Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 36 (2005) 214–223
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

A Bayesian approach to the estimation of ancestral genome arrangements

Bret Larget [a,*], Joseph B. Kadane [b], Donald L. Simon [c]

[a] Departments of Botany and Statistics, University of Wisconsin, Madison, USA
[b] Department of Statistics, Carnegie Mellon University, USA
[c] Department of Mathematics and Computer Science, Duquesne University, USA

Received 27 March 2004; revised 1 March 2005
Available online 11 May 2005

Abstract

We describe a Bayesian approach to estimate phylogeny and ancestral genome arrangements on the basis of genome arrangement data using a model in which gene inversion is the sole mechanism of change. While we have described a similar method to estimate phylogenetic relationships in the statistics literature, the novel contribution of the present work is the description of a method to compute probability distributions of ancestral genome arrangements. We assess the robustness of posterior distributions to different specifications of prior distributions and provide an empirical means to selecting a prior distribution. We note that parsimony approaches to ancestral reconstruction in the literature focus on the development of computationally efficient algorithms for searching for optimal ancestral genome arrangements, but, unlike Bayesian approaches, do not include assessment of uncertainty in these estimates. We compare and contrast a Bayesian approach with a parsimony approach to infer phylogenies and ancestral arrangements from genome arrangement data by reanalyzing a number of previously published data sets.

Keywords: Gene inversion; Gene order; Genome arrangement; Genome rearrangement; Markov chain Monte Carlo; Phylogeny; Sorting by reversal

1. Introduction

Phylogenetic inference on the basis of molecular sequence data is an active area of research with a long history. Felsenstein (2004) is a recent book which includes a description of the history of the field, describes most common methods of phylogenetic inference, and contains an extensive list of references for further exploration on many related topics. Huelsenbeck et al. (2001) contains a better description of the impact of Bayesian methods in the field.

Methods for estimating phylogeny and ancestral genome arrangements from genome arrangement data are much newer than methods for sequence data and the relevant literature is much smaller. These methods are of increasing importance because of the rapidly growing amount of complete genomic data available for analysis. Furthermore, since processes that rearrange entire genomes are thought to be much rarer than processes that affect genetic data at the sequence level, genome arrangements may be more informative about deep evolutionary relationships than sequence data.

There are two fundamentally different approaches to analyzing genome arrangement data. One point of view, which we call the *optimal phylogeny viewpoint*, frames phylogeny reconstruction in the following way. Each tree may be scored for a given data set by counting some measure of genomic rearrangement with trees requiring the least amount of rearrangement seen as optimal. The logic is that optimal trees are most likely (in some sense) to be correct. With the optimal phylogeny viewpoint, a phylogenetic tree reconstruction method is good if: (1) it is computationally efficient, (2) it is accurate at search-

* Corresponding author. Fax: +1 608 262 7509.
E-mail address: brlarget@wisc.edu (B. Larget).

1055-7903/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2005.03.026

ing tree space to find optimal trees, and (3) simulation studies show that trees found by the method are likely to be close to the correct tree.

The *statistical phylogeny viewpoint* emphasizes different criteria. From the statistical point of view, a phylogenetic method is good if: (1) it is likely to produce accurate estimates; (2) the estimates are associated with measures of uncertainty for which there is a theoretical basis of interpretation; (3) measures of uncertainty are accurate; (4) the method is robust to erroneous assumptions; (5) and the method may be implemented efficiently. The most important difference between the optimal phylogeny viewpoint and the statistical phylogeny viewpoint is the emphasis in the latter on measures of uncertainty. While we are motivated to produce good statistical methods, we will show that the methods described in this paper have good optimal phylogeny properties as well.

1.1. Genome arrangements

Genome arrangements are represented abstractly as signed permutations, where each permutation element represents either a gene or a block of genes. Elements of the same sign correspond to genes located on the same strand. Gene inversions are rearrangement events that correspond to reversals of signed permutations, where the reversal changes both the order and the signs of the affected elements. Circular genomes with $n+1$ gene blocks may be represented as signed permutations of length n by choosing an arbitrary reference gene and reading the remaining genes around the circle.

There are several possible ways to measure the distance between two arrangements. (See Pevzner, 2000, for example.) The breakpoint distance between two genome arrangements counts the number of adjacent pairs of genes in one arrangement that are not present in the other. This distance is not directly a function of any presumed mechanism for rearrangement. The reversal distance counts the minimal number of gene inversions necessary to transform one arrangement into another. Additional distances are defined by allowing other types of rearrangement, such as gene transposition. These distances can also be extended to genome arrangements on multiple chromosomes if we consider rearrangement events that affect more than one chromosome. However, in the present work we restrict consideration to unichromosomal genome arrangements and processes that rearrange genomes on a single chromosome.

The most straightforward analysis attempts to reconstruct the genome rearrangement events that separate two genome arrangements. Hannenhalli and Pevzner (1995) found the first (exceedingly clever) polynomial time algorithm for computing the reversal distance between any two arrangements. Kaplan et al. (1999) and Bader et al. (2001) simplified and improved the method. More recent work seeks to estimate phylogeny and ancestral genome arrangements among three or more species. The most studied approach is based on the principle of maximum parsimony: reconstructions that involve the smallest possible number of genome rearrangements are sought.

There are several genome arrangement publications that take the optimal phylogeny viewpoint. Cosner et al. (2000b) describes the Maximum Parsimony for Rearranged Genomes Problem as the search for a tree and genome arrangements at the internal nodes to minimize the sum of the pairwise distances over edges of the tree. If the distance measure counts breakpoints, an optimal tree is called a minimum-breakpoint tree. Sankoff and Blanchette (1998) and Blanchette et al. (1999) describe a computational method to search for minimum-breakpoint trees. Cosner et al. (2000b), Moret et al. (2001, 2002a,b), and Tang and Moret (2003) (and further references therein) describe subsequent improvements to this approach which substantially increase the speed of finding minimum-breakpoint trees, and also allow searches for most parsimonious trees that minimize the total number of gene inversions. The Multiple Genome Rearrangement Problem (Bourque and Pevzner, 2002) is the same problem in the special case where gene inversions are the only rearrangement mechanism. Solutions to this problem are most parsimonious in that they require the smallest number of total changes, or the smallest number of rearrangement events when the distance measure counts rearrangements.

There are now a few publications that describe the statistical phylogeny viewpoint. In previous work, we have approached the problem of phylogenetic inference from genome arrangements this point of view. Simon and Larget (2001) describe a Bayesian approach to the problem that was limited to small simulated data sets. Larget et al. (2002) solves the computational difficulties that limited our previous approach and describes a Bayesian method of inference that is computationally tractable for genuine data sets. York et al. (2002) and Miklós (2003) also use a Bayesian approach to reconstructing genome rearrangement histories, but restrict attention to two-taxon trees. More recently, we describe computational advances that allow us to analyze the complete mitochondrial genome arrangements of 87 metazoan taxa (Larget et al., 2005a) and we compare our software BADGER (Simon and Larget, 2004) with the program GRAPPA (Bader et al., 2002) as a tool for finding most parsimonious reconstructions (Larget et al., 2005b).

The remainder of this paper compares a Bayesian approach with maximum parsimony as applied to several example data sets. The types of inference possible in a Bayesian analysis are very different from those made within the maximum parsimony framework. Specifically, our analyses include calculations of uncertainty

in both the estimated ancestral sequences and the phylogeny.

2. Methods

2.1. Model

We assume a very simple model with gene inversion as the sole mechanism of genome rearrangement. 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 and assume that all unrooted tree topologies are equally likely. Edges of the unrooted tree have independent lengths selected from a Gamma distribution. Given an edge length, a Poisson number of gene inversions with this mean are realized, so that the unconditional distribution of the number of events per edge is negative binomial. The event locations on each edge are independent and uniformly distributed. Given that a gene inversion occurs, we assume that all possible gene inversions are equally likely. The observed data are completely determined by the tree topology, edge lengths, and inversion scenario. We are able to integrate out analytically the specific dependence on the edge lengths and the absolute locations of the gene inversions and so can evaluate the joint posterior distribution of the tree topology and the ordered sequence of specific gene inversion events on each edge up to a normalizing constant. See Larget et al. (2002) for further details.

2.2. Prior distribution

This model contains two hyperparameters, the shape and scale parameters (α and λ, respectively), of the Gamma distribution for edge lengths. The prior probability of x inversions on an edge is

$$P(x \mid \alpha, \lambda) = \frac{\Gamma(x+\alpha)}{\Gamma(\alpha)x!}\left(\frac{\lambda}{1+\lambda}\right)^{\alpha}\left(\frac{1}{1+\lambda}\right)^{x}, \quad x = 0, 1, 2, \ldots \tag{1}$$

The reparameterization $\mu = \alpha/\lambda$ and $\psi = (1 + 1/\lambda)$ is easier to interpret. The prior distribution for the number of inversions on an edge becomes

$$P(x \mid \mu, \psi) = \left(\frac{\Gamma((\mu/(\psi-1))+x)}{x!\Gamma(\mu/(\psi-1))}\right)\left(\frac{1}{\psi}\right)^{\mu/(\psi-1)}\left(\frac{\psi-1}{\psi}\right)^{x}, \quad x = 0, 1, 2, \ldots$$

with mean and variance μ and $\mu\psi$, respectively. Fig. 1 shows this prior distribution for two different choices of the hyperparameters. The graphs in Fig. 1 show distributions for various values of the hyperparameters.

2.3. Markov chain Monte Carlo

The state space for our Markov chain consists of the tree topology, the gene inversion counts on each edge, and the relative order in which the specific inversions occur constrained to be consistent with the observed arrangements. We refer to the sequence of gene inversions on an edge as its *history*. Larget et al. (2002) describes a method to sample from this state space by cycling through three different updates. In a single update, a (possibly different) tree topology and set of histories are proposed. This proposal is either accepted or rejected by Metropolis-Hastings. (See Gelman et al., 1995, for example.) If rejected, the current state is sampled again. The resulting (dependent) random sample of trees and histories are distributed according to the Markov chain's stationary distribution, which is the desired Bayesian posterior distribution.

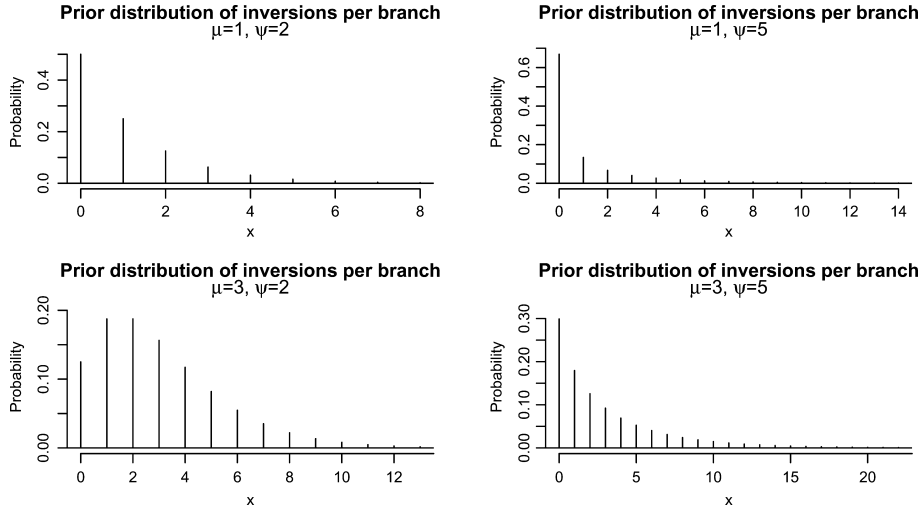

Fig. 1. Prior probability distributions for choices of hyperparameters used in the virus and Campanulaceae examples.

Table 1
Brief descriptions of update methods

Method	Description
Update 1	Slide an internal node between the histories on the two edges connecting two of its three neighbors. Update the history on the edge to the third neighbor
Update 2	Update the history on a single edge
Update 2x	Update part of the history on a single edge
Update 3	Swap the locations of two edges each adjacent to a randomly selected internal edge. Update the histories on the moved edges
Update 4	Similar to tree-bisection–reconnection (TBR). Remove an internal edge, reconnect the two unrooted subtrees at a random location, and create a new history on the new edge
Update 5	Similar to Update 3, except that only one of the two possible edges is moved

Updates 1, 2, and 2x do not affect the tree topology, but modify only the histories on one or more edges in the same proximity of the tree. Updates 3, 4, and 5 have the potential to change the tree topology as well as the histories on a set of proximate edges.

The present paper uses Updates 1, 2, 2x, 3, 4, and 5 from Simon and Larget (2004) summarized briefly in Table 1 and described more completely in Larget et al. (2005a). The additional update methods improve the mixing properties of the Markov chains. Updating a history on a single edge consists of selecting at random a sequence of inversions that rearrange a known arrangement at one end of an edge to a known arrangement on the other end. At each step of generating this sequence, the set of all possible inversions are divided into three classes (good, not-so-good, and bad). A single inversion is selected with uniform probability within each class where these class probabilities decrease from good to bad. The definitions of the classes are slightly different in this paper than what we used in Larget et al. (2002). We now include inversions on two breakpoints in the same cycle of a hurdle to be good rather than not-so-good. (See Larget et al., 2002, for definitions of these terms.) Update 2x is similar to the update method in York et al. (2002).

3. Results

We first show that arrangements that are closer in reversal distance are not necessarily more likely. Assume that we have a small artificial genome with nine genes arranged in a circle, so the arrangements are represented by signed permutations of size eight. Consider these two examples:

$$p_1 = (8, 3, 7, 1, -5, -4, -6, 2) \quad \text{and}$$
$$p_2 = (2, 3, 4, 5, 6, 8, 1, 7). \tag{2}$$

The first permutation requires four reversals to sort, the second five. While it might be supposed that the first permutation would be more likely than the second if a random number of random reversals with mean equal to the actual distance of the first permutation from the identity (i.e., four) were applied to the identity permutation, this turns out not to be the case. Applying a Poisson(4) distributed number of random reversals to the identity permutation with all possible reversals being equally likely, the second arrangement is more than twice as likely as the first. The reason is that there is but a single sequence of four reversals that sorts the first permutation while there are 200 sequences of reversals of length five that sort the second. Table 2 contains counts of the number of short sorting sequences for the two permutations.

Table 2
Numbers of sorting sequences

Permutation	Distance	Number of sorting sequences			
		4	5	6	7
p_1	4	1	8	791	9,918
p_2	5	0	200	2,668	147,282

For the signed permutations in Eq. (2), the second column lists the minimal number of reversals to sort, and the remaining columns contain the number of distinct sorting sequences by length.

There are a total of 36 possible reversals for permutations of length eight. The probability of achieving these permutations after applying a Poisson(4) distributed number of random reversals to the identity permutation may be calculated by conditioning on the realized number of reversals.

$$\begin{aligned} &P(\text{identity to } p) \\ &= \sum_{k=0}^{\infty} \{P(\text{exactly } k \text{ reversals}) \\ &\quad \times (\#\text{ of sorting sequences of } p \text{ of length } k)\}/ \\ &\quad (\text{total } \#\text{ of sequences of length } k). \end{aligned} \tag{3}$$

Truncating this sum at $k = 7$, the probability of p_1 is approximately $2.8 \times \exp(-4 \times 4^4/(36)^4)$ while the probability of p_2 is approximately $6.5 \times \exp(-4 \times 4^4/(36)^4)$, more than twice as large. This indicates that the most parsimonious reconstructions may not be the most likely, even if the mean of the prior probability distribution is selected to maximize the posterior probability of that reconstruction.

3.1. Real data examples

We analyze three sets of data that have appeared previously in the literature in order to compare our results

with those from previous analyses. Our strategy is the same for each set of data. In order to assess the robustness of the results to prior specification, for each data set we made calculations for five different sets of hyperparameter values. Four of the sets consist of all combinations of a high and low value for μ and ψ. In the fifth set, we use an empirical Bayes approach and estimate values for μ and ψ from the edge lengths of the neighbor-joining tree formed from the observed pairwise reversal distances. We estimate μ as the mean of the neighbor-joining edge lengths and ψ as the maximum of 1.1 and the ratio of the variance to the mean of the neighbor-joining edge lengths. (The parameter ψ cannot be one or less.)

We replicated each set of runs four times using different initial states and sequences of pseudo-random numbers. In all cases we find that differences in calculations made with the same set of hyperparameters are consistent with Monte Carlo sampling error and there is no indication of potential convergence problems. We combine the four samples in each case prior to analysis.

The first two real data examples have only three taxa. For these two examples we consider the posterior distributions of the genome arrangement of the internal node and of the total number of inversions on the tree. To obtain results in each of these analyses, we cycle through four separate MCMC update methods that modify the inversion history only to sample the possible rearrangement scenarios.

The third data example is a set of chloroplast arrangements from 13 taxa. For this analysis we examine the posterior distributions of clades as well as the total tree length.

Table 3
Viral genome arrangements

Virus	Arrangement
HSV	(1–16)($\overline{19–17}$)(20–23)($\overline{25–24}$)
EBV	(1–16)($\overline{20–17}$)(21–25)
CMV	(1–11)($\overline{13–12}$)($\overline{16–14}$)($\overline{25–24}$)(17–23)

The notation (20–23) stands for the sequence 20, 21, 22, and 23 while the notation ($\overline{19–17}$) represents the sequence −19, −18, −17, and so on.

3.2. Herpes virus example

Bourque and Pevzner (2002) reanalyzes a small virus data set studied in Hannenhalli et al. (1995) with Herpes simplex virus (HSV), Epstein–Barr virus (EBV), and Cytomegalovirus (CMV). The viral genome arrangements are displayed in Table 3. The unrooted tree relating these viruses contains a single internal ancestral node with edges to each of the three leaves. Hannenhalli et al. (1995) reduce the gene arrangements to signed permutations of seven gene blocks and find two most parsimonious rearrangement scenarios that each require seven total rearrangements. Bourque and Pevzner (2002) do not block the genes with common arrangements in the three viruses, and analyze three signed permutations of length 25, reporting a single rearrangement scenario with seven total rearrangements.

Our results are based on simulation runs of one million cycles of Updates 1, 2, and 2x, subsampled every 100 cycles. For this very small data set, our initial tree is not distinguishable from the other trees we sample. Burn-in is essentially immediate and we do not discard any sample points. Each set of runs results in a combined sample of 40,000 trees and histories. The first four sets of runs use all combinations of the hyperparameter values $\mu = 1$, 3 and $\psi = 2$, 5. Fig. 1 shows the induced prior distributions on the number of inversions per edge. The fifth set of runs used the estimates from the neighbor-joining tree of $\mu = 2.167$ and $\psi = 1.1$.

All five sets of hyperparameters put most of the posterior probability on the same set of six ancestral arrangements, shown in Table 4. The results are quite consistent for these various prior distributions—calculated probabilities differ by a few percentage points over the different priors. The two most probable ancestral arrangements are the only two consistent with a most parsimonious reconstruction. Notice, however, that one of these two arrangements is about three times as probable as the other consistently for different priors.

Table 5 shows the posterior distribution of the total number of inversions. It is quite probable under a range of priors that there are only seven total gene inversions on the tree. However, even the empirical prior indicates

Table 4
Posterior probabilities of ancestral arrangements for various priors

Arrangement	$\mu = 1$ $\psi = 2$	$\mu = 1$ $\psi = 5$	$\mu = 3$ $\psi = 2$	$\mu = 3$ $\psi = 5$	$\mu = 2.167$ $\psi = 1.1$
(1–25)	0.660	0.601	0.654	0.640	0.668
(1–23)($\overline{25–24}$)	0.287	0.262	0.285	0.277	0.285
(1–16)($\overline{20–17}$)(21–25)	0.017	0.061	0.013	0.030	0.006
(1–16)($\overline{19–17}$)(20–23)($\overline{25–24}$)	0.015	0.055	0.011	0.026	0.005
(1–16)($\overline{19–17}$)(20–25)	0.006	0.007	0.007	0.008	0.005
(1–16)($\overline{20–17}$)(21–23)($\overline{25–24}$)	0.004	0.006	0.007	0.006	0.004
All others	0.011	0.008	0.023	0.013	0.027

Table 5
Distribution of total number of gene inversions in the virus example

Total	$\mu = 1$ $\psi = 2$	$\mu = 1$ $\psi = 5$	$\mu = 3$ $\psi = 2$	$\mu = 3$ $\psi = 5$	$\mu = 2.167$ $\psi = 1.1$
7	0.923	0.826	0.891	0.870	0.925
8	0.060	0.140	0.070	0.090	0.053
9	0.015	0.027	0.035	0.034	0.021
10	0.002	0.006	0.003	0.005	0.001
11+	0.000	0.001	0.000	0.001	0.000

a probability of 7.5% that the actual inversion history is not one of the most parsimonious reconstructions.

In this example, the optimal phylogeny point of view analysis provides no means to estimate the uncertainty in a most parsimonious reconstruction. Our analysis quantifies the uncertainty in a fashion that is somewhat robust to prior assumptions.

3.3. Human, fruit fly, and sea urchin mitochondrial arrangements

Sankoff et al. (1996) and Bourque and Pevzner (2002) analyze the mitochondrial genome arrangements of human, sea urchin, and fruit fly. These authors blocked some genes to find shorter permutations of length 33. Bourque and Pevzner (2002) report a single best reconstruction that requires a total of 39 reversals. However, these analyses do not address the question of uncertainty in the reconstructions.

We use the full mitochondrial arrangements with 37 genes in a circular genome, displayed in Table 6. Our results are based on simulation runs of one hundred million cycles of Updates 1, 2, and 2x, subsampled every 1000 cycles. The greater distances between taxa in this example require longer simulations for accurate calculation than the previous example. Burn-in is essentially immediate and we do not discard any sample points. Each set of runs results in a combined sample of 400,000 trees and histories. The first four sets of runs use all combinations of the hyperparameter values $\mu = 5$, 8 and $\psi = 2$, 5, reflecting that we expect greater distances among these distantly related taxa. The fifth set of runs used the estimates from the neighbor-joining tree of $\mu = 12.5$ and $\psi = 2.293$.

There is far greater uncertainty in the ancestral arrangement in this example as compared to the previous example where most of the posterior probability was on a handful of arrangements. Also, in contrast to the previous example where the results were quite robust to differences in the values of the hyperparameters, the sizes of credible regions are quite different depending on the choice of prior. (See Table 7.) Prior distributions with means much smaller than the observed tend to concentrate posterior probability more heavily on arrangements consistent with the observed minimum of 39 total gene inversions. But the estimates based on the neighbor-joining hyperparameter estimates are spread the most (as expected since the mean is so much higher in this run as compared to the others). With this empirically determined prior, over 100,000 arrangements are necessary to create a 90% credible region for the ancestral arrangement.

With such a diffuse posterior distribution, a list of arrangements and their probabilities is an insufficient summary. We can, however, partially summarize the distribution in a manner similar to that used for summarizing large sets of trees with a majority rule consensus tree. Instead of searching for clades, we can summarize a distribution on arrangements by listing the maximal subsequences that have posterior probability greater than 0.5. Table 8 displays these common subsequences.

For each of the priors we examined, the probability that the total number of inversions on the tree exceeds

Table 7
Human, fruit fly, and sea urchin ancestral arrangement credible region sizes

Probability	$\mu = 5$ $\psi = 2$	$\mu = 5$ $\psi = 5$	$\mu = 8$ $\psi = 2$	$\mu = 8$ $\psi = 5$	$\mu = 12.5$ $\psi = 2.293$
0.50	500	92	2,034	327	7,610
0.75	9,703	2,925	24,692	8,457	44,819
0.90	42,593	20,663	72,293	37,382	103,387
0.95	62,593	39,218	92,293	57,382	123,387

Each count indicates the size of the smallest set of arrangements as calculated by MCMC that are necessary to achieve a posterior probability of a given probability for each of several prior distributions.

Table 6
Mitochondrial genome arrangements

Taxon	Arrangement
Human	(1–36)
Fruit fly	(26)(3–4)(2)(5–9)($\overline{-33}$)(10)($\overline{-34}$)(14)($\overline{18}$)($\overline{22}$)($\overline{16}$)($\overline{13–11}$)(20–21)($\overline{17}$)(19)($\overline{1}$)($\overline{27}$)($\overline{15}$)($\overline{25–23}$)(28–32)(35–36)
Sea urchin	(10–11)(3–7)(1)(9)(12–14)(16–17)(19)(22–23)($\overline{18}$)(20)($\overline{21}$)(29)($\overline{34}$)(15)(33)(32)($\overline{35}$)($\overline{24}$)(30)($\overline{2}$)($\overline{36}$)(8)(26–28)(31)(25)

Human, fruit fly, and sea urchin mitochondrial genome arrangements expressed as signed permutations of size 36 relative to the human arrangement with cox1 as the reference gene. $-$S2 = 1, D = 2, cox2 = 3, K = 4, atp8 = 5, atp6 = 6, cox3 = 7, G = 8, nad3 = 9, R = 10, nad4L = 11, nad4 = 12, H = 13, S1 = 14, L1 = 15, nad5 = 16, $-$nad6 = 17, $-$E = 18, cob = 19, T = 20, $-$P = 21, F = 22, rns = 23, V = 24, rnl = 25, L2 = 26, nad1 = 27, I = 28, $-$Q = 29, M = 30, nad2 = 31, W = 32, $-$A = 33, $-$N = 34, $-$C = 35, and $-$Y = 36.

Table 8
Common partial arrangements from the human, fruit fly, and sea urchin example

Partial arrangement	Posterior probability
(35–36) (0–2)	0.955
(3–9)	0.924
(10–15)	0.558
(16–17)	0.972
(19–21)	0.519
(22–32)	0.650

Each displayed arrangement has posterior probability greater than 0.5. All subarrangements of the displayed partial arrangements are also seen in a majority of sampled histories.

the minimum observed value of 39 ranges from about 93% ($\mu = 5$, $\psi = 2$) to over 99% ($\mu = 5$, $\psi = 5$). An analysis that focuses on arrangements consistent with a most parsimonious reconstruction is an inadequate summary of the inherent uncertainty in the ancestral arrangement, especially if a single arrangement is reported. For example, in our combined samples from different priors, we found 127 unique arrangements consistent with most parsimonious reconstructions and have no reason to believe that we have found them all. Even so, the combined posterior probability for the entire collection of most parsimonious reconstructions is small under each prior we have considered including the empirically determined one.

3.4. Campanulaceae chloroplast genome arrangements

Cosner et al. (2000a,b), Moret et al. (2001, 2002b), and Bourque and Pevzner (2002) analyze a data set of chloroplast genome arrangements with 105 markers from 12 Campanulaceae genera plus the outgroup tobacco. These arrangements are in Table 9. (Cosner et al. (2000a) and Cosner et al. (2000b) contain several typographical errors in reporting these genome arrangements. The arrangements in Table 9 are consistent with the data set available on the Web site of one of the authors of these papers.)

This example is more complicated than the previous examples because there is considerable uncertainty in the true phylogeny as well as in the ancestral arrangements. For 13 taxa, there are 13,749,310,575 possible unrooted binary trees. Based on a heuristic search of part of the tree space, Moret et al. (2001) finds 216 different trees that require only 67 total gene inversions. Subsequent improvements in their algorithm reduced the scores for some of these trees to 64 (Moret et al., 2002b). Using a different heuristic search method, Bourque and Pevzner (2002) also reports a single tree with 65 total gene inversions.

We again made five sets of runs with different prior distributions. The first four sets of runs use all combinations of the hyperparameter values $\mu = 1$, 3 and $\psi = 2$, 5, reflecting that we expect greater distances among

Table 9
Campanulaceae arrangements

Label	Genera	Arrangement
1	Trachelium	(1–15)(76–56)(53–49)(37–40)(35–26)(44–41)(45–48)(36)(25–16)(90–84)(77–83)(91–96)(55–54)(105–97)
2	Campanula	(1–15)(76–56)(53–49)(39–37)(40)(35–26)(44–41)(45–48)(36)(25–16)(90–84)(77–83)(91–96)(55–54)(105–97)
3	Adenophora	(1–15)(76–56)(53–49)(39–37)(28–35)(40)(26–27)(44–41)(45–48)(36)(25–16)(90–84)(77–83)(91–96)(55–54)(105–97)
4	Symphyandra	(1–15)(76–56)(39–37)(49–53)(40)(35–26)(44–41)(45–48)(36)(25–16)(90–84)(77–83)(91–96)(55–54)(105–97)
5	Legousia	(1–4)(9–15)(76–56)(27–26)(44–41)(45–48)(36–35)(25–16)(90–84)(77–83)(91–96)(5–8)(55–53)(105–98)(28–34)(40–37)(49–52)(97)
6	Asyneuma	(1–15)(76–61)(56–53)(60–57)(27–26)(44–41)(45–48)(36–35)(25–16)(89–84)(77–83)(90–96)(105–98)(28–34)(40–37)(49–52)(97)
7	Triodanus	(1–15)(76–56)(27–26)(44–41)(45–48)(36–35)(25–16)(89–84)(77–83)(90–96)(55–53)(105–98)(28–34)(40–37)(49–52)(97)
8	Wahlenbergia	(1–11)(60–56)(53–49)(37–40)(35–28)(12–15)(76–61)(27–26)(44–41)(45–48)(36)(54)(25–16)(90–84)(77–83)(91–96)(55)(105–97)
9	Merciera	(1–10)(49–53)(28–35)(40–37)(60–56)(11–15)(76–61)(27–26)(44–41)(45–48)(36)(54)(25–16)(90–85)(77–84)(91–96)(55)(105–97)
10	Codonopsis	(1–8)(36–18)(15–9)(40)(56–60)(37–39)(44–41)(45–53)(16–17)(54–55)(61–76)(96–77)(105–97)
11	Cyananthus	(1–8)(28)(36–29)(27–26)(40)(56–60)(37–39)(25–9)(44–41)(45–48)(55–49)(61–96)(105–97)
12	Platycodon	(1)(8)(2–5)(29–36)(56–50)(28–26)(9)(49–45)(41–44)(37–40)(16–25)(10–15)(57–59)(6–7)(60–96)(105–97)
13	Tobacco	(1–105)

Chloroplast genome arrangements of 12 genera of Campanulaceae and the outgroup tobacco are displayed in maximal gene blocks relative to the outgroup tobacco. These data are at http://www.cs.utexas.edu/users/stacia/ismb2000/.

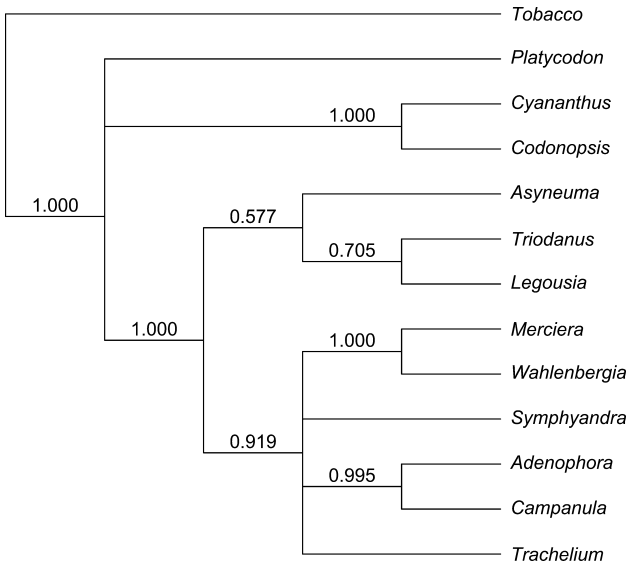

Fig. 2. Consensus tree of Bayesian posterior sample for the Campanulaceae example.

these distantly related taxa. The fifth set of runs used the estimates from the neighbor-joining tree of $\mu = 2.448$ and $\psi = 2.004$. Each run was for 100,000,000 cycles, subsampled every 100. We removed the initial 10% of each run as burn-in.

Fig. 2 shows the majority rule consensus tree of the trees sampled using the empirical prior distribution. The number of different tree topologies in credible regions of 50, 75, 90, and 95% are about 50, 105, 205, and 305, respectively. The majority rule consensus tree is the same for all of the priors we consider, except for the resolution of the taxa *Legousia*, *Asyneuma*, and *Triodanus* which differed from that shown in Fig. 2 using the prior with $\mu = 1$ and $\psi = 5$. This prior placed the smallest posterior probability on the set of trees consistent with most parsimonious reconstructions among the set of prior we considered. Probabilities of other common clades vary by a few percentage points for the alternative priors we considered. Using the empirical prior distribution, the posterior probabilities for the total number of gene inversions are, respectively, 64 (70.4%), 65 (24.0%), 66 (4.7%), 67 or more (0.9%). We found a total of 180 trees that required only 64 inversions.

4. Discussion

Comparisons between different phylogenetic methods are often controversial, perhaps in part because different authors making the comparisons use different criteria to evaluate the effectiveness of the methods. The comparisons we make in this paper between our method and the methods of other authors will be made more clear by making the criteria of the comparisons explicit.

One point of view, which we call the *optimal phylogeny viewpoint*, frames phylogeny reconstruction in the following way. Each tree may be scored when evaluated using a given data set. The scoring function is chosen so that the tree that scores best is likely (in some sense) to be the correct tree. With the optimal phylogeny viewpoint, a phylogenetic tree reconstruction method is good if it is computationally efficient and accurate at searching tree space to find optimal trees. Furthermore, it is good if simulation studies show that trees found by the method are likely to be close to the correct tree.

The *statistical phylogeny viewpoint* emphasizes different criteria. From the statistical point of view, a phylogenetic method is good if: (1) it is likely to produce accurate estimates; (2) the estimates are associated with measures of uncertainty for which there is a theoretical basis of interpretation; (3) the method is robust to erroneous assumptions; (4) the method uses available data efficiently, in the sense that measures of uncertainty are accurate; and (5) the method may be implemented efficiently. The most important difference between the optimal phylogeny viewpoint and the statistical phylogeny viewpoint is the emphasis in the latter on measures of uncertainty.

From the optimal phylogeny point of view, the method we present here is competitive with methods that search for maximum parsimony reconstructions, at least on the examples in the manuscript. We have not yet made a thorough comparison of the computational efficiencies of the various approaches. From the statistical phylogeny point of view our approach has advantages because the other approaches do not address the evaluation of the uncertainty in the estimates. We think that the type of summaries of the posterior distribution on trees and on ancestral arrangements in the manuscript are a richer description of the information in a data set than that available from a parsimony analysis.

A most parsimonious reconstruction must always be a lower bound on the actual number of genome rearrangement events. The best case for maximum parsimony methods is in the case in which the most parsimonious reconstruction is very likely to be correct. Then a biologist interpreting the results has a good basis from which to start. For example, in the herpes virus example, one ancestral arrangement has a substantial amount of posterior probability and is not too bad of a summary by itself. But if individual most parsimonious reconstructions are very unlikely, there is a high degree of uncertainty about which reconstruction is correct. In the human, fruit fly, and sea urchin example, there is considerable uncertainty in the ancestral arrangement. To report a single ancestral arrangement in this case is highly misleading. The real difficulty is that maximum parsimony methods provide no warning

when the single reconstruction selected has low probability of being correct.

By contrast, Bayesian methods report a full posterior distribution on the space of possible trees and arrangements. If one of those is very likely (whether it is most parsimonious or not), that fact will be evident from the distribution. If there are many, roughly equally likely trees or ancestral arrangements, that also will be evident.

The Bayesian analyses have other virtues as well. Because the Markov chain Monte Carlo sampler typically spends the bulk of its time on trees of high probability, it coincidentally can find better maximum parsimony trees than found by other computational approaches for some data sets. For example, in the Campanulaceae data set, we found 180 different trees with 64 inversions. We expect that other researchers interested in finding most parsimonious reconstructions may find stochastic search based on MCMC to be more efficient than current heuristic optimization methods, at least as part of an initial search strategy to find a good starting point for a heuristic search. Bourque and Pevzner (2002) describe the Campanulaceae data set with its 13 taxa as "one of the most challenging genome rearrangement data sets." Larget et al. (2002) successfully applies the Bayesian approach used in this paper to a data set with 19 taxa, a problem in which the tree space is more than 460 million times as large.

A Bayesian approach has other benefits. First, it is possible to incorporate gracefully other sources of information. This information may come from previous studies on other data. Furthermore, it is straightforward in principle to extend our current model by adding other mechanisms of genome rearrangement or to use prior information about inversion hot spots to remove the assumption that all possible inversions are equally likely. Extending the approach to the multichromosomal data sets described in Bourque and Pevzner (2002) should also be possible.

A common criticism of Bayesian methods is the choice of prior distribution. Ideally, an individual researcher will specify a prior that is an accurate description of his or her prior belief. In a field as new as reconstructing evolutionary histories of genome rearrangement where there are minimal examples by which to form prior opinion, we expect that some readers will prefer an empirical means to specify a prior distribution. This paper suggests one method for doing this by estimating the values of the hyperparameters using the neighbor-joining tree.

Acknowledgment

All three authors were supported in part by NIH Grant R01 GM068950-01.

References

Bader, D.A., Moret, B.M., Warnow, T., Wyman, S.K., Yan, M., Tang, J., Siepel, A.C., Caprara, A., 2002. GRAPPA, version 1.6. Available from: <http://www.cs.unm.edu/moret/GRAPPA/2b>.

Bader, D.A., Moret, B.M.E., Yan, M., 2001. A linear-time algorithm for computing inversion distances between signed permutations with an experimental study. Journal Computational Biology 8, 483–491.

Blanchette, M., Kunisawa, T., Sankoff, D., 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. Journal of Molecular Evolution 49, 193–203.

Bourque, G., Pevzner, P., 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Research 12, 26–36.

Cosner, M.E., Jansen, R.K., Moret, B.M.E., Raubeson, L.A., Wang, L.-S., Warnow, T., Wyman, S., 2000a. An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In: Comparative Genomics (DCAF-2000). Kluwer Academic Publishers, Montreal, Canada, pp. 99–121.

Cosner, M.E., Jansen, R.K., Moret, B.M.E., Raubeson, L.A., Wang, L.-S., Warnow, T., Wyman, S., 2000b. A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB-2000). AAAI Press, Menlo Park, CA, pp. 104–115.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton.

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

Hannenhalli, S., Pevzner, P., 1995. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In: Proceedings of the Twenty-seventh Annual ACM-SIAM Symposium on the Theory of Computing. ACM Press, New York, pp. 178–189.

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

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

Larget, B., Simon, D.L., Kadane, J.B., 2002. Bayesian phylogenetic inference from animal mitochondrial genome arrangements (with discussion). Journal of the Royal Statistical Society, Series B 64, 681–693.

Larget, B., Simon, D.L., Kadane, J.B., Sweet, D., 2005a. A Bayesian analysis of metazoan mitochondrial genome arrangements. Molecular Biology and Evolution 22, 486–495.

Larget, B., Simon, D.L., Sohn, S., 2005b. A comparison between BADGER and GRAPPA. Bioinformatics. In press.

Miklós, I., 2003. MCMC genome rearrangement. Bioinformatics 19 (Suppl. 2), ii130–ii137.

Moret, B., Tang, J., Wang, L., Warnow, T., 2002a. Steps toward accurate reconstruction of phylogenies from gene-order data. J. Comput. Syst. Sci. 65, 508–525.

Moret, B.M.E., Siepel, A.C., Tang, J., Liu, T., 2002b. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In: Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI'02), Rome, September 2002.

Moret, B.M.E., Wang, L., Warnow, T., Wyman, S., 2001. New approaches for reconstructing phylogenies from gene order data.

In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB-2001), pp. 165–173.

Pevzner, P., 2000. Computational Molecular Biology—An Algorithmic Approach. The MIT Press, Cambridge, MA (Chapter 10).

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

Sankoff, D., Sundaram, G., Kececioglu, J., 1996. Steiner points in the space of genome rearrangements. International Journal of the Foundation of Computer Science 7, 1–9.

Simon, D.L., Larget, B., 2001. Phylogenetic inference from mitochondrial genome arrangement data. In: Alexandrov, V., Dongarra, J., Juliano, B., Renner, R., Tan, C. (Eds.), Computational Science—ICCS 2001, Lecture Notes in Computer Science, vol. 2074. Springer-Verlag, Berlin, pp. 1022–1028.

Simon, D.L., Larget, B., 2004. BADGER, version 1.02b. Department of Mathematics and Computer Science, Duquesne University. Available from: <http://badger.duq.edu/>.

Tang, J., Moret, B., 2003. Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. In: Proceedings of the 8th Workshop on Algorithms and Data Structures (WADS'03), Lecture Notes in Computer Science, vol. 2748. Springer-Verlag, Berlin, pp. 37–46.

York, T., Durrett, R., Nielsen, R., 2002. Bayesian estimation of the number of inversions in the history of two chromosomes. Journal of Computational Biology 9, 805–818.