EXPLORING THE COMBINATORICS OF PROTEIN COMPLEXES

Jung Hun Ohn [1]) Jihun Kim [2]) and Ju Han Kim [3])*

Abstract

Protein complexes are made up of different protein subunits and each protein participates in a variety of protein complexes. We explored this combinatorial property of protein complexes based on a proteome-wide experiment to purify and identify protein complexes by mass spectrometry. We quantified the degree to which a given complex is made up of exclusively participating protein members with what is called 'Isolation Index'. Each protein complex belongs to one of nine functional groups and Isolation Indices are enumerated both within and beyond the functional group, termed intra-category Isolation Index and whole category Isolation Index, respectively. The scatter plot of the two Isolation Indices reveals two distinct patterns of protein complexes: transcription, DNA maintenance, chromatin structure, RNA metabolism, protein synthesis and turnover and intermediate and energy metabolism associated protein complexes share large numbers of components with other protein complexes irrespective of functional groups, while cell cycle and signaling related protein complexes have little components in common with complexes of the same functional group. The former protein complexes occupy core positions in a higher order map of nine functional groups based on protein component sharing but the latter complexes are placed in the periphery in contrast to the map of functional groups based on binary protein interactions.

1. Introduction

Proteins seldom act alone; their spatial and temporal assemblies and disassemblies form the basis of cellular functions. These concerted interactions like a sophisticated machine gave rise to the description, 'molecular machines.' [1] Numerous biochemical researches have described large molecular machines like spliceosome, cyclosome, proteasome and nuclear pore complexes. [2][3][4][5]

The molecular machines are described from the study of protein complexes comprising two or more subunit proteins, which may or may not be identical. Although we have extended our knowledge of two-protein complexes through profiling of binary interactions by systematic experiments like yeast two-hybrid, the composition of a larger protein complex is not easily inferred from pair-wise interaction datasets.

[1] Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea email: jhohn2@snu.ac.kr

[2] Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea email: djdoc@snu.ac.kr

[3] Seoul National University Biomedical Informatics (SNUBI) and Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea email: juhan@snu.ac.kr

*Corresponding author

Recently, there have been attempts to systematically purify protein complexes in physiological conditions. Gavin *et al.* used tandem-affinity purification (TAP) and mass spectrometry on a large-scale to isolate and identify protein complexes in yeast. [6] Ho *et al.* also applied what they called high-throughput mass spectrometric protein complex identification (HMS-PCI) to identify protein complexes on a proteome-wide scale. [7]

It is proposed that protein complexes are composed of subunit proteins that share similar expression patterns, functional classifications, and cellular localizations and reflect a higher level of molecular organization. [8] Therefore, when investigated on a systematic scale, protein complex may work as a reasonable unit in deciphering the overall organization of the proteome.

Meanwhile, a particular complex is not necessarily composed of invariable protein members nor is any constituting molecule involved uniquely in that specific complex. Gavin *et al.* illustrated this combinatorial aspect by linking complexes that share components and derived a higher order network of protein complexes. [6] In a lower level of molecular organizations, a protein molecule is itself made up of conserved domains or motifs that also constitute other kinds of protein molecules and endow them with a different functionality. In a likely manner, different molecular machines often use the same protein subunits to exert different functions. [9] This multi-level combinatorial utilization of molecules ends at the level of organelles or cells.

We investigate this combinatorial nature at the level of protein complexes by quantifying the degree to which a given complex is made up of exclusively participating protein subunits both all through the functional groups and within each functional group. We present a higher order organization map of nine functional groups based on protein component sharing.

2. Data and Methods

2.1. Dataset

Gavin *et al.* used tandem-affinity purification (TAP) and mass spectrometry on a large scale and identified 232 distinct protein complexes in yeast. [6] A total of 1353 proteins constitute the complexes. The 232 protein complexes are roughly assigned into nine functional groups according to YPD and by literature mining in the original article. The numbers in the parentheses are the number of protein complexes within each functional group.

1. Cell cycle (13)
2. Polarity and structure (8)
3. Intermediate and energy metabolism (43)
4. Membrane biogenesis and traffic (20)
5. Protein synthesis and turnover (33)
6. Protein/RNA transport (12)
7. RNA metabolism (28)
8. Signaling (20)
9. Transcription/DNA maintenance/chromatin structure (55)

2.2. Isolation Index

A protein complex is defined to be 'isolated' if its composing subunit proteins participate 'exclusively' in the specific complex. A protein complex is less isolated if its building blocks are parts of other complexes as well.

To quantify the degree to which a protein complex is isolated, the 'Isolation Index' is assigned for each complex t.

For a complex t there exists a set of genes, Ct = {gene | gene is a sub-component of protein complex t} and the 232 protein complexes are partitioned into nine subsets of different functional groups, Si, i=1,2,…,9 as depicted above.

First of all, from the original protein complex dataset we generate a binary 1353-by-232 matrix A = <*aij*>,

$$a_{ij} = \begin{cases} 1 \text{ , if the gene i participates in the complex j.} \\ 0 \text{ , if the gene i does not participate in the complex j.} \end{cases}$$

2.2.1. Whole category Isolation Index (Iw,t)

For a protein complex t, Dt is the sub-matrix of A, where Dt = <d*ij*>, d*ij* = a*ij* for i ∈ Ct and j = 1, 2,…, 232. The whole category Isolation Index is defined as the similarity between the data matrix Dt and the ideal pattern matrix Pt = <p*ij*>,

$$p_{ij} = \begin{cases} 1 \text{ , } i \in C_t \text{ and } j = t. \\ 0 \text{ , } i \in C_t \text{ and } j \neq t. \end{cases}$$

, where Pt represents a data matrix of an ideal protein complex whose components are parts of only the protein complex out of 232 complexes.

The cosine index is used as the similarity measure and the whole category Isolation Index (Iw,t) is defined as the following;

$$I_{w,t} = \frac{\sum_{i,j}(d_{ij} \cdot p_{ij})}{\sqrt{\sum_{i,j} d_{ij}^2} \sqrt{\sum_{i,j} p_{ij}^2}}$$

It ranges from 0 to 1 and the value closer to 1 signifies that the specific complex is more isolated among the 232 protein complexes.

2.2.2. Intra-category Isolation Index (Ii,t)

In contrast to the whole category Isolation Index, intra-category Isolation Index is defined considering the group membership of each complex. Assuming a protein complex t belongs to the functional group k, the data matrix Dt′ is the sub-matrix of A, where Dt′ = <d′*ij*>, d′*ij* = aij for i ∈ Ct and j ∈ Sk. The ideal pattern matrix is Pt′ = <p′*ij*>,

$$p'_{ij} = \begin{cases} 1 \text{ , } i \in C_t \text{ and } j = t. \\ 0 \text{ , } i \in C_t \text{ and } j \neq t \text{ and } j \in S_k. \end{cases}$$

The cosine index is used as the similarity measure and the intra-category Isolation Index (Ii,t) is defined as the similarity between the matrices Dt′ and Pt′;

$$I_{i,t} = \frac{\sum_{i,j}(d'_{ij} \cdot p'_{ij})}{\sqrt{\sum_{i,j} d'^{2}_{ij}} \sqrt{\sum_{i,j} p'^{2}_{ij}}}$$

The index closer to 1 means that the protein complex is isolated and has less subunit proteins in common with other protein complexes within the same functional category.

3. Results and Discussion

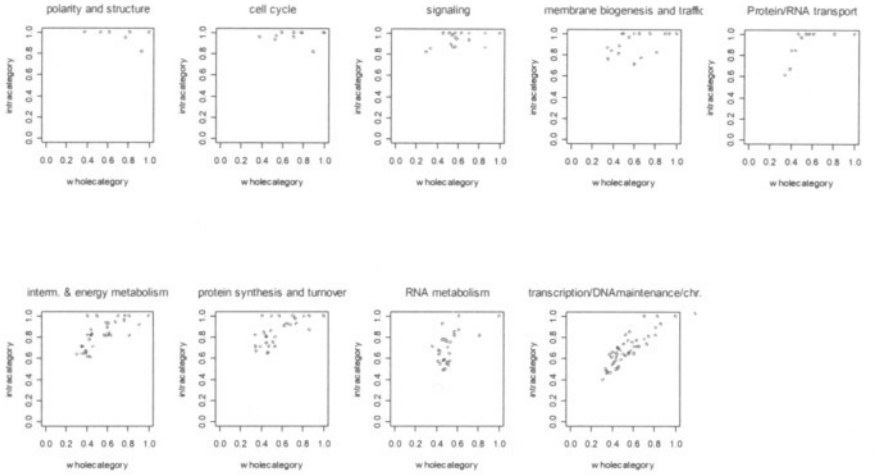

Fig. 1. Scatter plots of whole category (Iw,t) and intra-category (Ii,t) isolation indices for each complex according to nine functional groups

Figure 1 shows the scatter plots of Iw,t and Ii,t for each protein complex according to its functional group. Interestingly, we can observe two different kinds of complexes. Some protein complexes have low whole category Isolation Indices but high intra-category Isolation Indices and are scattered in left upper panel in each plot. Other complexes have little difference in the two indices and are spread around the Y = X axis.

The former pattern of complexes are mostly found in processes like polarity and structure, cell cycle and signaling and shown in the upper panel, while the latter pattern of complexes are the norms in the processes of transcription /DNA maintenance/ chromatin structure and RNA metabolism in the lower panel.

Cell cycle and signaling associated proteins generally form small complexes of regulatory units. Their ultimate goal is to preserve, transfer or amplify signals. The enrichment of these processes in complexes that have significantly higher intra-category Isolation Indices than the whole category Isolation Indices gives the lesson; though the same protein component seems to be used in different

complexes with other functional annotations, the sharing of components with complexes within the same functional category is supposedly deterred to maximize and ensure the specificity of signaling.

On the other hand, the processes like transcription, DNA maintenance, chromatin structure and RNA metabolism are carried out by large molecular machines of various transcriptional and translational complexes. These molecular machines carry out large numbers of functions in concert both in time and space. For example, in RNA polymerase II complex (reflected in a complex of 9th functional group whose whole category and intra category Isolation Indices are 0.746267 and 0.805473 respectively), different combinations of five protein components build the structures called jaw, clamp, cleft, shelf and funnel that works for its attachment to DNA, the maintenance of the RNA-DNA duplex, the access of template strand to the active site and its translocation along the strand. [9][10]

From the point of parsimony or considering the diverse number of functions to be carried out with the limited number of protein components, it is of course waste of resources to invent different exclusively composed machines to perform so many different functions. Instead, the cell seems to have evolved to utilize different combinations of protein components to perform various tasks to keep its life go on.

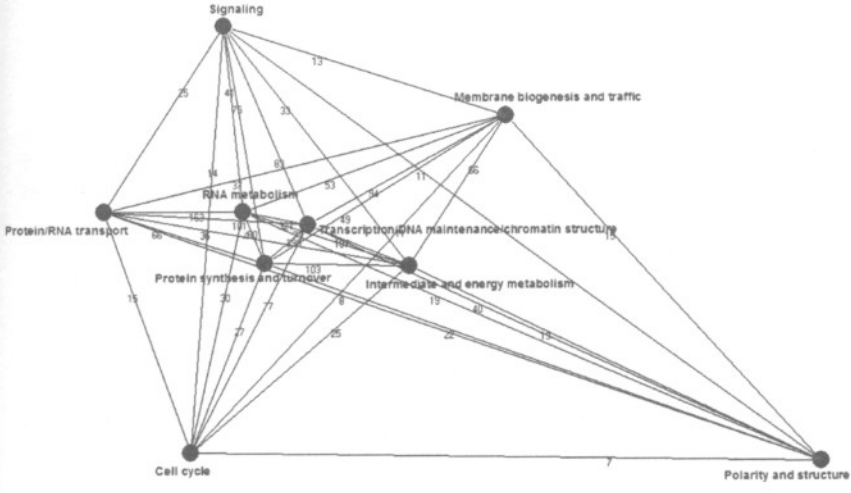

Fig. 2. A higher order map of nine functional groups. The number in each edge is the number of shared protein components between the two functional groups

Figure 2 gives a higher order map of nine functional groups. As Gavin *et al.* presents in the article, a network of 232 complexes is generated by taking the number of protein co-memberships between complexes as the edge weight. [6] We further drew a map of nine functional groups with the number of protein co-memberships between functional groups as edge weight because each complex belongs to a functional group. In this map, the closer the groups are, the more protein components are held in common.

The four functional groups of Transcription/DNA maintenance/chromatin structure, RNA metabolism, protein synthesis and turnover and intermediate and energy metabolism are strongly interconnected from the point of shared protein components and occupy core positions. They are

functional groups whose complexes have similar whole category and intra-category Isolation Indices and presented in the lower panel of figure 1. But the processes with the significantly different whole category and intra category Isolation Indices like polarity and structure, cell cycle and signaling are placed in the periphery in figure 2. It is interesting that we have parallel results through different methodologies.

And this map of functional groups is in contrast with its counterpart map of functional groups based on 2,358 direct physical interactions among 1,548 proteins based on biochemical experiments and *yeast two-hybrid* studies suggested by Schwikowski *et al.* [11] They found cell cycle control process shows the most interactions with other functional groups and is placed in a core position in the network of different functional groups. This discordance seems to arise because the map in figure 2 is considering protein component co-membership between functional groups and formed from the dataset of protein complexes where specific interactions between components are not known and a protein complex is a likely candidate of a molecular machine. Cell cycle associated proteins may have various one-to-one interactions with other proteins from different functional groups. But, when they are incorporated into a protein complex and work as a component of a specialized unit, they are uniquely involved in the specific complex.

Quantifying the combinatorial property of gene groups or protein groups such as clusters from microarray experiments or pathways with the above-introduced method will help in entangling the basically combinatorial phenomenon of life, the 'pleiotropy'.

Acknowledgements

We appreciate Dr. Gavin A.C. who kindly provided the original dataset of genes and complexes. This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Science and Technology, Republic of Korea (2005-00162). J. H. Ohn's training is supported by a grant from Ministry of Health & Welfare, Korea (0412-MI01-0416-0002).

References

[1] Alberts, B.: The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell. 92 (1998) 291-294

[2] Rout, M.P., Aitchison, J.D., Suprapto, A., Hjertaas, K., Zhao, Y, Chait, B.T.: The yeast nuclear pore complex: composition, architecture, and transport mechanism. J. Cell Biol. 148(4) (2000) 635-651

[3] Zachariae, W., Shin, T.H., Galova, M., Obermaier, B., Nasmyth, K.: Identification of subunits of the anaphase-promoting complex of Saccharomyces cerevisiae. Science. 274(5290) (1996) 1201-1204

[4] Neubauer, G., Gottschalk, A., Fabrizio, P., Seraphin, B., Luhrmann, R., Mann, M.: Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. Proc Natl Acad Sci U S A. 94(2) (1997) 385-390

[5] Verma, R., Chen, S., Feldman, R., Schieltz, D., Yates, J., Dohmen, J., Deshaies,R.J.: Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. Mol Biol Cell. 11(10) (2000) 3425-3439

[6] Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster,

B., Neubauer, G., Superti-Furga,G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 415(6868) (2002) 141-147

[7] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., Tyers, M.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. 415(6868) (2002) 180-183

[8] Dezso, Z., Oltvai, Z.N., Barabasi, A.L.: Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. Genome Res. 13(11) (2003) 2450-2454

[9] Gavin, A.C., Superti-Furga, G.: Protein complexes and proteome organization from yeast to man. Curr Opin Chem Biol. 7(1) (2003) 21-27

[10] Cramer, P., Bushnell, D.A., Kornberg, R.D.: Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science. 292(5523) (2001) 1863-1876

[11] Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nature Biotechnol. 18(12) (2000) 1257-1261