Phylogenetic Invariants for Metazoan Mitochondrial Genome Evolution

David Sankoff [1] **Mathieu Blanchette** [2]

sankoff@ere.umontreal.ca blanchem@cs.washington.edu

[1] Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-Ville, Montréal, Québec H3C 3J7

[2] Department of Computer Science, University of Washington, Seattle, WA 98195

Abstract

The method of phylogenetic invariants was developed to apply to aligned sequence data generated, according to a stochastic substitution model, for N species related through an unknown phylogenetic tree. The invariants are functions of the probabilities of the observable N-tuples, which are identically zero, over all choices of branch length, for some trees. Evaluating the invariants associated with all possible trees, using observed N-tuple frequencies over all sequence positions, enables us to rapidly infer the generating tree.

An aspect of evolution at the genomic level much studied recently is the rearrangements of gene order along the chromosome from one species to another. Instead of the substitutions responsible for sequence evolution, we examine the non-local processes responsible for genome rearrangements such as inversion of arbitrarily long segments of chromosomes. By treating the potential adjacency of each possible pair of genes as a "position", an appropriate "substitution" model can be recognized as governing the rearrangement process, and a probabilistically principled phylogenetic inference can be set up. We calculate the invariants for this process for $N = 5$, and apply them to mitochondrial genome data from coelomate metazoans, showing how they resolve key aspects of branching order.

1 Introduction

The use of gene order data for finding globally optimal phylogenetic trees is inherently difficult. Not only are some measures of genomic distance computationally complex [2], but more important, the extension of any of them, even the reversals-distance for signed genomes (quadratic complexity [8]), or the breakpoint distance (linear complexity [18]), to three or more genomes — multiple genome rearrangement — is NP-hard [3, 12]. This holds even for the smallest example, the "median" problem: find the "ancestor" genome which is closest to three given genomes.

In contrast to reversal distance and related edit-distances, for which there are not even any good heuristics, the breakpoint distance — essentially the number of pairs of adjacent genes in one genome which are not adjacent in the other — does have a simple reduction to the Traveling Salesman Problem and can benefit from efficient software available for the latter to find the median of three moderate-sized genomes. This can be incorporated into a heuristic for the optimization of fixed-topology phylogenies, and ultimately to the search for optimal topologies [13].

In this kind of phylogenetic inference, however, breakpoint distance is used as a parsimony criterion. And like all parsimony criteria, under the simplest probabilistic models of mutation, there is a range of tree topologies, especially trees with some very short branches and some very long branches, which breakpoint minimization may reconstruct incorrectly.

An approach to tree reconstruction under probabilistic models which is designed to be insensitive to branch lengths is that of phylogenetic invariants. When applied to N aligned nucleotide sequences, the invariants associated with a specific tree topology are precomputed functions of the probabilities of the N-tuples at each sequence position. They are identically zero (i.e. independent of the tree branch lengths) when these N-tuples are generated under a given mutational model over that specific tree. We can rapidly evaluate the functions specific to each tree, using the N-tuple frequencies as estimates of the probabilities, and estimate the correct tree as the one for which the invariant values are closest to zero.

The various sets of breakpoints in a multi-genome comparison, however, do not resemble a multiple alignment of sequences in any way, so that the phylogenetic invariants developed in the context of nucleotide sequence data are not applicable. Stochastic models for reversals and other genome rearrangement processes proposed in analogy to probability models for nucleotide substitution do not generally have the semigroup property necessary to develop a theory of invariants [14]. In this paper, 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 $N = 5$.

The mitochondrial genome of many metazoans has been completely sequenced and the genes they contain identified. Mitochondrial gene order has proven useful for the inference of metazoan phylogeny [1]. The conservatism of some of the genomes and the extreme divergence of others, i.e. the presence of both short and long branches, is the chief difficulty in the reconstruction of this phylogeny. Here we apply our new method of breakpoint invariants to explore three problems in coelomate metazoan phylogeny: the protostome-deuterostome split, the internal structure of the protostomes, and the internal branching order of the deuterostomes.

2 The method of invariants for models of sequence evolution.

Consider the 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 $\mathbf{T} = (V, E)$. 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 $|E|$ (unknown) 4×4 Markov matrices with positive determinant all belonging to a (known) semigroup. (Reference [13] cites most of the semigroups which have been proposed in this context.) 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.

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 (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. Much of the subsequent research into linear and polynomial invariants is cited in reference [13]. In Section 4, we will derive all linear invariants for unrooted binary trees for five species using a method of Fu [5] applied to a semigroup of matrices much larger than 4×4.

3 Metazoan phylogeny

Aspects of coelomate metazoan phylogeny are controversial (cf [6, 14]; among the taxa analyzed here (Table 1), only the split between deuterostomes and protostomes seems undisputed. Many scholars would group annelids and molluscs as sister taxa, with arthropods related to these at a deeper level. But there are proponents of a grouping (*Articulata*) of annelids and arthropods as sister taxa. Hemichordates have traditionally been grouped with the chordates, but recent evidence has led many to group them closer to the echinoderms. See Figure 1.

	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 1: Coelomate mitochondrial genomes compared in this investigation, with higher taxonomic levels. Reference [13] cites the original sources for the mitochondrial sequences.

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. [15], an artifact of the latter being highly divergent, the former two relatively conservative.

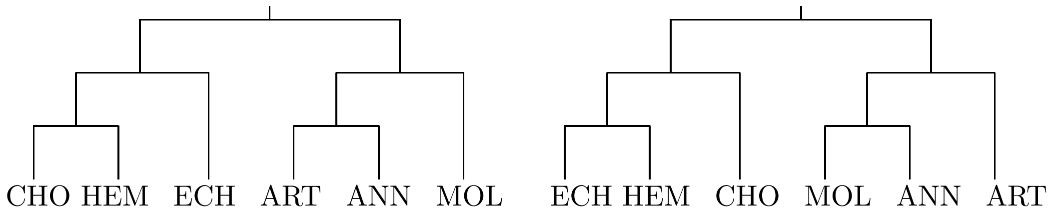

Figure 1: Two views of coelomate evolution. The phylogeny on the left, "TOL", represents the "Tree of Life" (Maddison and Maddison 1995). The other, "CAL", is from the University of California Museum of Paleontology (Valentine, n.d.).

How were the species in Table 1 chosen? As of August 1998, mitochondrial gene order was known for 59 metazoan species, including 39 chordates, one hemichordate, seven arthropods, one annelid, five echinoderms, three molluscs and three nematodes.

As will be explained in Section 4, the linear invariant methodology is appropriate to at most five species when the number of genes is only 37. Previous studies [1] indicate that the nematode lineage (pseudocoelomates) has diverged almost to the point where the mitochondrial genomes of these species have as many breakpoints with the remaining genomes (all coelomates) as do pairs of random genomes. We thus concentrated on the coelomate metazoans. To focus on the earliest divergences we picked relatively conservative members of each of the major groupings; e.g. two highly divergent snail genomes were bypassed in favour of the conservative *Katharina tunicata*. This resulted in the six species in Table 1. The same set of thirty-seven genes is present in the mitochondrial genome of all of them.

4 Extended Jukes-Cantor model for breakpoint data.

In developing a model of the evolution of the breakpoint set between two circular genomes such as that of the mitochondrion, we first consider the effect of a typical rearrangement event, an inversion (or reversal), on the set of gene adjacencies of a given genome. Recall that the n genes are not only ordered along the genome, but they are also "signed" (+ or -) to indicate their reading (transcription) direction or, equivalently, on which of the two DNA strands the gene is located. Inversion of a fragment changes the sign of each gene in that fragment, i.e. changes its "strandedness" and reading direction. Also, reversing the order and signs of all n genes in the entire genome simply provides a different representation of the same genome, since neither strand of DNA has a particular status, and the circular genome has no natural starting point.

The inversion disrupts two adjacencies, say between genes f and g and between h and k, where $g \cdots h$ is the inverted fragment. Two new adjacencies are created, between f and $-h$ and between $-g$ and k. (Note that for $n > 2$ either $g = h$ or $f = k$ is possible but not both). In the initial state, g succeeds f; equivalently, $-f$ succeeds $-g$ in the inverted representation of the genome. In the final state $-h$ succeeds f, or $-f$ succeeds h.

For a random inversion, defined by two adjacencies chosen at random among the n in the circular genome, the probability that any adjacency fg is replaced by $f, -h$, where h is any gene in the genome other than f, is $\binom{n}{2}^{-1}$. This probability is independent of f, g and h.

The model we construct, however, will not assume that inversion is the only mechanism of genome rearrangement, nor will it assume that only one rearrangement event has occurred in a given time interval. We will nonetheless carry over the notion of a change parameter that is the same for all f, g and h. But we will not assume that only $-h$ can replace g, where h and not $-h$ appears in the original genome, as in the single inversions case. Transpositions, multiple inversions, single-gene movements could also play a role, in unknown proportions. Thus, 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 $1-(2n-3)\alpha > \alpha$ 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 length of the time interval t (in arbitrary units) is related to α by $t = -\log \det M(\alpha)$. The Jukes-Cantor model is 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 them, since we are interested in finding the correct tree topology in a way which is insensitive to these lengths.

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. Note that g and $-g$ occurring in the same N-tuple will warrant 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.

5 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)). The term "complete" is used in the sense that these eleven invariants form a basis for the ideal of invariants. We use the configuration label as a shorthand for the configuration probability normalized by the number of N-tuples it represents.

$$xyzyx - xyzyw - xyzzx + xyzzw$$

$$xyzyz - xyzyx - xyzwz + xyzwx$$

$$xyzxy - xyzxw - xyzzy + xyzzw$$

$$xyzxz - xyzxy - xyzwz + xyzwy$$

$$xyzzx - xyzzy - xyzwx + xyzwy$$

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

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

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

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

$$\begin{aligned} &xyxxy - xyxxz - xyxyy + xyxzz - xyyyx + xyyyz \\ &+xyyxx - xyyzz + xyzyy - xyzxx - xyzwy + xyzwx \end{aligned}$$

$$\begin{aligned} &xyxyy - xyxzz - xyyxx + xyyzz - xyzyy + xyzxx \\ &+k(xyxzy - xyxzw - xyyzx + xyyzw - xyzwy + xyzwx) \end{aligned}$$

In our context, $k = 2n - 2 = 72$. There are other invariants, but they are not *phylogenetic*, i.e. they are zero for all trees.

5.1 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. The situation would be much worse for $N = 6$ with 203 configurations, one of the reasons for not proceeding beyond $N = 5$ here.

5.2 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 [1, 13] 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 much longer than others), nonetheless the 11 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.

6 Results.

6.1 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 [15] and [1]. 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 1, and the third was the artifactual tree in that figure. Thus our method succeeded in correctly grouping CHO and ECH, despite the discordance of branch lengths which defeat distance-matrix-based attempts [1]. And it also confirmed the ANN+MOL grouping in CAL versus the TOL grouping of ANN+ART.

6.2 What is the sensitivity of our method with small genomes?

A more clear-cut result of our method would see the tree **T** emerge with no invariant scoring less than Ψ and all other trees scoring less than Ψ (i.e. "significant") on at least one invariant, for some threshold Ψ. We simulated $N = 5$ data for a range of genome sizes, from $n = 8$ to $n = 140$, with $n/4$ random inversions disrupting gene order on each branch of the tree, with 2000 repetitions of the experiment for each n. (Recall from Section 6.1 that for $n = 37$, the minimum breakpoint tree warrants $R = 10$, approximately $n/4$.)

The form of the invariants ensure that they each converge to a limit, zero for **T** and non-zero for each of several other trees. If Ψ is small enough, and n is large enough, only the invariants for **T** will be below the significance level. Results of our simulations in Figure 2 indicate that for Ψ small enough to exclude all trees except **T** — a "true" threshold, we require $n = 140$ at least. For smaller n, the tree **T** is also likely to be "rejected" by at least one invariant. For $n = 37$, **T** is the sole tree to pass the tests of all 11 invariants with $\Psi = 0.01$ only 15 % of the time. If Ψ is relaxed to a value that will maximize acceptance of **T** only, say $\Psi = 0.1$, only 40 % can be attained for $n = 37$. This explains our recourse to more equivocal, statistical criteria in our selection of the CAL tree in Section 6.1. Another set of simulations tested the performance of our method on $N = 5, n = 37$ data for a range of branch lengths, the same on each branch. We used 10,000 simulations per branch length. As can be seen in Figure 3, the rate of success drops off rapidly with branch length and decreasing Ψ so that with 10 random inversions per branch, successful discrimination in favour of **T** is 40 % for a threshold value of $\Psi = 0.1$, and for 20 inversions it is only 25%; with 10 inversions and $\Psi = 0.01$, the success rate is only 17%.

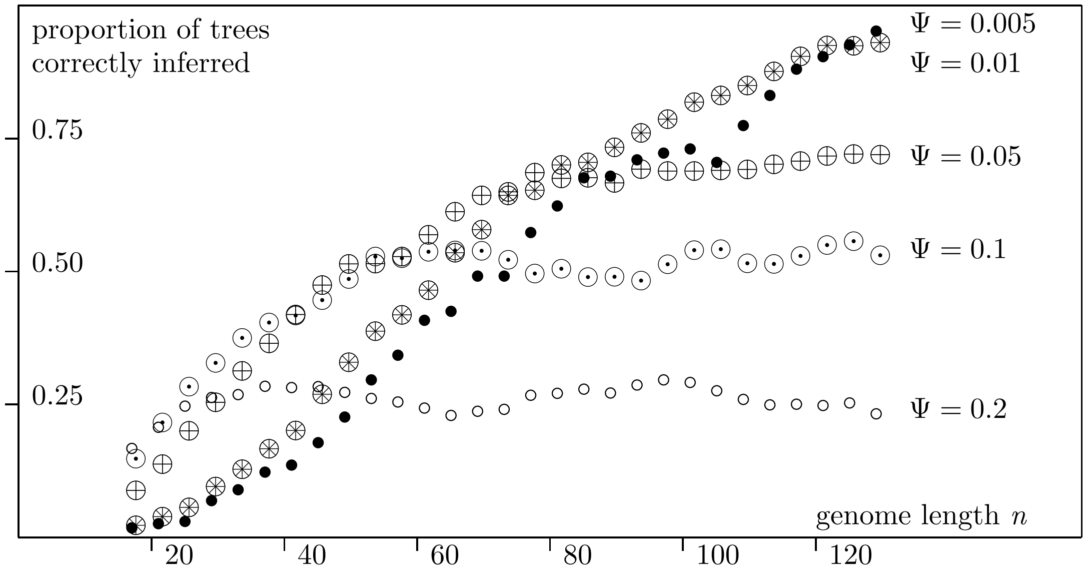

Figure 2: Proportion of trees correctly inferred as a function of genome length and Ψ. Curves smoothed using a window size of three data points.

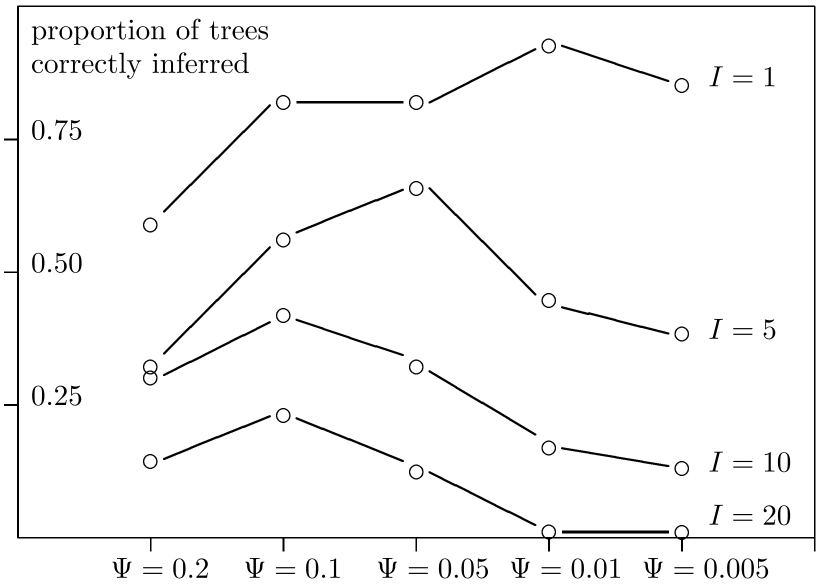

Figure 3: Proportion of trees correctly inferred as a function of Ψ and number of rearrangements I.

6.3 The Balanoglossus data.

The recently sequenced mitochondrial genome *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 $10^{-7}, 10^{-7}, 6\times10^{-8}$, 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 1, while the CAL tree is third best.

Conclusions and further work.

Perhaps the most promising direction for this work lies towards larger genome size — plastids, prokaryotes and, when more eukaryotes are completely sequenced, nuclear genomes. Increasing n only linearly increases the time to compute configuration frequencies, which is extremely rapid in any case, and *does not at all change the time required to compute the invariants*. Our simulations indicate that the method should be able to identify the true tree with a high degree of accuracy for large genomes.

Indeed, there are a growing number of genomes with hundreds and thousands of genes for which the DNA sequence is known, and this presents an ideal opportunity for our method. There are, however, several problems with larger genomes. The first is that one-to-one homology identifications, which are "given" in the case of mitochondria, tend to be very uncertain for a large proportion of genes, especially those in gene families involving two, three or many similar genes in each genome. Second, one constraint on our method is that it applies only to genomes having the same genes. Short of further mathematical improvements, this means we would have to apply our method to the intersection of the gene sets of all the genomes, entailing the loss of much of the data. Third, there is the possibility that undetected horizontal transfer of genes or operons might introduce a variety of errors into the analysis.

Multichromosomal genomes are handled as easily as single-chromosome ones in our analysis, since the model pertains to single breakpoints and not to whole fragments, which behave differently in inversions, transpositions and reciprocal translocations. 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 [1], 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 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 6.1 and 6.3, might be systematized.

It is important to note that the Jukes-Cantor matrix does not derive directly from any probabilistic model for inversions of signed genomes or for transpositions, although such a matrix can be derived for the case of inversions on unsigned genomes [14]. The problem of finding some invariant analysis, perhaps not based on breakpoints considered in isolation, of the pure inversion problem with signed genomes is mathematically of great interest. Of some interest as well is whether the theory developed here can be extended in the direction of other semigroups. Linear invariant theory is well-developed, for the Kimura models [16] and others, and biological interpretation in the breakpoint context is possible.

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 neither of these is unequivocal. Our method clearly distinguishes between deuterostomes and protostomes, which is not always the case with other approaches using rearrangement data.

Acknowledgments

Research supported by grants to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Genome Analysis and Technology program, and a NSERC fellowship for graduate studies to MB. DS is a Fellow of the Canadian Institute for Advanced Research.

References

[1] Blanchette, M., Kunisawa, T., Sankoff, D., Gene order breakpoint evidence in animal mitochondrial phylogeny, manuscript, Centre de recherches mathématiques, 1998.

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

[3] Caprara, A., Formulations and complexity of multiple sorting by reversals, manuscript, University of Bologna, 1998.

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

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

[6] 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*, 9:481–488, 1998.

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

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

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

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

[11] Maddsion, D. and Maddison, W., Metazoa *In: Tree of Life* website, http://phylogeny.arizona.edu/tree/eukaryotes/animals/animals.html, 1995.

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

[13] Sankoff, D., and Blanchette, M., Multiple genome rearrangement and breakpoint phylogeny, *Journal of Computational Biology*, 5:555–570, 1998.

[14] Sankoff, D., and Blanchette, M., Comparative genomics via phylogenetic invariants for Jukes-Cantor semigroups, to appear *In: Proceedings of the International Conference on Stochastic Models*, L. Gorostiza and G. Ivanoff, eds., Conference Proceedings Series, Canadian Mathematical Society, in press, 1998.

[15] 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 (USA)*, 89:6575–6579, 1992.

[16] 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*, 31:289–296, 1993.

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

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