BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 22 2012, pages 2994–2995
doi:10.1093/bioinformatics/bts548

Phylogenetics

Advance Access publication September 14, 2012

DLRS: gene tree evolution in light of a species tree

Joel Sjöstrand[1,2,*], Bengt Sennblad[1,3], Lars Arvestad[1,2,4] and Jens Lagergren[1,5,*]

[1]Science for Life Laboratory, Stockholm University, SE-17121, Solna, [2]Department of Numerical Analysis and Computer Science, Stockholm University, SE-100 44, Stockholm, [3]Atherosclerosis Research Unit, Department of Medicine, Karolinska Institutet, SE-171 77, Stockholm, [4]Swedish e-Science Research Centre and [5]School of Computer Science and Communications, KTH Royal Institute of Technology, SE-100 44, Stockholm, Sweden

Associate Editor: David Posada

ABSTRACT

Summary: PrIME-DLRS (or colloquially: 'Delirious') is a phylogenetic software tool to simultaneously infer and reconcile a gene tree given a species tree. It accounts for duplication and loss events, a relaxed molecular clock and is intended for the study of homologous gene families, for example in a comparative genomics setting involving multiple species. PrIME-DLRS uses a Bayesian MCMC framework, where the input is a known species tree with divergence times and a multiple sequence alignment, and the output is a posterior distribution over gene trees and model parameters.

Availability and implementation: PrIME-DLRS is available for Java SE 6+ under the New BSD License, and JAR files and source code can be downloaded from http://code.google.com/p/jprime/. There is also a slightly older C++ version available as a binary package for Ubuntu, with download instructions at http://prime.sbc.su.se. The C++ source code is available upon request.

Contact: joel.sjostrand@scilifelab.se or jens.lagergren@scilifelab.se.

Supplementary Information: PrIME-DLRS is based on a sound probabilistic model (Åkerborg *et al.*, 2009) and has been thoroughly validated on synthetic and biological datasets (Supplementary Material online).

Received on June 10, 2012; revised on August 14, 2012; accepted on September 3, 2012

1 INTRODUCTION

The surge in Bayesian phylogenetic methods over the last decade can be attributed to their ability to achieve more realistic models as well as a solid track record for resolving outstanding biological issues (Huelsenbeck *et al.*, 2001). Recently, the interplay between gene family evolution and species phylogenies has gained particular attention, partly owing to the information that gene trees may confer on unresolved species relationships and also because it can shed light on orthology and paralogy of genes. With sophisticated conventional phylogenetic models burgeoning, it is therefore not surprising that there is a growing interest to extend them to a setting entailing both gene trees and species trees.

*To whom correspondence should be addressed.

2 MODEL

In Arvestad *et al.* (2009), we presented a probabilistic model for gene family evolution. This work was extended in Åkerborg *et al.* (2009) to create the model *DLRS* (formerly *GSR*), in which the evolutionary process inheritantly relies on a species tree while allowing for a relaxed molecular clock. The acronym captures the key concepts underlying the model:

- Duplication events.
- Loss events.
- Rate heterogeneity.
- Sequence evolution.

Figure 1D illustrates the model: starting with a single lineage at the most ancient point in the species tree, a gene family evolves down by means of duplication, loss and speciation events. Duplications and losses are governed by duplication rate δ and loss rate μ—akin to a birth–death process—whereas lineages branch deterministically at speciations. Lineages that fail to reach the leaves of the species tree are pruned away. Consequently, the model properly accounts for species evolution to guide a gene family's evolution in a way that, for instance, a Yule process or a uniform distribution for gene tree divergence times do not. The resulting clock-like gene tree is then relaxed by applying *iid* rates on each branch in order to generate non-clock-like branch lengths on which a substitution model of choice acts to produce family sequences.

3 FEATURES

To accompany the DLRS model, we have developed the PrIME-DLRS software package as a part of the PrIME suite of phylogenetics tools (http://prime.sbc.su.se). PrIME-DLRS uses a Bayesian MCMC framework to—given a species tree S with divergence times t and a multiple sequence alignment D—sample from the posterior distribution of gene trees G, branch lengths l and model parameters θ (encompassing δ, μ, etc.). Using the Metropolis–Hastings algorithm and the factorization,

$$p[G, l, \theta | D, S, t] = \frac{P[D|G, l]p[G, l|\theta, S, t]p[\theta]}{P[D|S, t]},$$

it suffices to compute the right-hand side numerator. $P[D|G, l]$ can be computed with Felsenstein's pruning algorithm (Felsenstein, 1981), whereas PrIME-DLRS uses a dynamic

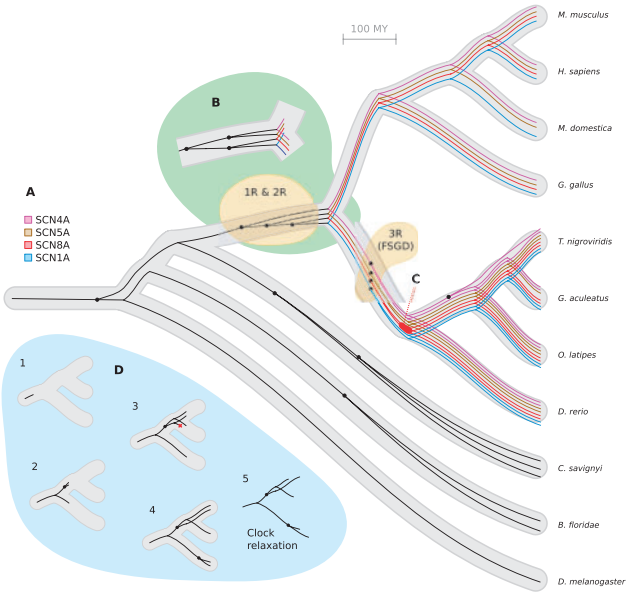

Fig. 1. (**A**) A reconciliation of an inferred gene tree for an SCN alpha subfamily obtained from Widmark *et al.* (2011). Note, however, that during an MCMC iteration, PrIME-DLRS will integrate over all possible reconciliations of the gene tree. (**B**) and (**C**) Ambiguous resolution of the gene tree. Posterior probability may aid in shedding light on such, more uncertain, relationships. This property will be exploited further in future versions of PrIME-DLRS. (**A**)–(**C**) See further Supplementary Material. (**D**) The underlying generative model of gene family evolution. Sequence evolution occurs over the relaxed final gene tree

programming approach on a discretization of S to calculate $p[G, l|\theta, S, t]$, and uniform priors are employed for $p[\theta]$. It should be noted that this approach avoids potential perils associated with including divergence times and substitution rates in the state space, instead of algorithmically integrating these into the branch lengths l.

PrIME-DLRS provides a variety of user options, including some popular predefined substitution models (e.g. Jones *et al.*, 1992; Jukes and Cantor, 1969) as well as allowing for user-defined fixed-parameter substitution models for nucleotide, protein and codon data. A gamma distribution is currently used for the *iid* edge rates. The implementation also supports modelling rate variation over sites using discrete gamma rate categories (Yang, 1994). Output consists of a tab-delimited file of samples from the posterior and can be analysed in any suitable tool, e.g. the coda package for CRAN R (Plummer *et al.*, 2006). PrIME-DLRS has been implemented in Java and C++. Compared to the original version published with Åkerborg *et al.* (2009), the newer versions are more robust and typically 4–20 times quicker (see further Supplementary Material).

4 CONCLUSION

By incorporating information on the species tree, PrIME-DLRS significantly improves on conventional species tree unaware approaches when inferring gene trees (Rasmussen and Kellis, 2010). In Åkerborg *et al.* (2009), the method was shown to be self-consistent on artificial data generated in accordance with the model. In the same article, the significance of species tree guidance and a relaxed molecular clock was demonstrated on biological data utilizing a predicted whole-genome duplication in yeast. These findings have been further corroborated by the ability to classify syntenic one-to-one orthologs in *Drosophila* (Supplementary Material). Unlike certain similar approaches that take the species tree into consideration (e.g. Rasmussen and Kellis, 2010), PrIME-DLRS requires no *a priori* derivation of clock relaxation parameters and can be run on individual gene families. Moreover, as demonstrated in the Supplementary Material, by inferring δ and μ, the model can provide estimates of genome-wide evolutionary duplication and loss rates.

The Bayesian property of yielding a posterior distribution not only implicitly brings forth a confidence measure on gene trees but also provides an important step toward probabilistic orthology analysis when considering the reconciliation of the gene trees with the species tree (Fig. 1 and Supplementary Material).

ACKNOWLEDGEMENT

We thank Dan Larhammar and colleagues for assistance in obtaining the SCN alpha dataset.

Funding: The Swedish Research Council (2010-4757 and 2010-4634 to J.L. and L.A.); Karolinska Institutet Distinguished Professor Grant to A. Hamsten (to B.S.).

Conflict of Interest: none declared.

REFERENCES

Åkerborg,Ö. *et al.* (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. USA*, **106**, 5714–5719.

Arvestad,L. *et al.* (2009) The gene evolution model and computing its associated probabilities. *J. ACM*, **56**, 1–44.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Huelsenbeck,J.P. *et al.* (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Jones,D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, **8**, 275–282.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.) *Mammalian Protein Metabolism*. Vol. III, Academic Press, New York, pp. 21–123.

Plummer,M. *et al.* (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.

Rasmussen,M.D. and Kellis,M. (2010) A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol.*, **28**, 273–290.

Widmark,J. *et al.* (2011) Differential evolution of voltage-gated sodium channels in tetrapods and teleost fishes. *Mol. Biol. Evol.*, **28**, 859–871.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.