BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 22 2010, pages 2910–2911
doi:10.1093/bioinformatics/btq539

Phylogenetics

Advance Access publication September 21, 2010

BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis

Bret R. Larget[1,2], Satish K. Kotha[3], Colin N. Dewey[3,4] and Cécile Ané[1,2,*]

[1]Department of Statistics, [2]Department of Botany, [3]Department of Computer Sciences and [4]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI 53706, USA

Associate Editor: David Posada

ABSTRACT

Motivation: BUCKy is a C++ program that implements Bayesian concordance analysis. The method uses a non-parametric clustering of genes with compatible trees, and reconstructs the primary concordance tree from clades supported by the largest proportions of genes. A population tree with branch lengths in coalescent units is estimated from quartet concordance factors.

Availability: BUCKy is open source and distributed under the GNU general public license at www.stat.wisc.edu/~ane/bucky/.

Contact: ane@stat.wisc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 23, 2010; revised on September 14, 2010; accepted on September 17, 2010

1 INTRODUCTION

As sequencing costs continue to drop and multiple orthologous genes become easily available for a given set of individuals, phylogenetic trees are now commonly inferred from multiple loci at once. However, trees inferred from different loci are very often incongruent with each other. While some of this discordance might be explained by stochastic and technical errors (undetected paralogy or model misspecification), it has become obvious that biological processes are often at the heart of the discordance, including incomplete lineage sorting (ILS), horizontal gene transfers or hybridization.

Concatenation of all loci is known to be powerful in some cases, but also known to report inflated support values or to be misleading in other cases (Kubatko and Degnan, 2007). Several approaches now specifically account for the discordance between genes, such as MDC (Maddison, 1997), STEM (Kubatko *et al.*, 2009), BEST (Liu, 2008) and *BEAST (Heled and Drummond, 2010). The former two methods assume that each gene tree is inferred without error, while the latter two methods integrate uncertainty in gene tree estimation. These methods assume that the sole reason for discordance is ILS as modeled by the coalescent (Kingman, 1982).

The Bayesian concordance approach (BCA; Ané *et al.*, 2007) is an alternative method that integrates over gene tree uncertainty and does not make any particular assumption regarding the reason for discordance. It assumes no recombination within loci and free recombination between loci. BCA uses a non-parametric clustering of genes with information sharing across compatible genes. Its primary goal is to estimate the concordance factor (CF) of each clade, i.e. the proportion of genes that truly have the clade (Baum, 2007). The primary concordance tree is reconstructed from the clades with the largest CFs, in order to capture the main vertical phylogenetic signal (Galtier and Daubin, 2008). CFs measure the genomic support of each clade and summarize the horizontal signal: clades with moderately low CFs display relationships that are not in the primary concordance tree, but that are still true for a minority of the genome. In this note we describe BUCKy, a program that implements BCA. BUCKy version 1.4.0 includes new features added to the version used in Ané *et al.* (2007), including the estimation of a population tree with branch lengths measured in coalescent units.

*To whom correspondence should be addressed.

2 PROGRAM DESCRIPTION

BUCKy takes as input the complete tree files generated by the Bayesian analysis of each individual locus, in the format generated by MrBayes (Ronquist and Huelsenbeck, 2003). BUCKy's output from the Bayesian analysis consists of a sample of gene trees from their joint distribution, from which CFs are estimated with credibility intervals. Finally, these CFs are used (as described below) to produce the main output: a concordance tree and a population tree. BUCKy can also generate a pairwise similarity measure between loci: the posterior probability that two loci share the same tree. This matrix can help reconstruct gene clusters and detect outlier loci. An option allows the user to run the analysis on a subset of taxa, bypassing the need to re-run the lengthy Bayesian analysis of individual loci pruned to the desired taxon subset. The user may also easily skip loci that are missing one or more taxa in the desired list, and analyze only the remaining loci. For datasets with many loci and many taxa, an option allows the use of a sparse data structure to reduce the space requirement (see Supplementary Material).

The concordance and population trees: the primary concordance tree features relationships inferred to be true for a large proportion of genes. It is built as a greedy consensus: clades are ranked by their estimated CFs and included in the concordance tree one by one as long as they do not contradict a clade with a higher CF already in the tree. Degnan *et al.* (2009) showed that this greedy consensus provides an inconsistent estimate of the population tree when discordance is caused by the coalescent and if the population tree belongs to a region they called the 'too-greedy' zone. They also showed that a method based on rooted triples, called R*-consensus, is consistent.

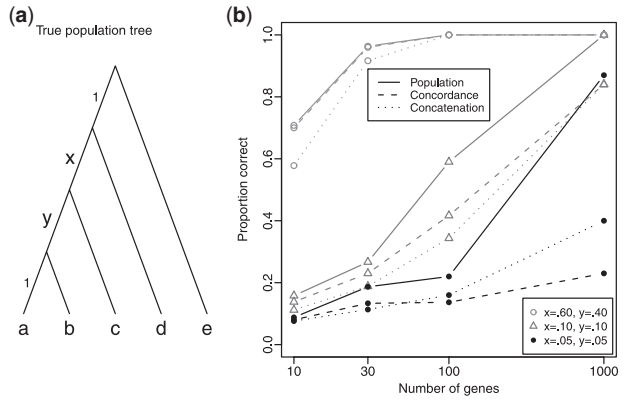

Fig. 1. (**a**) Population tree used in simulations with branch lengths in coalescent units. (**b**) Proportion of times estimated trees matched the true population tree: population tree from BUCKy (solid line), concordance tree from BUCKy (dashed line) and consensus tree from MrBayes after concatenation (dotted line).

To estimate the population tree, BUCKy implements a consensus method similar to the R*-consensus, based on unrooted quartets and which consistently identifies the species tree (Allman *et al.*, 2010). CFs are estimated from the full taxon set alignment and quartets are considered only afterwards. The posterior mean CF of each quartet is computed and transformed to an integer weight as follows, to ensure consistency. For each set of four taxa, the quartet with the largest estimated CF is favored and given weight 1, while the other two conflicting quartets are given weight 0. If CFs are estimated accurately and if the coalescent is solely responsible for the discordance, then all quartets with weight 1 must be compatible and identify the true population tree. In practice, incompatibilities between the favored quartets are resolved with the quartet-joining algorithm described by Xin *et al.* (2007), which starts from the star tree and progressively joins pairs of nodes. BUCKy currently outputs a population tree with the same set of leaves as in the gene trees (but see Supplementary Material). To estimate the coalescent units u on a branch of the population tree, BUCKy first calculates $\hat{p}$, the average posterior mean CF of all quartets defining the branch. Under the coalescent model, any such quartet has a CF of $1-2/3\exp(-u)$ (Allman *et al.*, 2010). Therefore, the branch length is estimated as $\hat{u}=-\log(3/2(1-\hat{p}))$. Note that large coalescent units are difficult to estimate numerically when the quartet CF approaches 1. In practice, for any $\hat{p}>0.99997$, $\hat{u}$ is set to a maximum of 10.

Method accuracy: from the population tree in Figure 1a, the coalescent was used to generate gene trees along which sequences of length 500 were generated with a mutation rate $\theta=0.01$. BUCKy was then used to estimate the concordance and population trees (Supplementary Material). The gene sets were also concatenated to obtain a consensus tree with MrBayes. Figure 1b shows the proportion of times these three trees matched the true population tree. All tested sets of coalescent branch lengths (x,y) correspond to a fair amount of discordance. With $x=0.6$ and $y=0.4$, the population tree is outside the 'anomaly zone' described by Degnan and Rosenberg (2006) . With $x=y=0.1$, the tree is in the anomaly zone but outside the too-greedy zone. An even greater level of discordance is obtained using $x=y=0.05$ with a tree in the too-greedy zone. In this case, BUCKy's estimated population tree is consistent but the concordance tree and the concatenation method do not estimate the true population tree consistently. Coalescent branch length estimates had a positive bias but became more accurate with more loci (Supplementary Material).

3 CONCLUSION

BUCKy is a program to combine multiple orthologous loci with potential conflict between their phylogenetic trees. The estimated primary concordance tree summarizes the vertical phylogenetic signal shared by the largest proportion of loci, while estimated CFs provide information about the horizontal signal. Although BUCKy makes no assumption regarding the reason for discordance when reconstructing gene trees and CFs, a population tree with branch lengths in coalescent units is estimated from CFs. This population tree estimation is based on and consistent under the coalescent model. The user may chose to prefer the estimated concordance tree when forces other than ILS are believed to be at work, and to prefer the estimated population tree otherwise.

ACKNOWLEDGEMENTS

We thank David Baum for helpful discussions, and Sarah Friedrich for her skilled help with website design.

Funding: National Science Foundation (DEB-0949121); WARF grant from the University of Wisconsin–Madison.

Conflict of Interest: none declared.

REFERENCES

Allman,E.S. *et al.* (2010) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.*, [Epub ahead of print; doi: 10.1007/s00285-010-0355-7, July 23, 2010].

Ané,C. *et al.* (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.*, **24**, 412–426.

Baum,D.A. (2007) Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*, **56**, 417–426.

Degnan,J.H. and Rosenberg,N.A. (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet.*, **2**, e68.

Degnan,J.H. *et al.* (2009) Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol.*, **58**, 35–54.

Galtier,N. and Daubin,V. (2008) Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. B*, **363**, 4023–4029.

Heled,J. and Drummond,A.J. (2010) Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, **27**, 570–580.

Kingman,J.F.C. (1982) The coalescent. *Stoch. Proc. Appl.*, **13**, 235–248.

Kubatko,L.S. and Degnan,J.H. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.*, **56**, 17.

Kubatko,L.S. *et al.* (2009) Stem: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, **25**, 971–973.

Liu,L. (2008) Best: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, **24**, 2542–2543.

Maddison,W.P. (1997) Gene trees in species trees. *Syst. Biol.*, **46**, 523–536.

Ronquist,F. and Huelsenbeck,J.P. (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Xin,L. *et al.* (2007) A new quartet approach for reconstructing phylogenetic trees: quartet joining method. In Lin,G. (ed) *Computing and Combinatorics*, Vol. 4598 of *Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg, pp. 40–50.