6.047/6.878 Lecture 21: Phylogenomics II

Guest Lecture by
Matt Rasmussen
Orit Giguzinsky and Ethan Sherbondy

November 14, 2012

Contents

List of Figures

1 Introduction

In the previous chapter, we covered techniques for reasoning about evolution in terms of trees of descent. The algorithms we covered for tree-building, UPGMA and neighbor-joining, assumed that we were comparing fully aligned sections of sequences.

In this section, we present additional models for using phylogenetic trees in different contexts. Here we clarify the differences between species and gene trees. We then cover a framework called reconciliation which lets us effectively combine the two by mapping gene trees onto species trees. This mapping gives us a means of inferring gene duplication and loss events.

We will also present a phylogenetic perspective for reasoning about population genetics. Since population genetics deals with relatively recent mutation events, we offer the Wright-Fisher model as a tool for representing changes in whole populations. Unfortunately, when dealing with real-world data, we usually are only able to sequence genes from the current living descendants of a group. As a remedy to this shortcoming, we cover the Coalescent model, which you can think of as a time-reversed Wright-Fischer analog.

By using coalescence, we gain a new means for estimating divergence times and population sizes across multiple species. At the end of the chapter, we touch briefly on the challenges of using trees to model recombination events and summarize recent work in the field along with frontiers open for exploration.

2 Inferring Orthologs/Paralogs, Gene Duplication and Loss

There are two commonly used trees, Species tree and Gene tree.

2.1 Species Tree

Species trees that show how different species evolved from one another. These trees are created using morphological characters, fossil evidence, etc. The leaves of each tree are labeled as species and the rest of the tree shows how these species are related. An example of a species tree is shown in Figure 1.

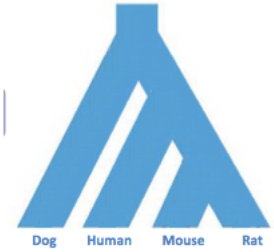


Figure 1: Species Tree

2.2 Gene Tree

Gene trees are trees that look at specific genes in different species (leaves are genes). The leaves of gene trees are labled with gene sequences or gene ids associated with specific sequences. Figure 2 shows an example of a gene tree that has 4 genes (leaves). The sequences associated with each gene are presented on the right side of Figure 2.

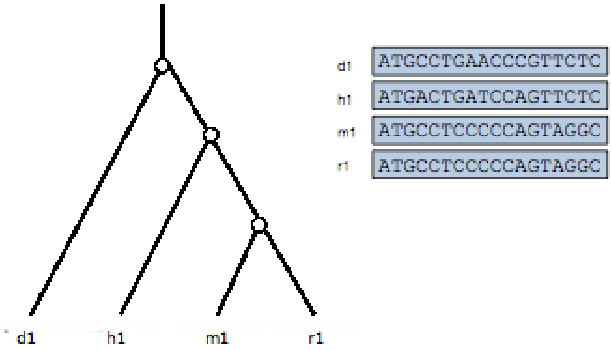


Figure 2: Gene Tree

2.3 Gene Family Evolution

Gene trees evolve inside a species tree. This section explains how we can fit gene trees inside a species trees. An example of a gene tree contained in a species tree is shown in Figure 3 below.

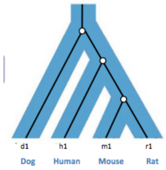


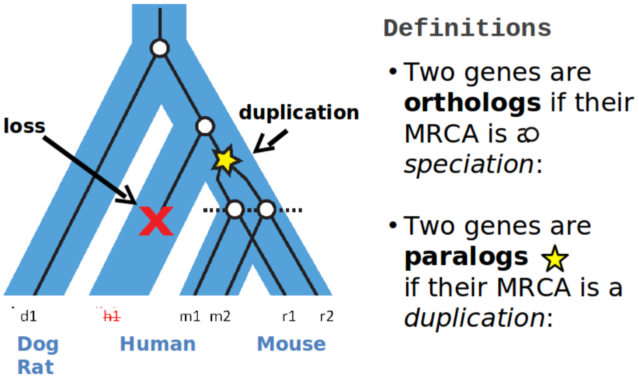
Figure 3: Gene Tree Inside a Species Tree

2.3.1 Definitions

Two genes are **orthologs** if their recent common ancestor (MRCA) is a speciation (splitting into different species).

Paralogs are genes whose MRCA is a duplication.

Figure 4 below illustrates how these types of genes can be represented in a gene tree. The tree below has 4 speciation nodes, one duplication and one loss.



Definitions

- Two genes are **orthologs** if their MRCA is a *speciation*:
- Two genes are **paralogs** ☆ if their MRCA is a *duplication*:

Figure 4: Gene Family Evolution: Gene Trees and Species Trees

2.3.2 Reconciliation

Reconciliation is an algorithm that helps determine how a gene tree fits inside a species tree by mapping the vertices in the gene tree to vertices in the species tree.

Maximum Parsimony Reconciliation (MPR) algorithm

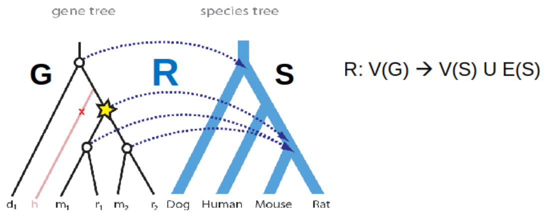
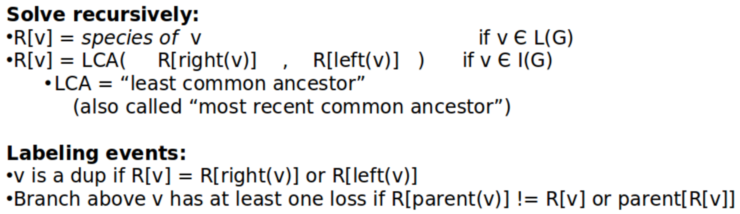


Figure 5: Maximum Parsimony Reconciliation (MPR)

Given a gene tree and a species tree, the algorithm finds the reconciliation that minimizes the number of duplications and deletions. Figure 5 above shows an example of a possible mapping from a gene tree to a species tree. Figure 6 presents the pseudocode for the MPR algorithm.



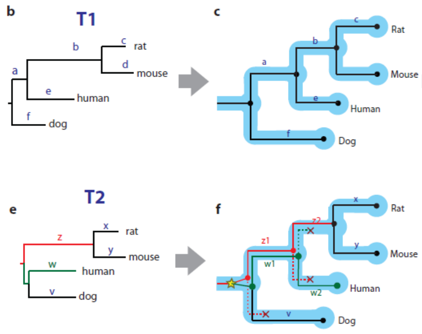
```
Solve recursively:
•R[v] = species of  v                          if v ∈ L(G)
•R[v] = LCA(   R[right(v)]   ,   R[left(v)]   )      if v ∈ I(G)
    •LCA = "least common ancestor"
      (also called "most recent common ancestor")

Labeling events:
•v is a dup if R[v] = R[right(v)] or R[left(v)]
•Branch above v has at least one loss if R[parent(v)] != R[v] or parent[R[v]]
```

Figure 6: Maximum Parsimony Reconciliation Recursive Algorithm

Duplication events map to the same as both of its children. Loss event maps to gap in the mapping. Gene tree accuracy is important; even one branch misplaced can dramatically increase error.

3 Learning Across Gene Trees



•Previous algorithms treated

- ignored what species they came from

•**Insight:** If we know the

Figure 7: Using species trees to improve gene tree reconstruction.

If we knew the species tree we could know beforehand that we expect the branch to be longer. We can develop a model for what kind of branch lengths we can expect. We can use conserved gene order to tell orthologs and build trees.

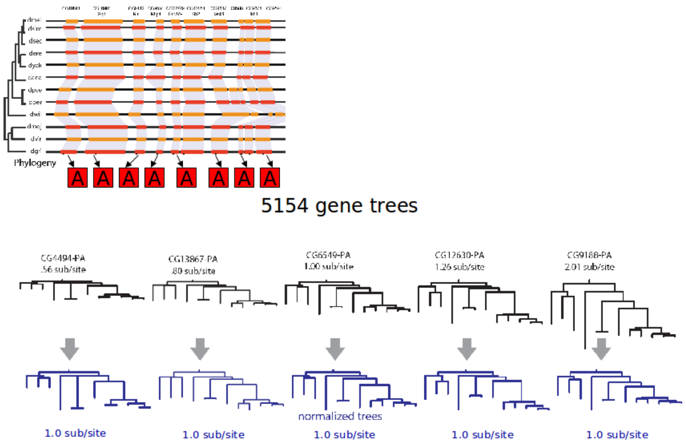


Figure 8: We can develop a model for what kind of branch lengths we can expect. We can use conserved gene order to tell orthologs and build trees.

When gene is fast evolving in one species, it is fast evolving in all species. We can model a branch length as two different rate components. One is gene specific(present across all species) and a species specific which is customized to a specific species.
This method greatly improves reconstruction accuracy.

4 Modeling Population and Allele Frequencies

With the advent of next-gen sequencing, it is becoming economical to sequence the genomes of many individuals within a population. In order to make sense of how alleles spread through a population, it's helpful to have a model to compare data against. The **Wright-Fisher** reproduction model has filled this role for the past 70 years.

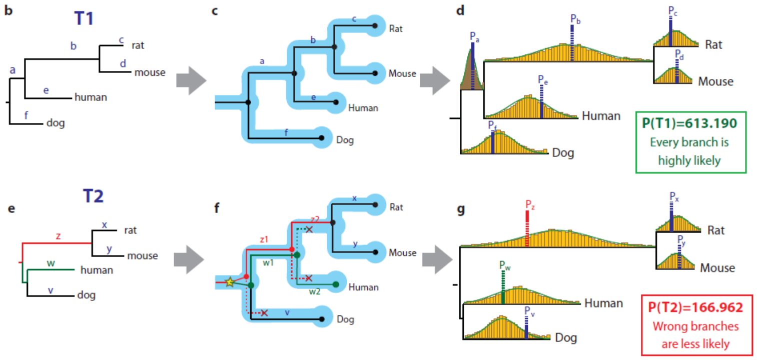


Figure 9: Branch length can be modeled as two different rate components: gene specific and species specific.

4.1 Wright-Fisher

Like HMMs, Wright-Fischer is a Markov process: at each step, the system randomly progresses, and the current state of the system depends only on the previous state. In this case, state transitions represent reproduction. By modeling the transmission of chromosomes to offspring, we can study genetic drift.

The model makes a number of simplifying assumptions:

1. Population size, **N**, is constant at each generation.

2. Only members of the same generation reproduce (no overlap)

3. Reproduction occurs at random

4. The gene being modeled only has 2 alleles

5. Genes are neutrally selected for

Note that Wright-Fisher is not an appropriate choice if you're trying to model the change in frequency of a gene that is positively or negatively selected for. If we use Wright-Fisher to model the chromosomes of diploid individuals, the population size of the model becomes **2N**.

In English, here's how Wright-Fisher works:

At every generation, for each child, we randomly select from the parents (with replaccement). The allele of the child becomes that of the randomly selected parent.

Repeat this process for many generation, with the children serving as the new parents, ignoring the ordering of chromosomes.

It really is that simple. To determine the probability of k copies of an allele existing in the child generation when it had a frequency of p in the parent generation, we can use this formula:

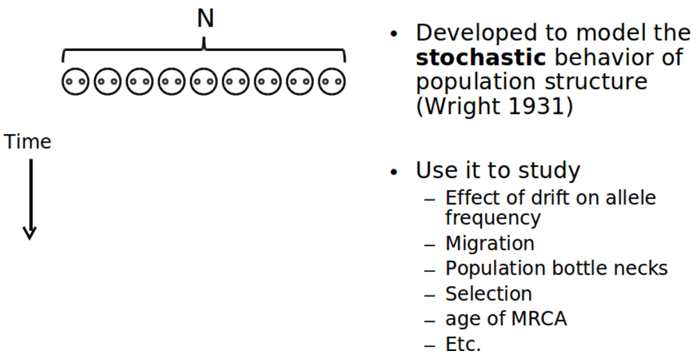
$$\binom{2N}{k} p^k q^{2N-k} \tag{1}$$

Continue for many generations and ignore ordering of chromosomes.

The coalescent model only focuses on the genealogy. It only is concerned about the lineages we have sequences for; do not have to worry about others. It is a probabilistic model that works backwards in time to find when they have common ancestors.

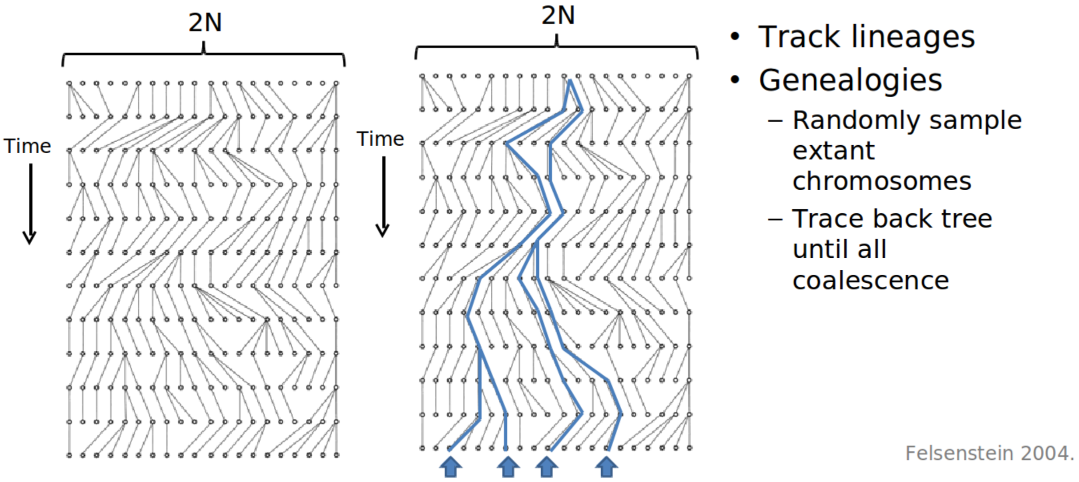
Say we have $2N$ individuals, what is the probability that k lineages do not have any coalescent events in parental generation? What is the probability that the first coalescent of k lineages is at t generations? This process can be seen as a geometric distribution.

Can repeat to find when all individuals coalesce. Each branch of species tree can be seen as having its own Wright-Fisher inside of it.



- Developed to model the **stochastic** behavior of population structure (Wright 1931)
- Use it to study
  - Effect of drift on allele frequency
  - Migration
  - Population bottle necks
  - Selection
  - age of MRCA
  - Etc.

Figure 10: The Wright-Fisher model



- Track lineages
- Genealogies
  - Randomly sample extant chromosomes
  - Trace back tree until all coalescence

Felsenstein 2004.

Figure 11: The Wright-Fisher model continued over many generations and ignoring the ordering of chromosomes.

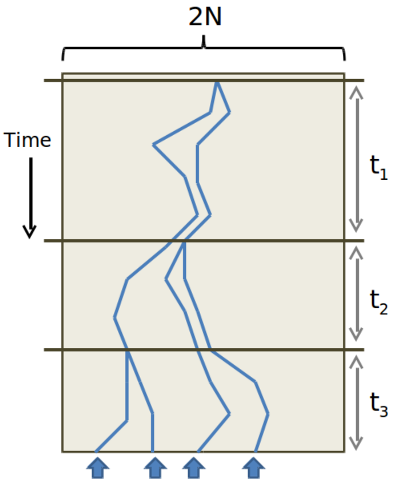
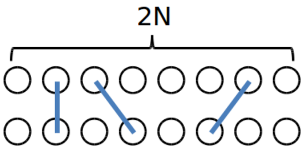


Figure 12: The coalescent model.



$$\frac{(2N-1)}{2N}\frac{(2N-2)}{2N}\cdots\frac{(2N-k+1)}{2N}=\prod_{i=1}^{k-1}\left(1-\frac{i}{2N}\right)$$

$$=1-\sum_{i=1}^{k-1}\frac{j}{2N}+O\left(\frac{1}{N^2}\right)=\boxed{1-\binom{k}{2}\frac{1}{2N}+O\left(\frac{1}{N^2}\right),}$$

For k<<N, O(1/N^2) is very small

Figure 13: Geometric probability distribution for coalescent events in k lineages.

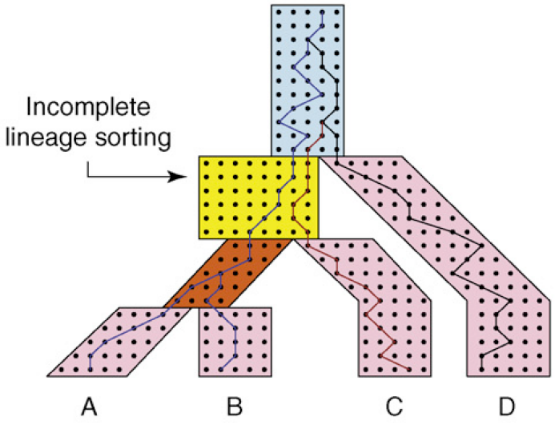


Figure 14: Multispecies Coalescent Model. Leaf branches track one lineage. There is a lag time from when population separated and when two actual gene lineages find a common ancestor. The rate of coalescent slows down as N gets bigger and for short branches. Deep coalescent is depicted in light blue for three lineages. The species and gene tree are incongruent since C and D are sisters in gene tree but not the species tree. There is a $\frac{2}{3}$ chance that incongruence will occur because once we get to the light blue section, the Wright-fisher is memory less and there is only $\frac{1}{3}$ chance that it will be congruent. Effect of incongruence is called incomplete lineage sorting.

5 SPIDIR:Background

As presented in the supplementary information for SPIDIR, a gene family is the set of genes that are descendents of a single gene in the most recent common ancestor (MRCA) of all species under consideration. Furthermore, genetic sequences undergo evolution at multiple scales, namely at the level of base pairs, and at the level of genes. In the context of this lecture, two genes are orthologs if their MRCA is a speciation event; two genes are paralogs if their MRCA is a duplication event.

In the genomic era, the species of a modern genes is often known; ancestral genes can be inferred by reconciling gene- and species-trees. A reconciliation maps every gene-tree node to a species-tree node. A common technique is to perform Maximum Parsimony Reconciliation (MPR), which finds the reconciliation R implying the fewest number of duplications or losses using the recursion over inner nodes v of a gene tree G. MPR fist maps each leaf of the gene tree to the corresponding species leaf of the species tree. Then the internal nodes of G are mapped recursively:

$R(v) = MRCA(R(right(v)), R(left(v)))$

If a speciation event and its ancestral node are mapped to the same node on the species tree. Then the ancestral node must be an duplication event.

Using MPR, the accuracy of the gene tree is crucial. Suboptimal gene trees may lead to an excess of loss and duplication events. For example, if just one branch is misplaced (as in ??) then reconciliation infers 3 losses and 1 duplication event. In [6], the authors show that the contemporaneous current gene tree methods perform poorly (60% accuracy) on single genes. But if we have longer concatenated genes, then accuracy may go up towards 100%. Furthermore, very quickly or slowly evolving genes carry less information as compared with moderately diverging sequences (40-50% sequence identity), and perform correspondingly worse. As corroborated by simulations, single genes lack sufficient information to reproduce the correct species tree. Average genes are too short and contains too few phylogenetically informative characters. While many early gene tree construction algorithms ignored species information, algorithms like SPIDIR capitalize on the insight that the species tree can provide additional information which can be leveraged for gene tree construction. Synteny can be used to independently test the relative accuracy of different gene tree reconstructions. This is because syntenic blocks are regions of the genome where recently diverged organisms have the same gene order, and contain much more information than single genes.

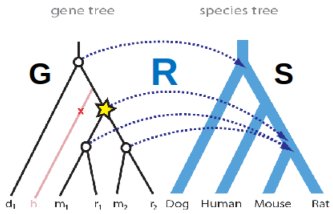


Figure 15: MPR reconciliation of genes and species tree.

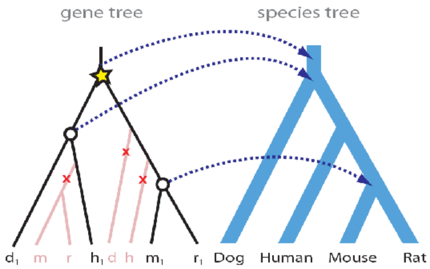


Figure 16: Inaccuracies in gene tree.

There have been a number of recent phylogenomic algorithms including: RIO [2], which uses neighbor joining (NJ) and bootstrapping to deal with incogruencies, Orthostrapper [7], which uses NJ and reconciles to a vague species tree, TreeFAM [3], which uses human curation of gene trees as well as many others. A number of algorithms take a more similar track to SPIDIR [6], including [4], a probabilistic reconciliation algorithm [8], a Bayesian method with a clock,[9],and parsimony method using species tree , as well as more recent developments: [1] a Bayesian method with relaxed clock and [5], a Bayesian method with gene and species specific relaxed rates (an extension to SPIDIR) .

6 SPIDIR: Method and Model

SPIDIR exemplifies an iterative algorithm for gene tree construction using the species tree. In SPIDIR, the authors define a generative model for gene-tree evolution. This consists of a prior for gene-tree topology and branch lengths. SPIDIR uses a birth and death process to model duplications and losses (which informs the prior on topology) and then then learns gene-specific and species-specific substitution rates (which inform the prior on branch lengths). SPIDIR is a *Maximum a posteriori (MAP)* method, and, as such, enjoys several nice optimality criteria.

In terms of the estimation problem, the full SPIDIR model appears as follows:

$argmaxL, T, RP(L, T, R|D, S, \Theta) = argmaxL, T, RP(D|T, L)P(L|T, R, S, \Theta)P(T, R|S, \Theta)$

The parameters in the above equation are: D = alignment data , L = branch length T = gene tree topology , R = reconciliation , S = species tree (expressed in times) , Θ = (gene and species specific parameters [estimated using EM training], λ, μ dup/loss parameters)). This model can be understood through the three terms in the right hand expression, namely:

1. the sequence model– $P(D|T, L)$. The authors used the common HKY model for sequence substitutions, which unifies Kimura's two parameter model for transitions and transversions with Felsenstein's model where substitution rate depends upon nucleotide equilibrium frequency.

2. the first prior term, for the rates model– $P(L|T, R, S, \Theta)$, which the authors compute numerically after learning species and gene specific rates.

3. the second prior term, for the duplication/loss model– $P(T, R|S, \Theta)$, which the authors describe using a birth and death process.

Having a rates model is very rates model very useful, since mutation rates are quite variable across genes. In the lecture, we saw how rates were well described by a decomposition into gene and species specific rates. In lecture we saw that an inverse gamma distribution appears to parametrize the gene specific substitution rates, and we were told that a gamma distribution apparently captures species specific substitution rates. Accounting for gene and species specific rates allows SPIDIR to build gene trees more accurately than previous methods. A training set for learning rate parameters can be chosen from gene trees which are congruent to the species tree. An important algorithmic concern for gene tree reconstructions is devising a fast tree search method. In lecture, we saw how the tree search could be sped up by only computing the full $argmaxL, T, RP(L, T, R|D, S, \Theta)$ for trees with high prior probabilites. This is accomplished through a computational pipeline where in each iteration 100s of trees are proposed by some heuristic. The topology prior $P(T, R|D, S, \Theta)$ can be computed quickly. This is used as a filter where only the topologies with high prior probabilities are selected as candidates for the full likelihood computation.

The performance of SPIDIR was tested on a real dataset of 21 fungi. SPIDER recovered over 96% of the synteny orthologs while other algorithms found less than 65%. As a result, SPIDER invoked much fewer number of duplications and losses.

7 Conclusion

Incorporating species tree information into the gene tree building process via introducing separate gene and species substitution rates allows for accurate parsimonious gene tree reconstructions. Previous gene tree reconstructions probably vastly overestimated the number of duplication and loss events. Reconstructing gene trees for large families remains a challenging problem.

8 Current Research Directions

9 Further Reading

10 Tools and Techniques

11 What Have We Learned?

References

[1] O. Akerborg, B. Sennblad, L. Arvestad, and J. Lagergren. Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci*, 106(14):5714–5719, Apr 2009.

[2] Zmasek C.M. and Eddy S.R. Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3(14), 2002.

[3] Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, DEhal P, Wang J, and Durbin R. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, 34, 2006.

[4] Arvestad L., Berglund A., Lagergren J., and Sennblad B. Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics*, 19 Suppl 1, 2003.

[5] M. D. Rasmussen and M. Kellis. A bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol*, 28(1):273290, Jan 2011.

[6] Matthew D. Rasmussen and Manolis Kellis. Accurate gene-tree reconstruction by learning gene and species-specific substitution rates across multiple complete genomes. *Genome Res*, 17(12):1932–1942, Dec 2007.

[7] C.E.V. Storm and E.L.L. Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99, Jan 2002.

[8] Hollich V., Milchert L., Arvestad L., and Sonnhammer E. Assessment of protein distance measures and tree-building methods for phylogenetic tree reconstruction. *Mol Biol Evol*, 22:2257–2264, 2005.

[9] Wapinski, I. A. Pfeffer, N. Friedman, and A. Regev. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, 23(13):i549–i558, 2007.