Abstract

Bioinformatics houses a rich collection of research related to comparative genomics. This thesis addresses some of these issues and present several new theoretical findings, efficient algorithms, and computational aspects concerning the aforementioned research problems. This thesis first proposes a novel metric, called deformity index, to compute the correctness of a phylogenetic tree based on the existing biological knowledge of the clades. The strength of deformity index is that it does not require any complete reference tree. Therefore, this method can easily adapt itself with the present knowledge about the evolutionary relations among a set of organisms. Next, this thesis presents a novel technique viz. Genomic Footprint to represent a genomic sequence on a two-dimensional space. It follows a random walk model where the number of steps at each point is determined by different biological properties and sequence composition. Apart from that, this thesis demonstrates a novel multidimensional feature vector, called GRAFeat, and a distance function, callled GRADiF. Various statistical tests are performed to demonstrate that GRAFeat feature vector can extract evolutionary signatures from a genomic sequence and GRADiF distance function is suitable for using GRAFeat to compute distance between two sequences. Apart from that, the dimension of GRAFeat is significantly smaller than those of the state-of-the-art methods. Hence, it can be utilized for the large-scale genomic studies. To examine its performance on large-scale genomic studies, this feature vector is utilize for classifying organisms based on their taxonomy ranks. Since computation of GRAFeat feature vector requires various hyper-parameters, this thesis first suitably determines these hyper-parameters from the sequence, dynamically. Finally, various supervised machine learning-based classifiers are applied on GRAFeat feature vector. Experimental results exhibit that the performance of the feature vector is significantly better than those of the existing methods. Finally, this thesis contributes a novel deep learning framework in phylogenetic study. To address this issue, it exploits Genomic Footprint and exhibits a novel convolutional neural network (CNN)-based model for inferring phylogenetic tree. Since the availability of true phylogenetic relationships are very limited, this thesis demonstrates a novel technique to utilize the simulated data for training and validation of the model. It is observed that the trained model can infer phylogenetic trees more accurately than that of the existing methods. Experimental results and detailed analyses have been furnished to present the usefulness and effectiveness of the proposed techniques.