

# Interaction Generality, a Measurement to Assess the Reliability of a Protein-Protein Interaction

Rintaro Saito<sup>1,2</sup>

rsaito@sfc.keio.ac.jp

Harukazu Suzuki<sup>1</sup>

harukazu@gsc.riken.go.jp

Masaru Tomita<sup>2</sup>

mt@sfc.keio.ac.jp

Yoshihide Hayashizaki<sup>1</sup>

yoshihide@gsc.riken.go.jp

<sup>1</sup> Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan, and Genome Science Laboratory, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

<sup>2</sup> Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, Tsuruoka, Yamagata 997-0035, Japan

**Keywords:** protein-protein interaction, false positives, principal component analysis

## 1 Introduction

As whole-genome and complete cDNA sequences became available for numerous organisms, the focus of many research efforts is shifting rapidly from genomics to proteomics. One of the most important approaches in proteomics is the large-scale analysis of protein-protein interactions, because most proteins work as complexes to regulate biological processes in cells and even the whole body. High-throughput genome-wide screening of protein-protein interactions has been carried out in yeast, *Caenorhabditis elegans*, and higher organisms, such as mouse. Several successful computational analyses of interaction data also have been completed [1, 4].

On the other hand, it is widely accepted that the publicly available protein-protein interaction data, especially those obtained from two-hybrid systems, contain many false-positive interactions [2]. Mering *et al.* [5] estimate that as much as 50% of interactions obtained from yeast-two hybrid are false-positives according to the certain criteria. Thus a method to assess reliability of each protein-protein interaction is necessary. Here we define Interaction Generality (IG) measure and show that it can assess reliability of protein-protein interactions.

## 2 Method and Results

The interaction generality (IG) for target interacting pair A-B was defined by the following procedure. First, the protein C that interacts directly with the target interacting pair A-B is classified into one of 5 groups (*a1*, *a2*, *l*, *f*, *d*) according to its topological properties of interaction. When C interacts with both A and B, it is classified as “*a1*”. When C interacts with A but not B, and C has another interacting protein that interacts with B, it is classified as “*a2*”. When C is not classified as “*a2*”, interacts with A but not B, and has at least one interacting protein that interact with A, it is classified as “*l*”. If C does not meet these 3 conditions and interacts with another protein, it is classified as “*f*”. If C does not interact with any proteins except for A or B, it is classified as “*d*”. Then numbers of proteins belonging to each class are counted as  $n = (Na1, Na2, Nl, Nf, Nd)$ . The IG values are calculated as  $IG = np + C$ , where  $p' = (Pa1, Pa2, Pl, Pf, Pd)$  is a parameter for each class and C is a constant. The parameters and the constant, which were determined by principal component analysis of set of  $n$ 's, are as follows;  $p' = (-0.057, 0.0963, 0.179, 0.920, 0.331)$  and  $C = -5.603$ .

To investigate whether IG value is efficient for the assessment of reliability, distributions of IG values for reproducible interactions and non-reproducible interactions were calculated. As shown in Fig. 1, IG values for reproducible interactions are significantly lower than those of non-reproducible ones, suggesting that IG value can be used to select reliable interactions.

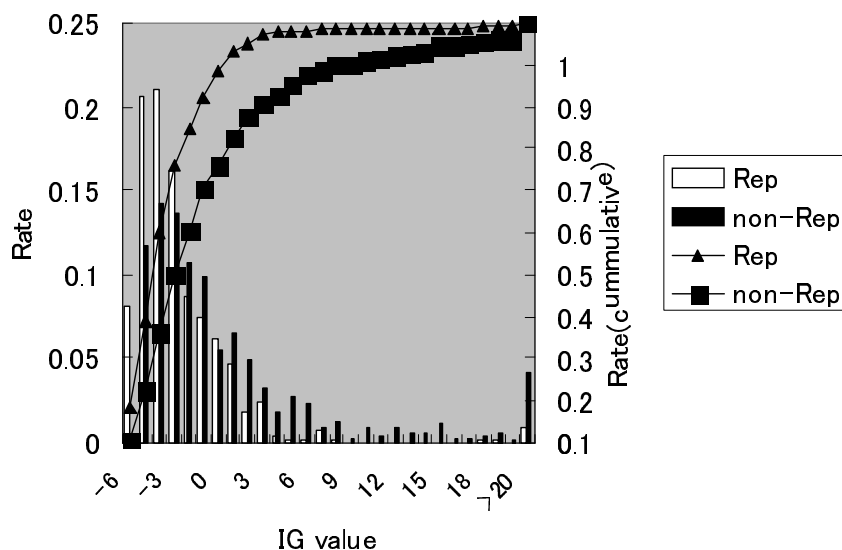


Figure 1: Distribution of IG values for reproducible and non-reproducible interactions. Rates of interactions having their IG values  $-7 < IG \leq -6$ ,  $-6 < IG \leq -5$ , ...,  $19 < IG \leq 20$  and  $> 20$  are shown. “Rep” and “non-Rep” indicate reproducible and non reproducible interactions respectively. Rates and their cumulative values of their corresponding IG values are shown as bars and lines respectively.

### 3 Discussion

In defining the IG, we incorporated principal component analysis where 5 parameters ( $Pa1$ ,  $Pa2$ ,  $Pl$ ,  $Pf$ ,  $Pd$ ) for topological properties ( $a1$ ,  $a2$ ,  $l$ ,  $f$  and  $d$ ) were set. With principle component analysis, one can determine the parameters and the constant without requiring knowledge of whether each interaction is true-positive or false-positive. This is a great advantage since we do not know which of the non-reproducible interactions are indeed false positives. Using the IG described in this paper, we obtained more reliable dataset than those in our original report [3]. We believe that the approach we describe here is applicable to higher organisms such as human and mouse whose large amount of protein-protein interaction data will soon become available.

### References

- [1] Fellenberg, M., Albermann, K., Zollner, A., Mewes, H.W., and Hani, J., Integrative analysis of protein interaction data, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:152–161, 2000.
- [2] Legrain, P., Wojcik, J., and Gauthier, J.M., Protein–protein interaction maps: A lead towards cellular functions, *Trends Genet.*, 17(6):346–352, 2001.
- [3] Saito, R., Suzuki, H., and Hayashizaki, Y., Interaction generality, a measurement to assess the reliability of a protein-protein interaction, *Nucleic Acids Res.*, 30(5):1163–1168, 2002.
- [4] Schwikowski, B., Uetz, P., and Fields, S., A network of protein-protein interactions in yeast, *Nat. Biotechnol.*, 18(12):1257–1261, 2000.
- [5] Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P., Comparative assessment of large-scale data sets of protein protein interactions, *Nature*, 417(6887):399–403, 2002.