A Bayesian Analysis of Metazoan Mitochondrial Genome Arrangements

Bret Larget, Donald L. Simon,† Joseph B. Kadane,‡ and Deborah Sweet†*[1]

*Departments of Botany and of Statistics, University of Wisconsin at Madison; †Department of Mathematics and Computer Science, Duquesne University, and ‡Department of Statistics, Carnegie Mellon University

Genome arrangements are a potentially powerful source of information to infer evolutionary relationships among distantly related taxa. Mitochondrial genome arrangements may be especially informative about metazoan evolutionary relationships because (1) nearly all animals have the same set of definitively homologous mitochondrial genes, (2) mitochondrial genome rearrangement events are rare relative to changes in sequences, and (3) the number of possible mitochondrial genome arrangements is huge, making convergent evolution of genome arrangements appear highly unlikely. In previous studies, phylogenetic evidence in genome arrangement data is nearly always used in a qualitative fashion—the support in favor of clades with similar or identical genome arrangements is considered to be quite strong, but is not quantified. The purpose of this article is to quantify the uncertainty among the relationships of metazoan phyla on the basis of mitochondrial genome arrangements while incorporating prior knowledge of the monophyly of various groups from other sources. The work we present here differs from our previous work in the statistics literature in that (1) we incorporate prior information on classifications of metazoans at the phylum level, (2) we describe several advances in our computational approach, and (3) we analyze a much larger data set (87 taxa) that consists of each unique, complete mitochondrial genome arrangement with a full complement of 37 genes that were present in the NCBI (National Center for Biotechnology Information) database at a recent date. In addition, we analyze a subset of 28 of these 87 taxa for which the non-tRNA mitochondrial genomes are unique where the assumption of our inversion-only model of rearrangement is more plausible. We present summaries of Bayesian posterior distributions of tree topology on the basis of these two data sets.

Introduction

The relationships among several major groups of animals is uncertain. Authors relying on different data and different methods of analysis often reach quite disparate conclusions about evolutionary hypotheses. For example, the trees showing relationships among various metazoan phyla on page 4 of Nielsen (2001) and on page 849 of De Rosa (2001) contain significant inconsistencies, especially in the placement of the phyla Brachiopoda and Arthropoda. Nielsen supports his conclusions primarily on the basis of morphology, whereas De Rosa justifies his conclusions through genomic arrangements of the *Hox* and mitochondrial genes.

Several additional authors have argued that genome arrangement data are potentially highly informative for inferring deep evolutionary relationships. (See Smith *et al.* 1993; Boore *et al.* 1995; Boore and Brown 1998, for example.) These papers make inferences from genome arrangement data in an informal manner. They argue that since the number of possible arrangements is so large and the rate of rearrangement so slow, common parts of genome arrangements in different species are much more likely to be present because of a lack of rearrangement since the time of the common ancestor rather than because of the common partial arrangement arising more than once independently.

Other researchers have developed quantitative methods to determine phylogenies from genome arrangements. In these methods, genome arrangements are represented mathematically as *signed permutations*. One represents circular genome arrangements of $n + 1$ genes and linear genome arrangements of n genes as signed permutations of size n. The permutation describes the relative positions of the genes, while the sign indicates the strand. Biological events that rearrange genomes correspond to operations that modify permutations. For example, a gene inversion corresponds to the reversal of a portion of a permutation where both the order and the signs of the affected permutation elements are switched. A single reversal on elements two through four changes the permutation 1, 2, 3, 4, 5, 6 to 1, −4, −3, −2, 5, 6 whereas the change from 1, 2, 3, 4, 5, 6 to 1, 4, 2, 3, 5, 6 can be explained by either three reversals or a single transposition.

A Parsimony Approach

Algorithms to find the minimal number of gene inversions to explain a rearrangement history for two-taxon trees have been known since the mid 1990s. (See Pevzner [2000, chapter 10 and references therein].) Several authors have extended this maximum parsimony approach to trees with more than two taxa, although there is no tractable algorithm guaranteed to find the correct solution even in the three-taxon tree case. Sankoff and Blanchette (1998), Cosner *et al.* (2000), Moret *et al.* (2002), Tang and Moret (2003), and Bourque and Pevzner (2002) have used a parsimony approach and cite references to several other articles that do the same. The software GRAPPA (Bader *et al.* 2002) searches for most parsimonious solutions and is freely available.

The authors of these articles frame the problem as one of optimization, where the objective is to design algorithms that find most parsimonious solutions efficiently. However, these authors do not consider the statistical question of uncertainty assessment. How strong is the evidence that the relationships in the optimal tree are, in fact, correct? The conventional method of assessing reliability of trees

[1] Present address: Department of Chemical Engineering, University of Maryland at College Park.

Key words: Bayesian phylogenetics, gene inversion, gene order, genome arrangements, mitochondrial genomes.

E-mail: larget@stat.wisc.edu.

Mol. Biol. Evol. 22(3):486–495. 2005
doi:10.1093/molbev/msi032
Advance Access publication November 3, 2004

estimated by parsimony from aligned sequences is the bootstrap, relying on the assumption of independence among sites. However, sites in genome arrangements are highly dependent, so it is unclear how the bootstrap could be applied.

A Bayesian Approach

In contrast to the parsimony approach, a Bayesian approach to the problem of estimating phylogenies from genome arrangement data offers a complete statistical framework for evaluating evolutionary hypotheses. In previous work (Larget, Simon, and Kadane 2002), we described a method to quantify the uncertainty in evolutionary hypotheses on the basis of genome arrangements. Larget, Simon, and Kadane (2002) examine the complete mitochondrial genome arrangements known at the time. We concluded that the data were consistent with the Lophotrochozoa/Ecdysozoa hypothesis (that brachipods are protostomes and that molting animals are a monophyletic group) and that there was overwhelming evidence against the placement of brachiopods and arthropods in a tree such as that in Nielsen (2001). The analysis in Larget, Simon, and Kadane (2002), however, was weakened by its reliance on a non-informative prior distribution that placed prior odds of the correct classification of the 19 taxa analyzed into their respective eight phyla at about 300 billion to one against. The result was a posterior distribution with most weight on biologically implausible trees. We were able to draw conclusions only by pruning taxa whose placement was highly inconsistent with widely held belief. We included only rudimentary summaries of the posterior distribution.

In this article we improve upon our previous approach by incorporating a limited amount of prior information, essentially by restricting attention to trees that maintain several monophyletic groups for which there is strong support in the literature. In the time since publication of our previous work, there has been an increase in the number of known complete mitochondrial genome arrangements. We use this additional information in the calculations reported here. Furthermore, we have made substantial software improvements so that we are able to analyze much larger data sets. We briefly describe the computational innovations in this article.

Ultimately, we are interested in quantifying the uncertainty in phylogenetic relationships using all available molecular data, including both genome arrangements and molecular sequences, while also incorporating morphological information through a prior distribution. This approach would require the joint modeling of all of the mechanisms of molecular evolution. We have several technical challenges to overcome before we can carry out such an analysis. One challenge is to model jointly several simultaneous methods of genome rearrangement, such as gene inversion, gene transposition, and gene duplication and random deletion. The success of other authors in developing models for two-taxon trees that incorporate alternative rearrangement mechanisms (York, Durrett, and Nielsen 2002; Miklós 2003) makes us optimistic that similar approaches will work on larger trees.

Mechanisms of Genome Rearrangement

Boore and Brown (1998) describe several mechanisms of genome rearrangement and Smith *et al*. (1993) provide evidence that multiple rearrangements might have acted to rearrange the tRNA genes in metazoan mitochondrial genomes. Despite this evidence, in this article we will restrict attention to gene inversion as the sole mechanism of genome rearrangement. If gene inversion is the primary mechanism of metazoan mitochondrial genome rearrangement, our model will yield reasonable results, but we must be cautious in interpretation of our findings because of unknown effects of model misspecification.

NCBI Data Sets

The National Center for Biotechnology Information (NCBI) maintains a list of sequenced complete mitochondrial genomes. On a recent date (April 20, 2004), this list contained sequences from 443 metazoan taxa. We created our first data set by first eliminating metazoan genomes that did not contain the typical collection of 37 genes (leaving 385 taxa), and then eliminated duplicates, resulting in 87 complete mitochondrial genome arrangements from a total of eight metazoan phyla. Counts of species per phylum are as follows: Annelida (2), Arthropoda (28), Brachiopoda (3), Chordata (40). Echinodermata (4), Hemichordata (1), Mollusca (8), and Nematoda (1).

Because mechanisms other than gene inversion might affect tRNAs more readily than the much larger 13 protein-coding and two-ribosomal genes, we also consider a subset of this data set where we ignore the positions of the tRNAs. Elimination of the 22 tRNA genes reduces the number of unique arrangements within each phylum to the following: Annelida (1), Arthropoda (8), Brachiopoda (3), Chordata (4). Echinodermata (4), Hemichordata (1), Mollusca (6), and Nematoda (1). We will refer to the first data set as *meta87* and the second as *meta28*. The mitochondrial genome arrangements and the accession numbers for the files from which we obtained this information are included with the *Supplementary Material*.

Materials and Methods

Bayesian Inference

The nature of Bayesian inference is to describe uncertainty with a probability distribution which changes in response to new information. Early papers in this area include those by Yang and Rannala (1997), Mau, Newton, and Larget (1999), and Larget and Simon (1999). Huelsenbeck *et al*. (2001) and Holder and Lewis (2003) provide more recent reviews of Bayesian phylogenetic methods.

Bayesian estimation of phylogeny on the basis of genome arrangement data is fundamentally similar to the now familiar case of aligned sequence data. We will describe a model for genome rearrangement, analogous to a model of nucleotide substitution, and a method of employing Markov chain Monte Carlo (MCMC) to sample from the posterior distribution of tree topologies. However, genome arrangement data differ from aligned sequence data in important ways so that the models and the subsequent techniques of

computation are substantially different. The major difference is that in the case of sequence data, we can assume independence between sites and explicitly compute the Markov transition probabilities from one nucleotide (or amino acid or codon) to another along a single edge.

The model for sequence change is nicely decomposed into a collection of models for nucleotide substitution at each site. The likelihood sums over all possible histories of nucleotide substitution that are consistent with the observed data. The same basic approaches extend to models for amino acids or codons. With genome arrangements, however, we cannot assume such independence among sites. As a consequence, we are not able to model rearrangement of genomes as a collection of independent, smaller models. The vast number of possible arrangements (in metazoan mitochondrial genomes, $2^{36} \times 36!$ which is much, much larger than 4, or 20, or 61) makes it practically impossible to compute rapidly the transition probabilities between any two different arrangements. To overcome this computational difficulty, we compute likelihoods of specific histories of genome rearrangement and rely on MCMC to sum over the various possible histories.

In Larget, Simon, and Kadane (2002) we described an MCMC method that cycled through three different update methods. This computational approach was wholly inadequate to correctly compute posterior probabilities for the *meta87* data set. To overcome the poor mixing of our old computational method on this new large data set, we have incorporated several major changes to the MCMC approach including adding three novel MCMC update procedures to our previous three, creating variants for several of the updates, and using Metropolis-coupled MCMC (MCMCMC) to speed mixing through running several chains simultaneously. Our software Bayesian Analysis to Describe Genomic Evolution by Rearrangement (BADGER), includes all of the improvements. The Appendix provides a detailed description of each update method and a summary of the respective acceptance probabilities. Updates 4–6 and variants for all methods are new.

Model for Genome Rearrangement

The likelihood model we describe here is nearly identical to the model in Larget, Simon, and Kadane (2002), although we introduce new notation to describe the distribution of the number of gene inversions per edge. Simply stated, the model we adopt selects a tree topology uniformly from a set of unrooted tree topologies, adds independent and identically distributed edge lengths, places gene inversion events on the tree according to a Poisson process, and selects the realized gene inversions independently and uniformly at random from all possible gene inversions. We find it useful to describe further this model with a step-by-step description of how we could simulate data with it.

1. Select a tree topology τ uniformly at random from a set $\mathcal{T}_n$ of unrooted tree topologies with n taxa.
2. Select independent edge lengths $t = \{t_i\}$, $i = 1,\ldots, E$, from a Gamma distribution with shape parameter α and scale parameter λ where $E = 2n - 3$ is the number of edges in the tree.
3. Select independent counts of realized gene inversions $x = \{x_i\}$, $i = 1,\ldots, E$, from Poisson distributions where the mean of the ith distribution is t_i.
4. For each edge, independently and uniformly at random select a set of x_i reversals $r_i = \{r_{ij}\}$, $j = 1,\ldots, x_i$, from the set of all possible reversals. If there are g genes in a circular genome, the size of this set is $R = g(g - 1)/2$. Let $r = \{r_i\}$, $i = 1,\ldots, E$, be the collection of reversals from all the edges.
5. Place these reversals on the corresponding edge at locations $s_i = \{s_{ij}\}$, $j = 1,\ldots, x_i$, that are independently and uniformly distributed and let $s = \{s_i\}$, $i = 1,\ldots, E$, be the collection of reversal locations on each edge.
6. Give an arbitrary node on the tree an arbitrary labeling (such as the identity) for its arrangement. Determine the arrangements of the leaf nodes to be consistent with the complete generated arrangement history.

The unnormalized posterior distribution of the parameters given observed arrangements y takes this form, accounting for various conditional independence relationships:

$$\mathrm{P}(\tau, t, x, r, s \mid y, \alpha, \lambda) \propto \mathrm{P}(y \mid \tau, x, r)\mathrm{P}(\tau)\mathrm{P}(t \mid \alpha, \lambda)\mathrm{P}(x \mid t) \mathrm{P}(r \mid x)\mathrm{P}(s \mid t, x). \quad (1)$$

We are able to integrate out the continuous nuisance parameters t and s, leaving this expression:

$$\mathrm{P}(\tau, x, r, \mid y, \alpha, \lambda) \propto \mathrm{P}(y \mid \tau, x, r)\mathrm{P}(\tau) \mathrm{P}(x \mid \alpha, \lambda)\mathrm{P}(r \mid x). \quad (2)$$

On the right hand side of equation (2), the first factor is one when the observed leaf arrangements y are consistent with the reversal history. Since we assume a uniform distribution over tree topologies in the set $\mathcal{T}_n$, the second factor is a constant. The third factor is a product of negative binomial probabilities, each of which arises from the gamma mixture of Poisson distributions, and is expressed as

$$\mathrm{P}(x \mid \alpha, \lambda) = \prod_i \left[\left(\frac{\Gamma(\alpha + x_i)}{x_i!\Gamma(\alpha)} \right) \left(\frac{\lambda}{1+\lambda} \right)^{\alpha} \left(\frac{1}{1+\lambda} \right)^{x_i} \right] \quad (3)$$

$$= \left(\frac{\lambda}{1+\lambda} \right)^{E\alpha} \left(\frac{1}{1+\lambda} \right)^{\sum_i x_i} \prod_i \left(\frac{\Gamma(\alpha + x_i)}{x_i!\Gamma(\alpha)} \right), \quad (4)$$

where i indexes the E edges in the unrooted tree topology. The uniform distribution on the R possible reversals implies that the fourth factor on the right-hand side of equation (2) is $R^{-\sum_i x_i}$.

The Poisson distribution has equal mean and variance. The effect of sampling edge lengths from a distribution and then sampling a realized number of inversions with a Poisson distribution whose mean is the edge length is to create a distribution for the number of inversions per edge that is over-dispersed relative to the Poisson. Our choice of the conjugate gamma prior leads to the negative binomial distribution whose mean is α/λ and variance is $(\alpha/\lambda)(1 + 1/\lambda)$. Because we find it easier to specify and interpret a prior distribution on the number of inversions per edge in terms of the mean and the over-dispersion relative to the Poisson distribution we re-parameterize the hyperparameters with $\mu = \alpha/\lambda$ and $\psi = (1 + 1/\lambda)$ so that the mean

and variance of the distribution of inversions per edge are μ and $\psi\mu$, respectively. With this re-parameterization, the equations above can be restated by letting $\alpha = \mu/(\psi - 1)$ and $\lambda = 1/(\psi - 1)$ so that the prior distribution for the number of inversions is

$$P(x \mid \mu, \psi) = \left(\frac{1}{\psi}\right)^{E\mu/(\psi-1)} \left(\frac{\psi-1}{\psi}\right)^{\sum_i x_i} \times \prod_i \left(\frac{\Gamma((\mu/(\psi-1)) + x_i)}{x_i!\Gamma(\mu/(\psi-1))}\right). \quad (5)$$

Prior Distributions on Tree Topology

The formulas that count the number of rooted and unrooted binary tree topologies with n taxa are well known (Felsenstein 1978). The number of unrooted binary trees is

$$u(n) = \begin{cases} 1 & \text{if } n = 1 \text{ or } n = 2 \\ (2n-5)!! = 1 \times 3 \times \cdots \times (2n-5) & \text{if } n \geq 3 \end{cases}, \quad (6)$$

and the number of rooted binary trees is $r(n) = u(n + 1)$. In light of prior information, we might also consider restricting attention to trees that maintain certain groups. If the n taxa are partitioned into k groups of sizes $n_1, n_2, \ldots, n_k$, where $\sum_{i=1}^{k} n_i = n$, the number of tree topologies with these groups as unrooted clades is

$$u(k) \times \prod_{i=1}^{k} r(n_i). \quad (7)$$

In this article, we will make the assumption that the partitioning of the taxa into their respective phyla with respect to the other taxa in the analysis is correct. This restricted prior distribution on tree topology distinguishes the model in this paper from the model in our earlier work (Larget, Simon, and Kadane 2002).

Were we instead to place a uniform prior distribution on all unrooted trees, the prior odds of trees that grouped animals into their respective phyla would be about 6.6×10^{49} against in the case of data set *meta87* and about 3.3×10^{18} against for data set *meta28*. We have a much stronger belief than this that grouping the animals in this study into phyla on the basis of morphology and other considerations is correct. BADGER allows grouping structure to be set as a run control. The MCMC updates that change the tree topology consider only those changes to tree topology which retain the grouping structure.

We also note that the prior probability on clades induced by a uniform prior probability on unrooted tree topology depends both on the number of taxa and the size of the clade. The prior for a specific edge that partitions an unrooted binary tree topology with n taxa into two groups of size j and $n - j$ is

$$r(j)r(n-j)/u(n). \quad (8)$$

In both data sets we consider in this article, our primary interest centers on the 10,395 possible unrooted trees that relate the eight metazoan phyla for which we have data. Any single 8-taxon binary unrooted tree topology has exactly five internal edges, each of which splits the taxa into two groups, of size 2/6, 3/5, or 4/4. There are 28 possible 2/6-splits, each with a prior probability of 1/11 $\doteq$ 0.0909, 56 possible 3/5-splits with prior probabilities 1/33 $\doteq$ 0.0303 each, and 70 possible 4/4-splits with prior probabilities 5/232 $\doteq$ 0.0216 each.

Clades for which the posterior probabilities exceed the prior probabilities have at least some level of support in the data, even if this support might be weak.

Prior Distribution on Edge Lengths

The posterior probability of tree topology given the likelihood model, observed arrangement data, and a uniform prior on tree topology is a function of the hyper-parameters μ and ψ that determine the prior distribution of the number of gene inversions per edge. In the absence of information other than the data on reasonable values to use for these hyper-parameters, we elect to take an empirical approach and use data-derived estimates. For each pair of taxa we can compute the observed inversion distance between their genomes (Pevzner 2000) using an improved algorithm (Bader, Moret, and Yan 2001). From this pairwise distance matrix we compute the Neighbor-Joining tree and then calculate the mean and variance of the edge lengths to estimate μ and $\psi\mu$, respectively. The total distance of the Neighbor-Joining tree is a lower bound on the total number of inversions necessary to reconstruct a complete inversion history consistent with the data, because the calculated pairwise distances are themselves minimum estimates. Furthermore, the edge lengths of the Neighbor-Joining tree are potentially continuous, whereas reconstructed histories must necessarily have integer-valued lengths per edge. Consequently, the prior we adopt has a bias toward most parsimonious reconstructions. If we find posterior distributions of total numbers of inversions with low probabilities on most parsimonious solutions, this will be due to the data and the likelihood model and not the prior distribution. The estimation of the hyper-parameters using edge lengths of a Neighbor-Joining tree is a small change from the approach we used in Larget, Simon, and Kadane (2002).

Calculation Details

BADGER allows a number of settings to control the running of the program. The *meta87* data set required substantially greater computational effort to produce the results we present here than did the *meta28* data set. For each data set, we began at randomly selected starting trees and inversion histories using a variation on the Neighbor-Joining algorithm. At each stage we use the Neighbor-Joining rule to connect the next two nodes, but select a random point on a random path between the two joined nodes to create an actual genome arrangement for the new internal node. We update the pairwise distance matrix using this arrangement rather than the usual Neighbor-Joining rule and iterate until the tree and inversion histories are complete. We do this first for each predetermined group of taxa, select a root for each, and then continue the

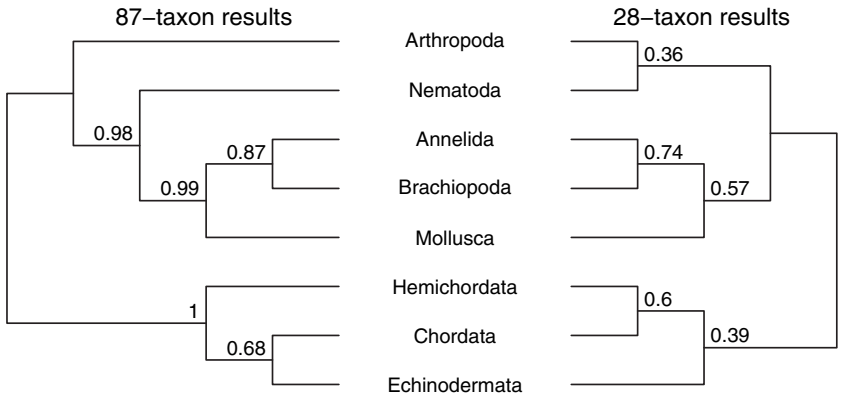

FIG. 1.—Posterior summaries of tree topology. The majority-rule consensus tree of the *meta87* data set appears on the left. The right tree summarizes the *meta28* posterior distribution showing the tree that contains the highest probability splits that do not conflict with splits with larger posterior probability.

process using the roots of the group trees to select an initial tree.

These initial trees tend to have some very long edges. We begin by cycling through all of the update methods ignoring the Hastings ratio correction for a number of *pre-burn-in* updates that rapidly drives the initial inversion history and tree closer to most parsimonious solutions where the posterior distribution is higher. We then use correct acceptance probabilities and cycle through all of our updates for very long runs, discarding the initial portion of these runs as burn-in.

For each data set we run ten simultaneous chains and swap states using MCMCMC. We repeat these analyses 20 times each, using different streams of pseudo-random numbers and compare the estimated probabilities for consistency. The posterior probabilities we report are averages over the 20 runs. The *meta28* data set used about 4 hours of computational time on a 2.8 GHz Pentium machine running Linux for each of the 20 independent analyses, whereas we ran the *meta87* data set for over 16 days of computational time per run on a set of similar machines. (We did the 20 separate runs in parallel, so the nearly one year of cumulative computational effort took one twentieth of the actual time to complete.) All calculated clade probabilities have Monte Carlo standard errors less than one percent. More details on run settings for these simulations are included in the supplementary material.

Results

Gene order information provides very little resolution at the individual taxon level for these data sets. The subsampled set of trees we analyze do not select the same exact tree topology twice for the *meta87* data set, for example. We do, however, find substantial posterior probabilities for relationships among animal phyla. The majority rule consensus tree summary of the *meta87* sample from the posterior is completely resolved at the phylum level, although there is considerable uncertainty remaining. Figure 1 compares the most probable relationships among metazoan phyla using each data set. Posterior probabilities for the *meta87* data set are higher, but the results might be more questionable because of the potential for serious model bias. The *meta28* data set is more poorly resolved, with only three phylum level splits having posterior probabilities greater than one half. Table 1 tabulates the posterior probabilities for all clades in each data set for which the posterior probabilities exceed the priors.

Discussion

Phylum-Level Relationships

In figure 1, we root the trees at a break between deuterostomes (Echinoderms, Hemichordates, and Chordates here) and protostomes. This is the most strongly supported split in the *meta87* data set. Both data sets support the group Lophotrochozoa (Annelida, Brachipoda, and Mollusca in this data set). It is interesting to note that marginal support for Ecdysozoa (Nematoda and Arthropoda in the present analysis) is stronger in the *meta28* analysis (probability equals 0.36) than it is in the *meta87* data set (probability equals 0.006). The placement of Arthropoda as outgroup to a group consisting of Nematoda and Lophotrochozoa with high posterior support in the *meta87* data set may very well be an artifact of model misspecification. There are 28 unique genome arrangements in Arthropoda in *meta87*, but only eight in *meta28*, where tRNA gene placement is ignored. The tRNA genes appear to be much more mobile and are highly likely to rearrange by mechanisms other than gene inversion, as we assume here.

It is apparent that complete mitochondrial genomes are far more informative than partial arrangements that ignore placement of tRNA genes. However, for this information to be used most effectively for phylogenetic inference leading to statistical estimates of uncertainty will require further development of models of genome rearrangement.

Limitations on Information in Gene Order

We note that the results of the analyses of both data sets are rather inconclusive as many possible trees retain non-neglible posterior probability. It is interesting to ask if mitochondrial gene order information alone is sufficient to resolve evolutionary realtionships among animals at the phylum level. The use of models that come closer to

Table 1
Posterior Probabilities of Common Clades

	Posterior		
	meta87	*meta28*	Split
1	1.000	0.391	{Nema,Arth,Ann,Moll,Brach}—{Echi,Hemi,Chor}
2	0.995	0.566	{Nema,Arth,Echi,Hemi,Chor}—{Ann,Moll,Brach}
3	0.980	0.083	{Arth,Echi,Hemi,Chor}—{Nema,Ann,Moll,Brach}
4	0.870	0.737	{Nema,Arth,Moll,Echi,Hemi,Chor}—{Ann,Brach}
5	0.681	0.216	{Nema,Arth,Ann,Moll,Brach,Hemi}—{Echi,Chor}
6	0.176	0.602	{Nema,Arth,Ann,Moll,Brach,Echi}—{Hemi,Chor}
7	0.143	0.050	{Nema,Arth,Ann,Moll,Brach,Chor}—{Echi,Hemi}
8	0.113	0.163	{Nema,Arth,Ann,Echi,Hemi,Chor}—{Moll,Brach}
9	0.006	0.356	{Nema,Arth}—{Ann,Moll,Brach,Echi,Hemi,Chor}
10	0.012	0.303	{Arth,Ann,Moll,Brach}—{Nema,Echi,Hemi,Chor}
11	0.000	0.249	{Arth,Ann,Moll,Brach,Hemi,Chor}—{Nema,Echi}
12	0.000	0.247	{Nema,Arth,Echi}—{Ann,Moll,Brach,Hemi,Chor}
13	0.113	0.163	{Nema,Arth,Ann,Echi,Hemi,Chor}—{Moll,Brach}
14	0.000	0.119	{Arth,Ann,Brach}—{Nema,Moll,Echi,Hemi,Chor}
15	0.000	0.115	{Arth,Moll}—{Nema,Ann,Brach,Echi,Hemi,Chor}
16	0.000	0.079	{Arth,Ann,Moll,Brach,Hemi}—{Nema,Echi,Chor}
17	0.000	0.066	{Nema,Arth,Ann,Brach}—{Moll,Echi,Hemi,Chor}
18	0.000	0.061	{Nema,Arth,Moll}—{Ann,Brach,Echi,Hemi,Chor}
19	0.000	0.060	{Nema,Arth,Echi,Chor}—{Ann,Moll,Brach,Hemi}
20	0.000	0.055	{Arth,Moll,Echi,Hemi,Chor}—{Nema,Ann,Brach}
21	0.000	0.039	{Arth,Echi}—{Nema,Ann,Moll,Brach,Hemi,Chor}
22	0.000	0.033	{Nema,Arth,Ann,Brach,Echi}—{Moll,Hemi,Chor}

NOTE.—The first eight splits are those with posterior probabilities larger than the prior found in the *meta87* data set. The remaining splits all have higher posteriors than priors in the *meta28* data set.

describing the actual mechanisms of mitochondrial genome rearrangement will help, as will the collection of more data, but there is reason to question whether these efforts alone will suffice. Clearly, an analysis that uses both arrangement data as well as sequence data has the potential to be far more informative. There are at least two avenues to pursue an analysis of combined sequence and arrangement data. One would be to develop an MCMC approach that models the two types of data simultaneously. A second possible approach would be to use the posterior of an analysis using one data type as an informative prior on trees for an analysis using the second data type. Doing this effectively might require changes in the standard software programs for Bayesian phylogenetics where prior specification is fairly limited.

The Restricted Prior Distribution

The results we present in this article are better than the quantitative results in Larget, Simon, and Kadane (2002), in which we did not sample any biologically plausible trees at all, in part because here we use a prior distribution that incorporates genuine biological prior information, albeit in a very simple way. The phylogenetic signal in genome arrangement data alone (observed with our present model) is fairly weak relative to what we have come to expect with sequence data, where the sequence data are often able to reject most biologically implausible trees without the aid of a restrictive prior on tree topology. On the other hand, posterior probabilities in analyses of distantly related taxa using sequence data rarely account for the uncertainty in alignment. We expect that genome arrangements will play an important and useful role in phylogenetic analyses, especially in situations where sequence alignment is highly uncertain.

Model Criticism

A limitation of the approach described in this article is the assumption that gene inversion is the sole mechanism of genome rearrangement. Although we have not carried out any formal goodness-of-fit tests on the suitability of the model, we expect that real data look different from data we could simulate under our model in important ways. Boore and Brown (1998) and Smith et al. (1993) describe several mechanisms of genome rearrangement including gene transposition, inverted gene transposition, and gene duplication and random deletion, and there may be others. One way to extend the methods in this article would be to incorporate several simultaneous mechanisms of genome rearrangement. Furthermore, our assumption that all possible gene inversions are equally likely could be relaxed. There might be important factors due to proximity to the origins of replication, the size of the affected fragments, and the size of noncoding gaps between genes. We expect that models that better incorporate biological understanding of processes of genomic rearrangement will provide more accurate quantitative measures of support for various phylogenetic hypotheses and will even allow us to make predictions about unseen arrangements. Furthermore, such models would provide a framework for inference about the actual biological mechanisms of genome rearrangement, which would be interesting in its own right.

A Comparison to Maximum Parsimony Methods

The published papers on a parsimony approach to the problem of reconstructing phylogenies on the basis of genome arrangement data have a different focus. From the perspective of these other authors, finding a tree or a set of trees with the minimal number of total inversions (or rearrangement events) is the goal, and the criterion for evaluating methods is the computational speed of the algorithms for finding the answers. Our focus is on assessing the uncertainty in the reconstructed trees and providing the scientist with an easily interpretable means of quantifying support for scientific questions of interest.

There is a connection between parsimony and Bayesian approaches to phylogeny reconstruction from genome arrangements. If we let the prior mean μ tend to 0, the Bayesian posterior distribution will be concentrated on those tree topologies and inversion histories that achieve the minimal possible number of total inversions. In this situation, solutions of the minimal total inversion trees could be good starting trees for our algorithms. If, on the other hand, we expect a priori that there may be many inversions on some edges, the set of most parsimonious reconstructions may have small posterior probability and will not be relevant in assessing uncertainty.

The only available software for seeking most parsimonious solutions is GRAPPA (Bader *et al.* 2002). GRAPPA carries out a branch-and-bound search over the entire tree space, and as such is limited to fairly small trees. On a single current desktop PC, analyses of data sets with

10 and fewer taxa take a minute or less to complete, but time requirements increase exponentially and problems with 15 or more taxa are not practical. Tang and Moret (2003) describe a method for using the Disc Covering Method (DCM) to extend GRAPPA to trees with 1000s of taxa, but this code is not available.

MCMC on Genome Arrangements

Our present MCMC approach to genome arrangement data is substantially more challenging than the MCMC methods developed for sequence data such as those in the program MrBayes (Huelsenbeck and Ronquist 2001). With sequence data, Felsenstein's pruning algorithm allows rapid calculation of the likelihood of the data by summing over all possible histories of point substitution consistent with the observed data. A consequence is that the state space of the Markov chain is the tree space and is wholly separated from the data. MCMC update methods are independent of the data except in the computation of likelihood ratios.

In contrast, the Markov chains we describe in this article for genome arrangements have as their state space the joint space of the tree and the complete history of genome rearrangement. The very large number of possible genome arrangements makes it impossible to compute complete probability transition matrices, and thus we are unable to compute the likelihood of the data by summing over all inversion histories directly. We solve this by augmenting the state space of the Markov chain with the genome arrangement history, in essence using MCMC rather than analysis to compute the likelihood of the data for a given tree. As a consequence, potential changes in the likelihood model (such as the incorporation of new mechanisms of rearrangement) would necessitate the development of new update methods for the MCMC approach, a challenging task that requires creativity, analysis to compute correct acceptance probabilities, careful implementation in code, and testing. The success by others in handling multiple mechanisms of rearrangement on two-taxon trees (York, Durrett, and Nielsen 2002; Miklós 2003) suggests, but in no way guarantees, that in the future we will be able to extend similar methods to large trees.

Direction of Future Work

The analytical approach we present here is the first rigorous statistical approach to phylogenetic inference from genome arrangement data that is computationally tractable for fairly large trees. In contrast, the maximum parsimony approach currently offers no assessment of uncertainty. As our best current inferences are fairly uncertain, we see estimation of uncertainty to be critical. We anticipate the development of more realistic models that include multiple mechanisms of rearrangement, of ways to do combined analysis of aligned sequence data with arrangement data and of more efficient computational approaches to be made by ourselves and others in the near future.

Supplementary Material

The genome arrangements used in this article are available as supplementary material in PDF format (meta87.pdf) and as plain text (meta87.txt and meta28.txt). The BADGER settings we used are in the file settings.pdf.

Acknowledgments

All four authors were supported in part by National Institutes of Health (NIH) grant R01 GM068950-01.

Appendix: Derivation of Acceptance Probabilities

Larget, Simon, and Kadane (2002) describe a procedure to generate at random a sequence of reversals that when applied in order to a source signed permutation result in a given target signed permutation. If the source and target are identical, the procedure stops with probability p_{stop}. Otherwise, at each stage the set of possible reversals is categorized as good, neutral, or bad. Some categories can be empty. We first pick a category according to the distribution in table 2 and then pick a reversal within the category uniformly at random. This reversal is applied to the source with the result becoming the new source. We continue this process until we stop with a complete sequence of reversals. Every possible sequence of reversals leading from the source to the target has a positive probability, and we compute the probability of the sequence we actually select.

When selecting the next reversal, ideally we would categorize reversals into those that decrease the inversion distance between source and target by one, those that leave the inversion distance unchanged, and those that increase the inversion distance by one. Instead, we use a categorization that is much faster to compute that is nearly, but not quite, identical. Please see Kaplan, Shamir, and Tarjan (1999) for the definitions of the following italicized terms. Our good category consists of *proper reversals* and reversals within the same *cycle* of an *unoriented connected component*. The neutral category consists of *improper reversals* within the same *cycle*. The bad category is all other reversals; namely, reversals that are not part of the same cycle. The categorization of reversals within the same *cycle* of an *unoriented connected component* as good rather than neutral is a change from Larget, Simon, and Kadane (2002).

In each of the following updates, one or two edges have all or part of their reversal histories replaced by the method briefly outlined in the preceding paragraph. We now describe the updates in turn. The first three updates (without variants) were first described in Larget, Simon, and Kadane (2002). The last three updates and variants of all updates are novel.

The state space consists of an unrooted leaf-labeled tree with n taxa and an associated ordered list of reversals on each edge that is consistent with the observed data. There are $2n - 3$ edges in total of which $n - 3$ are internal edges. There are $2n - 2$ nodes, of which $n - 2$ are internal nodes.

In the following descriptions, values in parentheses are the probabilities for each step in the proposal. Each proposal ratio (also called a Hastings ratio) is the probability of proposing the original state from proposed state divided by the probability of proposing the proposed state from the original state. The posterior ratio is the ratio of the posterior

Table 2
Reversal Generation Probabilities

				Category Probability	
#good	#neutral	#bad	P(good)	P(neutral)	P(bad)
+	+	+	p_{good}	$(1 - p_{good})p_{neutral}$	$(1 - p_{good})(1 - p_{neutral})$
+	+	0	p_{good}	$(1 - p_{good})$	0
+	0	+	p_{good}	0	$(1 - p_{good})$
+	0	0	1	0	0
0	+	+	0	$p_{neutral}$	$(1 - p_{neutral})$
0	+	0	0	1	0
0	0	+	0	0	1

NOTE.—At each stage in generating a sequence of reversals, all possible reversals are partitioned into three groups, good, neutral, and bad. The first three columns specify the possible cases where + indicates the category is not empty and 0 indicates the category is empty. Columns 4–6 are the probabilities of picking each category for each case.

probability of the proposed state over that of the current state. The acceptance probability is the minimum of one and the product of the posterior ratio and the proposal ratio.

Update 1

Select an internal node Z at random $(1/(n - 2))$. This node has three adjacent edges. Select an adjacent edge e_3 at random (1/3). This edge connects node Z to node C. The other two adjacent edges e_1 and e_2 connect Z to nodes A and B, respectively. Let x_i be the number of reversals on edge e_i. Pick a random location on the path from A to B $(1/(x_1 + x_2 + 1))$ and move Z to this new location. Edges e_1 and e_2 now have x_1' and x_2' reversals, but note that $x_1' + x_2' = x_1 + x_2$. Node Z has a new signed permutation determined by its new location on the path from A to B. Erase all reversals on e_3 and generate a new list with the new permutation at Z the source and the permutation at C the target (p_1).

To undo this proposal, we need to select the same internal node Z $(1/(n - 2))$ and random adjacent edge e_3 (1/3). We must propose the old location $(1/(x_1' + x_2' + 1))$ and generate the old sequence of reversals on e_3 (p_2). The proposal ratio is p_2/p_1.

In the variant **Update 1x**, everything is the same except that the target is the permutation at a random location on e_3 $(1/(x_3 + 1))$ rather than the permutation at node C. We leave the last partial sequence of reversals on e_3 unchanged. The complete length of the proposed history on e_3 is x_3'. The reverse proposal must update the same partial history on e_3 $(1/(x_3' + 1))$ and the proposal ratio is $(p_2/p_1) \times (x_3 + 1)/(x_3' + 1)$.

Update 2

Select a random edge $(1/(2n - 3))$. Randomly select one end point to be the source and the other to be the target (1/2). Replace the sequence from the source to the target (p_1). Compute the probability of having generated the same sequence of reversals from target to source (p_2). The proposal probability is $(p_1 + p_2)/(2(2n - 3))$.

To undo this proposal, we need to select the same edge $(1/(2n - 3))$, a random source and target (1/2), and the old list (p_3 or p_4, depending on designation of the source). The probability of the reverse proposal is $(p_3 + p_4)/(2(2n - 3))$, and the proposal ratio is $(p_3 + p_4)/(p_1 + p_2)$.

In the variant **Update 2x**, if the selected edge has x reversals, there are $x + 1$ locations. Two locations of the edge (possibly the same) are selected at random $(1/(x + 1)^2)$. The sequence between these locations is updated as in Update 2, and the other reversals remain unchanged. (The case when the two locations are the same is identical to updating an edge with no reversals.) If the proposed edge has a total of x' reversals, counting both unchanged and newly proposed reversals, the reverse proposal selects the same two locations with probability $(1/(x' + 1)^2)$ and the proposal ratio is $(p_3 + p_4)(x + 1)^2/((p_1 + p_2)(x' + 1)^2)$.

Update 3

Select an internal edge at random $(1/(n - 3))$. Call the two internal nodes joined by this edge Y and Z. Select one of the other edges adjacent to node Y (1/2) and one of the others adjacent to node Z (1/2) and label them e_1 and e_2, respectively. Edge e_1 connects node Y to node A while edge e_2 connects node Z to node B. Swap the end nodes of edges e_1 and e_2 so that Y is connected to B and Z is connected to A. Generate new reversals on edges e_1 and e_2 from source Y to target B (p_1) and from source Z to target A (p_2).

The reverse proposal must select the same internal edge $(1/(n - 3))$ and the same two edges to swap (1/4). We must then also generate the old sequences (p_3 and p_4). The proposal ratio is $(p_3p_4)/(p_1p_2)$.

Update 4

This update is essentially tree bisection-reconnection (TBR). Select an internal edge e at random $(1/(n - 3))$. Call the two internal nodes joined by this edge Y and Z. The other two edges adjacent to Y have x_1 and x_2 reversals while the other two edges adjacent to Z have x_3 and x_4 reversals. Remove edge e and nodes Y and Z from the tree, combining the two other edges into single edges for each Y and Z to form two new edges with $x_1 + x_2$ and $x_3 + x_4$ reversals, respectively. There are now two disconnected unrooted trees with n_1 and n_2 nodes. Pick a random edge e_5 on the first tree $(1/(2n_1 - 3))$ and a random edge e_6 on the second tree $(1/(2n_2 - 3))$. The number of reversals on these edges are x_5 and x_6. Pick a random location on each of these selected edges $(1/(x_5 + 1)$ and $1/(x_6 + 1))$, put new nodes at these locations, and connect them with a new edge. Pick a random direction (1/2) to generate new

reversals on this edge (p_1). The probability of generating the reversals in the other direction is (p_2).

To propose the original state, we must select the new edge to delete ($1/(n - 3)$), select the right edge from each subtree ($1/(2n_1 - 3)$ and $1/(2n_2 - 3)$), select the right locations on each of these edges ($1/(x_1 + x_2 + 1)$ and $1/(x_3 + x_4 + 1)$), pick a direction (1/2) to generate the original sequence (p_3), and compute the probability of generating the sequence in the other direction (p_4). The proposal ratio is $(p_3 + p_4)(x_5 + 1)(x_6 + 1)/((p_1 + p_2)(x_1 + x_2 + 1)(x_3 + x_4 + 1))$.

Update 5

Update 5 begins like Update 3 with the selection at random of an internal edge ($1/(n - 3)$) and one additional edge from each end point (1/4). Label the middle edge selected first e_2 and label one of the other edges e_1 and the other e_3 with each possibility equally likely (1/2). Edge e_i has x_i reversals for $i = 1, 2, 3$. Call the node adjacent to both edges e_1 and e_2 node Y, let e_4 be the other edge adjacent to node Y, and label the other node adjacent to e_4 node A. Pick a random location on e_3 ($1/(x_3 + 1)$) and move node Y to that location. Edges e_1 and e_2 are joined to become a common edge e_1' while edge e_3 is split into two edges e_2' and e_3'. Node Y now has a new permutation determined by its new location. Generate a new sequence of reversals on e_4 with the new permutation at Y the source and the permutation at A the target (p_1).

The reverse proposal must select edge e_2' as the initial internal edge ($1/(n - 3)$) and then select edges e_1' in the role as e_3 and e_3' in the role as e_1 (1/8). The same node Y is then moved to its initial location ($1/(x_1 + x_2 + 1)$) and the original sequence of reversals on edge e_4 is generated from node Y to node A (p_2). The proposal ratio is $p_2(x_3 + 1)/(p_1(x_1 + x_2 + 1))$.

The variant **Update 5x** is identical except that the target is a randomly selected location on edge e_4 ($1/(x_4 + 1)$) rather than at node A. The reversal sequence from the selected location to node A remains the same. If the new reversal sequence length is x_4', the reverse proposal must select the same location ($1/(x_4' + 1)$). The proposal ratio is $p_2(x_3 + 1)(x_4' + 1)/(p_1(x_1 + x_2 + 1)(x_4 + 1))$.

Update 6

Pick two distinct leaves, A and B, of the tree uniformly at random ($2/(n(n - 1))$). There are k internal nodes on the path between these two leaves. Pick one of these internal nodes ($1/k$) at random and call it Z. Two edges adjacent to Z are on the path; the other edge e is not on the path and connects Z to node C. Let y be the number of reversals on the path from A to B. This path also has $k - 1$ other internal nodes so that there are $y + k$ possible locations on the path to place node Z, including its current location. Pick one of these locations at random ($1/(y + k)$) and move node Z to this location. Node Z now has a new permutation. Generate a new sequence of reversals for edge e using node Z as the source and node C as the target (p_1).

The reverse proposal picks the same two leaves ($2/(n(n - 1))$), the same node Z ($1/k$), and the original location ($1/(y + k)$). We generate the original reversal sequence on edge e (p_2). The proposal ratio is p_2/p_1.

The variant **Update 6x** is identical except that we only regenerate part of the reversals on edge e. If edge e begins with x reversals and the proposed edge has x' reversals, the proposal ratio would be $p_2(x + 1)/(p_1(x' + 1))$.

BADGER cycles through updates 1, 1x, 2, 2x, 3, 4, 5, 5x, 6, and 6x in order. Updates 3, 4, 5, 5x, 6, and 6x can change the tree topology. In each of these cases, if we are using a restricted prior, we never propose a tree with prior probability zero. The proposal ratios are adjusted accordingly.

Literature Cited

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

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

Boore, J., T. Collins, D. Stanton, L. Daehler, and W. Brown. 1995. Deducing arthropod phylogeny from mitochondrial DNA rearrangments. Nature **376**:163–165.

Boore, J. L., and W. M. Brown. 1998. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr. Opin. Genet. Dev. **8**:668–674.

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

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

De Rosa, R. 2001. Molecular data indicate the protostome affinity of brachiopods. Syst. Biol. **50**:848–859.

Felsenstein, J. 1978. The number of evolutionary trees. Syst. Zool. **27**:27–33.

Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat. Rev. Genet. **4**:275–284.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754–755.

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

Kaplan, H., R. Shamir, and R. Tarjan. 1999. Faster and simpler algorithm for sorting signed permutations by reversals. SIAM J. Comput. **29**:880–892.

Larget, B., and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. and Evol. **16**:750–759.

Larget, B., D. L. Simon, and J. B. Kadane. 2002. Bayesian phylogenetic inference from animal mitochondrial genome arrangements (with discussion). J. R. Stat. Soc. Ser. B **64**:681–693.

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

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

Moret, B. M. E., A. C. Siepel, J. Tang, and T. Liu. 2002. 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.

Nielsen, C. 2001. *Animal Evolution: Interrelationships of the Living Phyla,* second edition. Oxford University Press, New York.

Pevzner, P. 2000. *Computational Molecular Biology—An Algorithmic Approach*, chapter 10. The MIT Press, Cambridge, Mars.

Sankoff, D., and M. Blanchette. 1998. Multiple genome rearrangement and breakpoint phylogeny. J. Comput. Biol. **5**:555–570.

Smith, M. J., A. Arndt, S. Gorski, and E. Fajber. 1993. The phylogeny of echinoderm classes based on mitochondrial gene arrangements. J. Mol. Evol. **36**:545–554.

Tang, J., and B. Moret. 2003. Scaling up accurate phylogenetic reconstruction from gene-order data. Bioinformatics **19**(**Suppl 1**):i305–i312.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14**:717–724.

York, T., R. Durrett, and R. Nielsen. 2002. Bayesian estimation of the number of inversions in the history of two chromosomes. J. Comput. Biol. **9**:805–818.

Nick Goldman, Associate Editor

Accepted October 26, 2004