Infer Species Phylogenies Using Self-Organizing Maps

Xiaoxu Han, Eastern Michigan University, USA

ABSTRACT

With rapid advances in genomics, phylogenetics has turned to phylogenomics due to the availability of large amounts of sequence and genome data. However, incongruence between species trees and gene trees remains a challenge in molecular phylogenetics for its biological and algorithmic complexities. A state-of-the-art gene concatenation approach was proposed to resolve this problem by inferring the species phylogeny using a random combination of widely distributed orthologous genes screened from genomes. However, such an approach may not be a robust solution to this problem because it ignores the fact that some genes are more informative than others in species inference. This paper presents a self-organizing map (SOM) based phylogeny inference method to overcome its weakness. The author’s proposed algorithm not only demonstrates its superiority to the original gene concatenation method by using same datasets, but also shows its advantages in generalization. This paper illustrates that data missing may not play a negative role in phylogeny inferring. This study presents a method to cluster multispecies genes, estimate multispecies gene entropy and visualize the species patterns through the self-organizing map mining.

Keywords: Data Mining, Entropy, Gene Trees, Phylogeny, Self-Organizing Maps, Species Trees

INTRODUCTION

With the surge of genomics, the availability of large amounts of sequence and genome data has revolutionarily changed traditional phylogenetics (Delsuc, Brinkmann, & Philippe, 2005). However, the gene tree and species tree problem still remains a fundamental problem in phylogenetics for its biological and algorithmic complexities (Ma, Li, & Zhang, 2000; Oliver, 2008; Rasmussen & Kellis, 2007). It means that the gene trees, which are phylogenies obtained from individual genes, are incongruent with the species tree. The incongruence between species trees and gene trees simply refers to the fact that they have different branch patterns in the tree topologies (Rasmussen & Kellis, 2007). The species phylogeny reflects the true evolution history of species where all lineages are results of speciation and divergence. Alternatively, a gene tree represents the evolutionary history of an individual gene for a set of organisms, which may not be the evolutionary history of species. Theoretically, it can be viewed as a hypothesis about how a gene evolves through gene duplication, gene loss and nucleotide substitution.

Although the incongruence between gene trees and species trees can be observed from all levels of taxa, it occurs more often for closely related species. For a group of species,
the probability of incongruence between a gene tree $\tau_g$ and species tree $\tau_s$ can be computed as

$$Pr(\tau_g \neq \tau_s) = \frac{2e^{-t/(2n)}}{3},$$

where $t$ is the speciation time and $n$ is the effective population size. For a fixed effective population size, the less related species have a relatively larger speciation time interval than the closely related species. Thus, the probability of incongruence between the gene trees and species tree is usually high for a set of closely related species (Nei & Kumar, 2000).

Such an incongruence may originate from biological and technical factors. Biologically, the factors like gene duplication, gene loss, lineage sorting or horizontal gene transfer would affect the evolutionary signature of a gene and cause a gene tree different from a species phylogeny (Nei & Kumar, 2000; Whitaker, McConkey, & Westhead, 2009). Technically, the factors like inappropriate substitution models, insufficient data sampling or artifacts in phylogenetic reconstruction methods also play important roles in inferring an incongruent gene tree (Huelsenbeck, 1995; Hahn, 2007). Due to its biological and algorithmic complexities, successful resolving incongruence between the gene trees and species trees may have fundamental impacts on molecular evolution and phylogenetics.

A few reconciliation models built on sequence and genome data have been proposed to resolve the gene tree and species tree problem (Liu et al., 2008; Thomson et al., 2008; Moret, Tang, & Warnow, 2005; Snel, Huynen, & Dutilh, 2005; Rokas, Williams, King, & Carroll, 2003). Two genome-based approaches among them have drawn investigators’ attentions. One is called a genome-tree approach that uses the complete genome data to build a phylogenetic tree and guarantees the final evolutionary tree is a species tree (Snel, Huynen, & Dutilh, 2005); Another is called a genome-based gene concatenation method that concatenates a set of widely distributed orthologous genes from the genomes of species interested (Rokas, Williams, King, & Carroll, 2003). Although both of them are promising in the problem-solving, they have their own limitations. The genome-tree method suffers from the relatively immature phylogenetic reconstruction models because classic evolution and phylogenetic models are build on sequence data instead of heterogeneous genome data. Moreover, although thousands of genomes are being sequenced, relatively a small number of relevant genomes are currently available for genome tree reconstruction (Delsuc, Brinkmann, & Philippe, 2005; Snel, Huynen, & Dutilh, 2005). Although the genome based gene concatenation method is welcomed by biologists for the sake of its “simplicity”, it has following limitations. Firstly, the gene concatenation number must be empirically determined for each dataset. Since building a phylogenetic tree is actually a NP-hard problem (Felsenstein, 2004; Chor & Tuller, 2006), the computational complexity for corresponding phylogenetic analysis would be prohibitive to determine the gene concatenation number, under the situation where large numbers of taxa are existed in the species phylogeny inference. Secondly, the random gene concatenation does not consider that some genes are more informative than others in the species phylogeny inference. It assumes all selected genes are equally informative in inferring a species tree. Thirdly, if the orthology criterion is not robust in data collection, then the genes used in phylogenetic reconstruction may have different evolutionary histories. The random gene concatenation under such a situation will bring noise in the phylogenetic tree reconstruction and the final tree inferred may not be robust or represent the true evolutionary relationships between organisms.

In this work, we propose a self-organizing map (SOM) (Haykin, 1999; Kohonen, 2001) based phylogeny inference algorithm, i.e., phylogenetic mining, to overcome the weakness of the genome based gene concatenation method. The idea of our algorithm is to employ a SOM mining to cluster the widely-distributed orthologous multi-species genes in the genomes interested to gain relationships between them. Then phylogenetically informative genes are selected from these clusters based on the maxi-
The Role of Big Data in Radiation Oncology: Challenges and Potentials
www.igi-global.com/chapter/the-role-of-big-data-in-radiation-oncology/121458?camid=4v1a