Abstract

Evolution in biology refers to changes in species and therefore of genes. Genomes of species are often evolving due to modifications, which might lead to extinctions of genes and the rise of new genes and therefore new species.

The evolutionary history of genes and species can be depicted as a tree, called a phylogenetic tree. In this tree one can observe the divergence of genes and often this tree will show the genetic events that caused this divergence. The extant genes cannot be completely independent from one another, instead, different instances of the "same gene" in different species - orthologous genes - can be used to explore their evolutionary history.

Orthologous genes are used for the reconstruction of gene family histories which has always been a challenge in phylogenetics. Orthology refers to the relationship between two genes that arose by a speciation event. Therefore, orthology detection is an important problem in comparative and evolutionary genomics and, consequently, a variety of orthology detection methods have been devised in recent years. Although many of these methods are dependent on generating gene and/or species trees, it has been shown that orthology can be estimated at acceptable levels of accuracy without having to infer gene trees and/or reconciling gene trees with species trees.

It makes sense, therefore, to look at the connection of trees and orthology from a different angle: *How much information about the gene tree, the species tree, and their reconciliation is already contained in the orthology relation between genes?*

Intriguingly, a solution to the first part of this question has already been given by [12] in a different context. In particular, they completely characterized maps of the form $(T;t)$ which they called *symbolic ultrametrics*. Moreover, in subsequent work [91] presented an algorithm that can be used to reconstruct T and t from any given symbolic ultrametric. In this thesis we investigate a new characterization of orthology relations in the field of graph theory for recovering the gene tree.

According to Fitch's definition [39], two genes are (co-)orthologous if their last common ancestor in the gene tree represents a speciation event. Otherwise, i.e., when their last common ancestor is

a duplication event, they are paralogs. The orthology relation on a set of genes is therefore determined by the gene tree T and an "event labeling" that identifies each interior vertex of T as either a duplication or a speciation event.

In the context of analyzing orthology data, the problem of reconciliating event-labeled gene trees with a species tree appears as a variant of the reconciliation problem where genes trees have no labels in their internal vertices.

When reconciling a gene tree with a species tree, it can be assumed that the species tree is correct or in the case of a unknown species tree, it can be inferred. Therefore one is interested in knowing for a given gene tree whether there exists a species tree. In this thesis we characterize event-labelled gene trees for which a species tree exists and species trees where event-labelled gene trees can be mapped.

However, reconciliation methods are not always the best options when detecting orthology. A fundamental problem is that, aside from multicellular eukaryotes, evolution does not seem to have conformed to the descent-with-modification model that gives rise to tree-like phylogenies, for example in the case of prokaryotes and viruses whose evolution involved horizontal gene transfer. To treat the problem of distinguishing orthology and paralogy within a more general framework, graph-based methods have been proposed to detect and differentiate among evolutionary relationships between genes in those organisms. In this work we introduce a measure of orthology that can be used to test graph-based methods and reconciliation methods that detect orthology.