Biological-based semi-supervised clustering algorithm to improve gene function prediction

Abstract:

Analysis of simultaneous clustering of gene expression with biological knowledge has now become an importanttechnique and standard practice to present a proper interpretation of the data and its underlying biology. However, commonclustering algorithms do not provide a comprehensive approach that look into the three categories of annotations; biologicalprocess, molecular function, and cellular component, and were not tested with different functional annotation database formats.Furthermore, the traditional clustering algorithms use random initialization which causes inconsistent cluster generation and areunable to determine the number of clusters involved. In this paper, we present a novel computational framework called CluFA(Clustering Functional Annotation) for semi-supervised clustering of gene expression data. The framework consists of threestages: (i) preparation of Gene Ontology (GO) datasets, functional annotation databases, and testing datasets, (ii) a fuzzy c -means clustering to find the optimal clusters; and (iii) analysis of computational evaluation and biological validation from theresults obtained. With combination of the three GO term categories (biological process, molecular function, and cellularcomponent) and functional annotation databases (Saccharomyces Genome Database (SGD), the Yeast Database at MunichInformation Centre for Protein Sequences (MIPS), and Entrez), the CluFA is able to determine the number of clusters andreduce random initialization. In addition, CluFA is more comprehensive in its capability to predict the functions of unknowngenes. We tested our new computational framework for semi-supervised clustering of yeast gene expression data based onmultiple functional annotation databases. Experimental results show that 76 clusters have been identified via GO slim dataset.By applying SGD, Entrez, and MIPS functional annotation database to reduce random initialization, performance on bothcomputational evaluation and biological validation were improved. By the usage of comprehensive GO term categories, thelowest compactness and separation values were achieved. Therefore, from this experiment, we can conclude that CluFA hadimproved the gene function prediction through the utilization of GO and gene expression values using the fuzzy c -meansclustering algorithm by cross referencing it with the latest SGD annotation.