Simultaneous Bayesian gene tree reconstruction and reconciliation analysis

Örjan Åkerborg[a], Bengt Sennblad[b], Lars Arvestad[a], and Jens Lagergren[a,1]

[a]School for Computer Science and Communication, Royal Institute of Technology, SE-10044 Stockholm, Sweden; and [b]Stockholm Bioinformatics Center, AlbaNova, Stockholm University, SE-10691 Stockholm, Sweden

Edited by Andrew Roger, Dalhousie University, Halifax, Canada, and accepted by the Editorial Board February 4, 2009 (received for review July 3, 2008)

We present GSR, a probabilistic model integrating gene duplication, sequence evolution, and a relaxed molecular clock for substitution rates, that enables genomewide analysis of gene families. The gene duplication and loss process is a major cause for incongruence between gene and species tree, and deterministic methods have been developed to explain such differences through tree reconciliations. Although probabilistic methods for phylogenetic inference have been around for decades, probabilistic reconciliation methods are far less established. Based on our model, we have implemented a Bayesian analysis tool, PrIME-GSR, for gene tree inference that takes a known species tree into account. Our implementation is sound and we demonstrate its utility for genomewide gene-family analysis by applying it to recently presented yeast data. We validate PrIME-GSR by comparing with previous analyses of these data that take advantage of gene order information. In a case study we apply our method to the ADH gene family and are able to draw biologically relevant conclusions concerning gene duplications creating key yeast phenotypes. On a higher level this shows the biological relevance of our method. The obtained results demonstrate the value of a relaxed molecular clock. Our good performance will extend to species where gene order conservation is insufficient.

comparative genomics | probabilistic modeling | genome evolution | yeast | ADH

Gene trees are fundamental to comparative genomics studies in a multispecies context. Several processes shape gene trees during evolution. It is well known that phenomena such as gene duplication and loss, lateral gene transfer, as well as incomplete allele sorting can result in incongruence between species trees and gene trees. The relative frequencies of these events vary across the tree of life, but it is clear that gene duplication and loss give rise to many incongruences among eukaryote gene trees, and that gene duplication is a major force in creating new genes among these organisms (1, 2). The first methods explaining incongruence between a species tree and a gene tree were based on the parsimony principle; Goodman et al. (3) pioneered the field by introducing the term *reconciliation* for the embedding of a gene tree into a species tree explaining the evolution of the former. They also provided an algorithm for computing the most parsimonious reconciliation (*MPR*). Several variations and formalizations of *MPR* have been provided (4–9).

It is interesting to compare this development of reconciliation methods with that of phylogenetic tree reconstruction methods. Although there has been an intense debate concerning the relative benefits of different methods for reconstructing phylogenetic trees, it is today common to describe the development as a progression starting in 1965 with parsimony methods (10–12), where an important step consisted of Maximum Likelihood (ML) methods (13), and where the most recent contribution is Bayesian methods (14). We believe it is time for reconciliation methods to progress to the probabilistic stage, partly because of inherent problems with *MPR*. Consider for example a very large gene family. It is clear that many gene duplications must have occurred during the evolution of this gene family and, hence, that the gene duplication rate in the family has been high. Knowing this, is it reasonable to believe that the reconciled tree best suited to explain the evolution of the gene family is the one that minimizes the number of duplications? We argue that one should also consider reconciliations other than *MPR* and let information about duplication and loss rates, which can be inferred from data, guide the selection of a reconciliation.

The *gene evolution model*, which was the first probabilistic model for how a gene family evolves with respect to duplications and losses, was presented in ref. 15, where an algorithm for computing the probability of a given reconciliation was also provided. In previous work, for example, refs. 16 and 17, evolution of gene count within a species tree has been modeled without reference to an explicit gene tree. The *gene sequence evolution model*, an integrated probabilistic model for gene duplication, gene loss, and sequence evolution under a molecular clock, was presented in ref. 18, together with a Markov chain Monte Carlo (MCMC) approach for estimating the posterior distribution over gene trees, or gene trees and reconciliations, after having observed sequence data for a gene family.

Several coalescent-based probabilistic methods for dealing with incongruence between a species tree and a gene tree caused by incomplete allele sorting have also been proposed (19–22). These authors use the coalescent model for modeling allele sorting only, but it is justified to ask whether the coalescent also is a good model for gene duplication and loss. However, although the coalescent has a very natural interpretation when modeling incomplete allele sorting, there is no natural interpretation of it when modeling gene duplications giving rise to multiple copies of a gene in individual genomes. Coalescent-based models also lack the concept of gene loss. A gene family affected by gene duplication and loss can, of course, also be affected by allele sorting. This implies that a model including all these events, generalizing our model as well as the coalescent model, may be desirable. From a purely practical point of view the framework we describe here can be applied to such gene families; however, it remains to be investigated how well our model serves as an approximation for the combined model. A method for genome-wide construction of gene trees was obtained in ref. 23 by combining, in an ad hoc manner, a neighbor-joining strategy (12) with a strategy to evaluate duplication frequencies and gene order information. This method was applied to gene family data from yeast, which is a group of taxa where gene order is strongly conserved (24). However, there is no reason to believe that gene order in general will benefit gene tree reconstruction in other groups, e.g., animals or plants (25, 26). Another gene tree reconstruction method that allows for species tree edge-specific as well as gene family-specific nucleotide substitution rates was presented and applied to sequenced fly genomes in ref. 27.

In this study, we present a new probabilistic model, GSR, for gene evolution that unifies sequence evolution models,

Author contributions: Ö.Å., B.S., L.A., and J.L. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.R. is a guest editor invited by the Editorial Board.

[1] To whom correspondence should be addressed. E-mail: jensl@csc.kth.se.

This article contains supporting information online at www.pnas.org/cgi/content/full/0806251106/DCSupplemental.

substitution rate evolution models, and a duplication-loss process that generates gene trees respecting the species tree. An algorithm for reconstructing a gene tree under the model is introduced. Our algorithm has been validated on synthetic data and evaluated on gene sequence data from yeast genomes. By evaluating the performance on a genomewide dataset, we have demonstrated that the algorithm is efficient enough to be applied in large-scale studies.

Model for Gene Tree and Sequence Evolution

The *gene sequence evolution model with iid rates across gene tree edges*, which we denote GSR, is a joint generalization of the model from ref. 18 and models used to obtain a relaxed molecular clock (i.e., substitution rates vary over the tree) (28, 29). GSR integrates the following probabilistic submodels:

1. A duplication-loss model describing gene evolution.
2. A substitution rate model describing rate variation over the gene tree.
3. A sequence evolution model describing substitution events.

Let the species tree S and the gene tree G be planted trees, i.e., trees with a root of degree one, with divergence times associated to their vertices. Because S and its divergence times are considered to be given, they will be omitted from our notation. The planted subtree of G containing u, its parent $p(u)$, and all descendants of u is denoted G^u. Gene tree vertices represent either a speciation or a duplication event; for speciation vertices the divergence time is given by the corresponding species tree vertex, and the divergence times for duplication vertices are given by the duplication-loss process. Divergence times associated with vertices of a tree induce, in the natural way, edge times. The purpose of the substitution rate model is to transform a dated tree, which typically is ultrametric (i.e., each root-to-leaf path has the same length), into a tree consistent with a relaxed molecular clock, thereby providing a biologically realistic prior distribution for *edge lengths*, i.e., the convolution of edge times and substitution rates conventionally used in substitution models. The substitution rate model also turns out to facilitate more efficient and more accurate gene tree reconstruction. Let l, r, and t denote functions associating an edge length, an edge-specific rate, and an edge time, respectively, to each edge of G so that, e.g., $l(u, v)$ is the edge length of edge $\langle u, v\rangle$. In the next three subsections, we briefly describe each of the GSR submodels.

Gene Duplication and Loss. The *gene evolution model* was described by Arvestad et al. in 2003 (15). Gene evolution is explicitly modeled as a gene tree G evolving inside a species tree S with given divergence times. Over any edge $\langle X, Y\rangle$ in the species tree, gene duplications and losses are modeled by a linear birth–death process with duplication rate λ and loss rate μ. The time for the process is given by the difference in divergence times between X and Y. Each gene lineage reaching a speciation vertex X in S splits into two independent processes. The process continues recursively down to the leaves where it stops. All gene lineages that do not reach a leaf in the species tree are pruned, leaving a *reconciled tree* comprising a binary gene tree and a reconciliation explaining how the gene tree has evolved. Computing the probability of a given reconciled tree under the model is nontrivial, but efficient algorithms were given in ref. 18. As will be described below, we will here view the outcome of the gene evolution model as a gene tree with divergence times, rather than as a reconciled gene tree, as in ref. 18; it is, however, clear that the former induce the latter.

Substitution Rates. The *gene sequence evolution model* of Arvestad et al. (18) assumed a molecular clock for substitution rates in the gene tree, i.e., the substitution rate was assumed to be constant over the tree. To allow more biologically realistic scenarios, we relax the molecular clock (28, 30–35). Our choice is the *iid*-Γ model (34, 36), where edge substitution rates are modeled as *independent and identically* Γ*-distributed* variables with mean m and variance ν.

Sequence Evolution. For each edge in the gene tree, we obtain an *edge length* by multiplying its edge time and edge rate. Sequence evolution over the tree can therefore be modeled by using any of the standard substitution models (13) used in phylogenetics. In this study, we have chosen to use the *JTT* amino acid model (37). The relation between lengths, rates, and times over all edges will be denoted by $l = rt$, or conversely $r = l/t$.

MCMC and Computation of Generation Probabilities

Our model has the following joint generation probability for sequence data D and a gene tree G

$$\Pr[D, G|S] = \int \Pr[D|G, l = rt]\, p[r|G] p[G, t|S] dt dr, \qquad [\mathbf{1}]$$

where $p[r|G]$ and $p[G, t|S]$ are prior probability density functions (PDFs) for rates and a gene tree with divergence times, respectively.

The central observation behind our MCMC algorithm for estimating the posterior of the GSR model is that Eq. **1** can be factorized into a likelihood function for the sequence evolution submodel and probabilities for the duplication and rate submodels. By using a Markov chain where the states, as usual in posterior estimation in phylogeny, consist of trees with edge lengths and additional parameters, the joint generation probability computations can then be performed by (*i*) a standard dynamic programming (DP) algorithm for sequence evolution (13) and (*ii*) a DP algorithm, described below, that estimates the duplication-loss and rate part of the likelihood function using a discretization methodology. Our framework also allows standard proposal functions to be used.

To simplify notation, let $\theta = (\lambda, \mu, m, \nu)$ denote the parameters of the gene evolution model and the substitution rate model. To determine acceptance probabilities in our Markov chain, probabilities $\Pr[G, l, \theta|D]$ need to be computed. These can be rewritten as

$$\Pr[G, l, \theta|D] = \frac{\Pr[D|G, l] \Pr[G, l|\theta] \Pr[\theta]}{\Pr[D]},$$

where the parameters θ are assigned independent priors (see Supporting Information (SI) *Appendix* for further information). As usual in MCMC estimation of posterior probabilities, the denominators will cancel in any ratio between two such probabilities. Moreover, the factor $\Pr[D|G, l]$ can be computed by using the standard DP algorithm introduced by Felsenstein (13). We will now provide the last component of our MCMC algorithm for estimating the posterior of the GSR model, by explaining how to estimate $\Pr[G, l|\theta]$. By definition

$$\Pr[G, l|\theta] = \int_t p[G, l, t|\theta] dt = \int_t p[(r = l/t)|m, \nu] p[G, t|\lambda, \mu] dt.$$

Consider a discretization S' of the species tree S, where the edges of S are augmented with additional vertices as follows. On every path from a leaf ℓ to the root there is a vertex at time $t_i = i\delta$ for $i \in \{0, .., d\}$, where δ is the length of the discretized interval. We will typically denote vertices of the gene tree by u, v, w and vertices of S' by x, y, z. When appropriate, vertices of S will be denoted X, Y, Z. An illustration with $d = 9$ is given in Fig. 1.

Let $\mathbf{t}$ be the set of all possible edge time vectors t corresponding to these discretized divergence times. This enables us to estimate $\Pr[G, l|\theta]$ by the following sum:

$$\sum_{t \in \mathbf{t}} \Pr[(r = l/t)|m, \nu] \Pr[G, t|\lambda, \mu]. \qquad [\mathbf{2}]$$

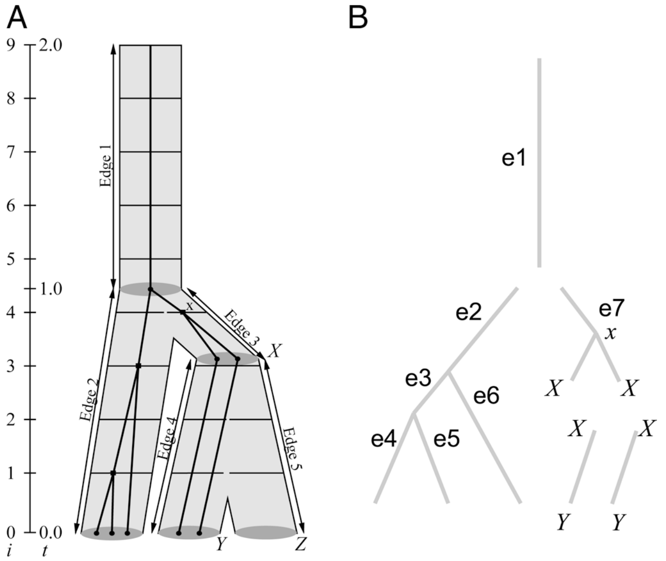

Fig. 1. The figure illustrates the example presented in the text. In *A*, a gene tree, thin black, is shown as evolving inside a discretized species tree. The edges discussed in the text are marked by arrows. Vertices in *S* are marked by gray ellipses, while the vertices augmented to form *S′* are marked by thin horizontal lines. Two gene tree vertices have been introduced in order to break gene tree edges that pass species tree vertex *X*. Each of these vertices are in the reconciliation placed on *X* and has a child placed on *Y* but not on *Z*. In both cases, this indicates a loss in the species lineage leading to *Z*. In *B*, the tree is cut into the *sliced subtrees* discussed in the text.

Although the above sum has exponentially many terms, it can be computed efficiently by using DP. We will devote the rest of this section to explain first, using an example, how to compute $\Pr[(r = l/t)|m,\nu]\Pr[G,t|\lambda,\mu]$ for a specific t, and then how to compute the sum Eq. **2**. We will consider increasingly more complex cases by varying the size and generality of the species tree. For fixed t, it is straightforward to compute the prior PDF for the rates, since

$$\Pr[(r = l/t)|m,\nu] = \prod_{e\in E(G)} \rho(r(e) = l(e)/t(e), m, \nu),$$

where $\rho(s,m,\nu)$ is the density function of the Γ-distribution and $E(G)$ is the set of edges in G. So, for now we will focus on the prior PDF of a gene tree whose vertices have been associated with discretization vertices in S'.

Define $p_{11}(x,y)$ as the probability that a single gene, starting at x, has $k+1$ descendants in y, for some k, of which one may or may not have descendants in the leaves of S while the remaining k go extinct before reaching the leaves of S, i.e,

$$p_{11}(x,y) = \sum_{k=0}^{\infty} \Pr[k \text{ births over } \langle x,y\rangle](k+1)\varepsilon(y)^k,$$

where $\varepsilon(y)$ is the probability that one gene starting in y has no descendants in the leaves of the species tree. For an edge $e = \langle u,v\rangle$ in G, where u and v are associated to x and y in S', respectively, we will use $p_{11}(e)$ to denote $p_{11}(x,y)$. The simplest and most straightforward case is obtained when the species tree consists of a single edge, i.e., the gene tree is simply evolving over a time interval. The sum for this simple case is completely analogous to the computational problem treated in ref. 36. In ref. 38, a PDF for a similar problem was derived, which in our case is equivalent to the following expression:

$$\Pr[G,t] = \prod_{e\in E(G)} 2\lambda p_{11}(e).$$

Notice that p_{11} has the Markovian property that, for subsequent vertices x,y, and z, $p_{11}(x,z) = p_{11}(x,y)p_{11}(y,z)$. Two additional issues must be dealt with: (*i*) how to handle the deterministic event of a speciation and (*ii*) how to handle implicit losses. To handle speciations we first break the gene tree wherever one of its vertices or edges passes a species tree vertex, the latter thereby implicitly indicates a gene loss. This decomposes the gene tree into sliced subtrees (Fig. 1). The implicit losses can be handled by including the probability for each such loss as a multiplicative factor in the PDF for the entire tree. We express the PDF for the entire gene tree as a product of multiplicative factors corresponding to these sliced subtrees. In our example the following factors are obtained over the edges of the species tree:

Edge 1: $p_{11}(e_1)$

Edge 2: $p_{11}(e_2)2\lambda p_{11}(e_3)2\lambda p_{11}(e_4)p_{11}(e_5)p_{11}(e_6)$

Edge 3: $p_{11}(e_7)2\lambda p_{11}(x,X)^2$

Edge 4: $p_{11}(X,Y)^2$

Edge 5: $\varepsilon(X,Z)^2$

We will now explain the DP algorithm for computing Eq. **2**. Since t is no longer fixed, the prior rate PDFs can no longer be handled separately. In the DP, we will consider subproblems where we want to estimate the probability of obtaining the planted subtree G^u and lengths l when (*i*) a single gene tree vertex starts to evolve at $x \in V(S')$ (i.e., the vertex set of S') and (*ii*) the event that creates u happens at $y \in V(S')$. We will denote this probability $s(x,y,u)$ and it can be computed by applying DP to the recursion defined by the cases listed below.

Let u be a vertex of G with children v and w. Moreover, we will use a function $\sigma(u)$, that represents *MPR*, in order to determine the most recent vertex in S' on which we can place u. The function σ is defined as follows: (*i*) for a leaf l of G, $\sigma(l)$ is the species for gene l and (*ii*) for a nonleaf vertex u of G with children v and w, $\sigma(u)$ is the most recent common ancestor of $\sigma(v)$ and $\sigma(w)$. Define S_x to be the subtree of S rooted at x, i.e., the part of S that consists of x and the descendants of x in S. The recursion consists of the following cases.

1. If u is a leaf of G corresponding to a gene found in the species $\sigma(u)$, then $s(\sigma(u),\sigma(u),u) = 1$.
2. If x is a speciation, i.e., $x \in V(S)$, and $x \neq \sigma(u)$, then we may assume that $\sigma(u)$ is a descendant of x in S', and let $s(x,x,u) = 0$.
3. If x is a speciation, i.e., $x \in V(S)$, and $x = \sigma(u)$, then
$$s(x,x,u) = \left(\sum_{y\in D_L(x)} s(x,y,v)\right)\left(\sum_{y\in D_R(x)} s(x,y,w)\right),$$
where $D_L(x)$ and $D_R(x)$ are the sets of descendants of the left and right child, respectively, of x in S'.
4. If x is a speciation, the leaves below u in G are genes found in the species that are leaves below y in S, and z is the child of x in S' that is above y in S', then
$$s(x,y,u) = p_{11}(x,z)\varepsilon(x,\bar{z})\frac{\rho(l(p(u),u)/t(x,y))}{\rho(l(p(u),u)/t(z,y))}s(z,y,u).$$
where $\varepsilon(x,\bar{z})$ is the probability that a single gene starting in x does not reach any leaf of $L(S_x)\setminus L(S_z)$ (since u only has descendant leaves below y, it must have been lost in

the lineage leading to the other child of x or below that child). If $y = z$ this simplifies to

$$s(x,y,u) = p_{11}(x,y)\varepsilon(x,\bar{y})\rho(l(p(u),u)/t(x,y)).$$

5. If x is a discretization vertex, i.e., $x \in V(S')\setminus V(S)$, and u thus corresponding to a duplication, then

$$s(x,x,u) = 2\lambda\left(\sum_{y\in D(x)\setminus\{x\}} s(x,y,v)\right)\left(\sum_{y\in D(x)\setminus\{x\}} s(x,y,w)\right),$$

where $D(x)$ are the descendants of x.

6. Last, if x is a discretization vertex, i.e., $x \in V(S')\setminus V(S)$, and, moreover, x has a child z that is above y in S', then

$$s(x,y,u) = p_{11}(x,z)\frac{\rho(l(p(u),u)/t(x,y))}{\rho(l(p(u),u)/t(z,y))}s(z,y,u).$$

Similar to case 4 above, when $y = z$, this simplifies to

$$s(x,y,u) = p_{11}(x,y)\rho(l(p(u),u)/t(x,y)).$$

The GSR model and the algorithm described above is implemented in a MCMC framework in the C++ program PrIME-GSR (for details, see *SI Appendix*).

Results and Discussion

Tests were conducted to choose the number of discretization steps, and the self-consistency of the MCMC implementation of our model was shown by using a generalization of the *90-percent test* previously described in ref. 15. Both self-consistency and discretization tests were performed on synthetic data (see *SI Appendix*).

Yeast Gene Family Analyses. We analyzed publicly available yeast datasets (39, 40) to compare our method's performance with previously published work on well studied gene families. A predicted whole genome duplication (WGD) occurring in the yeast lineage presents a challenge for our method, which models individual gene duplications rather than block or whole genome duplications.

The data comprise gene families sampled from 17 ascomycetes (see *SI Appendix*). To enhance comparisons, we used the same species tree of ascomycete fungi as in ref. 40 (*SI Appendix*, Fig. S2). The general structure of this tree is congruent with the results presented in e.g. ref. 41, but differs in the exact position of some of the species. We used our *MapDP* algorithm (36) to estimate divergence times. We ran three independent analyses and the result from these were in excellent agreement with each other. The obtained divergence times are shown in *SI Appendix*, Fig. S2 and scaled using the 400 Myr fossil dating of the *Ascomycete* root of ref. 42.

YGOB. The Yeast Gene Order Browser [YGOB version 1 (43), see also ref. 39] comprises gene family data from a subset of the species in *SI Appendix*, Fig. S2, and is based solely on gene order (*synteny*) information. Our comparison with YGOB was aimed at testing the effect of assuming a molecular clock. More specifically, we evaluated how often PrIME-GSR recovers two specific branching points predicted by YGOB, namely the vertex splitting pre-WGD and post-WGD species and the vertex corresponding to the WGD itself. We analyzed each gene family with PrIME-GSR by using both a fully relaxed clock (mode 1) and two constrained modes designed to be closer to a molecular clock. The constrained modes have varying strictness for stochastic variation in the substitution process: the variance was set to 0.001m in mode 2 and 0.0001m in mode 3, and the mean substitution rate m was inferred in the MCMC. The frequency of predictions of the pre/post-WGD vertex and the WGD vertex was recorded in all three modes.

The relaxed clock mode predicts the pre/post-WGD vertex in 167 (91%) gene families and the WGD itself in 121 (66%). In contrast, the corresponding numbers for mode 3 are only 72 (39.1%) and 14 (7.6%), respectively. Mode 2 performs slightly better, predicting the pre/post-WGD in 110 (59.8%) gene families and the WGD in 58 (31.5%), but the result is still considerably worse than for mode 1. For several of these gene families the difference in result was caused by substantially longer edge lengths for one of the post-WGD subtrees, probably due to relaxed selection following the duplication (44–46). Analyzing the data in molecular-clock mode resulted in different rooting of the same (unrooted) tree as in the relaxed clock analysis, and with this rooting the WGD vertex was not recovered. By allowing rate variation, i.e., the relaxed clock, the difference in edge length is accommodated. These results are also supported by the distribution of posterior estimates of coefficient of variation (CV) for the substitution rates, (see *SI Appendix*).

To assess the impact of including the species tree in the analysis, we ran MrBayes (14) on the same dataset by using the same amino acid replacement model (see *SI Appendix* for details). The pre/post-WGD vertex was found in 156 cases (78%) and the WGD in as few as 70 cases (35%). It is clear that PrIME-GSR has a strong advantage compared with a sequence-only model.

Orthogroups. Wapinski et al. (40) present a classification of fungal genes into 30,109 families, coined *orthogroups*, based on sequence and synteny data. The algorithm used to obtain the classification, SYNERGY, results in a single rooted gene tree for each orthogroup. We analyzed the orthogroup sequence data with PrIME-GSR and tested the agreement between our results and SYNERGY gene trees.

Following ref. 40 we partitioned the orthogroups into sets according to their size. The sets were *small*, *medium*, *large*, and *uniform*, including orthogroups with <5–10 genes, orthogroups with 10–17 genes, orthogroups with >17 genes, and orthogroups with exactly one gene for each of the 17 species, respectively. The uniform orthogroups were excluded from the medium set.

To obtain a posterior distribution over gene trees in each orthogroup, we ran PrIME-GSR analyses and the results are shown in Table 1. Overall we ranked the gene tree suggested by SYNERGY highest in 2,637 cases (54.8%). In a further 597 cases (12.4%) it showed up in our posterior distribution but was ranked in second place or lower. In 1,575 cases (32.8%) it did not show up in our posterior distribution. Our results and those obtained with SYNERGY are most in agreement when smaller orthogroups are analyzed. The SYNERGY tree for both the small and the medium set ranked first in most cases, 60.2% and 59.3%, respectively. For uniform orthogroups, which also can be expected to be relatively easy cases, we agree in 69.0% of the cases. However, we disagree on most of the large orthogroups. The gene tree suggested by SYNERGY ranked first in our posterior distribution in only 136 cases (17.6%), while in 548 cases (70.8%) it is not at all present in the posterior distribution.

This latter group was further investigated to verify that our results reflected real differences between the methods. For each of these 548 orthogroups, we reran PrIME-GSR with the tree suggested by SYNERGY held fixed, integrating over θ and the edge lengths of the tree. We then compared the unnormalized maximum a posteriori probabilities (MAP) for the two analyses. The SYNERGY tree MAP value was lower in 496 cases (90.5%), suggesting that our results are indeed due to differences in the compared models. Only in 52 cases (9.5%) could the results be attributed to convergence problems of the original MCMC analyses (see *SI Appendix*).

Case Study: ADH. In the study by Thompson et al. (47), six time-correlated duplications independent of the WGD were proposed to have affected a set of genes associated with the conversion of

Table 1. Number of orthogroups for which the gene tree suggested was ranked first, second, third or lower, and not ranked in the PrIME-GSR posterior distribution

Orthogroup		SYNERGY trees in posterior			
Size	Count	1st	2nd	3rd+	Outside
Small	1,477	889	131	87	370
		60.2%	8.9%	5.9%	25.1%
Medium	1,580	937	88	85	470
		59.3%	5.6%	5.4%	29.7%
Large	774	136	43	47	548
		17.6%	5.6%	6.1%	70.8%
Uniform	978	675	82	34	187
		69.0%	8.4%	3.4%	19.1%
Total	4,809	2,637	344	253	1,575
		54.8%	7.2%	5.3%	32.8%

glucose to ethanol, including the alcohol dehydrogenase (ADH) gene family. This proposal was based on a molecular clock argument. To evaluate these results, we analyzed the ADH orthogroup with PrIME-GSR. The posterior gene tree distribution is relatively flat, the best tree has a posterior probability of 0.10. However, most of the 763 trees sampled in the posterior differed only in a few subtrees and, as the majority rule consensus tree in *SI Appendix*, Fig. S4 shows, almost all clades in the consensus tree have a high posterior probability. This demonstrates the advantages of obtaining a posterior distribution compared with single point estimates. Our result shows that, while a large number of duplications have occurred in the ADH evolution in yeasts, most of these are not associated with the WGD (see *SI Appendix*). In particular, the latter result supports the prediction in ref. 47 for ADH1 and ADH2.

We also reconstructed the ADH tree using MrBayes (14), and this shows that there is a conflict between the information in sequence data and that given by the species tree constraints used by SYNERGY (see *SI Appendix*). The PrIME-GSR tree constitutes a intermediate between these extremes. By considering a posterior distribution of reconciliations rather than the single most parsimonious reconciliation only, it does a better job reconciling the incongruence between the sequence tree and the species tree, than does SYNERGY. PrIME-GSR does not use synteny data; it is therefore interesting to note that it identifies some relationships predicted by YGOB (39) based on synteny, but not present in the MrBayes consensus tree, and it calls other synteny predictions into question (see *SI Appendix*). We conclude that the posterior probabilities provided by PrIME-GSR are based on a sound, explicit probabilistic framework and that the posterior distribution provides an important support measure that is highly relevant to genome evolution analysis.

Concluding Remarks

The GSR model allows us to design a biologically realistic method to reconstruct a gene tree, in the sense that the processes that contributed to the evolution are considered when reconstructing the tree, i.e., reconstruction mirrors generation. In particular, we let the species tree and its divergence times constrain the gene tree. We implemented this method by using a MCMC framework in the program PrIME-GSR, and we have shown that the method is applicable to experimental data on a genomewide scale.

The PrIME-GSR implementation of the GSR model constitutes a major improvement compared with earlier similar methods. While a parsimony approach integrating evidence from the duplication-loss process and the substitution process has already been used in 1979 by Goodman et al. (3), the first probabilistic approach to integrate duplication-loss models and substitution models was used in an example application in 2003 (15), where posterior probabilities from traditional gene tree reconstructions using MrBayes (14) were combined with orthology probabilities to achieve weighted orthology probabilities. However, as pointed out in that study, such an approach rests on a flawed model, because the constraints induced by the species tree, particularly on gene tree edge lengths, are not considered in gene tree reconstruction. In a similar approach using the coalescent model, Liu and Pearl (22) attempted to accommodate this by a combination of additional MCMC and importance sampling analyses, as well as various approximation techniques. A mathematically correct way to combine the *gene evolution model* with traditional substitution models allowing *simultaneous* reconstruction and reconciliation of the gene tree was taken by Arvestad et al. in 2004 (18). However, in that model substitution rates follow a molecular clock. This is often an unrealistic assumption (48), and our results clearly indicate that a molecular clock model should be avoided when working with reconciled trees. Both yeast datasets show high estimated coefficient of variation (see *SI Appendix*) indicating deviation from a molecular clock, and our analysis of the YGOB data clearly shows that molecular clock assumptions affected performance negatively. A similar observation was made in ref. 22, which also assumed a molecular clock. The GSR model explicitly handles this problem by the inclusion of a relaxed clock submodel. Moreover, to handle problems with MCMC over reconciliations and divergence times, Arvestad et al. applied an approximation of the divergence time distribution that was based on a sampling strategy. It turned out that the precision of this approach was insufficient for larger problems, i.e., where many gene sequences are included. The approach of using MCMC to integrate over gene tree divergence times and substitution rates may at first sight appear attractive. However, from previous experience (36), we know that MCMC convergence over the time and rate space can be very slow. Moreover, there are technical problems related to branch-swapping on reconciled trees (see *SI Appendix*). The main technical improvements that enable our MCMC implementation are the factorization of generation probabilities and the discretization methodology. The fact that time can be integrated over Eq. **2** is important for performance.

Since Nei (49), the birth–death model has been an accepted model for gene evolution, but a robust justification of the model has not yet been demonstrated. Our model and tools make it possible to investigate the suitability of a simple duplication and loss model for genes.

Gene order, or synteny, information is another rich source of information for gene tree reconstruction in those organism groups where synteny conservation is high, e.g., between certain yeast species (24). In other groups where this is not the case, e.g., animals and plants (25, 26), its use is significantly more limited. We have in this article compared our results with a synteny-based gene database. The SYNERGY algorithm (23) combines gene sequence data (using a neighbor-joining framework), species trees, and synteny data in an ad hoc manner. In contrast to PrIME-GSR, the constraints imposed by the species divergence times on the gene-tree edge length are ignored, and, moreover, only a single best tree is output rather than a posterior distribution over trees. Moreover, the results from our yeast study may indicate that missing or weak conservation of gene order can mislead synteny-based methods. Nevertheless, inclusion of synteny information is clearly a priority target for future development of PrIME-GSR. Because of the complexity of the structural mutation process, it is not a trivial task. However, probabilistic models for gene inversions have been developed (50) and may provide a promising starting point.

ACKNOWLEDGMENTS. We thank Miranda McEvilly for reading earlier versions and suggesting improvements and anonymous reviewers for their constructive comments. This work was supported by the Swedish Research Council.

1. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302:1401–1404.
2. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res* 15:1153–1160.
3. Goodman M, Cselusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool* 28:132–168.
4. Page RD (1994) Maps between trees and cladistic-analysis of historical associations among genes, organisms, and areas. *Syst Biol* 43:58–77.
5. Page RD, Charleston MA (1997) From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol* 7:231–240.
6. Yuan YP, Eulenstein O, Vingron M, Bork P (1998) Towards detection of orthologues in sequence databases. *Bioinformatics* 14:285–289.
7. Page RD, Cotton JA (2002) *Vertebrate Phylogenomics: Reconciled Trees and Gene Duplications* (World Scientific, Teaneck, NJ), pp 536-547.
8. Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3:2.
9. Storm CE, Sonnhammer EL (2003) Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res* 13:2353–2362.
10. Camin JH, Sokal RR (1965) A method for reducing branching sequences in phylogeny. *Evolution (Lawrence, Kans)* 19:311–326.
11. Kluge AG, Farris JS (1969) Quantitative phyletics and the evolution of anurans. *Syst Zool* 18:1–32.
12. Felsenstein J (2003) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
13. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17:368–376.
14. Huelsenbeck JP, Ronquist F (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
15. Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19(Suppl 1): i7–i15.
16. Gu X (2000) *A Simple Evolutionary Model for Genome Phylogeny Based on Gene Content*, eds Sankoff D, Nadeau JH (Kluwer, Dordrecht, The Netherlands), pp 515–523.
17. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV (2002) Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evol Biol* 2:18.
18. Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2004) Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *Proceedings of RECOMB04* (ACM Press, New York), pp 326–335.
19. Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–1656.
20. Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. *Evolution (Lawrence, Kans)* 59:24–37.
21. Maddison W, Knowles L (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst Biol* 55:21–30.
22. Liu L, Pearl DK (2007) Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol* 56:504–514.
23. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* 23:i549–i558.
24. Fischer G, Rocha EPC, Brunet F, Vergassola M, Dujon B (2006) Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet* 2:e32.
25. McLysaght A, Enright AJ, Skrabanek L, Wolfe KH (2000) Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. *Yeast* 17:22–36.
26. Liu H, Sachidanandam R, Stein L (2001) Comparative genomics between rice and arabidopsis shows scant collinearity in gene order. *Genome Res* 11:2020–2026.
27. Rasmussen M, Kellis M (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res* 17:1932–1942.
28. Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol* 15:1647–1657.
29. Huelsenbeck JP, Larget B, Swofford D (2000) A compound poisson process for relaxing the molecular clock. *Genetics* 154:1879–1892.
30. Kishino H, Thorne JL, Bruno WJ (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol* 18:352–361.
31. Aris-Brosou S, Yang Z (2002) Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst Biol* 51:703–714.
32. Thorne JL, Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. *Syst Biol* 51:689–702.
33. Drummond AJ, Ho SY, Phillips MJ, Rambaut (2006) A relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
34. Lepage T, Bryant D, Philippe H, Lartillot N (2007) A general comparison of relaxed molecular clock models. *Mol Biol Evol* 24:2669–2680.
35. Rannala B, Yang Z (2007) Inferring speciation times under an episodic molecular clock. *Syst Biol* 56:453–466.
36. Åkerborg Ö, Sennblad B, Lagergren J (2008) Birth-death prior on phylogeny and speed dating. *BMC Evol Biol* 8:77.
37. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282.
38. Thompson EA (1975) *Human Evolutionary Trees* (Cambridge Univ Press, Cambridge, UK).
39. Byrne KP, Wolfe KH (2005) The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* 15:1456–1461.
40. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.
41. Kurtzman CP, Robnett CJ (2003) Phylogenetic relationships among yeasts of the 'saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res* 3:417–432.
42. Taylor T, Hass H, Kerp H (1999) The oldest fossil ascomycetes. *Nature* 399:648.
43. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440:341–345.
44. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
45. Seoighe C, Johnston CR, Shields DC (2003) Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Mol Biol Evol* 20:484–490.
46. Scannell D, Wolfe K (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res* 18:137–147.
47. Thomson JM, et al. (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat Genet* 37:630–635.
48. Bromham L, Penny D (2003) The modern molecular clock. *Nat Rev Genet* 4:216–224.
49. Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci USA* 94:7799–7806.
50. Larget B, Simon DL, Kadane JB (2002) Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *J R Stat Soc B* 64:681–693.