A Novel Approach for Prediction of Multi-Labeled Protein Subcellular Localization for Prokaryotic Bacteria

Chia-Yu Su, Allan Lo, Chin-Chin Lin, Fu Chang, Wen-Lian Hsu
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan
{cysu, allanlo, erikson, fchang, hsu}@iis.sinica.edu.tw

Abstract

We present a novel method to address multi-labeled protein subcellular localization prediction in Gram-negative bacteria using support vector machines (SVM) as classifiers. For a given protein sequence that may have more than one label, features are extracted from amino acid composition and molecular function related terms in Gene Ontology (GO) as input to SVM. We apply one-against-others SVM to proteins of Gram-negative bacteria in a 5-fold cross-validation. The results of the multi-labeled predictions are evaluated based on two criteria: class number and class category. For the first criterion, our method predicts the number of classes (class number) for each protein at an accuracy rate of 94.1%. For the second criterion, we compare the categories of the actual classes with the predicted classes proportionate to ranks, and obtain an accuracy of 83.2%. Our method is the first approach to predict and evaluate multi-labeled protein subcellular localization for prokaryotic bacteria and we demonstrate that it has a good predictive power.

1. Introduction

Advances in high-throughput genome sequencing and proteomics have generated very large amounts of uncharacterized gene and gene products. Given the sheer size and complexity of the datasets, development in automated systems is crucial in facilitating the functional annotation process. A key characteristic indicating a protein's function is its subcellular localization. Since many biological functions are limited to specific cellular compartments, knowledge of protein subcellular localization provides useful information for functional prediction of a protein. Such predictions serve to infer functions of the gene and gene products, inform and direct further experimental studies and assist in the drug discovery process.

A number of automated systems that specialize in the prediction of subcellular localization have been developed to-date. Early methods are based on the use of N-terminal sorting signals [1] and amino acid composition for feature extraction in conjunction with machine learning approaches such as neural network (NN) [2] and support vector machines (SVM) [3]. An integrative approach which combines several predictors for all localizations is PSORT [4]. The expert learning system can distinguish subcellular locations between eukaryotic and prokaryotic organisms (PSORT-B) [5]. The predicative quality of such integrative system is highly correlated with the improvement of individual predictors.

Nevertheless, there is a major limitation to the current approaches which only consider proteins with one and only one subcellular localization. Proteins that may occur in multiple localizations are not treated effectively. These proteins may simultaneously occur or move in between different cellular compartments and represent the class of proteins that are 'multi-labeled'. We describe a new approach for the prediction of multi-labeled localization for proteins of Gram-negative bacteria and a novel evaluation scheme.

2. Materials and Methods

2.1 Dataset

The set of proteins from Gram-negative bacteria used in PSORT-B is considered in this work. It is consisted of 1441 proteins with experimentally determined localizations, in which 1302 proteins are of single localization site: 248 cytoplasmic, 268 inner membrane, 244 periplasmic, 352 outer membrane and 190 extracellular; and 139 proteins of multiple localization sites: 14 cytoplasmic/inner membrane, 49 inner membrane/periplasmic and 76 outer membrane/extracellular.

2.2 Feature Extraction based on Amino Acid Composition

The coding scheme of protein sequences based on *k*-peptide compositions or their variations have been demonstrated effective in the predictions of protein subcellular localizations, in conjunction with the use of machine learning tools such as neural networks and support vector machines [6]. In order to code a sequence, a window with a length of *k* is moved along the sequence from the first amino acid to the *k*th amino acid from the end. We add symbol 'X' to the set of 20 symbols of amino acids since it may appear in some protein sequences. Therefore, a final vector of $21^2 =$ 441 dimensions are required respectively for di-peptide coding scheme.

2.3 Feature Extraction based on Gene Ontology (GO)

In order to fully utilize the core features of proteins related to their subcellular localizations, we define a protein based on the Gene Ontology (GO) [7]. In the GO database, gene products are organized into three criteria in a species-independent manner: cellular components, molecular function and biological process. We follow a similar approach as outlined by Chou and Cai to map InterPro entries to GO, but we only include molecular function related terms in our prediction [8]. The feature vector is consisted of entries with InterPro mappings to GO; if the entry is not present, a value of 0 is given.

3. Evaluation and Results

We have constructed five one-against-others SVM classifiers and applied 5-fold cross validation for testing to our datasets. For evaluation of multi-labeled localization predictions, we develop a new evaluation criterion proportionate to the ranks:

$$\textit{Category Score (CS)} = \frac{\sum_{i=1}^{n} P_i^c}{\sum_{i=1}^{n} R_i}, \qquad (1)$$

where n is the number of actual classes of a protein; P_i^c is the score given for a protein predicted as class c in the ith rank. In addition,

$$P_i^c = \begin{cases} Ri, & if\ c \in S \\ 0, & if\ c \notin S \end{cases} \quad \text{for } Ri = n\text{-}i\text{+}1 \text{ and } 1 \le i \le n \qquad (2)$$

where S is the set of actual classes of a protein; Ri is the score given for a correctly predicted protein in the ith rank. For example, a protein that has two localization sites A and B and predicted to be localized in sites A and C receives *Category Score (CS)* of (2+0) / (2+1) = 2/3. Our method successfully predicts the class number of multi-labeled proteins at an accuracy of 94.1%. Furthermore, an accuracy of 83.2% is achieved for category score of all multi-labeled proteins. For single-labeled predictions, we also achieve comparable results at the accuracy of 87.8%.

4. Conclusion

In this paper, a novel method for predicting multi-labeled protein subcellular localization in Gram-negative bacteria is introduced. We have combined feature extraction of protein sequences from amino acid composition and molecular function terms from Gene Ontology as input for SVM. The performance of our prediction is evaluated using a new multi-labeled metric and we achieve a good accuracy. Proteins with multi-labeled protein subcellular localization are a special class of protein localization prediction and further improvements in this area will lead to overall improvement in the predictive task.

5. References

[1] Nakai, K. "Protein sorting signals and prediction of subcellular localization", Adv. Protein Chem. (2000), 54:277-344.

[2] Reinhardt, A. and T. Hubbard, "Using neural networks for prediction of the subcellular location of proteins", Nucleic Acid Res. (1998), 26:2230-2236.

[3] Hua, S. and Z. Sun, "Support vector machine approach for protein subcellular localization prediction", Bioinformatics (2001), 17:721-728.

[4] Nakai, K. and M. Kanehisa, "Expert system for predicting protein localization sites in Gram-negative bacteria", Proteins (1991), 11:95-110.

[5]Gardy, J.L. et al. "PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria", Nucleic Acid Res. (2003), 31:3613-3617.

[6] Yu, C.S., C.J Lin. and J.K. Hwang, "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on *n*-peptide compositions", Protein Science (2004), 13:1402-1406

[7] Ashburner, M. et al. "Gene ontology: tool for the unification biology", Nat. Genet. (2000), 25:25-29.

[8] Chou, K.C. and Y.D Cai. "A new hybrid approach to predict subcellular localization of proteins by incorporating Gene ontology", Biochem. Biophys. Res. Commun. (2003), 311:743-747.