BEEP - Bayesian Estimator of Evolutionary Processes

Ann-Charlotte Berglund Sonnhammer[1], Lars Arvestad[2], Jens Lagergren[2], and Bengt Sennblad[3]

[1] Linnaeus Centre for Bioinformatics/UPPMAX, Uppsala University, Sweden, Ann-Chalotte.Sonnahammer@lcb.uu.se,
[2] Stockholm Bioinformatics Center/NADA, Royal Institute of Technology, Sweden,
[3] Stockholm Bioinformatics Center/DBB, Stockholm University, Sweden

Abstract. Orthology analysis provides the most fundamental correspondence between genes in different genomes. It is such intergenomic correspondences that provides comparative genomics with the power to translate information from one organism to another, e.g. from model organisms to human. Function prediction based on orthology is ubiquitously used by biologists. Gene tree and species tree reconstruction, as well as reconciliation are also problems important in multigenome-based comparative genomics and biology in general.
We provide tools with the capacity to perform practical orthology analysis, based on Fitch's original definition of orthology [6], and simultaneous reconstruction of gene trees and reconciliations. The tools [2, 3] use an integrated model of sequence evolution and gene duplication/gene loss, and are developed in a Bayesian framework that allows parameters to be specified by *a priori* distributions rather than by exact values. This is the first successful method ever that solves simultaneously for reconciliation and gene tree.

1 Integrated model of gene duplication/loss and sequence evolution

A reconciliation represents a hypothesis that explains how a gene tree has evolved with respect to a species tree. More specifically, a reconciliation is a mapping that associates vertices in the gene tree with the vertices in the species tree. Therefore, given such an association we can identify which genes are orthologs and which genes are paralogs under the current hypothesis. A key problem is that the number of possible hypothesis of how a gene tree has evolved with respect to a species tree is exponential. Traditionally, reconciliation-based orthology analysis has been performed in two steps. First the gene tree is reconstructed from sequence data, and then the gene tree is reconciled with the appropriate species tree such that the number of duplications is minimized, i.e., a parsimony criterion is applied. These parsimony-based methods have several weaknesses; firstly, they only consider a single gene tree, secondly, they only consider one reconciliation, the most parsimonious, and lastly, they fail to consider information

present in the species tree, e.g., divergence times. Efforts where bootstrapping have been applied to parsimony orthology analysis [8, 9], address the first of these weaknesses and corroborates the importance of expressing uncertainty in orthology analysis. Our tool, on the other hand, addresses all these problems, by combining the information in the species tree and the sequence data, not relying on any single gene tree. Moreover, our tool gives probabilities, i.e., measures of uncertainty, for orthology assignments based on all possible reconciliations.

Until now gene tree reconstruction has been performed without considering the species tree, although in many cases, reliable estimates of species trees with divergence times are available [1]. Recent advanced phylogeny programs [5] can propose alternative gene trees, but fail to account for constraints given by the species tree. In contrast, our model takes the species tree into account, and can naturally be extended to the estimation of a species tree given a selection of gene families.

In our probabilistic approach to gene tree reconstruction we use an integrated model of sequence evolution and duplication/loss. In the duplication/loss model we only consider three types of gene events: duplications, speciations and losses. By incorporating sequence evolution, we let the gene sequences evolve according to some standard sequence evolution model,see e.g. [4].

Our approach makes it possible to take into account, for example, the uncertainty due to incomplete sampling of genes. This and other strengths are illustrated with examples from real data, e.g., the histocompatibility complex (MHC) multigene family and the 60s ribosomal domain family.

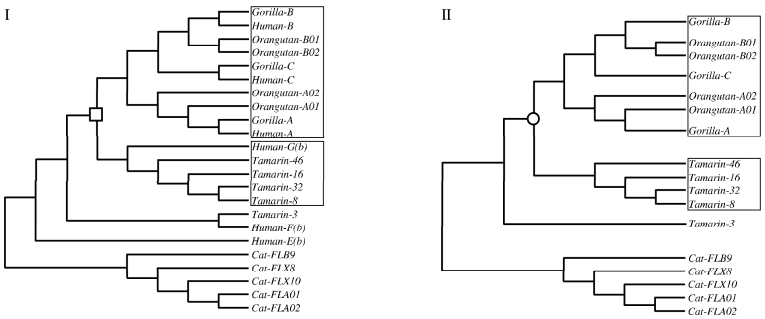

Fig. 1. The MHC class I gene trees for primate sequences extracted from [7]; the MHC class I genes for cat is included as an outgroup. The two homolog groups of interest are boxed and the status of the least common ancestor, v, of these two groups as interpreted by parsimony reconciliation is indicated. (I) The gene tree including all sequences from [7]. Parsimony reconciliation correctly identifies v as a duplication (indicated by a square). (II) The tree from (I), but with all human sequences removed, simulating that the human genome was not sampled. Parsimony reconciliation now erroneously identifies v as a speciation (indicated by a circle).

References

1. U. Arnason and A. Janke. Mitogenomic analyses of eutherian relationships. *Cytogenetic and Genome Research*, 96(1-4):20–32, 2002.
2. L. Arvestad, A.-C. Berglund, J. Lagergren, and B. Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics/ISMB'03*, 19:i7–i15, 2003.
3. L. Arvestad, A.-C. Berglund, J. Lagergren, and B. Sennblad. Gene Tree Reconstruction and Orthology Analysis Based on an Integrated Model for Duplications and Sequence Evolution. *Recomb'04*, 2004.
4. J. Felsenstein. Evolutionary trees from DNA-sequences - a maximum-likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.
5. J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, Aug 2001.
6. W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113., 1970.
7. M. Nei, X. Gu, and T. Sitnikova. Evolution by the birth-and-death process in multigene families of vertebrate immune system. *Proc. Natl. Acad. Sci. USA*, 1997.
8. C. E. Storm and E. L. Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99, Jan 2002.
9. C. M. Zmasek and Sean R. Eddy. Rio: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3(14), 2002.