Simulation of Gene Familiy Histories

Maribel Hernandez-Rosales[1,2], Nicolas Wieseke[2], Marc Hellmuth[2] and Peter F. Stadler[1,2]

[1]Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22, D-04103 Leipzig, Germany
[2]Department of Computer Science, Univ. Leipzig, Härtelstr. 16-18, D-04107 Leipzig, Germany
{maribel,wieseke,marc,studla}@bioinf.uni-leipzig.de

Abstract *The reconstruction of the evolutionary history of large gene families has remained a hard and complex problem, which amounts to disentangling speciation events from gene duplication events. The evaluation of reconstruction algorithms is hampered, furthermore, by the lack of well-studied cases that could serve as a gold standard. We present here a simulation environment designed to generate large gene families with complex duplication histories on which reconstruction algorithms can be tested and software tools can be benchmarked.*

Keywords phylogeny, reconstruction, gene family, simulation.

1 Introduction

The way gene families and genomes evolve can be understood in detailed only when the location of gene duplication episodes in the tree of life can be deciphered. Since most genes belong to larger gene families, the analysis of the gene family histories thus plays an important role in the study of genome evolution. Empirically, one frequently observes that the tree that describes the evolution of species, the species tree, is inconsistent with the tree that is obtained from a group of genes of a gene family (the gene tree). [1] deduced that this inconsistency might be the consequence of mistaking paralogs for orthologs. Orthologous genes refer to copies of genes that reveal the phylogeny of species, while paralogous genes have been created by duplication events. Phylogeny reconstruction can help to understand how gene families evolved and to identify the chronology of duplications within a gene family of a single species. Several software tools, including `GeneTree` [2], `DupTree` [3], `NOTUNG` [4], and `AUGIST` [5] have been developed for this task. There is, however, lack of both test data and evaluation procedure to test, compare, and benchmark their performance and results. Here we present a convenient method that simulates phylogenetic processes that fulfills those needs.

2 Methods

The simulation of gene family histories starts with the generation of *species trees*. Within these rooted bifurcating trees the nodes represent species and edges their relation. Specifically, internal nodes represent ancient species whereas leaf nodes represent extant species. Given a number of species N, we generate a random tree T under the Age Model described in [6]. This model starts with a rooted tree with two leaves. In an iterative process one of the leaves is selected and two new leaves are attached to it until the tree has N leaves. This model makes use of the idea that the longer a leaf has not been involved in a speciation, the less likely it will be in the future.

The user will introduce n number of genes (gene families), which will be placed at the root of the generated species tree T. T will then be traversed in a depth first order. For each visited edge a number of events is sampled from a stochastic Poisson Process $P_{\lambda,l}$ where λ is the probability of the event to happen and l the branch length. The process may generate none, one or a series of these events: one gene gets duplicated (gene duplication), a group of genes gets duplicated (cluster duplication), the whole group of genes gets duplicated (genome duplication) and one gene of the species gets lost (gene loss). Based on the fact that when there is a gene duplication, one of the copies might be lost or become nonfunctional [7], after each visited node, the copies that were generated might be lost with a probability $\theta = \ell_1 + M\,\ell_2$, where M is the size of the gene family and ℓ_1 and ℓ_2 are user-defined rates that allow the expansion or contraction of gene families. With

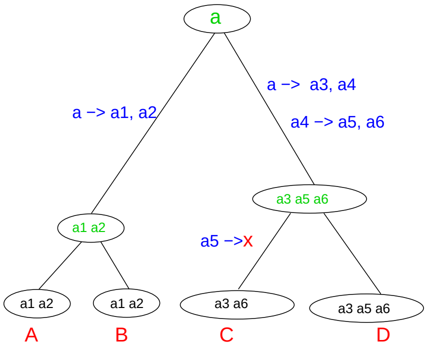

Figure 1. A one-gene family history: from a node parent to a node child, there can be both duplications and losses of genes.

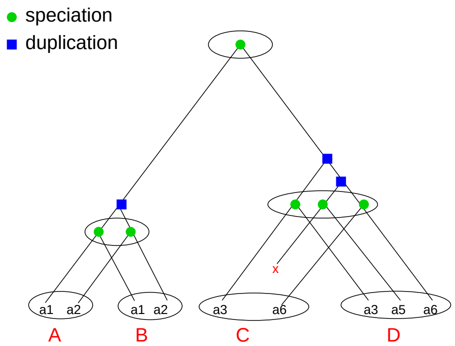

Figure 2. The reconciled tree: the gene tree embedded in the species tree. Each internal node represents an event, either an speciation or a gene duplication.

this definition of θ, large-scale duplications are considered since it is known that in the wake of multiple gene duplications and in particular for genome duplications we have to expect that many duplicated genes are rapidly lost again through the formation of pseudogenes [8]. A small example of a gene family history generated by our simulation is shown in Fig. 1. We also show the gene tree generated from the gene family history embedded in the species tree. Each leaf node represents a gene and each internal node represents an event (speciation or duplication). This tree is typically depicted as the reconciled tree as in Fig. 2.

Finally, the algorithm will generate one gene tree for each species, i.e. the pruned reconciled tree containing only genes of a certain species. Furthermore, for each gene family the orthology and homology matrices are computed. To generate the orthology matrix, we say that two genes are orthologous if their lowest common ancestor (LCA) in the reconciled tree represents a speciation event. To generate the homology matrix, a gene a from species i is homologous to gene b from species j if for every gene c from species i and every gene d from species j the $LCA(a,b) \leq LCA(c,b)$ and $LCA(a,b) \leq LCA(a,d)$.

3 Discussion

We propose an algorithm that simulates gene family histories akin to real data. This will allow reconstruction algorithms to measure their accuracy and performance. Given a certain reconstruction method one might ask if the orthology matrix could be deduced from the inferred reconciled tree or if the homology relation between the genes was predicted correctly. Furthermore it could be analysed if the method was able to infer the gene duplications and losses. A method that is able to detect large scale duplications will then identify the cluster and genome duplications generated by our algorithm.

References

[1] Goodman, M. and Czelusniak J. and Moore, G.W. and Romero-Herrera, A.E. and Matsuda, G., Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Systematic Zoology*, 28:132-163, 1979.

[2] R.D. Page,GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*,14:819–820, 1998.

[3] Wehe, A. and Bansal, M. S. and Burleigh, J. G. and Eulenstein, O., DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, 24:1540-1541,2008.

[4] Vernot, B. and Stolzer, M. and Goldman, A. and Durand, D., Reconciliation with non-binary species trees. *Comput Syst Bioinformatics Conf*, 6:441-452, 2007.

[5] Oliver, J. C., AUGIST: inferring species trees while accommodating gene tree uncertainty. *Bioinformatics*, 24:2932-2933, 2008.

[6] S. Keller-Schmidt, M. Tugrul, V. M. Eguiluz, E. Hernandez-Garcia, K. Klemm, An Age Dependent Branching Model for Macroevolution. (Submitted on 15 Dec 2010). arXiv:1012.3298v1

[7] S. Ohno, Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999, *Seminars in Cell and Developmental Biology*. 10:517-522, 1999.

[8] Sonja Prohaska, Claudia Fried, Christoph Flamm, Günter P. Wagner, Peter F. Stadler, Surveying Phylogenetic Footprints in Large Gene Clusters: Applications to Hox Cluster Duplications. *Mol.Evol.Phylog.*, 31: 581-604, 2004.