion and further work.

Though much probabilistic modeling of gene sequence change has been incorporated into phylogenetic analysis, very little research has gone into mathematical approaches to phylogenetics based on gene order, and virtually none, previous to the present undertaking, into probability models for the evolution of gene order on a phylogeny.

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 inversions only on signed data in terms of semigroups of matrices may not be possible, significantly better approximations than Jukes-Cantor may well be feasible.

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 (Sankoff and Blanchette 1998b) 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 (Blanchette *et al.* 1998), 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 Section 9, might be systematized.

Our results at the biological level include the early branching of arthropods within the protostomes, and the grouping of the hemichordates with the chordates, though the latter grouping is equivocal. Our method clearly distinguishes between deuterostomes and protostomes.

Acknowledgements

Research supported by grants from the Natural Sciences and Engineering Research Council of Canada 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.

References

[1] 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* 290, 457-465.

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

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

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

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

[6] Blanchette, M., Kunisawa, T. and Sankoff, D. (1998). Gene order breakpoint evidence in animal mitochondrial phylogeny. manuscript.

[7] 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.

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

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

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

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

[12] Caprara, A. (1997). Formulations and complexity of multiple sorting by reversals. manuscript, University of Bologna.

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

[14] Cavender, J.A. (1989). Mechanized derivation of linear invariants. *Molecular Biology and Evolution* 6, 301-316.

[15] Cavender, J.A. and Felsenstein, J. (1987). Invariants of phylogenies: Simple case with discrete states. *Journal of Classification* 4, 57-71.

[16] Clary, D.O. and Wolstenholme, D.R. (1985). The mitochondrial DNA molecular of Drosophila yakuba: nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution* 22, 252-271.

[17] Drolet, S. and Sankoff, D. (1990). Quadratic invariants for multivalued characters. *Journal of Theoretical Biology* 144, 117-129.

[18] 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. *Systematic Biology* 41, 305-330.

[19] Evans, S.N. and Speed, T.P. (1993). Invariants of some probability models used in phylogenetic inference. *Annals of Statistics* 21, 355-377.

[20] Evans, S.N. and Zhou, X. (1998). Constructing and counting phylogenetic invariants. *Journal of Computational Biology.* to appear.

[21] Felsenstein, J. (1991). Counting phylogenetic invariants in some simple cases. *Journal of Theoretical Biology* 152, 357-376.

[22] Ferretti, V., Lang, B.F. and Sankoff, D. (1994). Skewed base compositions, asymmetric transition matrices and phylogenetic invariants. *Journal of Computational Biology* 1, 77-92.

[23] Ferretti, V. and Sankoff, D. (1993). The empirical discovery of phylogenetic invariants. *Advances in Applied Probability* 25, 290-302.

[24] Ferretti V. and Sankoff D. (1995). Phylogenetic invariants for more general evolutionary models. *Journal of Theoretical Biology* 173, 147-162.

[25] Ferretti, V. and Sankoff, D. (1996). A remarkable nonlinear invariant for evolution with heterogeneous rates. *Mathematical Biociences* 134, 71-83.

[26] Fu Y. X. (1995). Linear invariants under Jukes' and Cantor's one-parameter model. *Journal of Theoretical Biology* 173, 339-352.

[27] Giribet, G. and Ribera, C. (1998). The position of Arthropods in the animal kingdom: A search for as reliable outgroup for internal arthropod phylogeny. *Molecular Phylogenetics and Evolution* 9, 481-488.

[28] Hagedorn, T.R. (1998). Determining the structure and number of phylogenetic invariants. manuscript, College of New Jersey.

[29] Hagedorn, T.R. and Landweber, L.F. (1998). Phylogenetic invariants and geometry. manuscript, College of New Jersey and Princeton University.

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

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

[32] 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. *Journal of Computational Biology* 3, 19-31.

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

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

[35] Kaplan, H., Shamir, R. and Tarjan, R.E. (1997). Faster and simpler algorithm for sorting signed permutations by reversals. In: *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* ACM, pp.344-351.

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

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

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

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

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

[41] Pe'er, I. and Shamir, R. (1998). The median problems for breakpoints are NP-complete. manuscript, University of Washington.

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

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

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

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

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

[47] Sankoff, D., and Blanchette, M. 1998b. Phylogenetic invariants for metazoan mitochondrial genome evolution. In: *Genome Informatics 1998* (Eds: S. Miyano and T. Takagi) Universal Academy Press, Tokyo, pp. 22-23.

[48] Sankoff, D., and Blanchette, M. (1999) Comparative genomics via phylogenetic invariants for Jukes-Cantor semigroups. In:*Proceedings of the International Conference on Stochastic Models* (Eds: L. Gorostiza and G.Ivanoff) Conference Proceedings Series, Canadian Mathematical Society, in press.

[49] 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. *Proceedings of the National Academy of Sciences USA* 89, 6575-6579.

[50] Sankoff, D., Parent, M.-N., Marchand, I. and Ferretti, V. (1997) On the Nadeau-Taylor theory of conserved chromosome segments. In: *Combinatorial Pattern Matching. Eighth Annual Symposium* (Eds: A. Apostolico and J. Hein) *Lecture Notes in Computer Science* 1264, Springer Verlag, pp. 262-274.

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

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

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

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

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

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

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

[58] Valentine, J.W. (nd) University of California Museum of Paleontology Metazoa Systematics Page, http://www.ucmp.berkeley.edu/phyla/metazoasy.html

[59] Wada, H., and N. Satoh. 1994. 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, 1801-1804.

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