Database and Prediction of Sequence Motifs on Protein Molecular Interactions

Kenji Suzuki suzuken@kuicr.kyoto-u.ac.jp

Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611, Japan

Abstract

We are developing a new prediction tool for protein-protein interactions based on sequence motifs. For this attempt, we have collected the information of the interaction data from existing databases and literatures, and arranged it into a new interaction sequence motif database. This paper describes the database and the prediction system for protein-protein interaction motifs.

1 Introduction

As technologies in the area of genome research advances, huge amount of genome and proteome data are accumulating. Utilizing the huge data, genome scientists are exploring higher biological systems based on molecular interactions that take parts in metabolic pathway and signaling system. Here, we focus on signalling networks, and describe our attempt to predict protein interactions based on sequence motifs. As for proteins in signaling networks, many signalling domains have been reported [1]. These domains include those for protein-protein interactions, e.g. SH2 and SH3 domains. Motifs recognized by the domains could be different in different groups, even if the recognizing domains belong to the same group. Thus, the precise knowledge of interaction pairs is necessary for developing better prediction methods. And, if we could predict these interaction specificity, the predicted results would be helpful to further analyzing of signal networks.

2 Method

2.1 Database

The protein-protein interaction data is composed of two parts. One is general information of interaction domains listed in literature [1] and the other is cross references to individual sequence entries.

2.2 Classification of specificity

Among proteins in signaling networks, SH2 domain and kinase motif have been well investigated and characterized by experimental work [2, 3]. According to Songyang [3], SH2 domains are classified into four groups. In the recognition of phosphotyrosine, each group of SH2 domain has different specificity to different sequence patterns around the phosphotyrosine. Our classification of SH2 domains are based on Songyang [2], and summarized as follows:

1. SH2 domain was extracted from SWISSPROT.
2. The sequence of SH2 domains from SWISSPROT were aligned against the SH2 profile data in PROSITE.
3. The residues contributing to recognition specificity were detected and SH2 domains were classified into four groups.

3 Results and Discussion

Table 1 shows an example of phosphotyrosine - SH2 domain protein pair data. Table 2 shows the classified SH2 domains. Motifs around phosphotyrosine in Table 2 is described in Songyang [2].

Table 1: Database example

phosphotyrosine		SH2 domain	
name	motif	name	domain region
Shc	**Y**VNV	Grb2	60-152
EGF receptor	**Y**SSD	Grb2	60-152
PDGF receptor	**Y**VPM	Nck	282-376
:	:	:	:

Table 2: Classification of SH2 domain

Group	count(all)	count(human)	(N.B. phosphotyrosine motif)
Group 1	119	32	pTyr - hydrophilic - hydrophilic - Ile/Pro
Group 1A	52	13	pTyr - Glu - Glu - Ile
Group 1B	67	19	
Group 2	4	2	pTyr - Met - Glu - Pro
Group 3	39	8	pTyr - hydrophobic - Xxx - hydrophobic
Group 4	11	9	(no defined)

These motifs are so short(about 4 residues), that a search a sequence database by the motifs would produce many false positives. For this reason, phosphotyrosine motifs must be more closely examined and refined.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science', from the Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Bork, P., Schultz, J., Ponting, C.P. "Cytoplasmic signalling domains: the next generation" *TiBS*, 22:296–298, 1997.

[2] Songyang, Z. et al. "A method to identify protein sequences that fold into a known three-dimensional structure" *Mol. Cell. Bio.*, 14:2777–2785, 1994.

[3] Songyang, Z. et al. "A Structural Basis for Substrate Specificities of Protein Ser/Thr Kinases: Primary Sequence Preference of Casein Kinases I and II, NIMA, Phosphorylase Kinase, Calmodulin-Dependent Kinase II, CDK5, and Erk1" *Mol. Cell. Bio.*, 16:6486–6493, 1996.