J Mol Evol (2006) 63:240–250
DOI: 10.1007/s00239-005-0096-1

JOURNAL OF MOLECULAR EVOLUTION

© Springer Science+Business Media, Inc. 2006

Optimal Gene Trees from Sequences and Species Trees Using a Soft Interpretation of Parsimony

Ann-Charlotte Berglund-Sonnhammer,[1,2]* Pär Steffansson,[1]* Matthew J. Betts,[3†] David A. Liberles[1,3††]

[1] Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden
[2] Linnaeus Centre for Bioinformatics, Uppsala University, BMC Box 598, 75124 Uppsala, Sweden
[3] Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway

Received: 22 April 2005 / Accepted: 15 April 2006 [*Reviewing Editor:* Dr. Rafael Zardoya]

Abstract. Gene duplication and gene loss as well as other biological events can result in multiple copies of genes in a given species. Because of these gene duplication and loss dynamics, in addition to variation in sequence evolution and other sources of uncertainty, different gene trees ultimately present different evolutionary histories. All of this together results in gene trees that give different topologies from each other, making consensus species trees ambiguous in places. Other sources of data to generate species trees are also unable to provide completely resolved binary species trees. However, in addition to gene duplication events, speciation events have provided some underlying phylogenetic signal, enabling development of algorithms to characterize these processes. Therefore, a soft parsimony algorithm has been developed that enables the mapping of gene trees onto species trees and modification of uncertain or weakly supported branches based on minimizing the number of gene duplication and loss events implied by the tree. The algorithm also allows for rooting of unrooted trees and for removal of inparalogues (lineage-specific duplicates and redundant sequences masquerading as such). The algorithm has also been made available for download as a software package, Softparsmap.

Key words: Parsimony — Phylogeny — Gene duplication/gene loss

Introduction

As species and their genomes diverge during evolutionary history, the sets of genes and their sequences also diverge. Gene duplication has been proposed as a crucial source of evolutionary innovation in organisms, like Eukaryotes, with small effective population sizes (Ohno 1970; Francino 2005). With duplication comes initial redundancy, followed by neofunctionalization, subfunctionalization, and, most commonly. pseudogenization (see Lynch et al. 2001; Rastogi and Liberles 2005). This differential retention of duplicate genes between species can result in a different phylogenetic tree for individual gene families than for the species as a whole. Further, differential parsing of shared ancestral gene and nucleotide polymorphism (see Blanchette et al. 2004) as well as uncertainty in tree calculation methodologies (especially for EST and partial sequences) can obfuscate the correlation between the evolutionary history of a gene and the species. Additionally, when using Genbank (Benson et al. 2005) or even draft genome sequences as the starting point for phylogenetic analysis, many species will be represented with

*These two authors contributed equally to this work.
†*Current address:* EMBL Heidelberg, Meyerhoffstrasse 1, 69117 Heidelberg, Germany
††*Current address:* Department of Molecular Biology, University of Wyoming, Laramie, Wyoming 82071, USA
Correspondence to: David A. Liberles, Department of Molecular Biology, University of Wyoming, Laramie, Wyoming 82071, USA; *email:* liberles@uwyo.edu

some genes, while others that are actually present in the species genomes will not be represented in the datasets and falsely appear to have been lost or alternatively falsely appear in many copies. On top of ambiguity at the gene tree level, many vertices in species trees are unresolved and are represented as nonbinary to reflect this species history ambiguity.

Goodman et al. (1979) first introduced a mapping, which was then formalized by Page (1994), to explain the difference between a gene tree and its species tree. Additional algorithms for computing these mappings have also been presented (Zhang 1997; Eulenstein et al. 1998; Zmasek and Eddy 2001b). Another approach, Notung (Chen et al. 2000; Durand et al. 2005), allows consideration of gene tree uncertainty through a bootstrap value threshold and implements a weighting of gene duplication and loss events.

The problem of reconciling a gene tree to a species tree is used to solve two opposite but connected problems. The first problem is to infer a species tree given a set of gene trees, where the gene trees have different evolutionary histories (Guigo et al. 1996; Page 2000; Ma et al. 2000; Page and Cotton 2000; Cotton and Page 2002; Hallett and Lagergren 2002). The other problem is to infer a gene tree or a set of gene trees given a trusted species tree (Arvestad et al. 2003, 2004). The reconciliation can also be used for locating duplication events with respect to a species tree (Guigo et al. 1996) and for orthology analysis (Zmasek and Eddy 2002; Arvestad et al. 2003, 2004). Recent work has also focused on extending this type of approach to differentiating gene duplication events from lateral transfer events (Hallet et al. 2004). In this paper we wish to infer a rooted binary gene tree given a rooted nonbinary species tree and an unrooted, binary, or nonbinary gene tree considering the process of gene duplication.

These previously described algorithms require (or infer) a binary species tree and the approach has been successfully applied on a large scale, when there are no ambiguities in the species tree (Koonin et al. 2004). However, the NCBI taxonomy database (which, while formally a taxonomy, is commonly used as a species tree) (Benson et al. 2005) and other reference species trees are not binary in many places due to uncertainties, between gene trees both from different genes and from morphological characters (for example, the resolution of eutherian mammals). This problem was solved by Koonin et al. (2004) by performing the calculation over their species tree twice for the species in their dataset (once for a topology consistent with a clade of Ecdysozoa and once for a topology consistent with a clade of Coelomata). Here, building on previous algorithmic work, we present a more general mapping using a parsimonious approach toward uncertain speciation events or soft polytomies (for the original definition of soft polytomies, see Maddison 1989). Because the method embraces soft polytomies, we term it soft parsimony, in contrast to previous work, which we term hard parsimony.

One alternative to gene tree-to-species tree mapping for rooting of unrooted trees is midpoint rooting, where the point that is farthest from any extant sequence is designated as the root. However, as heterotachy (different modes of evolution in different subtrees of gene family trees) does not appear to be uncommon, this can falsely assign a root to a more recent fast-evolving branch (see Galtier 2001; Lopez et al. 2002; Siltberg and Liberles 2002).

For the reasons listed above, in the development of a large-scale database for understanding species evolution through the evolution of gene families (Liberles et al. 2001; Roth et al. 2005), it has been necessary to develop a soft parsimony based approach to map gene trees onto species trees. In future implementations of The Adaptive Evolution Database (TAED), an analysis of gene content could be coupled to an analysis of sequence evolution, as lineage-specific duplication has been proposed to play a major role in lineage-specific organismal evolution (for an interesting discussion see Francino 2005). In addition to the bootstrap (or posterior probability) threshold also implemented in Notung, it has been necessary to implement some additional features driven by considerations in the starting dataset (Genbank). Because many species have sparse sampling of genes from their genomes, it has been necessary to minimize, first, gene duplications and, second, gene losses rather than minimizing them together and attributing (with a weight) the loss of a gene to an absence in the genome. Also, because of the redundancy in GenBank (GenBank is an uncurated depository for gene sequences and many genes including those with mutations, splice variants, sequencing errors, etc., appear as multiple independent entries), it has been necessary to treat in-paralogues (lineage-specific duplicates) as redundant entries in an effort to improve gene family signal. The algorithm, as an option, can seek to exclude in-paralogues, as these are then not counted as duplications and are filtered out. An algorithm is presented that enables a mapping with all of the above features using the flexible soft parsimony approach, together with a downloadable software package, Softparsmap.

Methods

Multiple sequence alignments (MSA) were calculated using POA (Grasso and Lee 2004) and phylogenetic trees were built using MrBayes (Huelsenbeck and Ronquist 2001). The parameters used for the tree calculations were as described by Roth et al. (2005). The NCBI taxonomy (Benson et al. 2005) was used as a species tree. The objective of our method is twofold. First, we

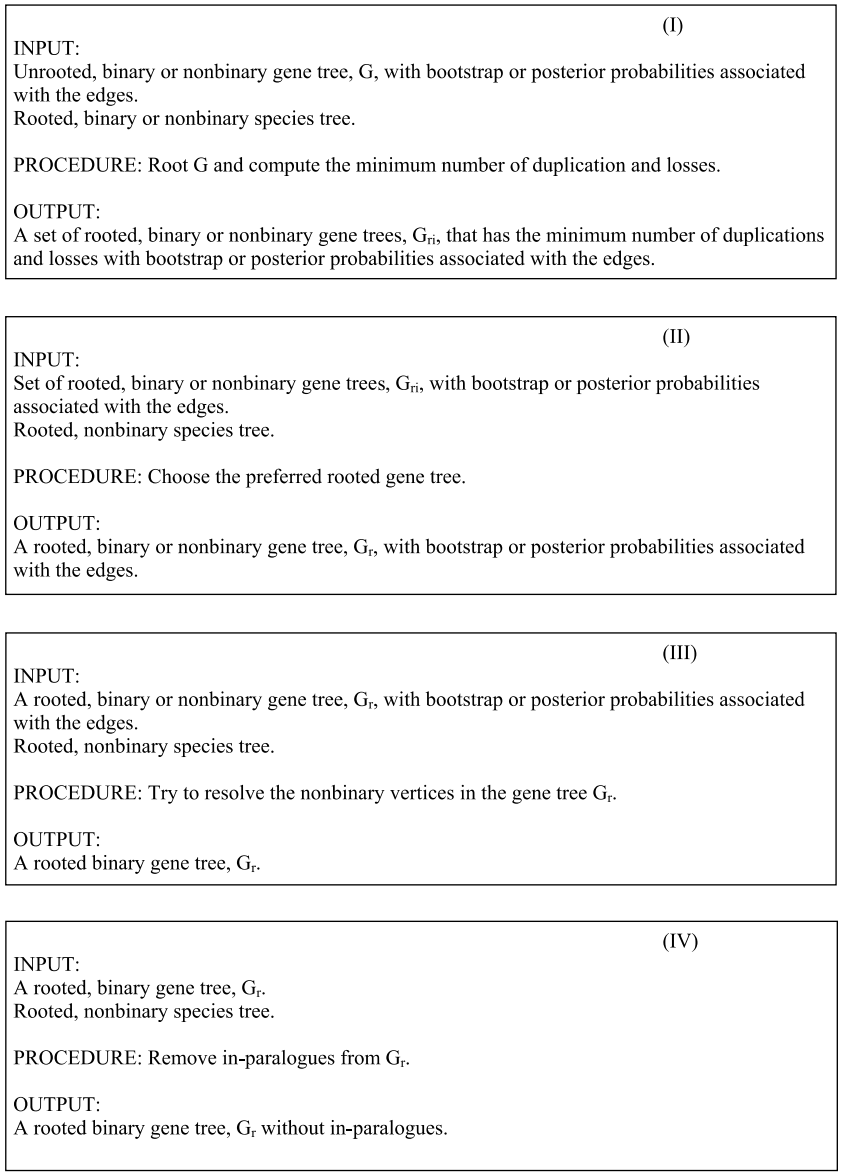

Fig. 1. A flowchart describes the algorithmic process of rooting an unrooted tree.

aim to root the unrooted phylogenetic tree from MrBayes, using the information from the corresponding species tree. Second, we aim to infer a topology of poorly resolved groups in the gene tree based on the species tree with a minimization of duplication and subsequently loss events as optimality criteria, detecting and filtering out redundant copies in the process. The flowchart for the method is illustrated in Fig. 1.

Our approach of rooting the gene tree follows that of Notung, but the methods differ in how the minimum number of duplications and losses are computed. First, the number of duplications is minimized and then the number of losses is minimized for the trees with the minimum number of duplications. Also, our method does not return all binary gene trees that have the minimum number of duplications and losses.

Algorithms

By mapping the vertices of a gene tree to the vertices of the corresponding species tree, each inner gene tree vertex can be labeled as being a duplication or speciation event. Hence, different mappings will describe different evolutionary scenarios. The *m*-mapping for mapping the vertices in the gene tree to the vertices in the species tree was introduced by Goodman et al. 1979. For any gene tree vertex g, $m(g)$ is the species to which genome g belongs. For our soft parsimony approach we defined another mapping, denoted M, for mapping gene tree vertices to species tree vertices. The two mappings are illustrated in Fig. 2, and the Appendix presents a formal definition of our M-mapping.

The objective of our method is to both root and resolve weakly supported edges of unrooted gene trees. This is done by finding the rooted gene trees corresponding to the unrooted gene tree that has the most parsimonious mapping, i.e., a mapping that results in the fewest duplications and losses.

Our method starts out with an unrooted, binary, or nonbinary gene tree, where all edges have bootstrap values or posterior probabilities attached to them. In the first step of our method a set of rooted gene trees is constructed by applying midedge rooting to each edge in the unrooted gene tree. Next the edges that have bootstrap values or posterior probabilities less then a predefined cutoff value are collapsed. From the resulting set of rooted gene trees with well-supported edges the following is performed. First,

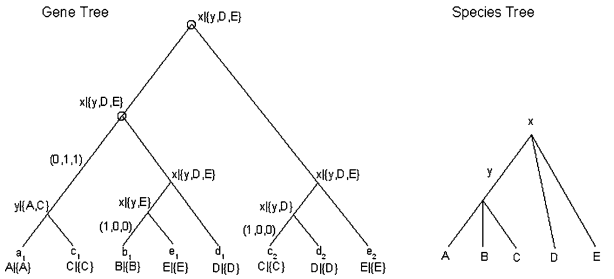

Fig. 2. At the left is a gene tree explaining the evolutionary relationship among the genes a_1, b_1, c_1, d_1, d_2, e_1, and e_2. The labels of the inner vertices have been omitted. The circles denote the duplication events detected by the soft parsimony approach. At the right is the species tree corresponding to the gene tree. The leaves of the species tree are labeled with the extant species *A, B, C, D*, and *E*, and gene a_1 belongs to the genome of species *A*, gene b_1 to genome *B*, and so on. The two inner vertices are labeled x and y. For every vertex g in the gene tree, the m- and M-mappings are given in the form of $m(g)|M(G)$. For the edges in the gene tree where the soft parsimony definition of loss detects gene losses, the losses are printed out as ($loss_1$, $loss_2$, $loss_3$). If the species tree is nonbinary, our soft parsimony definition infers a lesser or equal number of duplications compared to the definition introduced by Goodman et al. (1979), as well as a lesser or equal number of losses compared to the definition by Guigo et al. (1996). However, if the species tree is binary, the definitions are equivalent, and thus our approach will give the same number of duplications and losses in comparison with these other two approaches.

the minimum number of duplications is computed by summing over all gene tree vertices.

$$Dup^{(S,G)} = \sum_{g \in V(G) \setminus L(G)} dup^{(S,G)}(g) \tag{1}$$

where $dup^{(S,G)}(g)$ is the minimum number of duplications associated with gene tree vertex g (see below). Since the number of duplications associated with any gene tree vertex is independent of the number of duplications associated with any other gene tree vertex, the summation over the gene tree vertices can be done in any order. Only the rooted gene trees that minimize the number of duplications are kept, and this results in a subset to the original set of rooted gene trees. Of course this subset might be equal to the original set. Second, for this subset of rooted gene trees the minimum number of losses is computed by summing over all gene tree vertices:

$$Loss^{(S,G)} = \sum_{g \in V(G) \setminus L(G)} loss^{(S,G)}(g) \tag{2}$$

As in the previous step, only the rooted gene trees that meet the optimality criterion are kept. Here the optimality criterion is that the minimum number of losses should be minimized. Consequently, the resulting subset consists of rooted gene trees that all have the same number of duplications and losses. The weak edges that do not affect the number of duplications and losses are restored as they are encountered in the summation. Third, if the subset of rooted gene trees that minimize the number of duplications and losses has more than one member, the following procedure is applied to choose the preferred rooted gene tree. As soon as only one tree satisfies a criterion, the procedure stops. The first criterion is that the preferred tree should have the most internal vertices (i.e., the most nodes or branching points), the second criterion is that the preferred tree should have the least number of weak edges, and the third criterion is that the preferred tree should have the shortest root distance. However, it is not certain that a gene tree can be chosen from this procedure and in such cases our method returns any of the trees in the subset together with a warning. Fourth, the preferred rooted gene tree might not be binary, and thus the next step is to resolve the remaining collapsed edges. This is done by adding splits from the corresponding species tree and, if necessary, information from adding outgroups to the original unrooted gene tree (see the Appendix for more detail). Finally, in-paralogues are removed from the rooted gene tree by pairwise comparisons of the gene sequences of the in-paralogues. Given the sequences of two in-paralogues we choose to keep one of them according to the following criteria. First, if one of the sequences is complete while the other is only a fragment, the complete one is kept. Second, the longest sequence is kept. Third, the sequence with the highest GI number (most recent entry to GenBank) is kept.

Minimizing the Number of Duplications

As the leaves have no duplications associated with them, the minimum number of duplications for a given gene tree is computed by summing the minimum number of duplications for each inner gene tree vertex as shown in expression (1). Since the numbers of duplications associated with each gene tree vertex are independent, the computations can be done in any order. When the minimum number of duplication is computed for any inner gene tree vertex g to a gene tree G, the subtree rooted at g is considered, i.e., G_g.Given a binary inner gene tree vertex, the minimum number of duplications can readily be computed from

$$dup^{(S,G)}(g) = \begin{cases} 1 & \text{if the two children of } g \text{ have} \\ & \text{descendents within the same} \\ & \text{extant genome} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

For any nonbinary inner gene tree vertex g, the minimum number of duplications associated with g is calculated by partitioning the child vertices of g into sets, such that the members of any set do not have descendants in the same extant genome. The partitioning is done such that the number of sets is minimized. An example of how the minimum number of duplications is computed for a gene tree vertex is shown in Fig. 3a. The problem of computing the minimum number of duplications for nonbinary gene tree vertices in this way is NP-complete; see Theorem A4 in the Appendix.

Minimizing the Number of Losses

The minimum number of losses is computed after the minimum number of duplications has been computed. Moreover, the computations are only performed for the rooted gene trees that minimize the number of duplications, as gene loss is a secondary optimization criterion to gene duplication. When the minimum number of losses is computed for any inner gene tree vertex g, the subtree rooted at g with the children of g as leaves is considered. If the gene vertex is nonbinary, so is the subtree, and thus, a refinement of this nonbinary subtree is constructed before the minimum number of losses is computed. Our algorithm separates three types of loss that can occur in the planted subtree rooted at g and containing one of the two children of g. For each gene tree vertex we have to sum over these three types of loss:

$$\begin{aligned} loss^{(S,G)}(g) = \sum_{i=1}^{2} (&loss_1(g, c_i(g)) + \ loss_2(g, c_i(g)) \\ &+ loss_3(g, c_i(g))) \end{aligned} \tag{4}$$

For any binary gene tree vertex the minimum number of losses is computed directly using expression (4), but for nonbinary vertices in the gene tree another approach must be taken. The proposed algorithm for computing the minimum number of losses for a nonbinary gene tree vertex and the corresponding species tree is approximate, and it is presented in detail in the Appendix. For each nonbinary gene tree vertex g, the corresponding partitioning of the child vertices into sets, given from the minimization of duplications algorithm, is used to resolve the uncertainty, i.e., create a binary tree with the children of the current vertex as

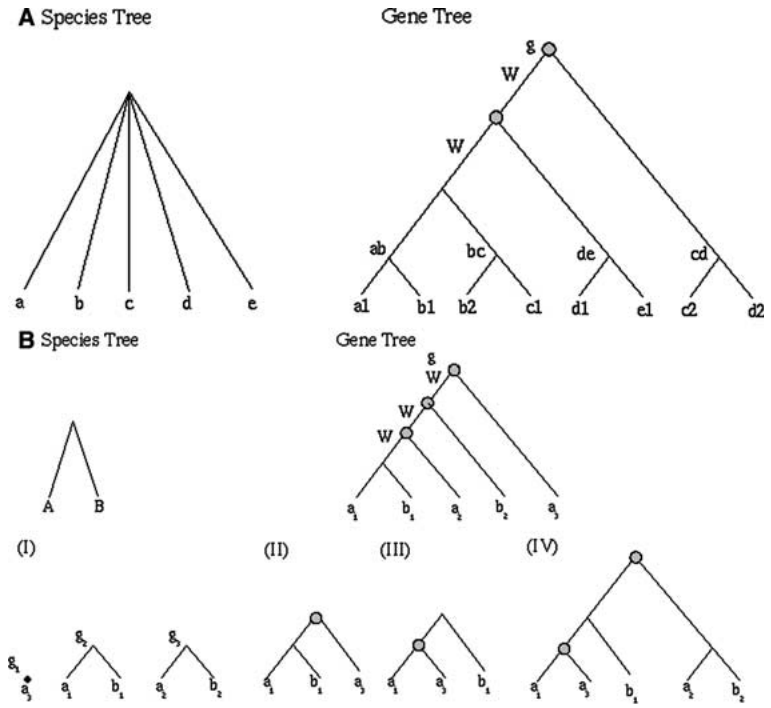

Fig. 3. **a** An overview of the gene duplication computation process is presented. In the upper left corner is the species tree corresponding to the rooted, binary gene tree in the upper right corner, for which we wish to compute the minimum number of duplications. The weak edges in the gene tree are labeled W. Duplication events are denoted by circles. Here we only show how to compute the minimum number of duplications for the gene tree vertex g. After the week edges have been collapsed, the child vertices of g are equal to the set $\{ab,bc,de,cd\}$ and the minimum partitioning of these vertices such that members of the same set do not have descendants in the same extant genome is $\{\{ab,cd\},\{bc,de\}\}$. The number of duplications associated with g is equal to the size of the minimum partition minus one. In this case the minimum number of duplications associated with g is one. **b** An overview of the gene loss computation process is presented. In the upper right corner is the rooted, binary gene tree for which we wish to compute the minimum number of losses, and in the upper left corner is the corresponding species tree. The gene tree leaves labeled a_1, a_2, and a_3 are genes present in the genome of the extant species A, and the gene tree leaves labeled b_1 and b_2 are genes present in the genome of the extant species B. The weak edges are labeled W, and duplications are denoted as circles. Here we only show how to compute the minimum number of losses for the gene tree vertex g. After collapsing the weak edges, the children of g are equal to the set $\{a_1,b_1,a_2,b_2,a_3\}$. From the duplication algorithm a minimum partitioning of these vertices into sets such that any members of the same set do not have descendants in the same extant genome is $\{\{a_1,b_1\},\{a_2,b_2\},\{a_3\}\}$, i.e., the minimum number of duplications associated with this gene tree vertex is equal to two. In the first step in the loss algorithm, a rooted tree is constructed for each set in the partition as illustrated in (I). Note that the tree constructed from the set $\{a_3\}$ is a single vertex. The first two trees in (I) are then combined as illustrated in (II). The duplication event associated with the root of this tree can be moved one step farther from the leaves by swapping b_1 and a_3, illustrated in (III), and the tree is kept for the subsequent steps of the algorithm. However, if the duplication event cannot be moved closer to the leaves, the first and third trees in (I) are combined and the resulting tree is tested to see if the duplication there can be moved closer to the leaves. If so, this tree is kept instead, but if the duplication event cannot be moved closer to the leaves in any of the trees constructed by combining the first tree in (I) with any other tree in (I), the tree constructed first would be chosen. Moreover, we continue to build trees from the remaining pairs of trees in (I), if any. The resulting tree(s) is(are) shown in (III). Next the tree(s) in (III) is(are) combined (if there is more than one) in the same way as in the previous step, and this procedure continues until we only have one tree as shown in (IV). Now the minimum number of losses for the gene vertex g can be computed, using expression (2), and in this example the minimum number of losses is equal to zero.

leaves. This tree is constructed such that the vertices labeled as duplications are as close to the leaves of this tree (as this will count redundant sequences as in-paralogues rather than out-paralogues, enabling them to be filtered from the duplication calculation). This tree is then used together with the corresponding species subtree in expression (2) to compute the minimum number of losses for the current gene vertex. In Fig. 3b, an example of how the minimum number of losses is computed for a gene tree vertex is presented.

Software

Softparsmap is available for download from http://www.ii.uib.no/~steffpar/softparsmap/. It is written in Java and requires JDK 1.4.2 or later.

Results and Discussion

The systematic application of a soft parsimony approach for analyzing gene trees in the context of species trees has been performed as part of The Adaptive Evolution Database (TAED) (Roth et al. 2005). Here, vertices with posterior probabilities of < 0.7 were collapsed to nonbinary trees. Then the NCBI taxonomy (Benson et al. 2005) was used as a species tree to minimize the number of gene duplication and loss events in the gene family tree to produce a binary result.

A total of 1217 of 11,704 rootable gene families trees in the Chordate half of TAED were modified

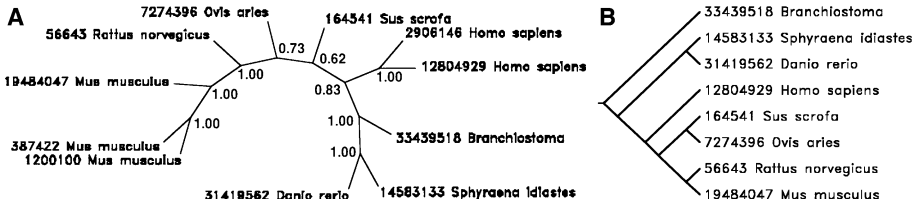

Fig. 4. **a** The unrooted tree for malate dehydrogenase from Mr. Bayes as calculated for TAED (Roth et al. 2005) before application of the soft parsimony algorithm. The leaf IDs are GenBank protein GIs. **b** The same tree has now been rooted and corrected using the soft parsimony algorithm. The Artiodactyls now form a single clade without implying a gene duplication event, and the in-paralogues along the human and mouse lineages have been filtered out because the original data set from GenBank contained redundancies. Trees here are visualized using ATV (Zmasek and Eddy 2001a).

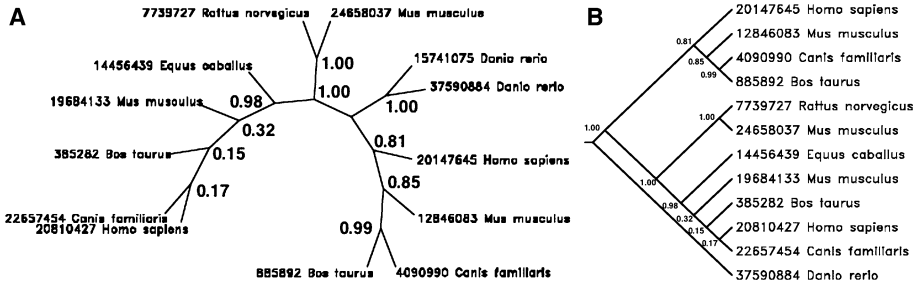

Fig. 5. **a** The unrooted tree for guanine nucleotide binding protein from Mr. Bayes as calculated for TAED (Roth et al. 2005) before application of the soft parsimony algorithm. **b** Following application of the soft parsimony algorithm, a tree with differential optimal branching of eutherian mammals in two independent clades after a more ancient gene duplication event is shown. While a zebrafish in-paralogue is removed, the ambiguous branching order of Primates, Rodents, Carnivores, and Artiodactyls is tolerated without imposing a solution from one clade to another or inferring any extra gene duplication events. Trees here are visualized using ATV (Zmasek and Eddy 2001a).

using this approach. This approach holds equal value for other gene family databases, where identification of orthologous genes is important. From TAED, a sample tree, where the algorithm corrects a tree in the expected manner, is shown in Figs. 4a and b. The example shown is malate dehydrogenase.

As calculated by Mr. Bayes (Huelsenbeck and Ronquist 2001), there is no root of the gene family tree that generates Artiodactyls (hoofed mammals with an even number of toes) as a monophyletic group without inferring a gene duplication and multiple selective loss events. Application of the soft parsimony algorithm results in the expected rooted tree, shown in Fig. 4b.

To walk through this, in the first step, the root is placed on the lineage separating *Branchiostoma* (amphioxus) from teleost (bony) fish. This results in one implied gene duplication event in eutherian (placental) mammals. Next the branch leading to *Sus scrofa* below the posterior probability threshold of 0.70 is collapsed. leading to possible nonbinary trees linking it with the *Ovis aries*/rodent vertex. Then the number of duplication events associated with different resolutions is assessed, and a resolution of the *Sus scrofa*/*Ovis aries* grouping as Artiodactyls with the root still on the branch separating amphioxus from teleosts now implies no gene duplication events.

Of course, gene duplication and loss events do occur. Bayesian approaches, which treat such events probabilistically, result in the explanation that such events are rarer and less likely to explain a tree such as that shown in Fig. 4a than, in this case, statistical uncertainty of branching (Arvestad et al. 2003).

Many of the families in TAED that have been corrected are multigene families, where the branching order is different following a gene duplication event and gene loss events. An example of this, where an ancient gene duplication event preceded the divergence of eutherian mammals and where the optimal tree shows a different eutherian mammal topology, is shown in Figs. 5a and b, from the guanine nucleotide binding protein gene family in TAED. This is an example of a tree where hard parsimony would force a less preferred topology on one of the postduplication clades. In this case, Primates are the outgroup to Rodents, Carnivores, and Artiodactyls on one half of the tree and Rodents are the outgroup to Primates, Carnivores, and Artiodactyls on the other half. Without a posterior probability threshold, the hard parsimony approach would infer a gene duplication event. With a posterior probability threshold (in this case), the tree would be corrected to one with even lower support to prevent inference of a gene duplication event using hard parsimony.

The algorithm presented under Methods and applied to TAED is available as software, called Softparsmap, and is available for download at http://www.ii.uib.no/~steffpar/softparsmap/. The optional functionality available in the software package includes tree rooting by minimization of gene duplication and loss events, removal of in-paralogues, removal of uncertainties or correction of weakly supported splits based on the reference species tree, gene tree-to-species tree mapping to allow identification of orthologues and paralogues, and comparison of tree topologies.

A fast, flexible, powerful approach is presented that allows a gene tree-to-species tree mapping using a soft parsimony algorithm. This approach has been applied systematically to gene families based on GenBank in TAED, modifying over 10% of gene families. The software package available from this method should be a valuable addition to the evolutionary bioinformatics toolbox.

Appendix: Mathematical Properties of the Soft Parsimony Algorithm

Definitions and Notation

A tree T is a connected graph with no cycles, and consists of a vertex set $V(T)$ and an edge set $E(T)$. The leaves of a tree are the vertices with degree 1, and the leaf set of T is denoted $L(T)$, which is a subset of $V(T)$. A tree T is *rooted* if there is exactly one distinguished vertex, the root, which is denoted $r(T)$. For a rooted tree T, a child of a vertex $u \in V(T)$ is denoted $c_i^T(u)$, and the set of children for a vertex u in T is denoted $C^T(u)$. The rooted subtree of T rooted at $u \in V(T)$ is denoted T_u. The planted subtree rooted at u containing only the child vertex $c_i(u)$ is denoted $T_{uc_i}(u)$. Further, for a nonroot vertex $v \in V(T)$, let $p^T(v)$ be the parent of v. For any $u, v \in V(T)$, let $v \leq^T u$ if $v \in V(T_u)$ and let $v <^T u$ if $v \in V(T_u)\backslash\{u\}$. A rooted tree T is *binary* if all interior vertices have two children and an unrooted tree is binary if all interior vertices are connected to three other vertices. For a tree T, $d = (L_1, L_2)$ is a *split* in T if there exists an edge $e \in E(T)$ such that L_1, L_2 are the two leaf sets of the trees formed when e is removed. Given a tree T and a set of vertices $V' \subseteq V(T)$, let $LCA^T(V')$ be the last common ancestor of the vertices in V'. Tree indexes are omitted whenever it is clear from the context, and trees are rooted if else is not specified.

A species tree S is a phylogenetic tree that describes the relationship between extant species. A gene tree G is a phylogenetic tree with a leaf labeling function $\sigma: L(G) \rightarrow L(S)$. The leaves in G represent genes and the leaves in S represent extant species, and thus a gene $g \in L(G)$ belongs to the genome of species $\sigma(g)$. We denote the set of all extant genomes that are represented by the leaves of the gene tree $\sum(G) = \cup_{g \in L(G)} \sigma(g)$. For two rooted gene trees G and G', the leaf $g_1 \in L(G)$ *equals* the leaf $g_2 \in L(G')$ if g_1 and g_2 are representing the same gene. The internal vertex $g_1 \in V(G)\backslash L(G)$ equals $g_2 \in V(G')\backslash L(G')$ if $L(G_{g_1}) = L(G'_{g_2})$. Further, G is said to *refine* G' if for every $g_2 \in V(G')$ there exists a $g_1 \in V(G)$ such that $g_1 = g_2$. Goodman et al. (1979) introduced a mapping m, mapping the vertices in the gene tree to the vertices in the species tree. More precisely, given a species tree S and a gene tree G, m is a surjective mapping such that for all, $g \in V(G), m(g) = LCA^S(\sum(G_g))$.

Gene Duplication and Gene Loss

Here we introduce a new soft parsimony mapping for uncertain species trees and binary gene trees. Further, we define how to compute gene duplication and gene loss events using this mapping.

Given a species tree S and a rooted binary gene tree G such that $\Sigma(G) = L(S)$, let M be the mapping M: $V(G) \rightarrow 2^{V(S)}$ such that

1. For any $g \in V(G)$ such that $m(g) \in L(S)$, $M(g) = \{m(g)\}$.
2. For any $g \in V(G)$ such that $m(g) \notin L(S)$, $M(g) = \{x \in C^S(m(g)) | \exists g' \in V(G), g' <^G g \wedge m(g') \leq^S x\}$ and define $Z(g) = \cup_{s \in M(g)} L(S_s)$.

Each mapping M for a given gene tree and species tree describes a hypothesis of how the gene tree evolved with respect to the species tree and, thus, which gene tree vertices are duplication events and which are speciation events.

For any interior vertex $g_1 \in V(G)\backslash L(G)$ the number of duplications associate with g is

$$dup^{(S,G)}(g) = \begin{cases} 1 & \text{if } |Z(c_1(g)) \cap Z(c_2(g))| > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{(A1)}$$

and the total number of duplications for G is

$$Dup^{(S,G)} = \sum\nolimits_{g \in V(G)\backslash L(G)} dup^{(S,G)}(g) \quad \text{(A2)}$$

The number of losses associated with a vertex $g \in V(G)$ and one of its two children $c_i(g) \in V(G)$ is defined as

$$loss_1(g, c_i(g)) = |\{s \in V(S)\} | m(c_i(g)) <^s s <^s m(g)\}| \quad \text{(A3)}$$

$$loss_2(g, c_i(g)) = \begin{cases} 1 & \text{if } m(c_i(g)) <^s m(g) \wedge |C^s \\ & (m(c_i(g)))\backslash M(c_i(g))| > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{(A4)}$$

$$loss_3(g, c_i(g)) = \begin{cases} 1 & \text{if } dup^{(S,G)}(g) = 1 \\ & \wedge M(g) \neq M(c_i(g)) \\ 0 & \text{otherwise} \end{cases} \quad \text{(A5)}$$

Then the number of losses assigned to an interior vertex $g \in V(G) \backslash L(G)$ and the total number of losses in the gene tree is

$$loss^{(S,G)}(g) = \sum_{i=1}^{2} (loss_1(g, c_i(g)) + loss_2(g, c_i(g)) + loss_3(g, c_i(g))) \quad \text{(A6)}$$

and

$$Loss^{(S,G)} = \sum_{g \in V(G) \backslash L(G)} loss^{(S,G)}(g) \quad \text{(A7)}$$

Uncertain Gene Trees

In many real gene trees, vertices with more then two children exist. Here we present a heuristic algorithm for computing the minimum number of duplications and losses over the set of all gene trees refining the given gene tree. Duplications are prioritized before losses.

Let G be any gene tree, and let S be a species tree, such that $\Sigma(G) = L(S)$. The set of binary gene trees that refine G is denoted BG^{G}. Moreover, the set of trees in BG^{G} that have the minimum number of duplications is denoted $BG_{min}^{(S,G)}$ and can be expressed as $BG_{min}^{(S,G)} = \{H' \in BG^{G} \mid \forall H \in BG^{G}, \quad Dup^{(S,H')} \leq Dup^{(S,H)}\}$.

The minimization of the number of losses is constrained by the minimum number of duplications. Thus, the numbers of duplications and losses are defined as

$Dup_{min}^{(S,G)} = Dup^{(S,H)}$ where $H \in BG^{(S,G)}$ and $Loss_{min}^{(S,G)} = Minimum_{H \in BG_{min}^{(S,G)}} Loss^{(S,H)}$

Computing the number of duplications $Dup_{min}^{(S,G)}$ is NP-complete (see Theorem A4), but for our heuristic approach, described below, duplications can be computed in polynomial time.

Algorithms

Since our model assumes that the number of duplications at a vertex $g \in V(G) \backslash L(G)$ is independent of the number of duplications at any other vertex $g' \in V(G) \backslash L(G)$, we can calculate the number of duplications at the vertices of G in any order. The algorithm computing duplications is based on Theorem A1.

Theorem A1. Let S be a species tree, let G be a gene tree such that $\Sigma(G) = L(S)$. For all $g \in V(G) \backslash L(G)$, denote $A(g)$ to be all possible partitions of the set $C^{G}(g) = \{g_1,\ldots,g_n\}$, let $A'(g) = \{PA \in A(g) \mid \forall D \in PA, \quad \forall g_i, g_j \in D;\ Z(g_i) \cap Z(g_j) = \phi \vee i = j\}$, and let $A'_{min}(g) = \{PA \in A'(g) \mid \forall PA' \in A'(g); PA \leq PA'\}$. Then $Dup_{min}^{(S,G)} = \sum_{g \in V(G) \backslash L(G)} |PA(g)| - 1$ where $PA(g) \in A'_{min}(g)$. In order to prove Theorem A1 the following lemmas are needed.

Lemma A2. Given a gene tree G and a species tree S such that $\Sigma(G) \subseteq L(S)$, then for all g, $g_1 = V(G)$, such that $g = p(g_1)$ it holds that $Z(g_1) \subseteq Z(g)$.

Proof. Let G be a gene tree and S a species tree such that $\Sigma(G) \subseteq L(S)$, and let $g, g_1 \in V(G)$, such that $g = p(g_1)$. Then the vertex set of G_{g_1} is a subset of the vertex set of G_g, which by the definition of the set Σ gives $\Sigma(G_{g_1}) \subseteq \Sigma(G_g)$, and further $m(g_1) \leq^{S} m(g)$. Consequently, for all $s \in M(g_1)$ there exists $s' \in M(g)$ for which $S \leq^{S} S'$, and thus $Z(g_1) \subseteq Z(g)$.

For a species tree S, a gene tree G, a gene tree $G' \in BG^{G}$, let the notation $(g_{11},g_{12}{:}g_1)g_2{:}g_0$ be an *up triplet* in G' if $g_0, g_1, g_2, g_{11}, g_{12} \in V(G')$, $g_1 = p(g_{11}) = p(g_{12})$, $g_0 = p(g_1) = p(g_2)$, $g_1 \in V(G') \backslash V(G)$, $dup^{(S,G')}(g_1) = 1$, and $dup^{(S,G')}(g_0) = 0$.

Lemma A3. Let S be a species tree and let G be a gene tree such that $\Sigma(G) = L(S)$, then for every gene tree $G' \in BG^{G}$ there exists a gene tree $G'' \in BG^{G}$ such that $Dup^{(S,G')} = Dup^{(S,G'')}$ and G'' do not contain any up triplets.

Proof. G'' is computed through following steps.

1. Let $G^n = G'$, where $n = 1$.
2. Find an up triplet $(g_{11}, g_{12}{:}g_1)\, g_2{:}g_0$ in G^n. If no up triplet exists, then $G'' = G^n$.
3. We know that $dup^{(S,G^n)}(g_1) = 1$, $dup^{(S,G^n)}(g_0) = 0$, and since $dup^{(S,G^n)}(g_1) = 1$ it follows that $|Z(g_{11}) \cap Z(g_{12})| > 0$. According to Lemma A2 $Z(g_{12}) \subseteq Z(g_1)$, and since $dup^{(S,G^n)}(g_0) = 0$, it follows that $0 = |Z(g_1) \cap Z(g_2)| \geq |Z(g_{12}) \cap Z(g_2)| = 0$. Consequently, it is possible to move the duplication associated with vertex g_1 one edge closer to the root of G^n by creating the following tree. Let G^{n+1} denote the tree obtained when we take G^n and exchange places with the rooted subtrees $G^n_{g_{11}}$ and $G^n_{g_2}$. Note that $Dup^{(S,G^n)} = Dup^{(S,Gn+1)}$. Set $n{:} = n + 1$ and resume at 2.

Theorem A1 can now be proven.

Proof. According to Lemma A3, we know that for any $G' \in BS_{min}^{(S,G)}$ it is possible to compute gene tree $G'' \in BS_{min}^{(S,G)}$ such that no up triplet exists in G''. Now for every gene vertex $g \in V(G)$, let $PA(g)$ be the partition computed using G'' in following steps.

1. Locate the gene vertices g', $g'_1, \ldots, g'_n \in N(G'')$ such that $g' = g$, $g'_1 = g_1, \ldots, g'_n = g_n$ where $C^G(g) = \{g_1, \ldots, g_n\}$.
2. If $dup^{(S,G'')}(g') = 0$, then let $PA(g) = \{g_1, \ldots, g_n\}$ and we are done.
3. Let L be list such that $L = \{g'_1, \ldots, g'_n\}$ and $PA'(g)$ be an empty set.
4. Set g'_t to be the first gene vertex in L.
5. If $dup^{(S,G'')}(p(g'_t)) = 0$, set g'_t to be $p(g'_t)$ and resume at 5.
6. Let $D'_i = \{g'' \in \{g'_1, \ldots, g'_n\} | g'' \leq g'_t\}$, add D'_i, to $PA'(g)$, and remove the gene vertices in D'_i from L. If $|L| > 0$, resume at 4.

From Lemma A3 nd the above steps it holds that $PA(g) \in A'(g)$. Furthermore, if $PA(g) \notin A'_{min}(g)$, then a binary gene tree with fewer duplications then G'' could be constructed, but $G'' \in BS^{(S,G)}_{min}$ and thus it must hold that $PA(g) \in A'_{min}(g)$. Consequently, for every gene tree $G' \in BS^{(S,G)}_{min}$ and for all $g \in N(G) \backslash L(G)$, there exists a minimum partition $PA(g) \in A'_{min}(g)$ such that $Dup^{(S,G)}_{min} = Dup^{(S,G')}_{min} = \sum_{g \in N(G) \backslash L(G)} |PA(g)| - 1$.

Inferring the number of duplications at a gene tree vertex.

Let G be gene tree and S species tree, such that $\Sigma(G) = L(S)$. Then for gene tree vertex $g \in V(G) \backslash L(G)$, let $A_g = C^G(g)$ and let P_g be an empty list of sets of gene tree vertices.

1. For all gene tree vertices in A_g,
2. Pick a gene tree vertex $g_i \in A_g$ such that for all $g_j \in A_g$ it holds that $|Z(g_j)| \leq |Z(g_i)|$.
   - 2.1. Put g_i in the first set $p_k \in P_g$ such that for all other gene tree vertices g_k in p_k it holds tha $Z(g_k) \cap Z(g_i) = \phi$.
   - 2.2. Let $A_g = A_g \backslash \{g_i\}$.
   - 2.3. If A_g is empty goto 4, otherwise goto 2.
3. Set $A_g = C^G(g)$ and P_g to an empty list of sets. For all gene tree vertices in A_g,
   - 3.1. pick a gene tree vertex $g_i \in A_g$ at random.
   - 3.2. Put g_i in the first set p_k of P_g such that for all other gene tree vertices g_k in p_k it holds that $Z(g_k) \cap Z(g_i) = \phi$.
4. Let $g_{i(k)}$ denote gene vertex g_i in set p_k, if $|\cap_k (\cup_i Z(g_{i(k)}))| \geq 1$ then a minimum partition is reached. The number of duplications equals the number of sets in P_g minus one.
5. If it is possible to create a set L_g of gene vertices by taking exactly one gene vertex from every set p_k in P_g such that for any two gene vertices $g_i, g_j \in L_g | Z(g_k) \cap Z(g_i)| \geq 1$, then a minimum partition is reached. The number of duplications equals the number of sets in P_g minus one.
6. If we have not reached the maximum number of rebuilds, goto 3. Otherwise the method has failed.

Inferring the number of losses at a gene tree vertex.

Let G be a gene tree and S a species tree, such that $\Sigma(G) = L(S)$. Then for a gene tree vertex $g \in V(G) \backslash L(G)$, let P_g be the partition as given from the duplication algorithm. The objective is to build a rooted binary gene tree G_{g_0} with $L(G_{g_0}) = C^G(g)$, such that for every $g_i \in L(G_{g_0})$ and its corresponding $g_j \in C^G(g)$, it holds that $M(g_i) = M(g_j)$. We wish to construct a gene tree G_{g_0} such that the number of losses is less then or equal to any other binary gene tree constructed with the vertices in $C^G(g)$ as leaves. However, the method proposed here is approximate, which means that there might exist another binary gene tree constructed with the same set of leaves that gives fewer losses. The gene tree G_{g_0} is then used in equations (A3)–(A5) to calculate the minimum number of losses associated with the gene tree vertex g. Let L_g be an empty sorted set of gene vertices such that for any two gene vertices $g_i, g_j, \in L_g$, g_i is before g_j if $|Z(g_i)||Z(g_j)|$.

1. For every set in P_g,
   - 1.1. build a gene tree G_{g_i} using the splits found in the species tree. Denote the root g_i and add g_i to L_g.
2. If there is more then one gene vertex in L_g resume at 3. Else,
   - 2.1. let g_0 be the gene vertex in L_g,
   - 2.2. set $G_{g_0} = G_{g_i}$ and find locally the mapping M that puts the duplications as close to the leaves as possible. Resume at 5.
3. Let $j = 1$.
   - 3.1. Let $g_0, g_j \in L_g$ be the roots of the first pair of subtrees for which a new gene tree G_{g_k} can be constructed by adding a gene tree vertex g_k and two edges such that $g_k = p(g_0) = p(g_j)$.
   - 3.2. For G_{g_k} find locally the mapping M that puts the duplications one step closer to the leaves. If no such mapping can be found, let $j = j + 1$ and resume at 3. l.
   - 3.3. Remove g_0, g_j from L_g and add g_k to L_g. Resume at 2.
4. Let $g_0, g_1 \in L_g$ be the roots of the first pair of subtrees in L_g. Construct a new gene tree G_{g_k} by adding a gene tree vertex g_k and two edges such that $g_k = p(g_0) = p(g_1)$.
   - 4.1. Remove g_0, g_1 from L_g and add g_k to L_g. Resume at 2.

5. Compute the losses of G_{g_0} by using equation (A7).

In different steps of the loss algorithm it says that we find locally the mapping M that puts duplications as close to the leaves as possible. That is done in the following way. For each triplet $(g_{11}, g_{12}{:}g_1)\ g_2{:}g_0$, i.e., rooted tree with three leaves, in G_{g_0} the mapping M is found such that duplications are assigned to vertices as close to the leaves as possible. This is successful if the root of the triplet is labeled as a duplication event and g_1 is not, and if we can swap g_{1i} and g_2 without increasing the number of duplication events occurring in the triplet.

NP-Completeness

The problem of computing the minimum number of duplications over the set of all gene trees refining a given gene tree using the soft parsimony mapping can be formulated as follows.

Uncertain Gene Tree (UGT). Instance: A species tree S, a gene tree G such that $\Sigma(G) = L(S)$, and an integer D. Question: Does a gene tree G' exist such that G' refines G and $Dup^{(S,G')} \leq D$?

Theorem A4. UGT is NP-complete.

Proof. UGT belong to NP since given an instance of the problem and a certificate in the form of a binary gene tree G', the answer of the question is $Dup^{(S,G')} \leq D$ which can be computed in polynomial time. To prove it to be NP-hard, we reduce the Partition into Cliques problem (Garey and Johnson 1979) to UGT.

The Partition into Cliques problem. Given an undirected graph $P = (V, E)$ and a positive integer $K \leq |V|$, can the vertices in P be partitioned into $k \leq K$ disjoint sets $V_1, V_2, \ldots, V_k$ such that for each V_i, $i = 1, 2, \ldots, k$, the subgraph $P_i = (V_i, E_i)$ induced by V_i is a clique (Garey and Johnson 1979)?

Let $P = (V, E)$ be any graph. Moreover, let S be a species tree and G a gene tree, for which there exists a bijection $\gamma{:}V \to C(r(G))$ such that for all $v_i, v_j \in V$, it holds that $\left|Z(\gamma(v_i)) \cap Z\left(\gamma\left(v_j\right)\right)\right| = 0$ if and only if the edge (v_i, v_j) exists in the edge set of P. A rooted gene tree G and a rooted species tree S that satisfy these conditions can be constructed as follows. Let $P' = (V', E')$ be the complement of P. Let the species tree S have one internal vertex, the root $s = r(S)$, such that the child set of s is the leaves of S, i.e., $C^S(s) = L(S)$, and define an injective function ρ that maps the edges in the graph P to the leaves of the species tree, i.e., $\rho{:}E' \to L(S)$. The gene tree G has root vertex $g = r(G)$, where $\left|C^G(g)\right| = |V'|$ and $\gamma(v_i) = g_i$ for all $g_i \in C^G(g)$. For every $g_i \in C^G(g)$ add a gene vertex g_{ir} for every edge $e_r = (v_i, v_j) \in E'$ and set $\sigma(g_{ir}) = \rho(e_r)$. For all $g_i \in C^G(g)$ such that $\left|C^G(g_i)\right| < 2$ add a gene vertex g_{ij} under g_i and a species vertex s_j under the root vertex s, and set $\sigma(g_{ij}) = s_j$. Furthermore, for any certificate $V_1, V_2, \ldots, V_k$, a certificate G' can be constructed such that $Dup^{(S,G')} + 1 \leq |\{V_1, V_2, \ldots, V_k\}|$. Construct a gene tree G'' refining G, by, for every $V_i \in \{V_1, V_2, \ldots, V_k\}$, adding a gene vertex g'_i under the root $g' = r(G'')$ and, for every $v_r \in V_i$, adding $G_{\gamma(v_r)}$ under g'_i. Then let G' be any binary gene tree refining G''. Thus, for any graph P and any positive integer $K \leq |V|$, an integer $D = K - 1$, a gene tree G, and a species tree S can be constructed in polynomial time such that UGT gives the same answer as Partition into Cliques.

Acknowledgments. We are grateful to Jens Lagergren for helpful discussions and also to three anonymous reviewers for their comments. This work was funded by Vetenskapsrådet, the Swedish Foundation for Strategic Research, and FUGE, the Norwegian national functional genomics platform.

References

Arvestad L, Berglund AC, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19:I7–I15

Arvestad L, Berglund AC, Lagergen J, Sennblad B (2004) Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence analysis. RECOMB 2004:326–335

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) Genbank. Nucleic Acids Res 33:D34–D38

Blanchette M, Green ED, Miller W, Haussler D (2004) Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res 14:2412–2423

Chen K, Durand D, Farach-Colton M (2000) Notung: a program for dating gene duplications and optimizing gene family trees. J Comput Biol 7:429–447

Cotton JA, Page RDM (2002) Going nuclear: Gene family evolution and vertebrate phylogeny reconciled. Proc Roy Soc London 269:1555–1561

Durand D, Halldorsson BV, Vernot B (2005) A hybrid micro-macroevolutionary approach to gene tree reconstruction. RECOMB 2005:250–264

Eulenstein O, Mirkin B, Vingron M (1998) Duplication-based measures of difference between gene and species trees. J Comput Biol 5:135–148

Francino MP (2005) An adaptive radiation model for the origin of new gene functions. Nature Genet 37:573–577

Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866–873

Garey MR, Johnson DS (1979) Computers and intractability, a guide to the theory of NP-completeness. Freeman, New York

Goodman M, Czelusniak 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 28:132–163

Grasso C, Lee C (2004) Combining partial order alignment and progressive multiple sequence alignment increases alignment

speed and scalability to very large alignment problems. Bioinformatics 20:1546–1556
Guigo R, Muchnik I, Smith TF (1996) Reconstruction of ancient molecular phylogeny. Mol Phylogenet Evol 6:189–213
Hallett MT, Lagergren J (2000) New algorithms for the duplication-loss model. RECOMB 2000:138–146
Hallet M, Lagergren J, Tofigh A (2004) Simultaneous identification of duplications and lateral transfers. RECOMB 2004:347–356
Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755
Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5(2):R7
Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA (2001) The Adaptive Evolution Database (TAED). Genome Biol 2(8):research0028.1-0028.6
Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1–7
Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. Genetics 159:1789–1804
Ma B, Li M, Zhang LX (2000) From gene trees to species trees. SIAM J Comput 30:729–752
Maddison WP (1989) Reconstructing character evolution on polytomous cladograms. Cladistics 5:365–377
Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York
Page RDM (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst Biol 43:58–77
Page RDM (2000) Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. Mol Phylogenet Evol 14:89–106
Page RDM, Cotton JC (2000) GeneTree: a tool for exploring gene family evolution. In: Sankoff D, Nadeau J (eds) Map alignment, and the evolution of gene families. Kluwer Academic, Dordrecht, pp 525–536
Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5:28
Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA (2005) The Adaptive Evolution Database (TAED): a phylogeny-based tool for comparative genomics. Nucleic Acids Res 33:D495–D497
Siltberg J, Liberles DA (2002) A simple covarion-based approach to analyse nucleotide substitution rates. J Evol Biol 15:588–594
Zhang LX (1997) On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. J Comput Biol 4:177–187
Zmasek CM, Eddy SR (2001a) ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics 17:383–384
Zmasek CM, Eddy SR (2001b) A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17:821–828
Zmasek CM, Eddy SR (2002) RIO: Analyzing proteomes by automated phylogenomics using resamples inference of orthologs. BMC Bioinform 3:14