Inferring Species Trees From Gene Duplication Episodes

J. Gordon Burleigh
University of Florida
Department of Biology
Gainesville, FL
gburleigh@ufl.edu

Mukul S. Bansal
The Blavatnik School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
bansal@tau.ac.il

Oliver Eulenstein
Iowa State University
Dept. of Computer Science
Ames, IA
oeulenst@cs.iastate.edu

Todd J. Vision
University of North Carolina
Dept. of Biology
Chapel Hill, NC
tjv@unc.edu

ABSTRACT
Gene tree parsimony, which infers a species tree that implies the fewest gene duplications across a collection of gene trees, is a method for inferring phylogenetic trees from paralogous genes. However, it assumes that all duplications are independent, and therefore, it does not account for large-scale gene duplication events like whole genome duplications. We describe two methods to infer species trees based on gene duplication events that may involve multiple genes. First, gene episode parsimony seeks the species tree that implies the fewest possible gene duplication episodes. Second, adjusted gene tree parsimony corrects the number of gene duplications at each node in the species tree by treating the largest possible gene duplication episode as a single duplication. We test both new methods, as well as gene tree parsimony, using 7,091 gene trees representing 7 plant taxa. Gene tree parsimony and adjusted gene tree parsimony both perform well, returning the species tree after an exhaustive search of the tree space. By contrast, gene episode parsimony fails to rank the true species tree within the top third of all possible topologies. Furthermore, gene trees with randomly permuted leaf labels can imply fewer duplication episodes than gene trees with the correct leaf labels. Adjusted gene tree parsimony reflects a potentially more realistic and, at least for small data sets, computationally feasible model for counting gene duplication events than treating each duplication independently or minimizing the number of possible duplication episodes.

Categories and Subject Descriptors
J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms
Experimentation

Keywords
Gene tree / species tree reconciliation, gene tree parsimony, gene duplication, duplication episode, whole genome duplication, phylogeny.

ACM-BCB'10, August 2-4, 2010, Niagra Falls, NY, USA.

1. INTRODUCTION
With the availability of tremendous amounts of genomic sequence data, there is an unprecedented opportunity to use data from paralogous genes for phylogenetic inference. One such approach is to infer the species tree that implies the minimum number of events that cause conflict among paralogous gene trees [e.g., 1-6]. Gene tree parsimony (GTP) infers a species tree that minimizes the reconciliation cost among all gene trees, where the reconciliation cost is defined as the number of gene duplications, or duplications and losses, implied by the species tree [e.g., 6]. With incomplete gene sampling, it is difficult or impossible to distinguish gene losses from missing gene sequences, and therefore, in this paper, we define reconciliation cost strictly in terms of gene duplications. The first GTP analyses obtained promising results from relatively small data sets for eukaryotes [1], snakes, [6], sharks [7], vertebrates [8, 9], *Drosophila* [10], and whales [11]. More recent work has renewed interest in GTP for large-scale genomic data sets. For example, Sanderson and McMahon [12] showed that GTP accurately infers a plant species tree from 557 gene trees. Although their study included only 7 taxa, it demonstrated that GTP can be an effective method for incorporating a wealth of express sequence tag (EST) data in phylogenetic inference. Algorithmic advances [13], implemented in the program DupTree [14], made it possible to perform truly genome-scale GTP analyses.

While GTP has provided promising results with a variety of data sets, there are numerous questions about its performance [e.g., 10, 12, 15, 16, 17]. One criticism is that, although multiple genes or even entire genomes can be duplicated in a single event, GTP counts all gene duplications independently. For example, a whole genome duplication involving 10,000 genes would be counted as 10,000 independent single gene duplications rather than one duplication event. This may be an especially relevant criticism in studies of plants, where changes in ploidy level are estimated to be associated with 15% of angiosperm speciation events [18] and over half of the duplicate genes retained in the *Arabidopsis* genome over the last 350 million years are the result of ancient genome duplications [19]. Several authors have suggested that GTP should seek the tree that implies the fewest gene duplication events that may involve multiple gene duplications rather than the tree that implies the fewest single-gene duplications [e.g., 8, 16]. However, such analyses have never been implemented.

In the following study, we examine two new methods to infer species trees based on the number of gene duplication events rather than the number of gene duplications, and we evaluate their

performance using a plant gene data set. The first method, which we call gene episode parsimony (GEP), seeks the species tree that implies the fewest gene duplication episodes [1, 2, 20]. The second method, which we call adjusted gene tree parsimony (adjusted GTP), adjusts the gene duplication score by assuming that only the largest possible duplication episode at each node in the species tree represents a large-scale gene duplication event, and all other duplications are treated as independent events.

2. METHODS

2.1 Definitions

In this study, all gene trees and species trees are rooted and binary. We count the number of gene duplications on a gene tree given a species tree using the gene duplication model of Goodman et al. [3], which explains incompatibilities between gene trees and a species tree through gene duplications. Under the gene duplication model, the minimum number of gene duplications that are necessary to reconcile the gene trees with the species tree can be inferred from the least common ancestor mapping (lca-mapping). In particular, a node in the gene tree can be interpreted as a *duplication* if it has a child with the same lca-mapping [13, 22]. The lca-mapping associates every node in the gene tree to the most recent node in the species tree that could have contained the pre-duplication ancestral gene; however, it is important to note that duplications often could have occurred prior to the lca in the species tree [1].

There are several ways to define the possible location(s) of a gene duplication on a species tree [e.g., 1, 2, 23, 24]. For our study, we follow Guigó, Muchnik, and Smith [1] and define the bounds of a duplication as a path between the most recent species that could have contained the duplication and its parent respectively. If there is no parent, the path runs between the most recent species for the duplication and the root of the species tree [see 1, 2, 20]. By convention, we consider a duplication to map to a node when it could have occurred on the branch subtending that node. Since there is often a range of possible mappings for each duplication, the number of possible mappings for the set of all duplications can be exponentially large in the size of the input trees. The challenge is to identify a mapping that minimizes the overall number of gene duplication events. This leads to the notion of gene duplication episodes [2, 20; Fig. 1]. Following the definition of Guigó, Muchnik, and Smith [1], any set of gene duplications, from the same or different gene trees, that occur on the same node in a species tree can be counted as a single gene duplication episode as long as none of the gene duplications in the set have an ancestor-descendant relationship with each other [Fig. 1]. Guigó, Muchnik, and Smith [1] and Page and Cotton [2] introduced heuristic approaches to estimate the minimum number of gene duplication episodes, but Bansal and Eulenstein [20] and Luo et al. [21] recently described and implemented exact and efficient solutions to find the minimum number of gene duplication episodes for a collection of gene trees on a species tree.

We also may want to infer something about the sizes of the episodes, or the number of gene duplications that can be assigned to (or explained by) that episode. The size of episodes can be counted in different ways. However, given any fixed mapping (like the lca-mapping or the mapping that minimizes the number of episodes), it is possible to compute the size of largest episodes at each species node. This can be done as follows. Observe that

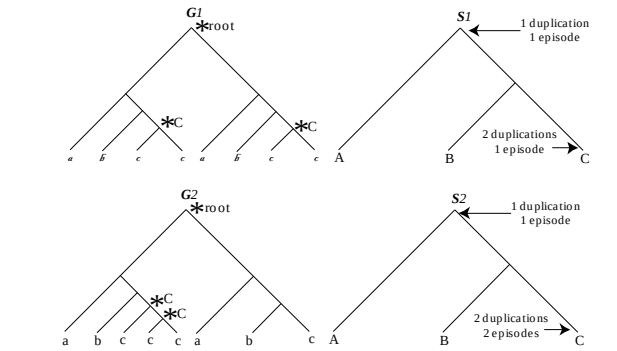

Fig. 1: Examples of gene duplication and episodes given rooted gene trees *G*1 and *G*2 and the corresponding species trees *S*1 and *S*2. Duplications in the gene trees are noted with a * followed by their location in the species tree. In both *G*1 and *G*2, there have been three gene duplications, one at the root node and two at species node *C*. In *G*1, the two duplications at species node *C* can be explained by a single duplication episode since neither is ancestral to the other. However, in *G*2, the two duplications are needed at species node *C* since one of these duplications is ancestral to the other.

the gene duplication nodes that map into the given species node induce a forest of trees based on their connectivity in the gene trees. For example, if two duplication nodes belong to the same gene tree and share a parent / child relationship then they must both be in the same tree in this forest. Every leaf node in each tree of this forest can be assigned to a single gene duplication episode. This means that the largest possible episode at that species node is the number of leaves in the forest at that species node.

For this study, we introduce two new variations of gene tree parsimony that attempt to account for gene duplication events rather than gene duplications. The first is *Gene Episode Parsimony (GEP)*. In gene episode parsimony, given a collection of gene trees, we seek the species tree that implies the fewest possible gene duplication episodes. The next variant is *adjusted Gene Tree Parsimony (adjusted GTP)*. In adjusted GTP, we adjust the GTP *reconciliation cost* (duplication score) by counting the largest possible episode at each species node as a single duplication event and counting all other duplications as single gene duplication events. For example, if there were 100 duplications at a species node, a minimum of 2 episodes, and the largest possible episode could have included 98 duplications, the reconciliation cost would be 100 for GTP, 2 for GEP, and 3 for adjusted GTP (because 98 duplications in the largest episode are counted as a single duplication event).

2.2 Gene Tree Data Set

Following Sanderson and McMahon [12], we tested the performance of the new phylogenetic methods using a tree of seven seed plant taxa, which is small enough to allow an exhaustive search of the possible tree space. In this way, we can observe the exact distribution of each objective function for all possible solutions. The taxa include one gymnosperm (*Pinus taeda*) and six angiosperms (*Oryza sativa, Solanum tuberosum, Arabidopsis thaliana, Glycine max, Lotus japonicus*, and *Medicago truculata*). These taxa also have much readily

available gene sequence data, are related by a well-accepted phylogeny [e.g., 25] that allows us to judge the accuracy of the results, and are known to have experienced several large-scale duplications since their common ancestor [e.g., 26, 27].

Amino acid alignments for gene families were obtained from Phytome v. 2, an online comparative genomics database based on publicly available sequence data [28]. The sequence assembly, clustering, and alignment protocols used by Phytome are described in detail in the online documentation (http://www.phytome.org). To ensure positional homology of columns in the alignments, alignment columns containing many gaps or very diverse amino acid sequences as well as sequences that had little overlap with other sequences were pruned from the full alignments using REAP [28, 29]. We selected all 7,091 gene family alignments with at least four sequences and sequences from at least three of the 7 seed plant taxa. These alignments included 95,589 gene sequences.

We performed maximum likelihood (ML) phylogenetic analyses on each of the masked gene family alignments using RAxML-VI-HPC version 2.2.3 [30]. The ML analyses used the JTT amino acid substitution model [31] using the default settings for the optimization of individual per-site substitution rates and classification of these rates into rate categories ["JTTMIX model"; see 30]. If the ML analysis identified multiple maximum likelihood trees, we selected the first tree that was saved.

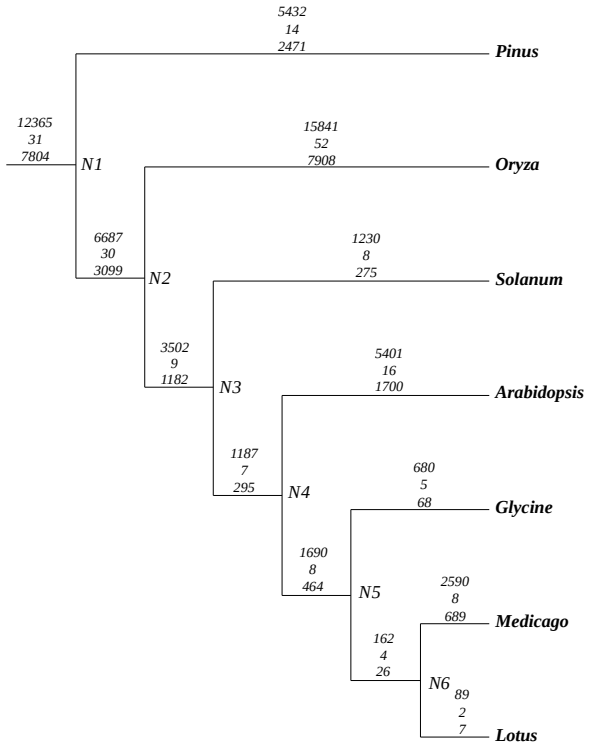

Fig. 2: The optimal species tree. Its internal nodes are labeled *N*1– *N*6. The three numbers on the branches leading to each node represent, from top to bottom, the minimum number of duplications, the minimum number of duplication episodes, and the adjusted duplication cost.

2.3 Phylogenetic Inference

GTP: GTP analyses require rooted gene trees as input. However, it is often difficult to determine the true root of a gene tree. Therefore, we examined every possible rooting of the gene trees for each species tree in order to find a rooting for each gene tree that minimizes the reconciliation cost [e.g., 12, 32]. We did this for all 10,395 possible 7-taxon rooted species trees using software that is now incorporated in DupTree [14].

GEP: We calculated the minimum number of episodes for each possible species tree using the program ExactMGD [20]. To make this analysis computationally feasible, we used a gene tree rooting that minimizes the number of duplications for each species tree. This gene tree rooting does not necessarily minimize the number of duplication episodes.

Adjusted GTP: We used the size of the largest possible gene duplication episode, calculated using ExactMGD, to determine the adjusted GTP reconciliation cost for each node in the species tree. Again, for the adjusted GTP analysis, we used gene tree rootings that minimize the total number of duplications.

2.4 Comparison to Randomized Gene Trees

To investigate the performance of the optimality scores in the absence of phylogenetic signal, we performed 1000 replicates in which we randomly permuted the leaf labels of the optimally rooted gene trees. We then calculated the number of duplications, episodes, and the adjusted duplication cost for each of the randomly permuted data sets with respect to the species tree.

3. RESULTS

3.1 Phylogenetic Inference

GTP: The conventional GTP analysis performed well; the tree with the minimum reconciliation cost (56,858 duplications) corresponds to the recognized species phylogeny [Fig. 2]. Among all possible rooted species trees, the reconciliation cost ranged from 56,858 to 65,367. The second best reconciliation cost was 43 duplications more than the optimal one. Only 8 trees had costs within 500 duplications of the optimal species tree, and only 14 had reconciliation costs within 1000 duplications.

GEP: By contrast, minimizing the number of episodes dramatically failed to return the true species tree. Among all possible species tree topologies, the minimum number of episodes ranged from 172 to 263. The true species tree implied a minimum of 194 episodes, and 3813 (36.7 %) of the 10,395 possible species trees implied fewer episodes than the true species tree. There was not an obvious relationship between the minimum number of duplications and episodes for a species tree [Fig. 3].

Adjusted GTP: In the adjusted GTP analysis, like the conventional GTP analysis, the true species tree had the minimum cost among all possible topologies [Fig. 4]. The adjusted GTP cost ranged from 25,988 to 35,795; thus, roughly half of the duplications could be clustered into multiple duplication episodes by selecting the largest episode at each node. The second best topology implied only 7 more duplication events than the optimal tree. Five trees were within 100 of the optimum, 23 trees within 500, and 34 trees within 1000. There appears to be a strongly positive and mildly nonlinear relationship between conventional GTP scores and adjusted GTP scores [Fig. 4]. Compared to the GTP analysis, top-ranked topologies under adjusted GTP are closer to the optimum and low-ranked topologies are further from

the optimum.

3.2 Comparison to Randomized Gene Trees

All 1,000 data sets in which the gene tree leaves were randomly permuted had a higher GTP score than the original data set [Table 1]. The average increase in reconciliation cost was approximately 3.2% of the optimum for GTP and 36.2% for adjusted GTP. These increases were due almost exclusively to the large increases in the number of duplications mapped to the deepest nodes [*N* 1 and *N* 2; Fig. 1]. Surprisingly, for GEP, the minimum number of duplication episodes was an average of 13.4% lower for randomized gene trees than for the original data set. While the minimum number of episodes did increase on the deepest nodes, as expected, it was more than compensated by the decrease in the number of episodes closer to the leaves.

Table 1. Effect of randomizing leaf labels on the gene trees on the inferred number of gene duplications and duplication episodes. The node names correspond to the nodes labeled on the species tree (Fig. 2). The numbers show the average difference (relative to the original unpermuted data) in the number of duplications, episodes, and adjusted duplications inferred at each node in the 1000 replicates in which the leaf labels of the gene trees were randomly permuted.

Node	Duplications	Episodes	Adjusted Duplications
N1	**20013**	**16**	**15925**
N2	**5932**	**2**	**3779**
N3	**-1570**	**4**	**-466**
N4	**28**	**3**	**59**
N5	**-1391**	**-2**	**-376**
N6	**-118**	**-2**	**-20**
Pinus	**-4388**	**-9**	**-2321**
Oryza	**-9109**	**-15**	**-4743**
Solanum	**-1020**	**-6**	**-266**
Arabidopsis	**-4078**	**-6**	**-1492**
Glycine	**-510**	**-3**	**-61**
Medicago	**-1919**	**-5**	**-631**
Lotus	**-51**	**-1**	**18**
Total	**1819**	**-23**	**9405**

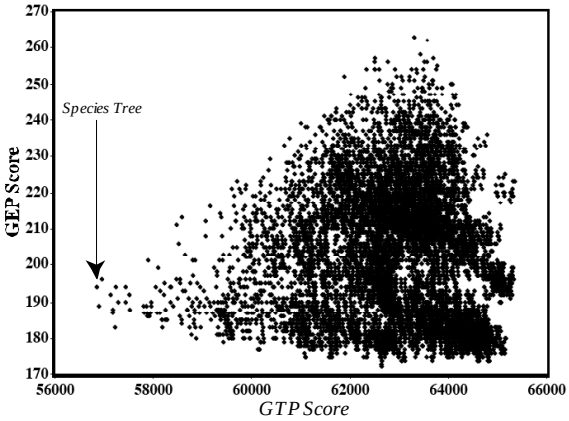

Fig. 3: Relationship between the score (reconciliation cost) inferred in the GTP and GEP analyses from each of the 7-taxon rooted species trees. The true species tree (Fig. 2) implies the lowest score for GTP, but 3813 species trees have a lower (better) score for GEP.

4. DISCUSSION

While the assumption of independence of gene duplications in gene tree-species tree reconciliation has long been recognized as problematic [e.g., 1], its practical consequences for phylogenetic inference in organisms that frequently undergo whole genome duplications, like plants, is not well understood. A host of studies in diverse sets of organisms have found that GTP, under the assumption of duplication independence, performs remarkably well [e.g., 7, 8, 9, 11, 12]. Our results reinforce the conclusion that GTP is robust to violations of the assumption of independence since it returned the true species tree as a unique solution despite strong evidence for multiple rounds of whole genome duplication in the history of seed plants [e.g., 26, 27].

In contrast, minimizing the number of gene duplication episodes [1, 2] performs poorly as an optimization criterion for species tree inference. Over a third of the possible species trees require fewer gene episodes than the true species tree [Fig. 2]. Perhaps even more troubling, one can improve the reconciliation cost (reduce the minimum number of episodes) by randomly permuting the leaf labels in the gene trees, essentially removing the phylogenetic signal from the input data. One reason for this result might be that the minimum number of episodes at each node in the species tree is the largest number of episodes at that node implied by any single gene tree. This will, in turn, be driven by a small number of large gene trees. While gene tree error results in over-counting duplications near the root of the species tree [33, Table 1], the large number of duplications that wrongly map near the root in small, error-prone gene families will not substantially elevate the GEP score. In fact, the addition of error can actually decrease the GEP score by reducing the number of episodes needed near the leaves.

Several other methods that, either directly or indirectly, attempt to determine the minimum number of gene duplication events across a collection of gene trees will likely suffer from some of the same weaknesses as GEP, as well as posing their own difficulties. One indirect approach is to minimize the number of locations (nodes) on the species tree where gene duplications occur [1, 2, 34]. As Bansal and Eulenstein [20] demonstrated, this does not necessarily minimize the number of duplication episodes. Furthermore, with large enough data sets, all nodes in the species tree likely will contain duplications, and thus all possible gene tree mappings will be equally optimal. Fellows, Hallet, and Stege [23] introduced another formulation of the episode problem in which there is a different range of possible mappings for each gene duplication, but this problem is intrinsically difficult to solve.

The failure of GEP leads us to question the biological relevance of inferring the minimum number of duplication

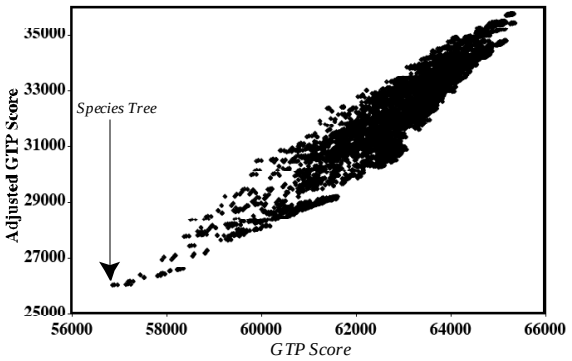

Fig. 4: Relationship between the score (reconciliation cost) in the GTP and adjusted GTP analyses inferred from each of the possible 7-taxon rooted species trees. The true species tree (Fig. 2) implies the lowest (optimal) score for both methods.

episodes. Although it is possible to calculate the minimum number of episodes needed to reconcile a set of gene trees with a species tree [20, 21], it likely has little or no relationship to the actual number of independent duplication events. For example, the 194 episodes required by our 7-taxon tree are likely more than an order of magnitude more whole genome duplications than have actually occurred [26]. Furthermore, among the more than 57,000 gene duplications required by the lca-mapping in our dataset, it would be very surprising if there were not many more than 194 independent gene duplication events.

The process of gene duplication likely consists of many events involving one or a few genes, and occasional, but much less frequent, events involving whole chromosomes or whole genomes. Under such a mixed model, we would expect that the number of duplication events is somewhere between the minimum number of episodes and the total number of duplications. If so, we would wish to cluster some, but not all duplications, into a relatively small number of episodes, each with a large number of duplications. The optimization criteria under this mixed model can be defined in a number of different ways. We explored only one simple, computationally feasible formulation, which we refer to as adjusted GTP, in which the largest episode at each node is counted once and all other duplications are assumed to represent independent events. Like GTP, adjusted GTP successfully identifies the correct tree as the optimum in our dataset.

The assumption underlying adjusted GTP is certainly an over-simplification, and we do not argue that the adjusted GTP score should be interpreted as the number of duplication events. However, it does provide a simple way to correct for the possible non-independence of duplications. Maere et al. [19] estimated that genome duplications accounted for 59% of the duplications in the genome of *A. thaliana*. By comparison, the adjusted GTP score is 54% of the conventional GTP score, suggesting that adjusted GTP results in a relatively realistic proportion of multiple duplication episodes to independent duplications. Interestingly, poorly ranked trees scored farther from the optimum in adjusted GTP than in conventional GTP, and adjusted GTP was by far the most sensitive of the three scoring criteria when applied to randomized input trees. These results suggest that adjusted GTP might provide better model discrimination than GTP when large-scale duplications are present, though further studies of this are warranted.

There are several other areas of future work on adjusted GTP that would be helpful. Most importantly, we need to learn if the current dataset is representative and whether there are circumstances under which adjusted GTP will outperform GTP. Even if adjusted GTP does sometimes outperform conventional GTP, the latter currently has some computational advantages. For one, recent algorithmic advances have made extremely large-scale conventional GTP feasible [14, 35, 36, 37], but currently no heuristics exist to estimate large species trees using adjusted GTP. In our study, we estimated the size of the largest episodes based on the mapping that required the fewest number of overall episodes across the tree [20]. We obtained similar results by calculating the adjusted GTP score using the episode mapping implied by the lca-mapping (data not shown), which suggests a potentially faster implementation. However, neither approach explicitly maximizes the size of the largest duplication episodes, which would be reasonable under the mixed duplication model. One can imagine further variants of adjusted GTP that relax the assumption of one episode per node. For example, one might only count episodes greater than some minimum size and allow any number of such episodes at a node.

5. ACKNOWLEDGMENTS

The authors would like to acknowledge funding from NSF EF-0334832, NSF DB-0227314, NSF DB-0830012, NIH R01-GM078991, and NESCent (NSF EF-0423641).

6. REFERENCES

[1] Guigó, R., Muchnik, I., and Smith, T.F. 1996 Reconstruction of ancient molecular phylogeny. Mol Phylogenet Evol. 6, 189–213.

[2] Page, R. D. M. and Cotton, J. A. 2002 Vertebrate phylogenomics: reconciled trees and gene duplications. Pacific Symposium on Biocomputing 536–547.

[3] Goodman, M., Czelusniak, J., Moore G. W., Romero-Herrera, A. E., and 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.

[4] Maddison, W. P. 1997 Gene trees in species trees. Syst Biol. 46, 523–536.

[5] Page, R. D. M. and Charleston, M. A. 1997 From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol. 7, 231-240.

[6] Slowinski, J. B., Knight, A., and Rooney, A. P. 1997. Inferring species trees from gene trees: A phylogenetic analysis of the elapidae (serpentes) based on the amino acid sequences of venom proteins. Mol Phylogenet Evol. 8, 349–362.

[7] Martin, A. P. and Burg, T. M. 2002 Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. Syst Biol. 41, 570–587.

[8] Page, R. D. M. 2000 Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. Mol Phylogenet Evol. 14, 89–106.

[9] Cotton, J. A. and Page, R. D. M. 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. P Roy Soc Lond B. Biol. 269, 1555-1561.

[10] Cotton, J. A. and Page, R. D. M. 2004. Tangled tales from multiple markers: reconciling conflict between phylogenies to build molecular supertrees. In: Bininda-Emonds ORP, editor. Phylogenetic supertrees: combining information to reveal the tree of life. Dordrecht, Netherlands: Springer-Verlag. p. 107-125.

[11] McGowen, M. R., Clark, C., and Gatesy, J. 2008 The vestigial olfactory receptor subgenome of odontocete whales: phylogenetic congruence between gene-tree reconciliation and supermatrix methods. Syst. Biol. 57, 574-590.

[12] Sanderson, M. J., and McMahon, M. M. 2007 Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol Biol. 7, S3.

[13] Bansal, M. S., Burleigh, J.G., Eulenstein, O., and Wehe, A. 2007 Heuristics for the gene-duplication problem: A $\theta(n)$ speed-up for the local search RECOMB 2007, LNCS 4453, 238-252.

[14] Wehe A, Bansal MS, Burleigh JG, Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. 24:1540-1541.

[15] Simmons, M. P. and Freudenstein, J. V. 2002 Uninode coding vs gene tree parsimony for phylogenetic reconstruction using duplicate genes. Mol Phylogenet Evol. 23, 481–498.

[16] Cotton, J. A. and Page, R. D. M. 2003 Gene tree parsimony vs. uninode coding for phylogenetic reconstruction. Mol Phylogenet Evol. 29, 298–308.

[17] Wilkinson, M., Cotton, J. A., Creevey, C., Eulenstein, O., Harris, S. R., Lapointe, F. J., Levasseur, C., McInerney, J. O., Pisani, D., and Thorley, J. L. 2005 The shape of supertrees to come: tree shape related properties of fourteen supertree methods. Syst Biol. 54, 419–431.

[18] Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L.H. 2009 The frequency of polyploidy speciation in vascular plants. Proc. Natl. Acad. Sci. USA 106, 13875-13879.

[19] Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. 2005 Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA. 102, 5454–5459.

[20] Bansal, M. S. and Eulenstein, O. 2008 The multiple gene duplication problem revisited. Bioinformatics. 24, i132-i138.

[21] Luo, C. W., Chen, M. C., Chen, Y. C., Yang, R. W. L., Liu, H. F., and Chao, K. M. 2009 Linear-time algorithms for the multiple gene duplication problems. IEEE/ACM Transactions on Computational Biology and Bioinformatics 99, 5555.

[22] Eulenstein, O. 1998 Predictions of gene-duplications and their phylogenetic development. PhD thesis, University of Bonn, Germany.

[23] Fellows, M., Hallet, M., and Stege, U. 1998 On the multiple gene duplication problem. ISAAC'98, LNCS 1533, 347-356.

[24] Doyon, J. P., Chauve, C., and Hamel, S. 2009 Space of gene/species trees reconciliations and parsimonious models. J. Comput. Biol. 16, 1399-1418.

[25] APG III. 2009 An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc. 161, 105-121.

[26] Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V. A., Ma, H., and dePamphilis, C. W. 2006 Widespread genome duplications throughout the history of flowering plants. Genome Res. 16, 738-749.

[27] Soltis, D. E., Albert, A. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., Sankoff, D., dePamphilis, C. W., Wall, P. K., and P. S. Soltis. 2009 Polyploidy and angiosperm diversification. Am. J. Bot. 96, 336-348.

[28] Hartmann, S., Lu, D., Phillips, J., and Vision, T. J. 2006 Phytome: a platform for plant comparative genomics. Nucleic Acids Res. 34, 724–730.

[29] Hartmann, S., and Vision, T. J. 2008. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol Biol. 8, 95.

[30] Stamatakis, A. 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22, 2688–2690.

[31] Jones, D.T., Taylor, W.R., and Thornton, J. M. 1992 The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.

[32] Górecki, P. and Tiuryn, J. 2007 Urec: a system for unrooted reconciliation. Bioinformatics. 23, 511–512.

[33] Hahn, M. 2007 Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol. 8, R141.

[34] Burleigh, J. G., Bansal, M. S., Wehe, A., and Eulenstein, O. 2009 Locating large-scale gene duplication events through reconciled trees: implications for identifying ancient polyploidy in plants. J. Comput. Biol. 16, 1071-1083.

[35] Bansal, M. S. and Eulenstein, O 2007. An $\Omega(n^2/\log n)$ speed-up of TBR heuristics for the gene-duplication problem. WABI 2007, LNCS 4645, 124-135.

[36] Bansal, M. S. and Eulenstein O. 2008 The gene-duplication problem: near-linear time algorithms for NNI based local searches. ISBRA 2008, LNCS 4983, 14–25.

[37] Wehe, A. and Burleigh, J. G. 2010 Scaling the gene duplication problem towards the tree of life: accelerating the rSPR heuristic search. BiCob 2010, LNCS, In press.