Abstract

Manal Helal

Linear normalised hash function for clustering gene sequences and identifying reference sequences from multiple sequence alignments

Background: Comparative genomics has put additional demands on the assessment of similarity between sequences and their clustering as means for classification. However, defining the optimal number of clusters, cluster density and boundaries for sets of potentially related sequences of genes with variable degrees of polymorphism remains a significant challenge. The aim of this study was to develop a method that would identify the cluster centroids and the optimal number of clusters for a given sensitivity level and could work equally well for the different sequence datasets. Results: A novel method that combines the linear mapping hash function and multiple sequence alignment (MSA) was developed. This method takes advantage of the already sorted by similarity sequences from the MSA output, and identifies the optimal number of clusters, clusters cut-offs, and clusters centroids that can represent reference gene vouchers for the different species. The linear mapping hash function can map an already ordered by similarity distance matrix to indices to reveal gaps in the values around which the optimal cut-offs of the different clusters can be identified. The method was evaluated using sets of closely related (16S rRNA gene sequences of Nocardia species) and highly variable (VP1 genomic region of Enterovirus 71) sequences and outperformed existing unsupervised machine learning clustering methods and dimensionality reduction methods. This method does not require prior knowledge of the number of clusters the distance between clusters, handles clusters of different sizes and shapes, and scales linearly with the dataset. Conclusions: The combination of MSA with the linear mapping hash function is a computationally efficient way of gene sequence clustering and can be a valuable tool for the assessment of similarity, clustering of different microbial genomes, identifying reference sequences, and for the study of evolution of bacteria and viruses.