BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 4 2007, pages 511–512
doi:10.1093/bioinformatics/btl634

Phylogenetics

URec: a system for unrooted reconciliation

Pawel Górecki[1,2,*] and Jerzy Tiuryn[1]

[1]Institute of Informatics, Warsaw University, Banacha 2, 02-678 Warsaw, Poland and
[2]Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received on November 7, 2006; accepted on December 11, 2006
Advance Access publication December 20, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: URec is a software based on a concept of unrooted reconciliation. It can be used to reconcile a set of unrooted gene trees with a rooted species tree or a set of rooted species trees. Moreover, it computes detailed distribution of gene duplications and gene losses in a species tree. It can be used to infer optimal species phylogenies for a given set of gene trees. URec is implemented in C++ and can be easily compiled under Unix and Windows systems.

Availability: Software is freely available for download from our website at http://bioputer.mimuw.edu.pl/~gorecki/urec. This webpage also contains Windows executables and a number of advanced examples with explanations.

Contact: gorecki@mimuw.edu.pl

1 INTRODUCTION

The relationships between species cannot always be inferred from a single gene family. This important property leads to the problem of reconstruction of the species tree from a set of gene families. However, it is even more complex than the task for a single family due to differences in gene families. They could be caused by gene duplications, gene losses, horizontal gene transfers, errors in sequencing or computational side-effects like determining correct parameters for alignment of gene sequences, gene tree inferring software, etc. Even if the gene family trees are reconstructed there is still the question: 'How to infer a species tree from a set of incongruent gene trees?'.

In most cases the general problem is transformed into a number of simple essential problems: 'compute (dis)similarity cost by comparing a species tree and a gene family tree'. Then, the final result is obtained by choosing the species tree with the optimal total cost.

In the fundamental step one can use a concept of '*reconciled tree*' which explains the differences between trees in terms of gene duplications and gene losses (Goodman *et al*., 1979; Guigo *et al*., 1996; Page, 1994; Page and Charleston, 1997). The model of reconciled trees (called sometimes duplication-loss model, or in short DL-model) is known to be both biologically consistent and mathematically well founded (Bonizzoni *et al*., 2003; Gorecki and Tiuryn, 2006a; Guigo *et al*., 1996; Mirkin *et al*., 1995). It seems to be promising but can be only applied to the rooted trees. Unfortunately, it is a serious limitation due to the fact that most of phylogenetic software produce unrooted gene trees (ML, MP, NJ, etc.) in which the common ancestor relation is undefined.

*To whom correspondence should be addressed.

Thus, we may consider the problem of '*unrooted reconciliation*', which can be stated as follows: 'reconcile rooted species trees with an unrooted gene family tree'. It was thoroughfully studied in our paper (Gorecki and Tiuryn, 2006b). Also, some rough ideas similar to this approach were presented in (Chen *et al*., 2000) (see Gorecki and Tiuryn, 2006b for a discussion). In Gorecki and Tiuryn, 2006b we introduced a dynamic programming algorithm for solving the problem of unrooted reconciliation. It is implemented in URec with the features allowing to reconcile sets of rooted species trees and unrooted gene trees.

The problem of reconstructing the species phylogeny for a set of gene trees in DL-model belongs to NP-complete complexity class (Ma *et al*., 2000) which suggests that the problem for unrooted variant is also very complex. In our experiments, we assume that the species trees are given. In simple cases (up to six species) it is enough to enumerate all species trees. In the next versions of this software we add features allowing to search in the species tree space (e.g. NNI approach).

2 THE PROGRAM AND ITS FEATURES

The program is implemented in C++ and can be easily compiled under Unix and Windows systems. Its time complexity is linear (in the size of input trees) for computing optimal rooting for reconciling unrooted gene tree with a rooted species tree. If the distributions are required then the complexity is quadratic.

(1) Input–output features:

- defining species or gene tree(s)—from a file, as a program argument or randomly generated,
- printing rooted variants of unrooted gene family trees and
- defining weights of gene duplications and gene losses.

(2) Main algorithms:

- finding a species tree with minimal cost for a set of unrooted gene family trees and
- finding a species tree by voting algorithm (see Gorecki and Tiuryn, 2006b for details).

(3) Computing summary of costs for every input species tree:

- total mutation cost (i.e. the total number of gene duplications and gene losses),
- total number of gene duplications or losses,

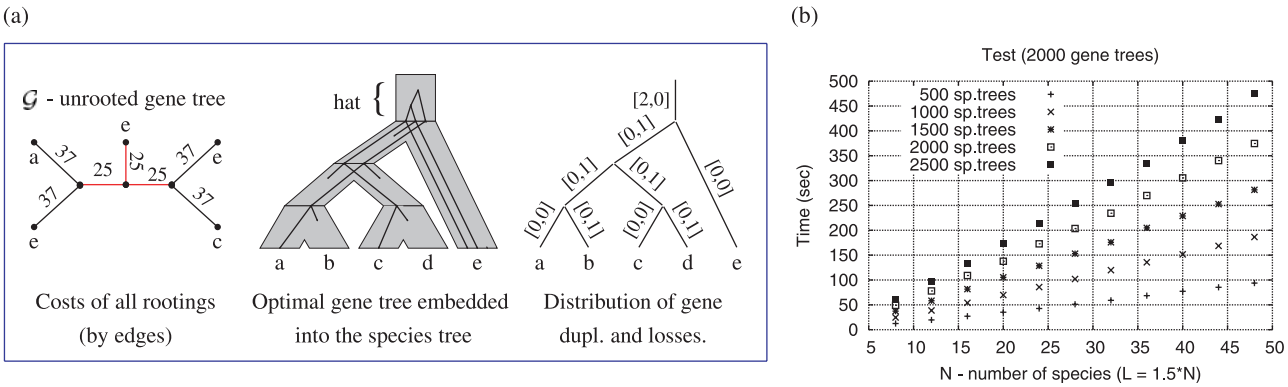

Fig. 1. (**a**) We present the fundamental step of unrooted reconciliation which is implemented in URec. The input is: an unrooted gene family tree $\mathcal{G}$ and a rooted species tree $\mathcal{S}$=(((*a*, *b*), (*c*, *d*)), *e*). The output can be: the optimal mutation cost, i.e. the total number of gene duplications and gene losses (here 7), a set of optimal rootings of $\mathcal{G}$ (the rootings are indicated by three internal edges marked by 7), an embedding for one of optimal rootings (the leftmost '7' in $\mathcal{G}$) and distributions of gene duplications and gene losses in the species tree (the numbers in square brackets denote counts of gene duplications and gene losses, respectively, associated to the lineages). (**b**) The diagram presents summary of reconciling random sets of species and gene trees, where *N* is the number of species, *L* is the number of leaves/genes in an unrooted gene tree. We computed the optimal total cost of reconciling each species tree with the set of gene trees.

- distributions of gene duplications and gene losses and
- printing distributions.

(4) For every reconciliation of an unrooted gene tree with a species tree (detailed results) computing:
- the optimal cost,
- an optimal rooted gene tree and
- a species tree with detailed evolutionary events.

3 EXAMPLES

The input trees are given in the standard nested parenthesis notation, for instance the species tree shown in Figure 1 is defined by (((*a*, *b*), (*c*, *d*)), *e*) and the unrooted gene tree by ((*a*, *e*), *e*, (*c*, e)). All input gene trees are unrooted but the binary notation is also allowed, for instance (*a*, *b*, *c*) is ((*a*, *b*,), *c*) in the context of gene trees. Some examples:

- *urec -b -O -g* '(*a*, *a*, *c*)' *-s* '((*a*, *b*), *c*)' *-cC* - reconcile ((*a*, *b*), *c*) (a species tree) with (*a*, *a*, *c*) (a gene tree); print an optimal rooted gene tree (-O); print the total cost (-c) and the numbers of gene duplications and losses (-C),
- *urec -bc -G gt.txt -S st.txt* - reconcile species trees defined in a text file *st.txt* with gene trees from file *gt.txt*; for each species tree: print the total cost (-c),
- *urec -bcd -G gt.txt -S st.txt* - reconcile species trees with gene trees; for each species tree: print the total cost (-c) and the distributions of gene duplications and gene losses in the species tree (-d),
- *urec -v -G gt.txt -S st.txt* - apply voting algorithm,
- *urec -E* 6 *-l* 1000 *-r abcd -p* - print 1000 random gene trees with six leaves with labels from {*a*, *b*, *c*, *d*}.

In Figure 1 we present a concept of unrooted reconciliation i.e. the fundamental step. From Gorecki and Tiuryn, 2006b we know that there can be more than one optimal rooting but all the optimal rootings have the same distribution of gene duplications and gene losses. Moreover, the optimal embeddings differ only in the rooted part of the tree (see 'hat' in Fig.1a).

Also, in this figure we present a test based on randomly generated gene and species trees. In this test we computed the optimal total cost of reconciling each species tree with the set of unrooted gene trees. We examined the performance of this software in order to study the dependence of run-time on the size of on input data. The experiments were performed on standard PC with Linux, AMD Athlon 64 3800+ and 1GB RAM. One easily notices the linear trends. Moreover, the whole process consists of many independent urooted reconciliations. We conclude that this algorithm can be effectively used in heuristic search or in parallel computing.

ACKNOWLEDGEMENTS

We thank Prof. Martin Vingron for constructive advise and Hannes Luz for helpful comments. Financial support is provided by KBN Grant 3 T11F 021 28.

Conflict of Interest: none declared.

REFERENCES

Bonizzoni,P., Vedova,G.D. and Dondi,R. (2003) Reconciling gene trees to a species tree. *Algorithms and Complexity*. In *Proceedings of the 5th Italian Conference (CIAC 2003)*, **2653**, 120–131.

Chen,K., Durand,D. and Farach-Colton,M. (2000) Notung: dating gene duplications using gene family trees. *RECOMB 2000*, pp. 96–106.

Goodman,M. *et al*. (1979) Fitting the gene lineage into its species lineage. A parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool*., **28**, 132–163.

Górecki,P. and Tiuryn,J. (2006) Dls-trees: a model of evolutionary scenarios. *Theoretical Computer Science*, **359**, 378–399.

Górecki,P. and Tiuryn,J. (2006) Inferring phylogeny from whole genomes. In *ECCB, 2006*, (to appear in Jan.'07).

Guigo,R. *et al*. (1996) Reconstruction of ancient molecular phylogeny. *Mol. Phy. Evol*., **6**, 189–213.

Ma,B. *et al*. (2000) From gene trees to species trees. *SIAM J. Comput*., **30**, 792–852.

Mirkin,B. *et al*. (1995) A biologically consistent model for comparing molecular phylogenies. *J. Comput. Biol*., **2**, 493–507.

Page,R.D.M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol*., **43**, 58–77.

Page,R.D.M. and Charleston,M.A. (1997) Reconciled trees and incogruent gene and species trees. *Mathematical Hierarchies and Biology, DIMACS Series in Mathematics and Theoretical Computer Science*, **37**, 57–70.