Phylogenetics

DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony

André Wehe¹, Mukul S. Bansal¹, J. Gordon Burleigh² and Oliver Eulenstein¹,*

¹Department of Computer Science, Iowa State University, Ames, IA 50011 and ²NESCent, Durham, NC 27705, USA

Received on November 5, 2007; revised on May 7, 2008; accepted on May 8, 2008

Abstract: DupTree is a new software program for inferring rooted species trees from collections of gene trees using the gene tree parsimony approach. The program implements a novel algorithm that significantly improves upon the run time of standard search heuristics for gene tree parsimony, and enables the first truly genome-scale phylogenetic analyses. In addition, DupTree allows users to examine alternate rootings and to weight the reconciliation costs for gene trees. DupTree is an open source project written in C++.

Availability: DupTree for Mac OS X, Windows, and Linux along with a sample dataset and an on-line manual are available at http://genome.cs.iastate.edu/CBL/DupTree

Contact: oeulensi@cs.iastate.edu

1 INTRODUCTION

With rapidly growing amounts of genomic sequence data, there is renewed interest in inferring species trees from gene trees with a history of duplication and loss (e.g. Sanderson and McMahon, 2007). There are several methods for inferring species trees from collections of potentially paralogous gene trees, including uninode coding (Simmons et al., 2000) and gene tree parsimony (GTP) (Maddison, 1997; Slowinski and Page, 1999). GTP takes a collection of rooted, binary gene trees and seeks a rooted, binary species tree with the minimum reconciliation cost for the corresponding taxa. The reconciliation cost can be the overall number of gene duplications, or gene duplications and losses, induced in the gene trees by the species tree. With incomplete gene sampling, it is difficult to distinguish gene loss from the absence of sequence data. Therefore, following Page and Charleston (1997), we define the reconciliation cost strictly in terms of gene duplications. The reconciliation cost for a collection of gene trees and a single species tree is linear time computable (Zhang, 1997). However, finding the minimum reconciliation cost across all possible species trees is an intrinsically difficult problem (Ma et al., 1998). Therefore, heuristics are commonly used to estimate a solution for GTP.

GTP has been successfully applied to phylogenetic inference in snakes (Slowinski et al., 1997), vertebrates (Page, 2000; Page and Cotton, 2002) and plants (Sanderson and McMahon, 2007). Yet the run time performance of previous GTP implementations is not suitable for large-scale studies. Currently, there are two programs, Gtp (Sanderson and McMahon, 2007) and GeneTree (Page, 1998), for the GTP approach. Gtp performs an exhaustive search of all possible species trees, and therefore, it is only suitable for analyses with very few taxa. GeneTree implements standard local search heuristics for gene tree parsimony based on the rooted Subtree Pruning and Regrafting (rSPR) tree edit operation (Semple and Steel, 2003). Due to the high time complexity of the underlying algorithm, GeneTree typically cannot infer large species trees. Furthermore, the GTP approach requires rooted input gene trees, and rooting gene trees can be difficult if there is a history of gene duplication and loss.

We introduce DupTree, a program that enables the first truly genome-scale phylogenetic analyses based on GTP, and also allows unrooted input gene trees. DupTree implements the same standard rSPR based heuristics that are found in GeneTree, but solves the local rSPR search problem much more efficiently than GeneTree using the algorithm from Bansal et al. (2007). Due to some random choices in the heuristic, the resulting species trees from DupTree and GeneTree may differ slightly. In practice, we noticed little or no difference.

Our efficiency study (Table 1) demonstrates the enormous speed-up of DupTree in comparison to GeneTree. For the study, we first generated sets of 20 random gene trees, each having the same set of taxa. Next, we conducted 6 runs, each with a different number of taxa in the 20 random gene trees (50, 100, 200, 400, 1000, 2000). Table 1 compares the resulting run times for GeneTree and DupTree. Additionally, we have run DupTree on datasets with (i) 3978 gene trees containing a total of 624 taxa (Bansal et al., 2007) and (ii) 18896 gene trees containing a total of 136 taxa. Neither of these datasets could be run with GeneTree. DupTree also implements novel features that can significantly enhance the quality of phylogenetic analyses. These features include methods to (i) examine alternate gene tree rootings, (ii) weight the reconciliation cost for each of the gene trees and (iii) define topological constraints for the resulting species tree.

In DupTree, the reconciliation cost is defined strictly in terms of gene duplications. Although there are highly efficient algorithms for search heuristics for the gene duplication objective (Bansal and Eulenstein, 2007; Bansal et al., 2007), such algorithms are unknown for objectives based on duplication loss and deep coalescence. Still, small studies under these objectives can be performed using heuristics implemented in GeneTree (Page, 2000).

2 FEATURES

The input for a DupTree analysis is a single file containing a list of the gene trees, all in Newick format. The output, in either...
Alternately, DupTree can examine the reconciliation cost of every gene tree and base its inference. Future work will include implementing a fast tree search to facilitate the incorporation of new genomic data into phylogenetic analyses. For example, there is a frequent interest in the reconciliation of each gene tree is multiplied by a 'weight' between 0 and 1, and the species tree is inferred based on the overall weighted reconciliation cost.

DupTree is designed to interact smoothly with complementary phylogenetic tools. For example, there is a frequent interest in determining where gene duplications and losses may have occurred on a species tree. The species tree from DupTree can be used, together with the given gene trees, as input for existing programs that can locate duplications and losses, such as COMPONENT (Page, 1995) and GeneTree (Page, 2003).

3 CONCLUSION

With DupTree, it is now possible to perform GTP analyses on large-scale genomic datasets across many taxa. Thus, DupTree can facilitate the incorporation of new genomic data into phylogenetic inference. Future work will include implementing a fast tree search based on 'tree bisection and reconnection (TBR)' (Bansal and Eulenstein, 2007), a parallel version of the rSPR tree search, as well as options for incorporating gene losses into the reconciliation cost.

ACKNOWLEDGEMENTS

We thank Mike Sanderson for providing extensive comments and testing of DupTree, and the reviewers for their suggestions. Funding: National Science Foundation (EF-0334832 to A.W., M.S.B. and O.E., and EF-0423641 to J.G.B.).

REFERENCES


