Protein Feature Prediction Using Chi-Square in PPI Networks

YongHo Lee
louse@etri.re.kr

JaeHun Choi
jhchoi@etri.re.kr

JongMin Park
jmpark@etri.re.kr

SunLee Bang
slbang@etri.re.kr

SooJun Park
psj@etri.re.kr

Bioinformatics Research Team, Electronics and Telecommunications Research Institute(ETRI), 161 Gajeong-dong, Yuseong-gu, Daejeon, 305-700, Korea

Keywords: protein feature prediction, Gene Ontology, PPI network, Chi-square algorithm

1 Introduction

PPI(Protein-Protein Interaction) networks are defined as the set of relationships among proteins. PPI data of the networks are very important, because they are used in predicting feature of unknown proteins. The prediction using PPI data has more reliability than other methods such as sequence or homology[1,3]. The significance of protein feature prediction is increasing more and more because they are effectively used in the high value-added bio-business field like new drug discovery. Picking out candidate proteins has to proceed for the purpose of reducing enormous expenses in biological experiments like new drug discovery.

For predicting feature of unknown proteins, we normalized Gene Ontology(GO)[5] terms and used Chi-square algorithm as a statistic method.

2 Method

Our proposed method consists of two parts. First part is preprocessing through normalization of GO terms, and second part is predicting feature of unknown proteins using Chi-square algorithm[4].

Each protein of PPI networks has GO term annotation information by biological experiments, but many proteins do not still have GO term annotations. GO term annotations could be useful to biologists by means of important information representing feature of proteins. We used Chi-square algorithm to predict this feature. This statistical method is known to be guaranteed reliability if plenty of data exist. However, GO term annotation information of proteins consists of biological process, cellular component, and molecular function, and each part is divided in diverse levels. Therefore, these raw data do not provide plenty of data for using the statistical method. For that reason, the normalization regularizing levels of GO term annotation information is preprocessed previous to the use of Chi-square algorithm. We normalized GO terms of known proteins for predicting feature of unknown proteins in PPI networks. This process changes different levels of GO terms into same levels of them by calculating similarity among terms. The similarity among GO terms is calculated by distance between two terms[2]. For instance, 4-level normalization makes different levels of GO terms into 4-levels of GO terms with calculated similarity. We assume that GO term A1 of protein P1 is level 4 and that GO term A2 of protein P2 is level 5 in 4-level normalization. After normalization, A1 changed A1', and its calculated similarity is 1. A2 changed A2', and the calculated similarity of A2 and A2' is K. Here, A2’ is 4-level GO term and a parent of A2 in GO term hierarchy. K is smaller value than 1.

We apply Chi-square algorithm to predicting protein feature. Chi-square procedure finds a global feature vector about proteins in total PPI networks and calculates global expectation numbers using the vector. After that, the algorithm gets local feature vectors about neighbor proteins which interact with each unknown protein, and calculates local expectation numbers using the vectors. Chi-square value is calculated by using global expectation numbers and local expectation numbers. If the value is bigger than threshold, we predict it as the feature of the unknown protein[4].
2.1 Chi-Square Value

The equation (1) calculates Chi-square value with global expectation numbers and local expectation numbers.

\[ \chi^2_k = \frac{(\pi_k - e_k)^2}{e_k} \]

(k = 1 ~ n, \( e_k \): Global Expectation Number, \( \pi_k \): Local Expectation Number)

2.2 System Flow

Figure 1 describes our proposed system flow. G-feature vector means global feature vector, and GEN means global expectation number. L-feature vector means local feature vector, and LEN means local expectation number.

![System Flow Diagram](image)

Figure 1: System Flow Diagram

3 Discussion

Our proposed protein feature prediction makes use of Chi-square algorithm about each feature within predefined interaction depth with direct or indirect interaction information among proteins. In addition, our system predicts a meaningful level in GO terms through normalization of GO terms. In the future, we will speed up GO term normalization process effectively and find an appropriate threshold for each Chi-square value.

References


