Abstract

Subcellular location of protein is constructive information in determining its function, screening for drug candidates, vaccine design, annotation of gene products and in selecting relevant proteins for further studies. The location of protein within the cell gives insight into its function and its biochemical activities. Computational prediction of subcellular localization deals with predicting the location of a protein from its amino acid sequence. The presence of N-terminal sorting signal, amino acid composition, physiochemical properties and sequence similarity are the biological features which are widely explored for predicting the location of the protein. Most of the existing prediction methods explore only N-terminal part of the amino acid sequence as the sorting signal is present there. For a computational localization prediction method to be more accurate, it should exploit all possible relevant biological features that contribute to the subcellular localization. To incorporate more biological information, the full length protein sequences are used in this study. A new biological feature, distribution of atomic composition is effectively used with multiple physiochemical properties, amino acid composition, three part amino acid composition and sequence similarity for predicting the subcellular location of the proteins.

This thesis proposes three methods, SLP System I, SLP System II and a hierarchical prediction method for predicting the location of proteins. In SLP System I, five biological features; atomic composition, physiochemical properties, amino acid composition, three part amino acid composition and sequence similarity of the amino acid sequences are probed for the prediction of the protein's location in the cell. Five modules are designed, each for the five biological features. Support Vector Machines are employed for all the modules except sequence similarity. A weighted voting system makes the final prediction from the predictions of the individual modules. The system exhibits accuracy of 100, 82.47 , 88.81 for self consistency test, jackknife test and independent data test respectively. Another system, SLP System II, which excludes sequence similarity module also tested out to understand the strength of the new features. This approach had accuracy of 100, 81.29, 84.60
for self consistency test, jackknife test and independent test respectively. The results provide evidence that the prediction based on the biological features derived from the full length amino acid sequence gives better accuracy than those derived from N-terminal alone. Considering the features as a distribution within the entire sequence will bring out underlying property distribution to a greater detail to enhance the prediction accuracy.

In the hierarchical subcellular localization prediction, a hierarchy is imposed among the subcellular locations based on various decision points on the pathways for protein translocation. The organelles whose proteins are translocated through Endoplasmic Reticulum are clubbed together and those translocated freely through the cytoplasm are clubbed as another group. Support Vector Machine based modules are implemented at each decision making point. Atomic composition, amino acid composition and three part amino acid composition are used as biological features in various modules. In independent data test, this method exhibits an accuracy of 86.21 which is higher than SLP System II. This indicates that implementing a logical and biologically supported hierarchy among the locations will enhance the prediction accuracy.

The proposed methods are able to give higher prediction accuracy because they explore more relevant biological features. The newly proposed biological feature, atomic composition, exhibits good prediction accuracy. Exploring the full length sequence also has contributed to the improved prediction accuracy. The atomic composition, full length sequence analysis and hierarchical classification are powerful resources for enhancing the subcellular localization prediction accuracy.

List of Publications

Journals

Conference proceedings


