Scalable and Robust Clustering and Visualization for Large-scale Bioinformatics Data

During the past few years, the advances in next generation sequencing techniques has enabled analysis directly on the genetic information within a whole microbial community, bypassing the culturing individual microbial species in the lab. These techniques have led to a proliferation of raw genomic data that enables an unprecedented opportunity for data mining. To analyze such large volume bioinformatics data, a pipeline called DACIDR has been proposed. It did taxonomy-independent analysis to group these sequences into operational taxonomic units (OTUs), referred as data clustering. It enables visualization of the clustering result by utilizing large scale of computation resources by leveraging the power of parallelization and multidimensional scaling (MDS) techniques. Many challenges emerged while processing such large-scale sequence data, therefore some innovative methods are proposed in DACIDR to overcome them. Firstly, in order to observe the proximity of the sequences in a lower dimension, sequence alignment technique is applied on each pair of sequences in order to generate similarity scores in high dimension. These scores need to be assigned with weights in order to achieve an accurate result in MDS. Therefore, a robust and scalable MDS algorithm using deterministic annealing technique is proposed here to solve problems with either missing distances or a non-trivial weight function. Secondly, the dataset with millions of sequences is usually divided into two parts: one part is processed with MDS that has quadratic space and time complexity, then the other part is interpolated that result with linear time complexity. This is also referred as interpolation. In order to achieve real time processing speed, a novel hierarchical approach for it has been proposed to reduce the time complexity of interpolation to sub-linear. Thirdly, phylogenetic tree is commonly used to observe the phylogeny and evolutionary path of various organisms. Traditional way of visualizing phylogenetic tree only preserves the correlations between ancestors and their descendants. By utilizing MDS and interpolation, an algorithm called interpolative joining has been proposed to display the tree with clustering result, where their correlations can be intuitively observed in a 3D tree diagram called Spherical Phylogram. The optimizations in these three areas greatly reduce the time complexity of visualizing sequence clustering result, and increase the accuracy of it.