BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–10

Bayesian Gene/Species Tree Reconciliation and Orthology Analysis Using MCMC

Lars Arvestad, Ann-Charlotte Berglund, Jens Lagergren and Bengt Sennblad

[1]SBC and Dept. of Numerical Analysis and Computing Science, KTH, SE-100 44, Stockholm, Sweden and [2]SBC and Center for Genomics and Bioinformatics, Karolinska Institutet, SE-171 77, Stockholm, Sweden

ABSTRACT

Comparative genomics in general and orthology analysis in particular are becoming increasingly important parts of gene function prediction. Although not yet being in its final form, our tool has the capacity to perform practical orthology analysis, based on Fitch's original definition, and more generally for reconciling pairs of gene and species trees.

We introduce a probabilistic gene evolution model based on a birth-death process in which a gene tree evolves "inside" a species tree. Our gene evolution model is biologically sound (Nei et al., 1997) and intuitively attractive. We develop a Bayesian analysis based on MCMC which facilitates approximation of an *a posteriori* distribution for reconciliations. That is, we can find the most probable reconciliations and estimate the probability of any reconciliation, given the observed gene tree. This also gives a way to estimate the probability that a pair of genes are orthologs. The main algorithmic contribution presented here consists of an algorithm for computing the likelihood of a given reconciliation. To the best of our knowledge, this is the first successful introduction of this type of probabilistic methods, which flourish in phylogeny analysis, into reconciliation and orthology analysis.

The MCMC algorithm has been implemented and tests show that it performs very well on synthetic as well as biological data. Using standard correspondences, our results carry over to allele trees as well as biogeography.

Contact: *E-mail*: {lottab,jensl}@nada.kth.se,
{bengt.sennblad,lars.arvestad}@sbc.su.se

INTRODUCTION

Orthology analysis provides the most fundamental correspondence between genes in different genomes. It is such intergenomic correspondences that provides comparative genomics with the power to translate information from one organism to another, e.g. from model organisms to human. Function prediction based on orthology is ubiquitously used by biologists. The concept of reconciliation is fundamental to orthology analysis. Together with a gene tree a reconciliation explains the evolution of a gene family in relation to a species tree.

We provide tools with the capacity to perform practical orthology analysis, based on Fitch's original definition of orthology (Fitch, 1970), and more generally for reconciliation of a given gene tree with a given species tree. The tools employ Bayesian inference and rest on a general and sound mathematical framework; our particular instance of this framework is mathematically non-trivial. The two most fundamental questions that our approach allows us to answer computationally are: (1) how many duplications and losses have occurred in the considered gene family? and (2) which genes are orthologs? In general, Bayesian analysis surpasses Maximum Likelihood by providing alternative solutions and the *a posteriori* distribution which yields the significance of the solutions.

There are a number of biological mechanisms with the capacity of causing a gene tree to disagree with a species tree. Among the examples of such mechanisms are: gene duplication, gene loss, and lateral gene transfers. Here the focus is on gene duplication and gene loss. In the probabilistic model we adopt, the gene tree evolves "inside" the species tree according to a birth-death process, where births model gene duplications and deaths model gene losses. Previously, in phylogeny analysis based on Maximum Likelihood as well as Bayesian statistics, birth-death processes have been used as *a priori* distributions for species trees (Rannala and Yang, 1996; Huelsenbeck et al., 2001). Nei et al. (1997) consider molecular data for the Major Histocompatibility Complex genes as well as the Immunoglobin genes, and conclude that the evolutionary patterns are in agreement with a birth-death process. There exists no good established knowledge of how common duplications and losses are, or how their frequencies relate. Nevertheless, such parameters are needed when trying to detect these types of genomic events. We evade this problem by applying a Bayesian approach that allows parameters to be specified by *a priori* distributions rather than exact values.

To the best of our knowledge, the problem studied here has previously only been studied with respect to deterministic parsimony models, see for instance (Goodman et al., 1979; Guigó et al., 1996; Koonin et al., 1998; Hallett and Lagergren, 2000a,b). Our probabilistic gene evolution model and our likelihood computation algorithm, enables us to apply Bayesian analysis, but also paves the way for Maximum Likelihood methods. It is clear that Maximum Likelihood methods have had an enormous impact on phylogeny. Also, Bayesian analysis performed using Markov Chain Monte Carlo (MCMC) techniques has with success previously been applied to phylogeny (Huelsenbeck et al., 2001). There is good reason to believe that probabilistic methods will, in a similar fashion, play a very significant role in reconciliation and orthology analysis. Recent efforts where bootstrapping were applied to parsimony methods for orthology analysis (Storm and Sonnhammer, 2002; Zmasek and Eddy, 2002) corroborates the importance of expressing uncertainty in orthology analysis. As always, uncertainty is best expressed in terms of probabilities.

The Bayesian analysis performed using MCMC techniques can be formulated as follows: Given a gene tree G and a species tree S, where a labeling of the leaves of G gives the association of genes to species, compute the *a posteriori* distribution on the set of reconciliations of G and S. Given a procedure for the likelihood calculation, applying the MCMC techniques is straightforward from a theoretical and algorithmic point of view, although in practice it requires craftsmanship and ingenuity. In our case, the likelihood computations are algorithmically intricate and the main theoretical contribution of this paper is the solution of this problem. *Notice that the MCMC framework together with the capacity to compute the likelihood and to sample from a, so called, proposal distribution gives our MCMC algorithm that estimates the* a posteriori *distribution of reconciliations.* The description of the proposal distribution is omitted from the present account.

The work presented here is a first step in a program that has the potential to offer tools for simultaneous analysis of multiple gene families. One computational problem of this type is that of, given sequence data for gene families, find a species tree and gene trees. The “goodness” of a solution should be based on how well the gene trees can be reconciled with the species tree as well as the likelihood of the sequence evolution induced by these reconciliations. The next step in this program is to introduce sequence evolution in the problem considered here. Our algorithms have been implemented and the experimental results are very positive. Already without sequence evolution our tool performs very well.

The rest of the paper is organized as follows. First, some (mostly standard) definitions are introduced, followed by a presentation of our probabilistic gene tree evolution model. MCMC techniques are then described briefly (a more detailed description can be found in Gilks et al. (1996), where also standard terms are defined) together with a brief review of the general framework as well as how it is instantiated in our case. Then follows a section where a recursive algorithm for computing the likelihood is given. Finally, experimental results are presented, followed by a discussion.

DEFINITIONS AND NOTATION

A *directed tree* T consists of a set of *vertices* $V(T)$ and a set of *arcs* $A(T)$. The set of *leaves* of T is denoted $L(T)$. The subforest of a directed tree T, induced by a subset U of $V(T)$ is the forest $T \backslash (V(T) \backslash U)$. In a *rooted* tree T all arcs are directed away from the root and the root is denoted $r(T)$. Such a tree is *binary* if each non-leaf has out degree two. For a directed rooted tree T and $u \in V(T)$, the subtree rooted at u, denoted T_u, is the subtree of T induced by all vertices reachable by directed paths from u in T. Moreover, for $\langle u, v \rangle \in A(T)$, the arc subtree of T for $\langle u, v \rangle$ is $T_v \cup \{\langle u, v \rangle\}$. A *species tree* S is a rooted directed arc-weighted binary tree S with weight function $w_S : A(S) \rightarrow R^+$. A gene tree will only be given w.r.t. a species tree S. A *gene tree* is a rooted directed binary tree given together with a leaf labeling function $\sigma : L(G) \rightarrow L(S)$. Intuitively, leaves in a gene tree G represent genes, leaves in a species tree S represent species, and the gene $l \in L(G)$ belongs to the genome of the species $\sigma(l)$. An *isomorphism* between two directed trees T and T' is a bijection $f : V(T) \rightarrow V(T')$ such that, $\langle x, y \rangle \in A(T) \iff \langle f(x), f(y) \rangle \in A(T')$ (i.e. there is an isomorphism between two trees if and only if one can be obtained by renaming the vertices of the other).

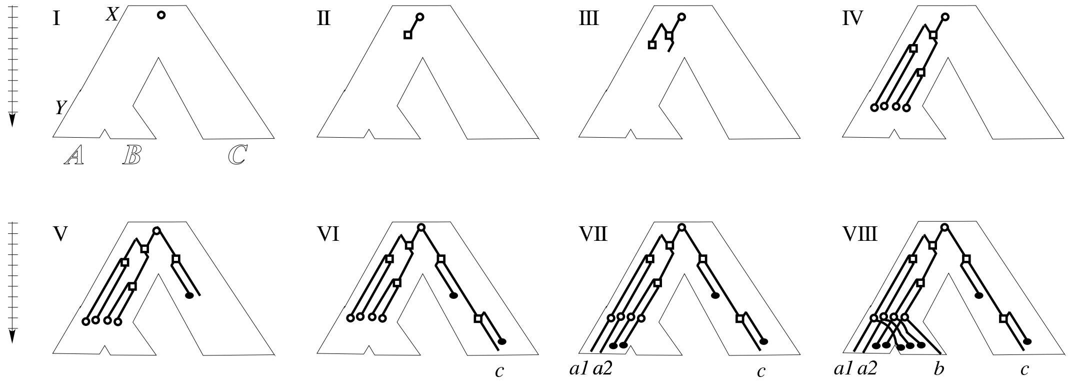

Fig. 1. Example of how a gene tree evolves inside a species tree. In (I), there is a species tree S (the same species tree as in Figure 2). The root node of S is X, the parent of A and B is Y, and there is a single gene in X. While the processes inside $\langle X, Y\rangle$ and $\langle X, C\rangle$ occur "simultaneously", they are independent and we will describe them separately. In (II), the birth-death process gives rise to a duplication, represented by a square vertex, from which two lineages start to evolve. In (III), another duplication occurs in the leftmost of these two lineages, followed by yet another duplication shown in (IV). In (V), we see a duplication followed by a gene loss of the leftmost lineage created in the duplication (losses are represented by filled circles). In (VI), another duplication occurs followed by the loss of the right lineage. In (VII), no duplication occurs, but the two rightmost lineages are lost. Finally, in (VIII), no duplication occurs, but the three leftmost lineages are lost. The resulting (pruned) gene tree is the tree in (II) of Figure 2. The reconciliation induced by the evolution is the one in (III) of Figure 2.

THE PROBABILISTIC GENE EVOLUTION MODEL

In this section, we describe the gene evolution model and introduce reconciliations. We will focus on how, given the parameters of the birth-death process, a gene tree evolves "inside" a given species tree S and how this induces a reconciliation. However, the *a priori* distributions for the parameters of the birth-death process are also a part of the model and we start by specifying those. Uniform distributions on intervals of bounded size are used as *a priori* distributions for the birth rate λ and death rate μ.

We will in the rest of this paper consider a fixed species tree S. The *gene evolution model* is described as follows. Again, the gene tree evolves "inside" the species tree according to a birth-death process. Over arcs of the species tree this is modeled by a standard birth-death process (Kendall, 1948; Nee et al., 1994) with birth rate λ and death rate μ. When the process reaches the end of an arc, i.e. a species tree vertex x, it is split into two identical copies. One of the processes evolves down the left outgoing arc of x and the other evolves down the right outgoing arc of x; moreover, this evolution continues recursively down towards the leaves of S where it stops. After the recursive process has stopped, the gene tree is pruned by removing vertices that have no descendants in the leaves of the species tree and finally by shortcutting vertices of degree two (i.e., removing the vertex and connecting its neighbors in the natural way). A leaf l of the resulting gene tree is labeled, in the natural way, i.e. with the unique leaf of S to which l reached in the evolution process. This ends the description of how the gene evolution model generates gene trees. In Figure 1, an example is provided of how a gene tree evolves inside a species tree. The resulting gene tree can be found in (II) of Figure 2.

To facilitate the introduction of reconciliations, some additional definitions will now be made. As mentioned, the gene evolution process also induce a reconciliation. This is explained after the definition. Let T be a rooted directed tree. For $u \in V(T)$, the *descendants* of u in T are the vertices of T_u (this means that u is a descendant of u). That v is a descendant of u in T is denoted $v \leq_T u$. A set $\mathcal{A} \subseteq V(T)$ is a $\leq_T$-antichain if for each pair $u, v \in \mathcal{A}$ it holds that $u \not\leq_T v$ and $v \not\leq_T u$, i.e., there are no two *different* members u and v of $\mathcal{A}$ such that v is a descendant of u. If T is the tree in (I) of Figure 2, then the $\leq_T$-antichains are: $\emptyset, \{X\}, \{Y\}, \{A\}, \{B\}, \{C\}, \{Y, C\}, \{A, B\}$ $\{A, C\}, \{B, C\}$, and $\{A, B, C\}$. Notice that, although the example provids antichains in a species tree, we will only use antichains in gene trees.

A *reconciliation* of a species tree S and a gene tree

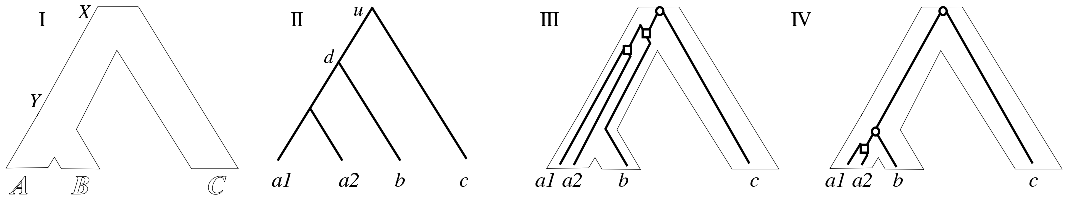

Fig. 2. Example of: (I) The species tree from panel (I) in Figure 1. (II) The pruned gene tree from (VIII) in Figure 1. The root is u. The parent of b is called d. (III) A possible reconciliation of the species tree and the gene tree (where $\gamma(X) = \{u\}$, $\gamma(Y) = \emptyset$, $\gamma(A) = \{a1, a2\}$, $\gamma(B) = \{b\}$, and $\gamma(C) = \{c\}$). One speciation is associated to the root of the species tree. Circles represent vertices of the gene tree that are speciations. (IV) The most parsimonious reconciliation of the species tree and the gene tree (i.e., the one explaining the disagreement between the gene and species tree using a minimum number of duplications, where $\gamma(X) = \{u\}$, $\gamma(Y) = \{d\}$, $\gamma(A) = \{a1, a2\}$, $\gamma(B) = \{b\}$, and $\gamma(C) = \{c\}$). Here also d is a speciation and it is also associated with Y.

G, with leaf labeling σ, is a function $\gamma : V(S) \to 2^{V(G)}$ such that:

1. $l \in \gamma(\sigma(l))$ for each $l \in L(G)$.
2. $\gamma(x)$ is a $\leq_G$-antichain, for any $x \in V(S)$.
3. If $u \leq_G v$, $u \in \gamma(x)$, and $v \in \gamma(y)$, then $x \leq_S y$.

In Figure 2, two examples of reconciliations can be found. The above conditions have the following intuitive explanations. The first condition demands that a gene should be associated with the species in which genome it can be found. The last two conditions demand that the gene tree should evolve “downwards” in the species tree.

The evolution of the gene tree inside the species tree also induces a reconciliation of the two trees as follows: $u \in \gamma(x)$ if and only if u is a “leaf” of the birth-death process when it reaches x, and u at the end of the process has a descendant in a leaf of the species tree below the left as well as the right child of x.

The reconciliation induced by the example of the gene evolution process in Figure 1 can be found in (III) of Figure 2.

In the general version of the process a part of the gene tree can have evolved *previous* to the speciation represented by the root of the gene tree. The description of the general version, as well as how to compute the likelihood in this case, is omitted because of space limitations.

MCMC GIVEN A PROCEDURE FOR LIKELIHOOD

This section contains a brief introduction to the MCMC framework as well as a brief description of how this framwork is applied in the present case. For a more thorough account of MCMC and standard MCMC terminology, we refer to the book by Gilks et al. (1996).

MCMC is a technique that facilitates estimation of the stationary distribution of a Markov Chain. It provides a uniform framework to design transition probabilities of a Markov Chain so that a sought stationary probability distribution is obtained. A random walk is performed in the Markov Chain according to the transition probabilities. In the present case, a state in the Markov chain is a triple (γ, λ, μ) where γ is a reconciliation, λ a birth rate, and μ is a death rate (in order to simplify the description, we describe the Markov chain as if λ and μ were discrete parameters). The sought stationary distribution is the *a posteriori* distribution

$$\begin{aligned}\Pr[\gamma, \lambda, \mu | G_{obs}] &= \frac{\Pr[G_{obs}, \gamma | \lambda, \mu] \Pr[\lambda, \mu]}{\Pr[G_{obs}]} \\ &= \frac{\Pr[G_{obs}, \gamma | \lambda, \mu]}{\sum_\gamma \sum_\lambda \sum_\mu \Pr[G_{obs}, \gamma | \lambda, \mu]},\end{aligned}$$

where the last equality holds, since uniform *a priori* distributions are used for λ, and μ. This distribution assigns to a state (γ, λ, μ) the probability that, in the gene evolution process: γ was the reconciliation, λ was the birth rate, and μ was the death rate, conditioned by the observed gene tree G_{obs}. In the limit, the fraction of visits to a state during the simulation, in relation to the total number of visits, *is* the stationary probability. In practice, frequencies are collected after a period of *burn in* (the time it takes for the chain to “forget” its starting state) and up to an estimated *stopping time* (sufficiently late to make the estimation of the stationary distribution reliable), see Gilks et al. (1996). The gene tree vertices associated with species tree vertices by a reconciliation are speciations; all other gene tree vertices are duplications. According to the original definition by Fitch (1970), two genes are orthologs if and only if their least common ancestor in the gene tree is a speciation. By summing the probability assigned to the triples with a reconciliation according to which gene x and y are orthologs, we can estimate the probability that x and y are orthologs.

In each iteration of an MCMC simulation a new state is proposed according to a specific proposal distribution. We omit the description of the proposal distribution that we use. The new state is accepted, i.e. becomes the current state, or rejected, in which case no change is made of the current state, according to an acceptance distribution. In the present case, a symmetric proposal distribution is used, which means that the algorithm proposed here is an instance of the Metropolis method (Gilks et al., 1996). Consequently, when the present state is (γ, λ, μ), the acceptance probability for a proposed state $(\gamma', \lambda', \mu')$ is $\alpha_{ij} = \frac{\Pr[G_{obs}, \gamma|\lambda, \mu]}{\Pr[G_{obs}, \gamma'|\lambda', \mu']}$.

Again, notice that the MCMC framework together with the capacity to compute the likelihood, $\Pr[G_{obs}, \gamma|\lambda, \mu]$*, and to sample from the proposal distribution gives our MCMC algorithm that estimates the* a posteriori *distribution of reconciliations.*

COMPUTING THE LIKELIHOOD

In this section, we briefly describe our main algorithmic contribution; namely, the algorithm that computes the likelihood

$$\Pr[\gamma, G|\lambda, \mu], \tag{1}$$

i.e., the likelihood of a reconciliation γ and a gene tree G, or equivalently the probability that the gene evolution model produces this pair for a birth rate λ and a death rate μ (recall that we are considering a fixed species tree S). We concentrate on an operational description, i.e., what is computed, and exclude motivations and proofs why these computations give the likelihood. The algorithm is recursive and runs in time $O(|V(S)| + |V(G)|^2)$. Its recursive structure follows that of the gene evolution model. We will describe the subproblems considered in the algorithm and how they are divided into smaller subproblems, but no pseudocode will be given. An example of how a subproblem can be divided into smaller subproblems will also be given.

To aid the recursive decomposition of G, we will, without loss of generality, assume that no arc of G stretches over more than one arc of S, w.r.t. γ; we formally express this as follows, if $\langle u, v\rangle \in A(G)$ and $u \in \gamma(x)$, then there is an arc $\langle x, y\rangle \in A(S)$ such that $v \in \gamma(y)$. In (III) of Figure 2 a pair without this property can be found and in (I) of Figure 3 a for our purposes equivalent pair with this property can be found.

Throughout this section, we will consider a fixed gene tree G, a fixed birth rate λ, and a fixed death rate μ. During the computation a table e_V is filled in. After the computation the sought likelihood, i.e., the value of (1), is given by $e_V(r(S), r(G))$. For the case where $\gamma(r(S))$ is a set, some extra steps have to be performed, but we will not describe those.

After some additional definitions, which are illustrated in Figure 3, have been made, formulae will be given which show how to compute e_V. If $u \in V(T)$ and $U \subseteq V(T)$, then $T_{u,U}$ denotes the subtree of T induced by $\{v : \exists u' \in U, u' \leq_T v \leq_T u\}$. We call this a *sliced subtree*. Let u be a vertex of G such that $u \in \gamma(x)$ and let $S^{x,y}$ be the arc subtree of S for $\langle x, y\rangle$. The tree $G_u^{x,y}$ is the subtree of G_u induced by $\{v \in V(G_u) : \exists z \in V(S^{x,y}), v \in \gamma(z)\}$. That is, the part of G_u that has evolved in the arc subtree of S for $\langle x, y\rangle$. Let $\gamma_u : V(S_x) \to 2^{V(G_u)}$ be defined by $\gamma_u(z) = \gamma(z) \cap V(G_u)$, for any $z \in V(S_x)$. Similarly, let $\gamma_u^{x,y} : V(S^{x,y}) \to 2^{V(G_u^{x,y})}$ be defined by $\gamma_u^{x,y}(z) = \gamma(z) \cap V(G_u^{x,y})$.

Define $e_V(x, u)$ as the probability that G_u and γ_u have evolved from u starting at x in S_x. Similarly, define $e_A(x, y, u)$ as the probability that $G_u^{x,y}$ and $\gamma_u^{x,y}$ has evolved from u starting at x in $S^{x,y}$. Assume that x has children y and y' in S. Obviously $e_V(x, u) = e_A(x, y, u)e_A(x, y', u)$; moreover, if v is a leaf of G and z a leaf of S such that $v \in \gamma(z)$, then $e_V(z, v) = 1$. Without loss of generality, assume that the set of descendants of u in $\gamma(y)$ is $[c]$ (for any integer c, we use $[c]$ to denote the set $\{1, \ldots, c\}$). We now give formulae that can be used to compute $e_A(x, y, u)$ (and analogously $e_A(x, y', u)$); afterwards we explain how to compute the factors occurring in the formula. For $c > 0$,

$$e_A(x, y, u) = AB, \tag{2}$$

where A equals

$$f_1(G_{u,\gamma(y)})f_2(G_{u,\gamma(y)}, [G_1, \gamma_1]_{\cong}, \ldots, [G_c, \gamma_c]_{\cong})$$

and B equals

$$\frac{P(t)(1-u_t)u_t^{c-1}}{(1-D_y u_t)^{c+1}} \prod_{i=1}^{c} e_V(y, i).$$

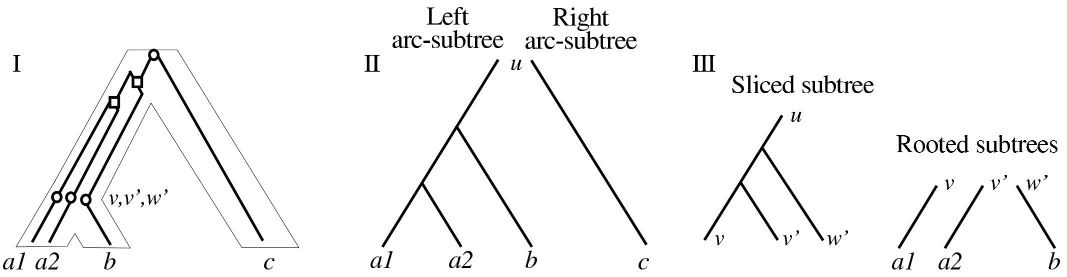

Fig. 3. (I) The reconciliation from a (II) Figure 2 with vertices added to the gene tree. The additional vertices are, from left to right, v, v', and w. (The names in the species trees are as in previous examples). The reconciliation is $\gamma(X) = \{u\}$, $\gamma(Y) = \{v, v', w'\}$, $\gamma(A) = \{a1, a2\}$, $\gamma(B) = \{b\}$, and $\gamma(C) = \{c\}$. (II) The trees $G_u^{X,Y}$ and and $G_u^{X,C}$, where G is the gene tree and S is the species tree of Figure 2 which are related by the reconciliation γ in (I). (III) $G_{u,\gamma(Y)}$ and $G_v, G_{v'}, G_{w'}$ for the same gene tree, species tree, and reconciliation as above.

The probability that a lineage starting in x does not reach any leaf in the species tree is denoted D_x and can be computed using the equality $D_x = e_A(x, y, u)e_A(x, y', u)$, where for y as well as y' it holds that $c = 0$ (D_y has an analogous meaning and can be computed analogously). For $c = 0$,

$$e_A(x, y, u) = 1 - P(t) + \frac{P(t)(1 - u_t)D_y}{1 - u_t D_y}. \quad (3)$$

Let t be the time for the arc $\langle x, y\rangle$, i.e., $t = w_S(x, y)$. The probability, $\Pr_1\{c, t\}$, of having c surviving lineages at time $t_0{+}t$ from the birth-death process, starting at time t_0, can be expressed by two functions of time, u_t and $P(t)$. Closed expressions for all three of these functions can be found in Nee et al. (1994). These results have previously been used to express the probability of generating a certain tree (Yang and Rannala, 1997). Using the same techniques we can express the probability that "the birth-death process starting with a vertex u produces a rooted directed tree T such that for a *certain* set of c leaves W it holds $T_{u,W} = G_{u,\gamma(y)}$" conditioned by that the "process yielded $c{+}d$ leaves in T", where $c, d \geq 0$.

For a tree T with leaves $[c]$ and labels $m_1, \ldots, m_c$, the factor $f_2(T, m_1, \ldots, m_c)$ is the number of different leaf labelings $\mathcal{L} : [c] \to \{m_1, \ldots, m_c\}$ such that there is an automorphism $f : V(T) \to V(T)$ (i.e., an isomorphism from T to itself) satisfying $m_i = \mathcal{L}(f(i))$. Also f_2 can be computed within the available time using a slight modification of the standard tree isomorphism algorithm. In the formula above the labels are isomorphism classes introduced below.

An isomorphism f between G_i and G_j is said to respect γ_i and γ_j if and only if, for any $u \in V(G_i)$ and $x \in V(S)$, it holds that $u \in \gamma_i(x) \Leftrightarrow f(u) \in \gamma_j(x)$. For any $i \in [c]$, let $[G_i, \gamma_i]_{\cong}$ denote the set of ordered pairs $\langle G_j, \gamma_j\rangle$ such that there is an isomorphism f between G_i and G_j that respects γ_i and γ_j (i.e., $[G_i, \gamma_i]_{\cong}$ is an isomorphism class). Whether $[G_i, \gamma_i]_{\cong}$ equals $[G_j, \gamma_j]_{\cong}$ can be computed within the available time using a slight modification of the standard tree isomorphism algorithm.

This completes the description of how to compute the likelihood, i.e., (2) and (3).

EXPERIMENTAL RESULTS

We have performed experiments to verify that the probabilistic gene duplication model is useful in the context of reconciliation and orthology analysis. That is, when combined with sequence evolution, it will contribute information not only by providing a translation of time in the species tree into time in the gene tree but also by its *a posteriori* distribution. The experiments clearly allow us to draw this conclusion.

To enable experiments a generator was implemented, which for a given species tree generates pairs consisting of a gene tree and a reconciliation according to the gene evolution model. We used "The 90%-test", i.e., statistics concerning the fraction of the total number of experiments in which the generated reconciliation was among best 90% of the reconciliations in the *a posteriori* distribution (ties were broken using a probabilistically sound method). For a sufficiently large number of experiments, the better the *a posteriori* distribution has been estimated the closer the *expected value* of this fraction will be to 0.9. Another important statistics is the fraction of experiments in which the generated reconciliations differ from the most

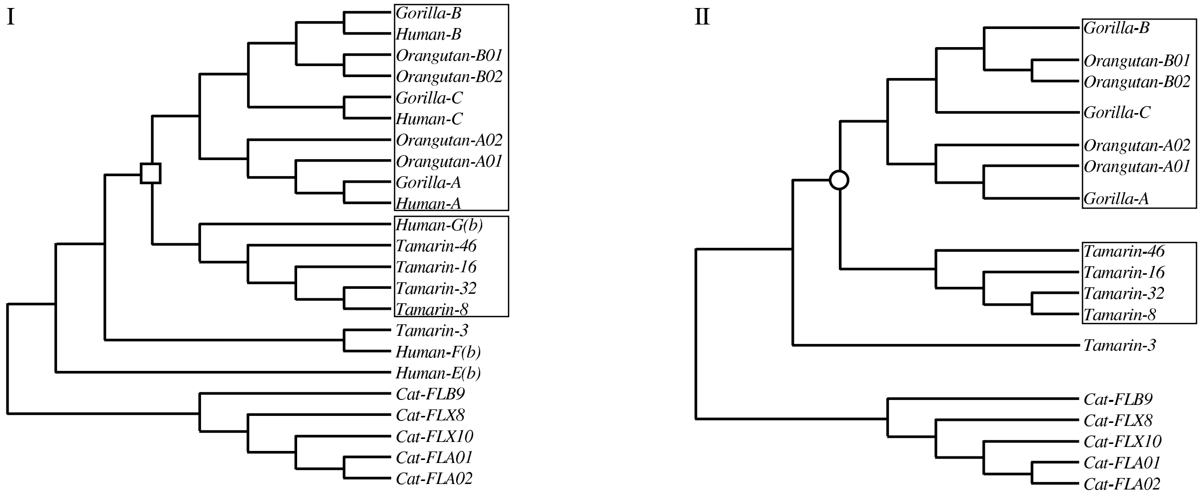

Fig. 4. The MHC class I gene trees for primate sequences extracted from Nei et al. (1997); the MHC class I genes for cat is included as an outgroup. The two homolog groups of interest are boxed and the status of the least common ancestor, v, of these two groups as interpreted by parsimony reconciliation is indicated. (I) The gene tree including all sequences from Nei et al. (1997). Parsimony reconciliation correctly identifies v as a duplication (indicated by a square). (II) The tree from (I), but with all human sequences removed, simulating that the human genome was not sampled. Parsimony reconciliation now erroneously identifies v as a speciation (indicated by a circle).

Table 1. Results from the 90% test.

λ	μ	$\Pr(\gamma_{\text{true}} \neq \gamma^*)$	Diagnostic
0.06	0.05	0.01	0.92
0.15	0.18	0.14	0.88

parsimonious. The species trees used here have 10 taxa, and branch lengths between 0.03 and 13.11. Various birth and death rates where used, which gave up to 42 genes (i.e., gene tree leaves). The experiments can be divided into two groups, one where the generated reconciliations almost always is the most parsimonious ($\gamma_{\text{true}} = \gamma^*$) and one where this quite often is not the case. For some values $\Pr[\gamma_{\text{true}} \neq \gamma^*]$ is as high as 0.4. One result from each group is described in Table 1. In both cases we are close to the expected result for the 90% test.

The histocompatibility complex (MHC) multigene family: The phylogenetic tree of MHC class I genes from Gorilla, Orangutan, Tamarin and Cat was used. This is a subtree of the gene tree described in Nei et al. (1997). For illustrative purposes we have removed the Human MHC class I genes from this subtree. When included, these reveal that pairs of genes that in the subtree can be paralogs or orthologs, are in fact paralogs (cf. Figure 4). Thus, it is interesting to see if our method, in contrast to the parsimony method, indicates that they may be paralogs. We used a species tree with divergence time estimates from Arnason and Janke (2002). The MCMC ran for 5000000 iterations with samples taken at every 1000th iteration. The Markov chain converged after approximately 500000 iterations, and the first 500 samples were discarded. The orthology probability for the pairs of genes is 0.89, allowing a significant fraction for the correct answer.

The 60s ribosomal domain family: As a case study of including information from sequence data with our method as well as considering alternative gene trees, the following test was conducted. The object was to estimate the posterior probability of orthology between two genes without being certain of the gene tree. The idea was to multiply the probability of a reconciliation with the probability of the gene tree it was based on. Although this approach is attractive, particularly with respect to its simplicity, it is unfortunately flawed. Posterior reconciliation probabilities are conditioned on a gene tree and a species tree, but species trees are disregarded in available software for computing

Table 2. Probabilities and bootstrap support values for RLA1 and RLA2 using three different methods.

	RLA1	RLA2
OrthoStrapper	0.35	0.46
RIO	0.12	0.21
Our method	0.88	0.91

gene-tree probabilities. Hence, we cannot view this technique as more than a first approximation of the true likelihood.

Both human and fly have several domains in the 60s ribosomal domain family, and we wanted to estimate the probability that the RLA1 and RLA2 domains from human were orthologous to their counterparts in fly. In a recent investigation (Christian Storm, personal communication) using OrthoStrapper (Storm and Sonnhammer, 2002) and RIO (Zmasek and Eddy, 2002), the orthology support was found low (Table 2), even though a phylogenetic analysis including related domains strongly suggests orthology (Figure 5).

All genes in the 60s ribosomal domain family (PF00428) for a set of five species (Figure 6) were taken from Pfam 7.2 Bateman et al. (2002) yielding 23 genes. MrBayes (Huelsenbeck and Ronquist, 2001) was used for the phylogenetic analysis. From an initial neighbour-joining tree, 10000 iterations of MCMC were run and every tenth tree was sampled. This run resulted in 89 sampled trees, of which the first three had 50% and the first 37 had 90% of the probability mass.

For each of the 89 trees sampled by MrBayes, 10^5 iterations of our MCMC algorithm were run, sampling reconciliations every 100 iterations and with the first 300 samples discarded as burn in. The posterior probabilities of orthology of the domains of interest were then registered and multiplied with the posterior probability of the gene tree. In a last step, the orthology probabilites from each gene tree were summed. The results are summarised in Table 2, and assigns a considerably higher probability for the expected orthologies than previously noted.

DISCUSSION AND FUTURE DIRECTIONS

The experimental results for our method are encouraging. The method shows self-consistent behaviour for synthetic data that are generated with the same model as that used in the analysis. It also produces reasonable results using biological data and clearly provides important additional information for orthology analysis to that of standard parsimony reconciliation.

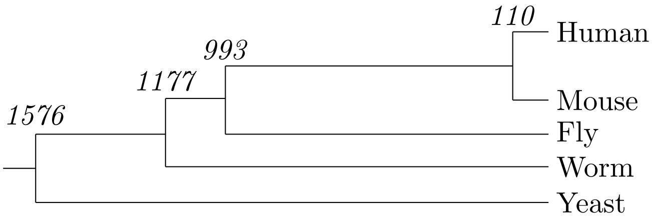

Fig. 6. Species tree for the *60s ribosomal* testcase. Numbers indicate million years ago and are taken from Wang et al. (1999).

In addition, our method performs well in terms of speed: The largest problem instance we have run the algorithm on consists of a species tree with 100 leaves and a gene tree with 228 leaves. 10^6 MCMC iterations, each requiring a likelihood computation (which has quadratic running time), were performed in 1686 CPU seconds on a 1100 MHz processor. We are currently porting the implementation to a cluster, which will allow more extensive experimental investigations.

The obvious next step is to add sequence evolution to the probabilistic model. The input would then be a species tree and a set of sequences representing a gene family, where each sequence is associated to a species.

In this setting, the aim would be to find pairs consisting of a gene tree and a reconciliation, or an *a posteriori* distribution over such pairs. The *a posteriori* probability of a solution should reflect how well the gene tree can be reconciled with the species tree according to our gene evolution model as well as the likelihood of the sequence evolution induced by the gene tree and the reconciliation. Our experimental results clearly show that the gene evolution model will have a significant and correct impact on this *a posteriori* distribution.

Lastly, it is worth noting that, using standard correspondences, our results also carry over to other areas of biology such as allele mapping and biogeography.

ACKNOWLEDGEMENTS

We thank Markus Arborelius for help with the initial implementation and the two anonymous referees for helpful comments. Funding for this project was in part from the Swedish Foundation for Strategic Research.

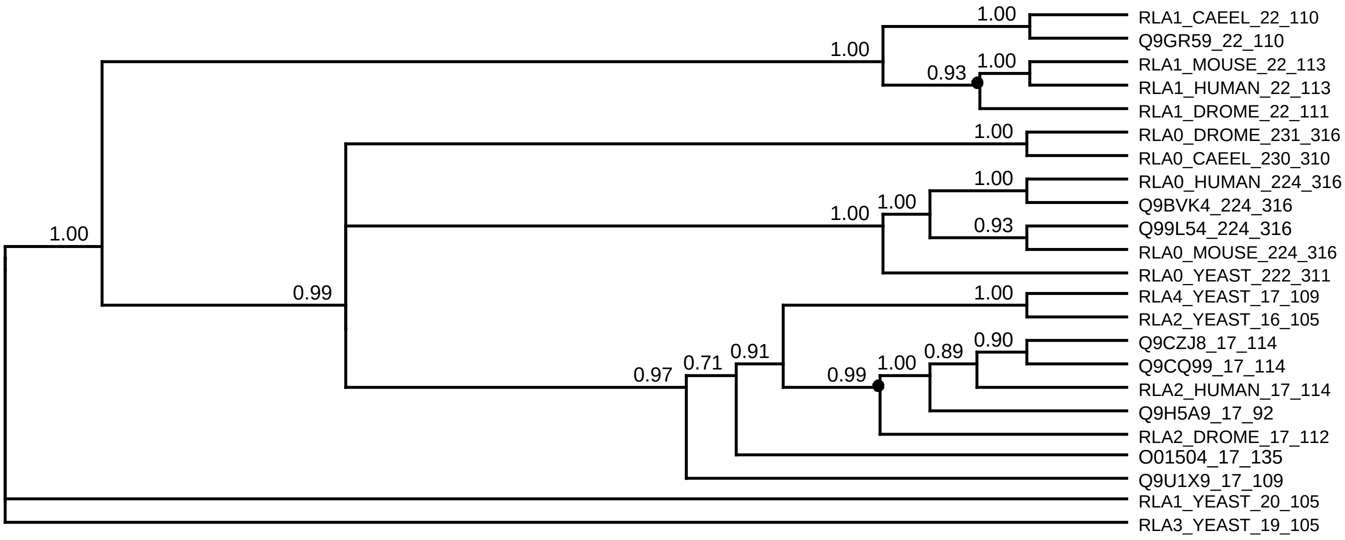

Fig. 5. Consensus tree for the 60s ribosomal protein domain family as computed by MrBayes (Huelsenbeck et al., 2001). The investigated nodes are marked by filled circles. The numbers indicate the posterior probability of branchings beeing correct.

REFERENCES

Arnason, U. and A. Janke (2002). Mitogenomic analyses of eutherian relationships. *Cytogenet. Genome Res. 96*(1-4), 20–32.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer (2002, Jan). The Pfam protein families database. *Nucleic Acids Res. 30*(1), 276–80.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool. 19*(2), 99–113.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Eds.) (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Goodman, M., J. Cselusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda (1979). Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool. 28*, 132–168.

Guigó, R., I. Muchnik, and T. F. Smith (1996). Reconstruction of ancient molecular phylogenies. *Mol. Phylogenet. Evol. 6*(2), 189–213.

Hallett, M. and J. Lagergren (2000a). Hunting for functionally analogous genes. In *FSTTCS00*, pp. 465–476.

Hallett, M. and J. Lagergren (2000b). New algorithms for the duplication-loss model. In 4^{th} *Annual RECOMB '00, Tokyo, Japan*, pp. 146–158.

Huelsenbeck, J. P. and F. Ronquist (2001, Aug). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics 17*(8), 754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback (2001, Dec). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science 294*(5550), 2310–2314.

Kendall, D. G. (1948). On the generalized "birth-and-death" process. *Ann. Math. Stat. 19*, 1–15.

Koonin, E., , R. L. Tatusov, and M. Y. Galperin (1998). Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol. 8*(3), 355–363.

Nee, S., R. M. May, and P. H. Harvey (1994). The reconstructed evolutionary process. *Phil. Trans. R. Soc. Lond. B 344*, 305–311.

Nei, M., X. Gu, and T. Sitnikova (1997, July). Evolution by the birth-and-death process in multigene families of vertebrate immune system. *Proc. Natl. Acad. Sci. USA 94*(15), 7799–7806.

Rannala, B. and Z. Yang (1996, Sep). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol.*

Evol. 43(3), 304–11.

Storm, C. E. and E. L. Sonnhammer (2002, Jan). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics 18*(1), 92–99.

Wang, D. Y., S. Kumar, and S. B. Hedges (1999, Jan). Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B. Biol. Sci. 266*(1415), 163–71.

Yang, Z. and B. Rannala (1997, Jul). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol. 14*(7), 717–24.

Zmasek, C. M. and S. R. Eddy (2002). RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics 3*(14).