

A Function Prediction of Un-annotated Proteins Based on Interaction Generality

Jae-Young Jung¹, Jae-Hun Choi², Jong-Min Park³, Seon-Hee Park⁴

Keywords: protein function prediction, protein interaction, computational model, false positive

1 Introduction.

A function prediction of un-annotated proteins is known as one of the most challenging problems. In assigning functions to proteins without known functions automatically, the use of protein-protein interaction data is considered as a more reliable method in proteomics compared to gene expression correlation or phenotype data.

Here, we propose a prediction model using protein-protein interaction data and present effectiveness in analyzing large-scale protein interaction data. The proposed approach is based on expression correlation and interaction generality to make prediction results more reliable.

2 Method and Results.

The protein-protein interaction data generated by two-hybrid systems might contain many false-positive interactions. Therefore, it is necessary to consider the assessment of these false-positive interactions in modeling a prediction method to assign functions to un-annotated proteins through protein-protein interaction data.

2.1 Interaction Generality

The proposed predictive model uses the concept of guilt by association for annotating functions to proteins without function through protein-protein interaction data. To assess the reliability of a certain protein-protein interaction between two proteins, interaction generality and expression correlation of interacting proteins are adopted^[1].

The interaction generality $IG(i, j)$ for target interacting pair P_i and P_j is given by

$$IG(i, j) = (|Ng(i)| + |Ng(j)| - 2) - \left(\sum_{\substack{P_\ell \in Ng(i) \\ P_\ell \neq P_j}} \delta(|Ng(\ell)|) + \sum_{\substack{P_m \in Ng(j) \\ P_m \neq P_i}} \delta(|Ng(m)|) \right)$$

where $Ng(i)$ is the neighbors of protein P_i , that is, the set of proteins interacting with protein P_i and delta function $\delta(x)$ assigns $\{1\}$ if its value > 1 , but assigns $\{0\}$ set to the others.

¹ Bioinformatics Research Team, Electronic and Telecommunication Research Institute, Daejeon, Korea. E-mail: jjy72@etri.re.kr

² Bioinformatics Research Team, Electronic and Telecommunication Research Institute, Daejeon, Korea. E-mail: jhchio@etri.re.kr

³ Bioinformatics Research Team, Electronic and Telecommunication Research Institute, Daejeon, Korea. E-mail: jmpark93@etri.re.kr

⁴ Bioinformatics Research Team, Electronic and Telecommunication Research Institute, Daejeon, Korea. E-mail: shp@etri.re.kr

2.2 Prediction Result

We simulated the proposed model with KDD Cup and MIPS Yeast data. The KDD Cup data has 1,243 proteins, of which 381 proteins are for test data [2]. The MIPS data of 3,510 proteins was divided into 3 test data sets by equal selection per function: 561 for set 1, 581 for set 2, and 574 for set 3. The proposed approach was tested to predict 13 functions for KDD Cup data and 19 functions for MIPS data. The prediction accuracy was between 83.0% and 90.0% as the interaction generality and the value of ℓ -largest frequency changed. The distributions of IG values on KDD cup data and protein interaction map are shown in Figures 1 and 2.

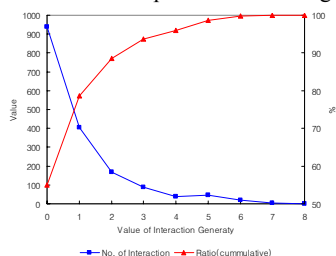


Figure 1: Distributions of IG values on KDD Cup data are shown. Values of interaction and cumulative values of their corresponding IG values are shown as rectangles and triangles respectively.

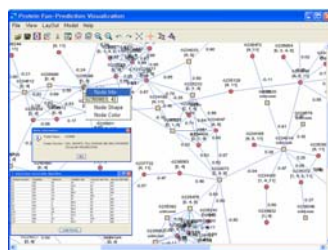


Figure 2: Screenshot showing protein-protein interaction map. Proteins in circles are annotated. The proteins in rectangle are unannotated. The result of function prediction is represented as dialog windows.

We execute function prediction only if expression correlation of interacting proteins is more than 0.5. Table 1 shows the result of prediction model in case the interaction generality is 1(5).

As shown in Table 1, the higher IG value of target interacting proteins, the lower accuracy of protein function is predicted.

ℓ	1	2	3	4	5	6	7
TP	435(542)	493(650)	527(1709)	537(726)	540(730)	541(732)	541(732)
TN	2487(3350)	2448(3265)	2362(3126)	2297(3018)	2283(2963)	2271(2931)	2260(2910)
FP	238(348)	180(240)	146(181)	136(164)	133(160)	132(158)	132(158)
FN	103(128)	142(213)	228(352)	292(460)	307(515)	319(547)	330(568)
Acc(%)	89.55(89.10)	90.13(89.63)	88.54(87.80)	86.85(85.71)	86.51(84.55)	86.18(83.86)	85.84(83.38)
No	251(336)	251(336)	251(336)	251(336)	251(336)	251(336)	251(336)

Table 1: The result of prediction model on KDD Cup data. The numbers in the parentheses are the result of prediction in case IG value is 5.

3 Discussion.

We proposed a model for protein function predictions in bioinformatics field. This model is using expression correlation and interaction generality information with guilt by association's concept to make prediction results more reliable.

4 References.

[2] Jie Cheng, Christos Hatzis, Hisashi Hayashi, Mark-A. Krogel, Shinichi Morishita, David Page, and Jun Sese, KDD cup 2001 report, SIGKDD Explorations, 3, Jan. 2002, 47-64.

[1] R. Saito et al., Interaction generality, a measurement to assess the reliability of a protein-protein interaction, Genome Informatics 13, 324-325, 2002 .