Bayesian Phylogenetic Inference from Animal Mitochondrial Genome Arrangements

Bret Larget[1], Donald L. Simon[1], and Joseph B. Kadane[2]

31 January, 2002

Summary

The determination of evolutionary relationships is a fundamental problem in evolutionary biology. Genome arrangement data is potentially more informative than DNA sequence data for inferring evolutionary relationships among distantly related taxa. We describe a Bayesian framework for phylogenetic inference from mitochondrial genome arrangement data using Markov chain Monte Carlo methods. We apply the method to assess evolutionary relationships among eight animal phyla.

Key words: Breakpoint graph, genome rearrangement, Markov chain Monte Carlo, phylogeny, signed permutations, sorting by reversal.

Corresponding Author:
Bret Larget
Dept. of Mathematics and Computer Science
Duquesne University
College Hall 440
Pittsburgh, PA, USA, 15282
`larget@mathcs.duq.edu`
(412) 396-6469

1. Department of Mathematics and Computer Science, Duquesne University

2. Department of Statistics, Carnegie Mellon University

1 Introduction

The idea that all living organisms are related through common descent is one of the fundamental organizing principles of modern biology. Consequently, the determination of evolutionary relationships is one of the most important activities that evolutionary biologists carry out. Prior to the wide-spread availability of massive collections of DNA sequence data freely available via computer networks and desktop computers with specialized software to analyze this data, most phylogenies, branching tree diagrams that display evolutionary relationships, were inferred by biologists on the basis of morphological data and characteristics. It is fairly common for a phylogeny that is strongly supported through an analysis of molecular data to be inconsistent with the traditional phylogeny based on morphological data. To complicate matters, biologists have developed a large number of methods for producing phylogenies from DNA sequence data, the results of which frequently conflict. Each method has its strong supporters and there is a lively debate in the biological literature arguing the relative merits of various methods of phylogenetic inference.

Very few methods for producing phylogenies from DNA sequence data have a statistical foundation that provides a framework for the assessment of uncertainty (Felsenstein, 1983). The maximum likelihood approach to phylogenetic inference (Felsenstein, 1981) is one notable exception. Swofford *et al.* (1996) provides an excellent overview of many commonly used methods for phylogenetic analysis from aligned DNA sequence data. More recently, several authors have developed Bayesian approaches to phylogenetic inference from DNA sequence data (Rannala and Yang, 1996; Yang and Rannala, 1997; Mau *et al.*, 1999; Newton *et al.*, 1999; Larget and Simon, 1999; Li *et al.*, 2000). Huelsenbeck *et al.* (2001) is a recent review article that addresses the recent impact of Bayesian methods on evolutionary biology.

There are, however, limitations to the usefulness of DNA sequence data to infer evolutionary relationships. Boore and Brown (1998) suggest several: selection, rapid rates of evolution, and alignment ambiguities. Under selection, nucleotide substitutions at homologous sites in different lineages could have different probabilities of propagating throughout a population. If the rate of evolution is very rapid, sequences may diverge so quickly that very little phylogenetic information may remain. If a large number of small-scale deletion and insertion events occur, there can be tremendous uncertainty in any attempt to align DNA sequences by homologous sites. Boore and Brown (1998) write

> . . . a single, completely resolved, unambiguous tree of life based on sequence comparisons seems unlikely to be realized.

Boore and Brown (1998) go on to suggest that gene order comparisons have several advantages, and that mitochondrial genomes are especially useful in inferring phylogeny among distantly related animals.

What is mitochondrial DNA? Mitochondria are small organelles found outside the cell nucleus in animals, plants, fungi, and protists. While most DNA in animals is located in chromosomes in the cell nucleus, the mitochondria contain a relatively small circular ring of DNA. Mitochondrial DNA is doubly stranded and the genes may be on either strand, although for some animals, all the genes are on the same strand. Animal mitochondrial DNA has several characteristics that are highly conserved. Most animal mitochondrial genomes contain about sixteen thousand nucleotide bases and contain the same 37 genes: 22 for transfer RNAs (tRNAs), two for ribosomal RNAs (large- and small-unit rRNA [*rrnL* and *rrnS*, respectively]), and thirteen for proteins (NADH dehydrogenase subunits 1–6 and 4L [*nad1–6* and *nad4L*], cytochrome oxidase subunits I–III [*cox1–cox3*], ATP synthase subunits 6 and 8 [*atp6* and *atp8*], and cytochrome *b* [*cob*]). There are a few known exceptions. Several nematodes and flatworms are missing the gene *atp8* in the mitochondria and have only 36 genes. The brown sea anemone and other individuals in the phylum Cnidaria have very unusual mitochondrial genomes, missing most of the tRNAs while some of the other genes are not contiguous. These exceptions aside, it is interesting and potentially informative that while the gene content is highly conserved, the order in which the mitochondrial genes are arranged can vary among different animal species. Unlike nuclear DNA, genes are tightly compact, meaning that there are very small regions of non-coding DNA between genes. All animal mitochondrial genomes contain one or more larger areas of non-coding DNA that is thought to be involved in the regulation of replication and transcription.

Why are mitochondrial genome arrangements useful for phylogenetic inference? Boore and Brown (1998) list several reasons why mitochondrial genome arrangement data has many advantages over other types of genetic data for phylogenetic inference among animals. These reasons include:

(1) mitochondrial gene content among all animals is nearly invariant; (2) there are a very large number of possible arrangements, so animals with shared arrangements are very likely to have common ancestry; (3) there is near certainty that homologous genes can be identified despite the substantial differences in the DNA sequences among homologous mitochondrial genes in different animal species; (4) mitochondrial genome arrangement probably does not affect selection; and (5) genome rearrangements are rare, even over long periods of evolutionary time.

In the early 1990s, complete mitochondrial genome arrangements were known for only about a dozen different species. Boore and Brown (1998) list 70 known arrangements in 1998. Helfenbein *et al.* (2001) report 127 known sequences in 2001. The most recent version of the Mitochondrial Gene Arrangement Source Guide (Boore, 2001) contains the complete mitochondrial genome arrangements of the 231 different species for which this was known and published by October 31, 2001. The rate at which new data is being collected is increasing rapidly.

What are the mechanisms of mitochondrial genome rearrangement? Boore and Brown (1998) describe several mechanisms of genome rearrangement. One mechanism is gene inversion. In a single gene inversion event, a sequence of consecutive genes is inverted which changes both the order of the genes and the strands on which the coding portions are located. Gene inversion is, perhaps, the primary mechanism by which the large non-tRNA coding genes rearrange with each other. A second mechanism for which there is evidence is a duplication/deletion sequence of events where several consecutive genes are duplicated followed by loss of function and subsequent deletion of a randomly chosen copy from each pair which may or may not lead to a different arrangement. This type of rearrangement may occur predominantly with tRNAs — genome arrangement differences between marsupials and other mammals can be explained by one such event. While there may be other mechanisms that move tRNAs to distant positions, these rearrangement mechanisms are not well understood.

A mathematical representation of a genome arrangement: We can mathematically represent a mitochondrial genome arrangement of $n+1$ genes as a signed permutation of size n. To do so, we select an arbitrary relabeling of the genes with the integers from 0 to n. Beginning at the reference gene in the direction of its transcription, the gene arrangement corresponds to a permutation of the integers from 1 to n. The signs of elements in the permutation are positive or negative depending on whether the genes are located on the same or different strand as the reference gene, respectively.

Because we do not understand or know how to model effectively all of the possible mechanisms that rearrange tRNAs, for the present study we consider only the mitochondrial genome arrangements of the non-tRNA genes and we assume that gene inversion is the sole mechanism by which these genes rearrange. A gene inversion manifests as a reversal in the signed permutation. Reversals change both the order and sign of the affected elements. For example, reversing the third through the eighth elements of the signed permutation $(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)$ results in the signed permutation $(1, 2, -8, -7, -6, -5, -4, -3, 9, 10, 11, 12, 13, 14)$.

There are complete mitochondrial genome arrangements known from individuals from nine separate phyla (Boore, 2001). Within each phylum, we selected an individual for each unique arrangement of non-tRNA coding genes that included the full complement of fifteen genes. The remaining data set, shown in Table 1, contains nineteen species from eight phyla. Eighteen of the genome arrangements are unique. One arrangement is common for birds and acorn worms. Table 2 displays the inversion distance between each pair of species. The phyla in this data set are Chordata (vertebrates), Hemichoradata (acorn worms), Echinodermata (sea stars, brittle stars, sand dollars, sea urchins, crinoids, and sea cucumbers), Brachiopoda (lamp shells), Mollusca (clams, snails, squids, and chitons), Annelida (segmented worms), Arthropoda (arachnids, crustaceans, and insects), and Nematoda (roundworms).

2 A model of genome rearrangement

We assume a very simple model of mitochondrial genome rearrangement, with gene inversion as the sole mechanism. We assume that the evolutionary relationships among the taxa in our analysis are described by a phylogeny in which each speciation event results in two lineages. We do not assume a molecular clock, so the overall rate of gene inversion may be different for different lineages. Our prior distribution is that all unrooted tree topologies are equally likely. Branches of the unrooted tree have independent lengths selected from a Gamma distribution. Given a branch length, a Poisson number of gene inversions with this mean are realized. Given that a gene inversion occurs, we assume that all possible gene inversions are equally likely. There are $1 \times 3 \times \cdots \times (2s-5)$ possible binary unrooted tree topologies (Felsenstein, 1978) to relate s taxa. In our present study for the nineteen taxa we consider, this count is over 6.3×10^{18}. Our inferences are based on the combination of ten independent samples of the posterior. Further details of this model important to understand the details of our computation are contained in the Appendix.

3 Example

The correct evolutionary relationships among several animal phyla are still unresolved. Several previous papers have used mitochondrial genome arrangement data to draw conclusions about evolutionary relationships among animal phyla that differ from previous conclusions based on shared morphological characteristics. Figure 1 is adapted from De Rosa (2001) and shows two competing versions of the evolutionary relationships among animal phyla. The left tree is a traditional viewpoint supported by shared morphological characteristics. The right tree has been hypothesized more recently and is supported by molecular evidence.

We focus our attention on two conflicting aspects of these trees. The traditional phylogeny places brachiopods closer to deuterostomes (echinoderms, hemichordates, and vertebrates) than to protostomes (arthropods, annelids, and molluscs) and places molluscs as an outgroup to arthropods and annelids (Hyman, 1940). In contrast, the new tree has brachiopods closer to the annelids and molluscs than the deuterostomes and places arthropods as more distantly related than annelids and molluscs (Halanych *et al.*, 1995; Aguinaldo *et al.*, 1997). We use the model described above to assess these aspects in conflict between these two trees. We do so by examining posterior probabilities of predicted clades. A set of taxa form a clade in a tree if they comprise a complete subtree.

Are brachiopods more closely related to deuterostomes or protostomes? De Rosa (2001) finds that a close relationship between brachiopods and protostomes is “most probable”, but “not definitely conclusive”. In our analysis, the deuterostomes appear together as a clade 39% of the time and the echinoderms appear together 98% of the time. We tend to put humans closer to echinoderms (68%) than to the domestic chicken and acorn worm (7%), but there is enough evidence in the data to substantially change a tiny prior probability that deuterostomes are a clade. However, we find no clades with posterior probability of at least 5% that include any single brachiopod species with some subset of the deuterostomes.

In contrast, we find many clades that include the brachiopods with the annelid and several mollusks. All of these taxa except for the squid appear together 18% of the time. (We have difficulty placing the squid. We place it with the fruit fly 13% of the time, with the hermit crab 14% of the time, and with the acorn worm and chicken 10% of the time.) There are many clades with all three brachiopods placed with the annelid and one or more molluscs. The brachiopods, the annelid, and the black chiton appear together 28% of the time.

None of these posterior probabilities are large, but this is because the completely uninformative prior we place on the tree topology includes very small prior probabilities that species from within the same phylum would form a clade. This being said, we find it to be very probable that brachiopods are more closely related to protostomes than to deuterostomes, adding evidence in favor of the new phylogeny.

Are annelids and molluscs more closely related than arthropods? Boore and Brown (2000) use mitochondrial genome arrangement data along with other evidence to conclude that molluscs and annelids are sister taxa with arthropods as an outgroup. In the part of their analysis that is based solely on gene arrangement data, they analyzed a single mollusc (*K. tunicata*), two annelids (the common earthworm and another for which the non-tRNA arrangement is identical), as well as single inferred ancestral sequences for chordates and arthropods. Using the minimum-breakpoint (Sankoff and Blanchette, 1998; M. Blanchette, 1999), they find a best tree with 76 breakpoints consistent with annelids and arthropods being most closely related. The next best tree has 80 breakpoints.

We do find several clades with relatively high posterior probability that include our single annelid with one or more molluscs, usually with brachiopods present as well. The clade of the brachiopods, the annelid, and the molluscs except for the squid appear together 18% of the time and a clade with two of the brachiopods (*Laqueus rubellus* and *Terebratalia transversa*) along with the annelid and the molluscs land snail and sea slug appear together 23% of the time. Common clades that appear at least 5% of the time that include arthropods along with the annelid invariably have brachiopods and one or more mollusc present as well. We find that the right tree in Figure 1 with a clade of brachiopods, annelids, and molluscs separate from arthropods to be more likely than one with annelids and arthropods as sister taxa, but that this conclusion is not as firm as the previous conclusion about the placement of brachiopods.

4 Discussion

Comparisons to other methods: Statistical methods for phylogenetic inference from genome arrangement data are in their infancy. The principle of parsimony says that the best tree is the one that requires the smallest number of genome rearrangement events. Most papers that include phylogenetic inferences from genome arrangement data use this principal in an informal manner, drawing conclusions on the basis of shared arrangements that are evaluated by eye.

Other methods are more formal. By using the fast algorithms for computing pair-wise reversal distances, it is possible to feed these genome-arrangement-based distance matrices into other methods that produce phylogenies from distance matrices to infer trees. Pevzner (2000) and the references within describe this approach. Sankoff and Blanchette (1998) describe a method that estimates phylogeny by searching for arrangements at internal nodes that minimize the changes in breakpoint distance, while Sankoff and Blanchette (1999) describe a method based on invariants of frequencies of site patterns. The latter two methods are not based on any mechanism of gene rearrangement. Mechanisms such as gene inversion, gene duplication/deletion, and gene transposition affect the breakpoint distance in different ways.

None of the alternative methods discussed here provide a framework for assessing uncertainty. The best tree is simply accepted as being the best. Clustering methods are prone to poor inferences because the sequence data is discarded — when two groups are joined, distance to other groups are not based on the likely gene arrangement at the ancestor of the new group. The method described in Sankoff and Blanchette (1999) uses all 37 genes, but is limited to five or fewer taxa, which greatly limits its usefulness. Table 2 shows how distant individual species from the same phylum can be apart from each other. Presumably the decision on which taxon to use to represent a particular phylum could greatly affect the inference.

To the best of our knowledge, the present work is the first to make phylogenetic inferences on the basis of gene arrangements that also provides assessment of uncertainty. Our earlier work on this problem (Simon and Larget, 2001) was limited to small artificial data sets. The computational approach described in this paper is not limited by the number of genes or taxa.

The Bayesian approach is very useful in this application, especially since the most likely tree is not very likely at all. A sample of trees drawn from the posterior distribution permits examination of which parts of the tree are well-established, and which parts are more uncertain. It also permits calculation of probabilities of biological hypotheses, such as those above.

Directions for further work: From a modeling perspective, the first extension of this work we would make is to include duplication/deletion and transposition as well as inversion. These additional mechanisms of rearrangement would require additional parameters for the relative speeds at which each occurs, leading to an interesting extension of this work. It would also be useful to use the tRNAs as well. A second modeling advance, to allow unequal probabilities for gene inversions of different lengths, must await further understanding of how gene inversion occurs at a molecular level to guide the development of a more realistic model.

This work may also be advanced by incorporating additional information. We could do this by jointly modeling gene arrangement processes with changes at the sequence level. We could also elicit more informative priors from experts in evolutionary biology.

Finally, we believe that advances in visualizing and summarizing samples of trees would help in this work. We should be able to infer ancestral genome arrangements, for example. This area is just beginning; there are many contributions statisticians can make.

5 Acknowledgements

The authors thank the referees and the editors whose comments led to substantial improvements in the paper, and to Jeff Boore for providing several references. B. L. was partially supported in this work by NSF grant DEB-0075406; J. B. K. by NSF grant DMS-9801401. B. L. thanks the Department of Statistics at Carnegie Mellon University where he was a visitor during the final stages of this work.

6 Appendix: Calculation Details

This appendix contains the mathematical description of the model we use, a derivation of the posterior distribution, descriptions of the Markov chain Monte Carlo proposals, and discussion of the mixing properties of the method.

A mathematical description of the model. The mathematical representation of an unrooted phylogeny for ℓ taxa includes an unrooted tree topology τ and a vector of branch lengths $\beta = \{\beta_i\}$, for $i = 1, \ldots, 2\ell - 3$. The unrooted tree topology is a connected acyclic graph with ℓ labeled leaf nodes (each of which is adjacent to one other node in the tree), $\ell - 2$ unlabeled internal nodes (each of which is adjacent to three other nodes), and a total of $2\ell - 3$ edges (branches). This type of tree results when the root is removed from a rooted binary tree. We let T_ℓ represent the set of all such possible unrooted tree topologies with ℓ leaves. Our prior is that the tree topology τ is chosen uniformly at random from T_ℓ and that the branch lengths β are independent and identically distributed from a Gamma(α, λ) distribution.

Each branch of the tree contains a list of reversals and their positions. The counts of reversals on the branches, $x = \{x_i\}$, $i = 1, \ldots, 2\ell - 3$, are independent Poisson random variables with means equal to the respective branch lengths. The jth reversal on the ith branch, r_{ij}, is located a distance u_{ij} from the beginning node, chosen uniformly at random along the branch, and results in the reversal of the interval from elements a_{ij} to b_{ij} in the signed permutation, where $1 \leq a_{ij} \leq b_{ij} \leq n$. The set M_n of possible reversals that act on permutations of size n has $\binom{n+1}{2} = n(n+1)/2$ elements. A reversal $(a, b) \in M_n$ acts as below.

$$(\pi_0, \ldots, \pi_{a-1}, \pi_a, \ldots, \pi_b, \pi_{b+1}, \ldots, \pi_{n-1}) \xrightarrow{(a,b)} (\pi_0, \ldots, \pi_{a-1}, -\pi_b, \ldots, -\pi_a, \pi_{b+1}, \ldots, \pi_{n-1}) \qquad (1)$$

Given (τ, x, r) and the permutation at one leaf of the tree, the remaining observable leaf permutations are determined. In fact, the permutations are determined at every point of the tree. Let D represent the observable data, an array of permutations indexed by the leaf nodes.

The prior distribution of these parameters is summarized here.

$$\begin{aligned}
\tau &\sim \text{Uniform}(T_\ell) && (2)\\
\beta_i &\sim \text{i.i.d. Gamma}(\alpha,\lambda) \quad \text{for } i=1,\dots,2\ell-3 && (3)\\
x_i|\beta_i &\sim \text{i.i.d. Poisson}(\beta_i) \quad \text{for } i=1,\dots,2\ell-3 && (4)\\
r_{ij}|x_i &\sim \text{i.i.d. Uniform}(M_n) \quad \text{for } i=1,\dots,2\ell-3,\ j=1,\dots,x_i && (5)\\
u_{ij}|\beta_i,x_i &\sim \text{i.i.d. Uniform}(0,\beta_i) \quad \text{for } i=1,\dots,2\ell-3,\ j=1,\dots,x_i && (6)
\end{aligned}$$

The joint prior on these parameters is

$$p(\tau,\beta,x,r,u)=\frac{1}{|T_\ell|}\prod_{i=1}^{2\ell-3} g(\beta_i)h(x_i|\beta_i)\left(\frac{1}{\beta_i}\right)^{x_i}\left(\frac{1}{|M_n|}\right)^{x_i} \qquad (7)$$

where

$$g(\beta_i)=\frac{\lambda^\alpha}{\Gamma(\alpha)}\beta_i^{\alpha-1}\mathrm{e}^{-\lambda\beta_i} \qquad (8)$$

and

$$h(x_i|\beta_i)=\frac{\mathrm{e}^{-\beta_i}\beta_i^{x_i}}{x_i!} \qquad (9)$$

The likelihood for D is an indicator if the observed data is consistent with the parameters and unobservable variables, $p(D|\tau,\beta,x,r,u)=1_{\{(\tau,x,r)\hookrightarrow D\}}$.

Derivation of the posterior distribution: We are primarily interested in evaluating the posterior distribution of the tree topology, $p(\tau|D)$. We begin by expressing the unnormalized joint posterior distribution of all of the parameters.

$$\begin{aligned}
p(\tau,\beta,x,r,u|D) &\propto p(\tau,\beta,x,r,u)p(D|\tau,\beta,x,r,u) && (10)\\
&= \frac{1}{|T_\ell|}\prod_{i=1}^{2\ell-3} g(\beta_i)h(x_i|\beta_i)\left(\frac{1}{\beta_i}\right)^{x_i}\left(\frac{1}{|M_n|}\right)^{x_i} 1_{\{(\tau,x,r)\hookrightarrow D\}} && (11)
\end{aligned}$$

To simplify this, we integrate out β and u analytically, suppressing most of the derivation. The remaining parameters are the tree topology and the ordered list of reversals on each branch.

$$\begin{aligned}
p(\tau,x,r|D) &\propto \frac{1}{|T_\ell|}\prod_{i=1}^{2\ell-3} 1_{\{(\tau,x,r)\hookrightarrow D\}}\int_{\beta_i=0}^{\infty} g(\beta_i)h(x_i|\beta_i)\left(\frac{1}{\beta_i|M_n|}\right)^{x_i}\prod_{j=1}^{x_i}\int_{u_{ij}=0}^{\beta_i} du_{ij}\,d\beta_i && (12)\\
&= \frac{1}{|T_\ell|}1_{\{(\tau,x,r)\hookrightarrow D\}}\left(\frac{1}{|M_n|(1+\lambda)}\right)^{\sum_{i=1}^{2\ell-3}x_i}\left(\frac{\lambda}{\lambda+1}\right)^{\alpha(2\ell-3)}\prod_{i=1}^{2\ell-3}\frac{\Gamma(\alpha+x_i)}{x_i!\Gamma(\alpha)} && (13)
\end{aligned}$$

Finally, we ignore some factors that do not depend on τ, x, or r.

$$p(\tau, x, r|D) \propto 1_{\{(\tau,x,r)\hookrightarrow D\}} \left(\frac{1}{|M_n|(1+\lambda)}\right)^{\sum_{i=1}^{2\ell-3} x_i} \prod_{i=1}^{2\ell-3} \frac{\Gamma(\alpha + x_i)}{x_i!\Gamma(\alpha)} \tag{14}$$

We sample from this unnormalized posterior $p(\tau, x, r|D)$ using Markov chain Monte Carlo (Metropolis *et al.*, 1953; Hastings, 1970) to calculate $p(\tau|D)$.

MCMC updates: We cycle through three updates, the first two of which leave the tree topology unchanged but modify the reversal histories, while the third changes the tree topology and modifies the reversal histories to remain consistent. The Updates 1 and 3 are displayed in Figure 2 which also defines the nomenclature used in the following description.

Update 1 begins by randomly picking an internal node of the tree and then randomly assigning labels to the three edges. If there are r reversals on the path from node A to node B, there are $r + 1$ ways to partition these reversals on edges 1 and 2 without changing their relative order. One of these partitions is chosen at random, which may change the induced signed permutation at node O. Then, any reversals on edge 3 are deleted and a new sequence of reversals is generated from node O to node C in the manner described below.

Update 2 begins by randomly picking any edge from the tree. The reversals on that edge are deleted and a new sequence of random reversals is generated for the edge in a randomly chosen direction as described below.

Update 3 begins by choosing an internal branch (*edge 3*) uniformly at random. Each adjacent node then picks at random one of its other edges (*edge 1* and *edge 4*). These edges and any subtree extending beyond nodes A and C are then swapped, resulting in a new tree topology. The reversals on *edges 2*, *3*, and *5* remain the same. Reversals on the two swapped edges are deleted. New reversal sequences are generated for *edge 4* from the signed permutation at node E to node C and for *edge 1* from the signed permutation at node F to node A.

The breakpoint graph: Our mechanism for proposing a sequence of reversals that change the source signed permutation s to the target t uses the breakpoint graph. We note that a set of reversals that acts on a signed permutation s to produce t will also act on st^{-1} to produce the identity permutation, so, without loss of generality, we can consider the problem of finding a sequence of reversals to sort a signed permutation. See Pevzner (2000) and Kaplan *et al.* (1999) and references therein for a more detailed description of the breakpoint graph and an algorithm to find a single minimal sequence of reversals to sort a signed permutation. We wish to be able to propose any sequence that sorts the signed permutation with minimal sequences being relatively likely.

Figure 3 shows an example of a breakpoint graph. We find it useful to represent a breakpoint graph as a circle. Pevzner (2000) draws the breakpoint graph with the nodes along a line. The definitions below are equivalent to those in Pevzner (2000), but are rephrased with the intention to add clarity.

The outer circle of numbers is a signed permutation where an element 0 has been added to connect the beginning to the end of the permutation. This mirrors a mitochondrial genome arrangement where 0 represents the reference gene. Figure 3 represents the arrangement of the crinoid relative to human. The inner circle of numbers is an unsigned circular permutation determined by the outer signed permutation. The element 0 is represented by $2n+1, 0$ where there are $n+1$ elements in the outer circle. For the rest, the label i corresponds to $2i-1, 2i$ if it has a positive sign and corresponds to $2|i|, 2|i|-1$ if it has a negative sign. Each element of the inner circle is a node in the breakpoint graph. In the example, the inner circle is an unsigned permutation of size 30 (of the integers from 0 to 29), twice as large as the number of genes we consider.

Breakpoints are represented by *black edges* along the inner circle that connect adjacent nodes that are out of sequence, and so differ in absolute value by more than one. The example has seven breakpoints. The *gray edges* in the interior of the breakpoint graph connect nodes with even values to nodes with values one larger when these nodes are not adjacent. The lines are *oriented* (solid) when they are separated along the inner circle by an even number of positions and are *unoriented* (dashed) otherwise. There are always equal numbers of black and gray edges. All nodes are either isolated or part of a cycle of alternating black and gray edges.

Two cycles are *connected* if a gray edge from one crosses a gray edge from the other. The cycles of the breakpoint graph are partitioned into *connected components.* A connected component is *unoriented* if all edges of all of its cycles are unoriented and is oriented otherwise. Each cycle must have an even number of oriented edges.

A *hurdle* is an unoriented connected component that does not separate two other unoriented connected components along the inner circle. A breakpoint graph contains a *fortress* if it has a special configuration of hurdles and other unoriented components that only arises in larger permutations than those in the present study. In the example, there are two connected components, one of which is unoriented. Each component is comprised of a single cycle. The cycle on the right is

$$(0 : 11 \text{ -u- } 10 : 13 \text{ -u- } 12 : 1 \text{ -u- }) \tag{15}$$

and the cycle on the left is

$$(20 : 22 \text{ -u- } 23 : 28 \text{ -o- } 29 : 25 \text{ -u- } 24 : 21 \text{ -o- }) \tag{16}$$

where black edges are represented by colons, and oriented and unoriented gray edges are represented by the symbols -o- and -u- respectively. The connected component on the right is a hurdle, while that on the left is not.

The minimal number of reversals to sort a signed permutation is a function of the number of breakpoints b, the number of cycles c, the number of hurdles h, and an indicator of a fortress f.

$$\text{minimum distance } = b - c + h + f \tag{17}$$

The example could be sorted by $7 - 2 + 1 + 0 = 6$ reversals.

A reversal is called *proper* if it reduces $b - c$ by one. However, all proper reversals do not reduce the distance by one because they could introduce a hurdle (or a fortress). A reversal will be proper if its end points are two breakpoints on the same cycle and these two breakpoints divide the oriented gray edges of the cycle so that there are an odd number in each semi-cycle. In the example, there are no proper reversals that act on the right cycle because any two breakpoints divide the cycle into two semi-cycles with zero oriented edges and zero is not odd. In the left cycle, there are six ways to choose two of the four breakpoints. Of these six reversals, the four proper reversals change: -11 to 11; -12 to 12; $-14, -13$ to $13, 14$; and $-11, -12, -14, -13$ to $13, 14, 12, 11$. The first three of these proper moves actually decrease the distance by one. The last adds a cycle but also adds a hurdle because the remaining gray edges all become unoriented, so the distance remains unchanged.

Proposing a sequence of reversals: Our basic approach is to iteratively add reversals until we have a list that changes the source to the target and decide to stop. If the signed permutation st^{-1} is the identity permutation itself, we quit and end the sequence of reversals with probability $q = 0.99$. Otherwise, we propose a random reversal from all possible. When we do not have the identity, we use the breakpoint graph of the signed permutation st^{-1} to partition the set of all possible reversals into three groups: proper reversals, improper reversals between breakpoints in the same cycle, and others. If there is at least one proper reversal, we choose one uniformly at random with probability $p = 0.99$. If there are no proper reversals or we have decided not to select one and there is at least one improper reversal within a cycle, we choose one of these with probability p. If we have not yet selected a reversal, we choose one of the others at random. We then iterate, adding another reversal to the sequence at each step, until we stop. Table 3 shows the probabilities that a reversal of a given type is the next one proposed. These probabilities are used in calculating acceptance ratios for the updates.

Computation details: We completed ten separate runs of 100,000,000 cycles through Updates 1, 2, and 3. The runs used different streams of pseudo-random numbers and began at different trees. In each run, we sampled every 500th tree topology, retaining 200,000 tree topologies. A single run required about three hours of CPU time on a machine with a 933 MHz Pentium III processor. We set the parameters $\alpha = 0.5$ and $\lambda = 0.25$ so that our prior had a mean of two gene inversions per branch with sufficient variance so that branches with ten or more gene inversions were not too unlikely.

Trace plots of the loglikelihoods indicate that burn-in was rapid in all runs. We discarded the initial 25% of each run and retained a combined total of 1.5 million tree topologies. Clade frequencies for most clades that appear relatively frequently are quite similar from run to run. For example, in the ten independent runs the calculated posterior probabilities that the three echinoderms form a clade have a mean of 0.978 and a standard deviation of 0.011, leading to an estimated Monte Carlo standard error of less than 0.004. Almost all clades have estimated Monte Carlo standard errors substantially less than one percent. There are a few exceptions. We expect that proposed changes in the tree topology that include branches with longer reversal lists mix more slowly. Estimates of clade probabilities for clades that include taxa that are a long distance from others had larger Monte Carlo errors. For example, the calculated posterior probabilities that the brachiopods *Laqueus rubellus* and *Terebratalia transversa* form a clade had a mean of 0.803 and an estimated Monte Carlo standard error of 0.03.

References

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

Boore, J. (2001) Mitochondrial gene arrangement source guide, version 6.0. http://www.jgi.doe.gov/Mitochondrial_Genomics.html. DOE Joint Genome Institute.

Boore, J. L. and Brown, W. M. (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Current Opinion in Genetics and Development*, **8**, 668–674.

Boore, J. L. and Brown, W. M. (2000) Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Molecular Biology and Evolution*, **17**, 87–106.

De Rosa, R. (2001) Molecular data indicate the protostome affinity of brachiopods. *Systematic Biology*, **50**, 848–859.

Felsenstein, J. (1978) The number of evolutionary trees. *Systematic Zoology*, **27**, 27–33.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Biology*, **17**, 368–376.

Felsenstein, J. (1983) Statistical inference of phylogenies. *Journal of the Royal Statistical Society, Series A*, **146**, 246–272.

Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M. A., Liva, S. M., Hillis, D. M. and Lake, J. A. (1995) Evidence from 18S ribosomal DNA that lophophorates are protostome animals. *Science*, **267**, 1641–1643.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Helfenbein, K. G., Brown, W. M. and Boore, J. L. (2001) The complete mitochondrial genome of the articulate brachiopod. *Molecular Biology and Evolution*, **18**, 1734–1744.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Hyman, L. H. (1940) *The invertebrates.* New York: McGraw-Hill.

Kaplan, H., Shamir, R. and Tarjan, R. (1999) Faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, **29**, 880–892.

Larget, B. and Simon, D. L. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, **16**, 750–759.

Li, S., Doss, H. and Pearl, D. (2000) Phylogenetic tree reconstruction using Markov chain Monte Carlo. *Journal of the American Statisical Society*, **95**, 493–508.

M. Blanchette, T. Kunisawa, D. S. (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *Jounral of Molecular Evolution*, **49**, 193–203.

Mau, B., Newton, M. A. and Larget, B. (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, **55**, 1–12.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Newton, M., Mau, B. and Larget, B. (1999) Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. In *Statistics in Molecular Biology and Genetics* (ed. F. Seillier-Moseiwitch), vol. 33 of *IMS Lecture Notes-Monograph Series*, 143–162.

Pevzner, P. (2000) *Computational Molecular Biology — An Algorithmic Approach*, chap. 10. The MIT Press.

Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, **43**, 304–311.

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

Sankoff, D. and Blanchette, M. (1999) Phylogentic invariants for genome rearrangements. *Journal of Computational Biology*, **6**, 431–445.

Simon, D. L. and Larget, B. (2001) Phylogenetic inference from mitochondrial genome arrangement data. In *Computational Science — ICCS 2001* (eds. V. Alexandrov, J. Dongarra, B. Juliano, R. Renner and C. Tan), no. 2074 in Lecture Notes in Computer Science, 1022–1028. Springer-Verlag.

Swofford, D., Olsen, G., Waddell, P. and Hillis, D. (1996) Phylogenetic inference. In *Molecular Systematics* (eds. D. M. Hillis, C. Moritz and B. K. Mable). Sinauer Associates.

Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.

Phylum	Species	Permutation
Chordata	Human	1 → 14
Chordata	Domestic Chicken	1 → 8, 10, 9, 11 → 14
Hemichordata	Acornworm	1 → 8, 10, 9, 11 → 14
Echinodermata	Sea star	6, 1 → 5, 7 → 11, −12, −14 → −13
Echinodermata	Sea urchin	6, 1 → 5, 7 → 11, 13 → 14, 12
Echinodermata	Crinoid	6, 1 → 5, 7 → 10, −11, −12, −14 → −13
Brachiopoda	*Laqueus rubellus*	10, 3, 8, −9, 5 → 6, 4, 2, 14, 1, 12, 11, 13, 7
Brachiopoda	*Terebratalia transversa*	10, 2, 4, 3, 8, 11, −9, 12 → 13, 7, 6, 1, 5, 14
Brachiopoda	*Terebratulina retusa*	1 → 3, 11 → 13, −9, 10, 6 → 8, 4 → 5, 14
Annelida	Common earthworm	1 → 2, 4, −9, 10, 3, 8, 6 → 7, 11 → 13, 5, 14
Arthropoda	Cattle tick	1 → 5, −13 → −11, −8 → −6, −9, 10, 14
Arthropoda	Fruit fly	1 → 5, −8 → −6, −9, 10, −13 → −11, 14
Arthropoda	Hermit crab	1, 5, 14, 2 → 4, −8 → −6, −9, 10, −13 → −11
Arthropoda	Wallaby louse	4, 13, 10, −7 → −6, 14, −8, 1, −5, −12 → −11, −3 → −2, −9
Mollusca	Squid	1, −8 → −6, 2 → 5, −10, 9, −13 → −11, 14
Mollusca	Black chiton	1 → 3, −8 → −6, −10, 9, −13 → −11, 4 → 5, 14
Mollusca	Land snail	12, −9, 8, 13, 6, 10, 1, −2, −3, -11, −5 → −4, 7, 14
Mollusca	Sea slug	12, −9, 8, 13, 6, 10, 1, −2, −3, −11, −5, 7, −4, 14
Nematoda	*Trichinella spirallis*	1, 13, −14, −8 → −6, −9, 10 → 12, 3 → 4, 2, 5

Table 1: **Mitochondrial genome arrangements of non-tRNA coding genes.** Each mitochondrial genome is recorded as a permutation relative to the gene order in humans beginning after the gene *cox1* in the direction of its transcription. The coding to signed permutations uses the following translation: *cox2*=1, *atp8*=2, *atp6*=3, *cox3*=4, *nad3*=5, *nad4L*=6, *nad4*=7, *nad5*=8, *nad6*=−9, *cob*=10, *rrnS*=11, *rrnL*=12, *nad1*=13, and *nad2*=14. Consecutive genes in the same order as in humans are listed as a range. For example, 1 → 3 means 1, 2, 3 and −8 → −6 means $-8, -7, -6$. The tick species here is *rhiphicephalus sanguineus.* All other known tick species non-tRNA coding gene arrangements are identical to that in fruit flies. The land snail is *Cepaea nemoralis.* The other known land snail non-tRNA coding gene arrangement is identical to the sea slug.

Human	0																		
Chicken	3	0																	
Acornworm	3	0	0																
Sea star	5	8	8	0															
Sea urchin	6	9	9	1	0														
Crinoid	6	8	8	1	2	0													
L. rubellus	13	13	13	14	13	13	0												
T. transversa	12	12	12	12	12	11	9	0											
T. retusa	5	5	5	9	9	10	13	11	0										
Earthworm	7	8	8	11	11	12	13	9	5	0									
Cattle tick	4	4	4	8	8	9	13	11	3	7	0								
Fruit fly	3	4	4	7	7	8	13	11	3	7	3	0							
Hermit crab	7	7	7	11	11	10	14	11	5	7	5	5	0						
Wallaby louse	8	10	10	12	12	13	13	14	9	11	10	10	10	0					
Squid	5	4	4	9	9	10	14	11	4	6	4	3	5	10	0				
Black chiton	5	5	5	9	9	10	13	10	1	5	3	4	6	10	5	0			
Land snail	12	12	12	12	13	12	13	12	12	10	11	12	13	14	12	11	0		
Sea slug	12	13	13	14	14	14	13	12	12	12	11	12	13	13	12	11	3	0	
T. spirallis	7	8	8	11	11	12	11	14	8	10	5	6	6	10	8	8	13	13	0

Table 2: **Inversion distances.** For each pair of taxa, the displayed count is the smallest number of gene inversions necessary to change the mitochondrial genome arrangement from one taxon to the other.

Case			Proper	Improper within a cycle	Others
$a > 0$	$b > 0$	$c > 0$	p/a	$(1-p)p/b$	$(1-p)^2/c$
		$c = 0$	p/a	$(1-p)/b$	—
	$b = 0$	$c > 0$	p/a	—	$(1-p)/c$
		$c = 0$	$1/a$	—	—
$a = 0$	$b > 0$	$c > 0$	—	p/b	$(1-p)/c$
		$c = 0$	—	$1/b$	—
	$b = 0$	$c > 0$	—	—	$1/c$

Table 3: **Reversal proposal probabilities.** In the table, a is the number of proper reversals, b is the number of improper reversals between breakpoints of the same cycle, and c is the number of others. The sum of these three is $|M_n| = 105$ in the present study. The parameter p is set to be 0.99. The expression in each cell is the probability that a specific reversal of the given type is proposed.

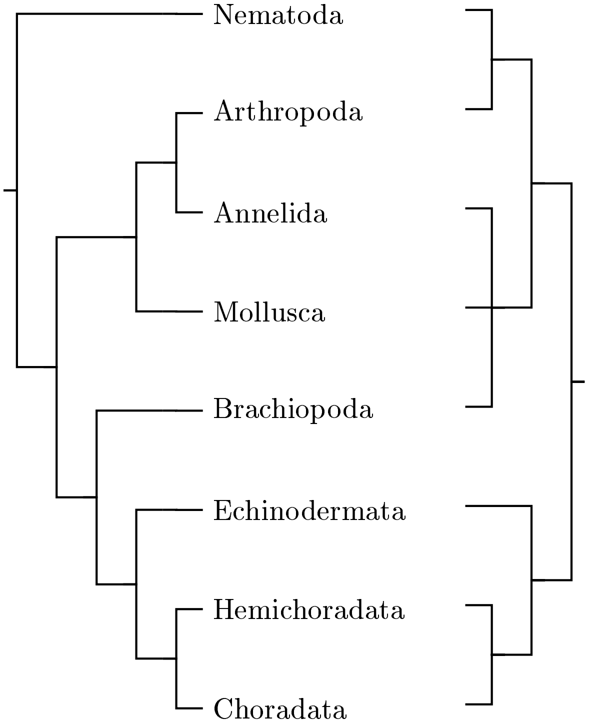
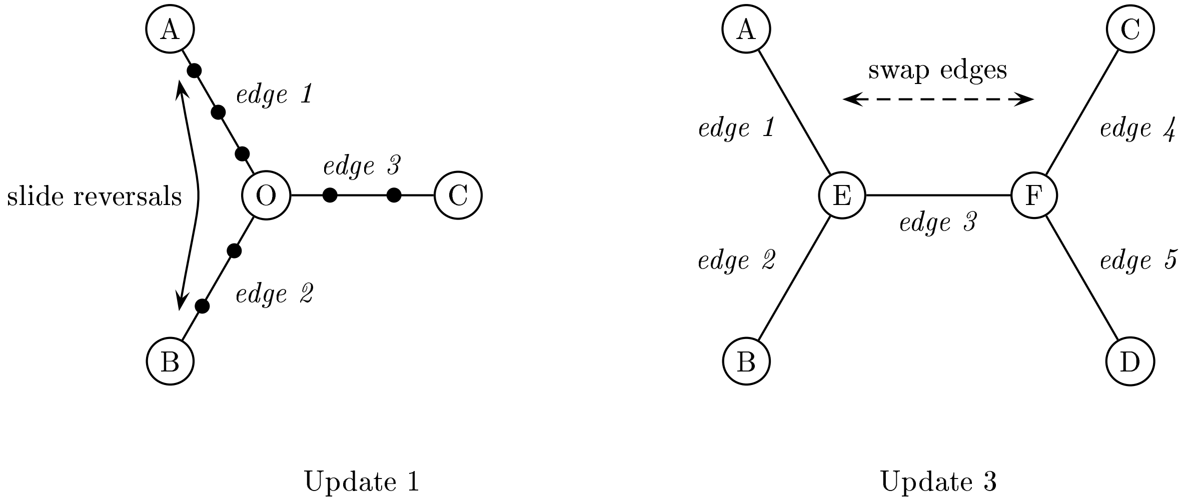


Figure 1: **Two competing simplified animal phylogenies.** The phylogeny on the left is a traditional phylogeny based on morphological characteristics. The phylogeny on the right has been proposed more recently on the basis of molecular data. The branching point that divides into three lineages, Annelida, Mollusca, and Brachiopoda, indicates that the three possible binary trees relating these three taxa are unresolved.



Update 1

Update 3

Figure 2: **Updates.** The graph on the left is related to Update 1. It is a subtree and node O has been randomly chosen. The other nodes may be either leaf nodes or internal nodes. Black dots on edges represent reversals. The graph on the right side is used in the explanation of Update 3. Update 2 is not pictured because it is trivial.

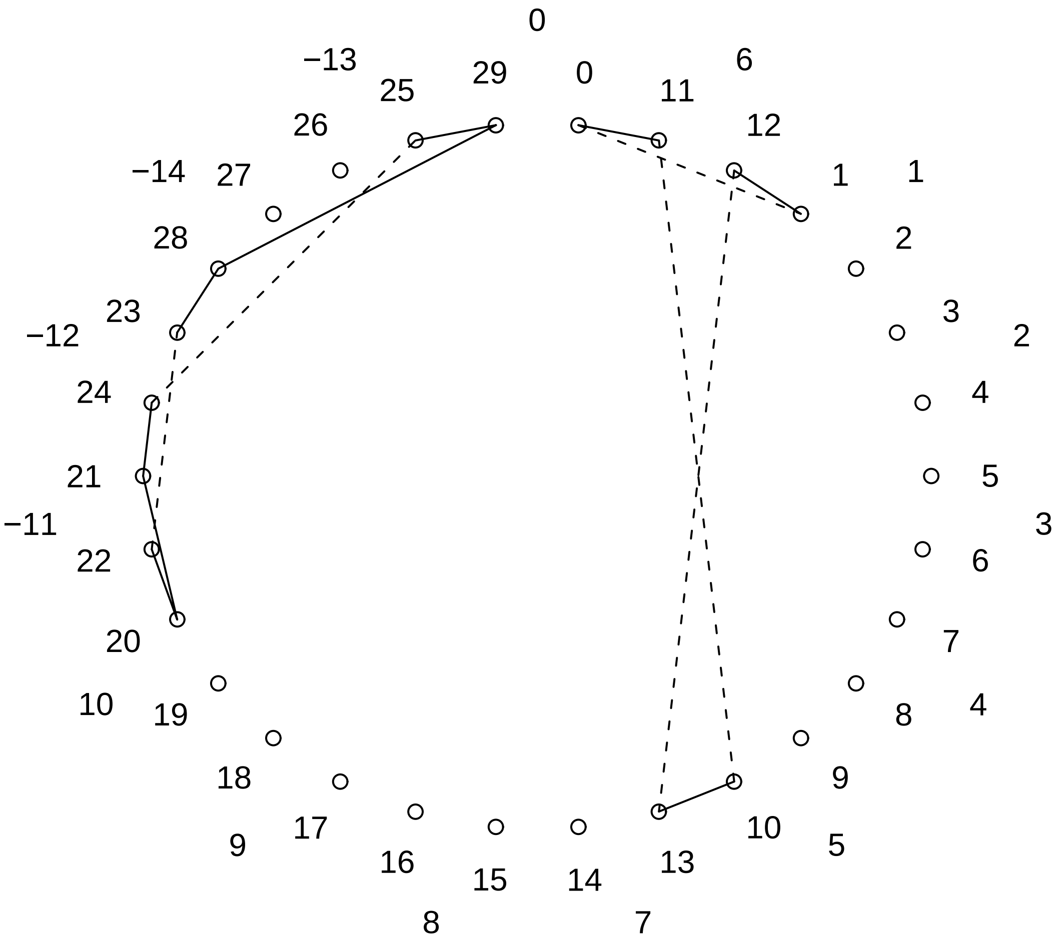


Figure 3: **Example breakpoint graph.** The breakpoint graph here is used to help explain the method for proposing reversals.