Title

Using Machine Learning to Detect Whole-Genome Duplications

Authors

Sofia Delgado, Michael T.W. McKibben and Michael Barker, Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, Arizona

Abstract

Whole genome duplications (WGD) are common amongst flowering plants, with each lineage having experienced on average ~3.5 in their history. However, their prevalence in the rest of the tree of life is poorly understood. Detecting WGDs is often accomplished using gene-age distributions represented as a histogram. WGDs can produce large peaks in these histograms, however the significance of these peaks is often subjectively interpreted by experts. Here, I developed an empirical approach for assessing gene-age distributions from WGD using machine learning. To generate training data for this approach, I simulated thousands of gene-trees across WGDs of different ages and rates of gene loss. I used this training data to train three machine learning models: Naive Bayes (NB), Support Vector Machine Learning (SVM) and Linear Support Vector Machine Learning (LSVM). Together, these models and the simulation framework are called Supervised Learning Estimation of Duplicated Genomes (SLEDGe). The performance of these models were then assessed using a combination of simulated data, and empirical data from land plants and insects. The best performing model was the SVM, identifying 88-93% of all the true positives depending on which dataset was used, however it detected about 70% of all true negatives. SLEDGE enables us to make rapid and accurate predictions of WGDs and can be used to explore the prevalence of WGDs in less-studied areas of the phylogeny.