Protein Interactions in the Whole Human Genome*

Kyungsook Han[1] khan@inha.ac.kr
Byunggyu Park[1] trf04@hotmail.com
Hyongguen Kim[1] khg0105@orgio.net
Hong-Jin Kim[2] hongjink@cau.ac.kr
Jong Park[3] jong@bio.cc

[1] Department of Computer Science and Engineering, Inha University, Inchon 402-751, Korea
[2] College of Pharmacy, Chung-Ang University, Seoul 156-756, Korea
[3] MRC-DUNN, Human Nutrition Unit, Hills Road,Cambridge, CB2 2XY UK

Keywords: protein-protein interaction, network, human interactome, homologous interaction

1 Introduction

As the genes of the human genome become known, one of the challenges is to identify all the interactions of the human genes. Protein interactions are known for several organisms such as yeast and *Helicobacter pylori* [2, 6] due to the recent improvements in protein interaction detection methods, but they are limited to the low-order species only. Direct determination of all the interactions between the whole human genes is still difficult even with current large-scale protein interaction detection methods.

This paper presents protein domain interactions for all whole human genes using the concept of 'homologous interaction'. We believe this is the first attempt to map a whole human interactome. A network of human protein interactions can define the size of the problem of mapping interactome and will assist many biomedical studies. It is also essential to the comparison of interactomic networks in evolutionary terms. It had been widely conjectured, but never shown, that core protein interactions are conserved among different organisms [5]. As our first step towards comparing the whole protein interactions across different organisms, we have built a network of human protein interactions bioinformatically and are currently comparing it with a network of yeast protein interactions.

2 Method and Results

The primary focus of the work described in this paper is to derive structural interactions of all human proteins. The 29,076 human protein sequences of the Ensembl release v4.28.1 [3] were compared with the PDB-intermediate sequence library (PDB-ISL) [7] using Gapped BLAST [1]. PDB-ISL contains 371,257 intermediate sequences for proteins of known structure, and thus genome sequences can be assigned protein domains by matching them to proteins of known structure. From the comparison of the human protein sequences to the intermediate sequences, 19,699 human protein sequences were assigned one or more protein domains and 9,377 human protein sequences were assigned no protein domain. The left table in Table 1 shows the number of domains assigned to human proteins.

After assigning protein domains to the human proteins, protein domain interactions in human proteins were determined based on PSIMAP [5]. PSIMAP classifies interactions between all known structural protein domains. There were total 25,706,777 protein domain interactions between the 29,076 human proteins. A network of human protein domain interactions contains 10,658 connected components in total. 10,589 out of 10,658 connected components were singleton sets. These singleton sets fall into three classes: (1) 9,377 human proteins with no domain assigned, (2) 1,182 human proteins with domains assigned but with no interaction with others, and (3) 31 human proteins that interact with themselves only (self-loops).

*This work was supported by the Ministry of Information and Communication of South Korea under grant IMT2000-C3-4.

Table 1: **Left table**: The number of domains assigned to 29,076 human proteins. **Right table**: The number of proteins in the connected components with at least 2 proteins. The number of interactions in the fourth column is the total number of interactions in all the connected components of specified size. 37 self-loops are not included in the total number of interactions in the right table.

#domains	#proteins	%
0	9,377	32.250
1	14,054	48.335
2	3,800	13.069
3	1,048	3.604
4	403	1.386
5	198	0.681
6	82	0.282
7	36	0.124
8	24	0.083
9	19	0.065
10	11	0.038
11	7	0.024
12	3	0.010
13	1	0.003
14	2	0.007
15	2	0.007
16	2	0.007
17	1	0.003
19	1	0.003
23	1	0.003
27	1	0.003
34	1	0.003
49	1	0.003
153	1	0.003
total	29,076	100.000

#proteins (A)	#conn. comp.(B)	$A \times B$	#interactions
18,113	1	18,113	25,700,387
59	1	59	3,481
18	1	18	324
16	1	16	256
14	1	14	547
12	1	12	168
11	1	11	285
10	2	20	133
9	1	9	162
8	1	8	64
7	7	49	358
6	1	6	48
5	8	40	205
4	8	32	124
3	12	36	105
2	22	44	93
total	69	18,487	25,706,740

The right table in Table 1 shows the number of domain interactions for the connected components with at least 2 proteins, sorted by their size. For example, there were 22 connected components that contain 2 proteins, and the total number of interactions in the 22 connected components was 93. It follows from this result that more than 62% of human proteins (18,113 proteins) and more than 99% of domain interactions (25,700,387 interactions) are contained in the single, largest connected component of a network. The largest connected component has a very high degree of interactions compared to the number of proteins.

The high degree of interactions might explain the reason that the number of human diseases exceeds the number of human genes despite the fact that the total number of human genes does not differ much from the number of genes of other organisms.

References

[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

[2] Gavin, A.-C. *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, 415:141–147, 2002.

[3] Hubbard, T. *et al.*, The ensembl genome datase project, *Nucleic Acids Res.*, 30(1):138–41, 2002.

[4] Park, J. and Bolster, D., Conservation of protein interaction network in evolution, *Genome Informatics*, 12:135–140, 2001.

[5] Park, J., Lappe, M., and Teichmann, S.A., Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and Yeast, *J. Mol. Biol.*, 307:929–938, 2001.

[6] Rain, J.-C. *et al.*, The protein-protein interaction map of *Helicobacter pylori*, *Nature*, 409:211–215, 2001.

[7] Teichmann, S.A., Chothia, C., Church, G.M., and Park, P., Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL, *Bioinformatics*, 16(2):117–124, 2000.