Modularity in the evolution of yeast protein interaction network

Soichi Ogishima[1]*, Hiroshi Tanaka[1,2] & Jun Nakaya[1]

[1]Department of Bioclinical Informatics, Tohoku Medical and Megabank Organization, Tohoku University, Seiryo-cho 4-1, Aoba-ku, Sendai-shi Miyagi 980-8575 Japan; [2]Department of Bioinformatics, Tokyo Medical and Dental University, Yushima 1-5-45, Bunkyo-ku, Tokyo 113-8510 Japan; Soichi Ogishima – Email: ogishima@sysmedbio.org; *Corresponding author

Abstract:
Protein interaction networks are known to exhibit remarkable structures: scale-free and small-world and modular structures. To explain the evolutionary processes of protein interaction networks possessing scale-free and small-world structures, preferential attachment and duplication-divergence models have been proposed as mathematical models. Protein interaction networks are also known to exhibit another remarkable structural characteristic, modular structure. How the protein interaction networks became to exhibit modularity in their evolution? Here, we propose a hypothesis of modularity in the evolution of yeast protein interaction network based on molecular evolutionary evidence. We assigned yeast proteins into six evolutionary ages by constructing a phylogenetic profile. We found that all the almost half of hub proteins are evolutionarily new. Examining the evolutionary processes of protein complexes, functional modules and topological modules, we also found that member proteins of these modules tend to appear in one or two evolutionary ages. Moreover, proteins in protein complexes and topological modules show significantly low evolutionary rates than those not in these modules. Our results suggest a hypothesis of modularity in the evolution of yeast protein interaction network as systems evolution.

Background:
Comprehensive data for protein interactions have been accumulating based on large-scale hybridization methods **[1,2]** and make it possible to understand the evolution of cellular networks at the system rather than the gene level. Protein interaction networks are known to show remarkable global structures, with scale-free and small-world properties. A scale-free structure is a network structure that exhibits power-law distributions of connectivity; most network components have a few connections while some components are extremely highly connected. In contrast, a small-world structure is a network structure that exhibits high clustering coefficients **[3]**. To explain the evolutionary processes of protein interaction networks possessing scale-free and small-world structures, preferential attachment and duplication-divergence models have been proposed as mathematical models. A preferential attachment model was proposed to generate a power-law distributed network of proteins **[4]**. In this model, new nodes are added to a pre-existing network, and are connected to each

of the pre-existing nodes with a probability proportional to the number of connections for each of the original nodes. This model has showed that hub proteins having a high connection degree are evolutionarily old **[5]**. Duplication-divergence models have also been proposed to generate a scale-free and small-world network of proteins **[6]** that assumes gene duplication plus re-wiring of the newly created proteins. In contrast, as molecular evolutionary analysis, yeast proteins have been classified into isotemporal categories according to their molecular evolutionary histories **[7]**. It showed that two proteins tend to interact with each other if they are in the same or similar categories, but otherwise to avoid each other. This observation suggests that synergistic selection is at work during network evolution and provides insights into the hierarchical "modularity" of cellular networks.

A modular structure is a third remarkable structural characteristic of a protein interaction network **[8, 9]**. Yook *et al.* showed yeast protein interaction networks exhibit scale-free

and hierarchical modularity, and suggested that modules should appear as distinct group of nodes that are highly interconnected with each other but have only a few links to nodes outside of the module **[8].** Fernandez showed the trend toward increasing modularity associated with evolutionary change in the yeast protein interaction network **[10].**We herein propose a hypothesis of modularity in the evolution of yeast protein interaction network based on an examination of relationships between the evolutionary ages of yeast proteins and their connection degrees.

Methodology:
Data collection
We collected yeast (*Saccharomyces cerevisiae*) protein interaction data from the MIPS (Munch Information Center for Protein Sequences) CYGD (Comprehensive Yeast Genome Database) at http://mips.gsf.de/genre/proj/yeast/index.jsp **[1],** which includes integrated data from Y2H (Yeast 2 hybrid), TAP (tandem affinity purification), and immunocoprecipitation experiments. We extracted maximum component of yeast protein interaction network which contains 3,801 proteins and 9,742 physical interactions. Functional annotations of yeast proteins based on the Gene Ontology (GO) were retrieved from the Saccharomyces Genome Database at http://www. yeastgenome.org/ (Oct 28, 2006). We collected amino acid sequences of 379 bacterias, *Arabidopsis thaliana*, *Schizosaccharomycespombe*, *Saccharomyces cerevisiae* (these are from the NCBI RefSeq Release 19 database at ftp. ncb .nih. gov / genomes), *Dictyosteliumdiscoideum* (from dictyBase at http:// dicty base .org/) and *Plasmodium falciparum* (from the PlasmoDB 5.0 at http://plasmodb.org /).

Phylogenetic profile
To assign yeast proteins into the evolutionary ages, we constructed a phylogenetic profile, which is a molecular evolutionary profile that indicates presence/absence of orthologous genes. To construct the phylogenetic profile, we first run a BLASTP **[11]** search with yeast proteins as query sequences against amino acid sequences of 379 bacterias, *Plasmodium falciparum* (plasmodium), *Arabidopsis thaliana* (plant), *Dictyosteliumdiscoideum* (social amoeba) and Schizosaccharomycespombe (fission yeast, fungi). *P. falciparum*, *A. thaliana*, *D. discoideum* and *S. pombe* are species whose genomes are completely sequenced between bacterias and yeast. The E-value threshold for screening was set to 1.0×10^{-10}. Second, the sequences of hit proteins were globally aligned using the ClustalW multiple alignment program **[12].** Among the aligned sequences, proteins with over 60% global similarities were finally identified as computational orthologues. Finally, we summarized the presence/absence of orthologous proteins to the correspondent yeast proteins as phylogenetic profiles.

Functional modules
We retrieved functional annotations of proteins from the Saccharomyces Genome Database at http:// www.yeastgenome. org based on the Gene Ontology (GO) biological processes. Based on these GO annotations, we identified 598 functional modules composed of interacting proteins for which the functional annotations are identical.
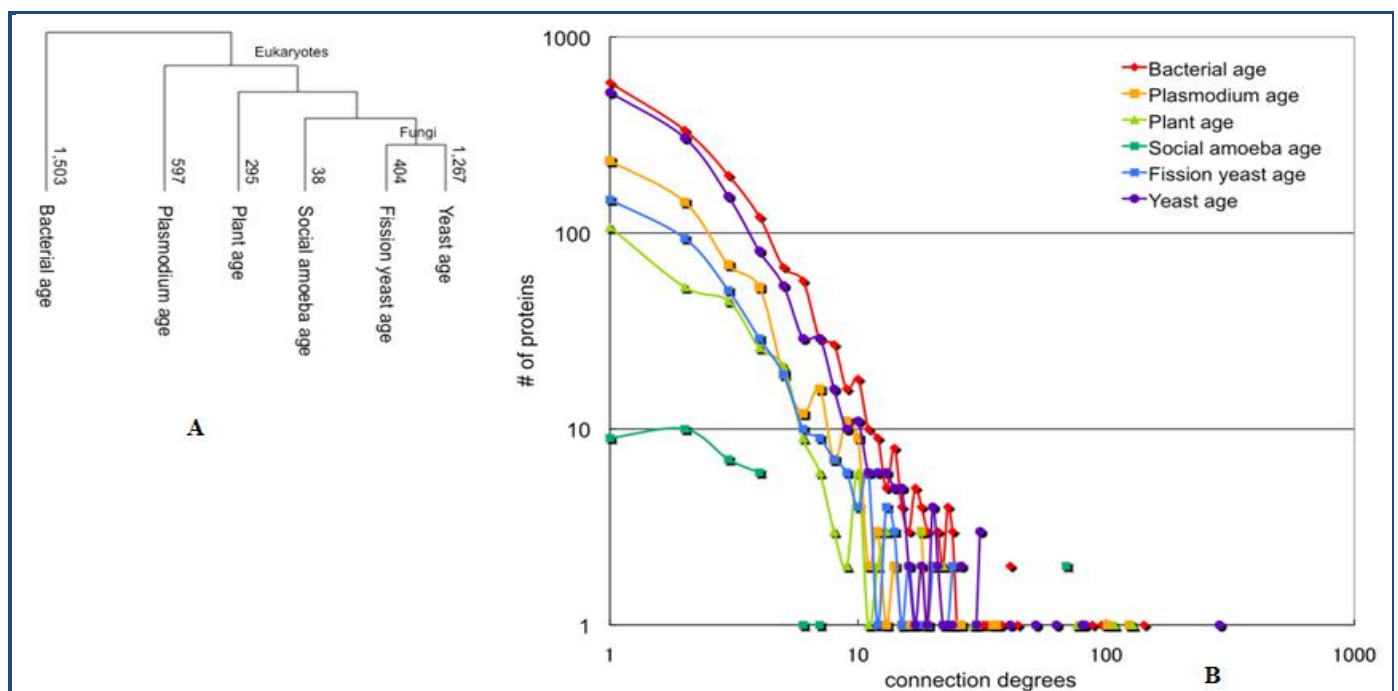

Figure 1: **A)** Six evolutionary ages (bacterias, plasmodium, plant, social amoeba, fission yeast and yeast)and number of proteins in each evolutionary age; **B)** Connection degree distributions of yeast proteins in six evolutionary ages (bacterias, plasmodium, plant, social amoeba, fission yeast and yeast).

Topological modules
We identified 43 topological modules whose maximum numbers of member are fewer than ten proteins, by cutting

interactions in decreasing order of shortest-path betweenness. A shortest path is a path between two nodes such that the sum of the hops of its constituent edges is minimized. Shortest-path

BIOINFORMATION

betweenness indicates the importance of interactions, which is the number of shortest paths between all pairs of nodes.

Evolutionary rates of proteins
We retrieved evolutionary rates of 3,035 yeast proteins (genes) from the work by Hirsh *et al* **[13],** which were obtained from

calculations of nonsynontmous (dN) and synonymous (dS) rates by comparison of orthologous gene sequences among four species of the genus Saccharomyces (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*).
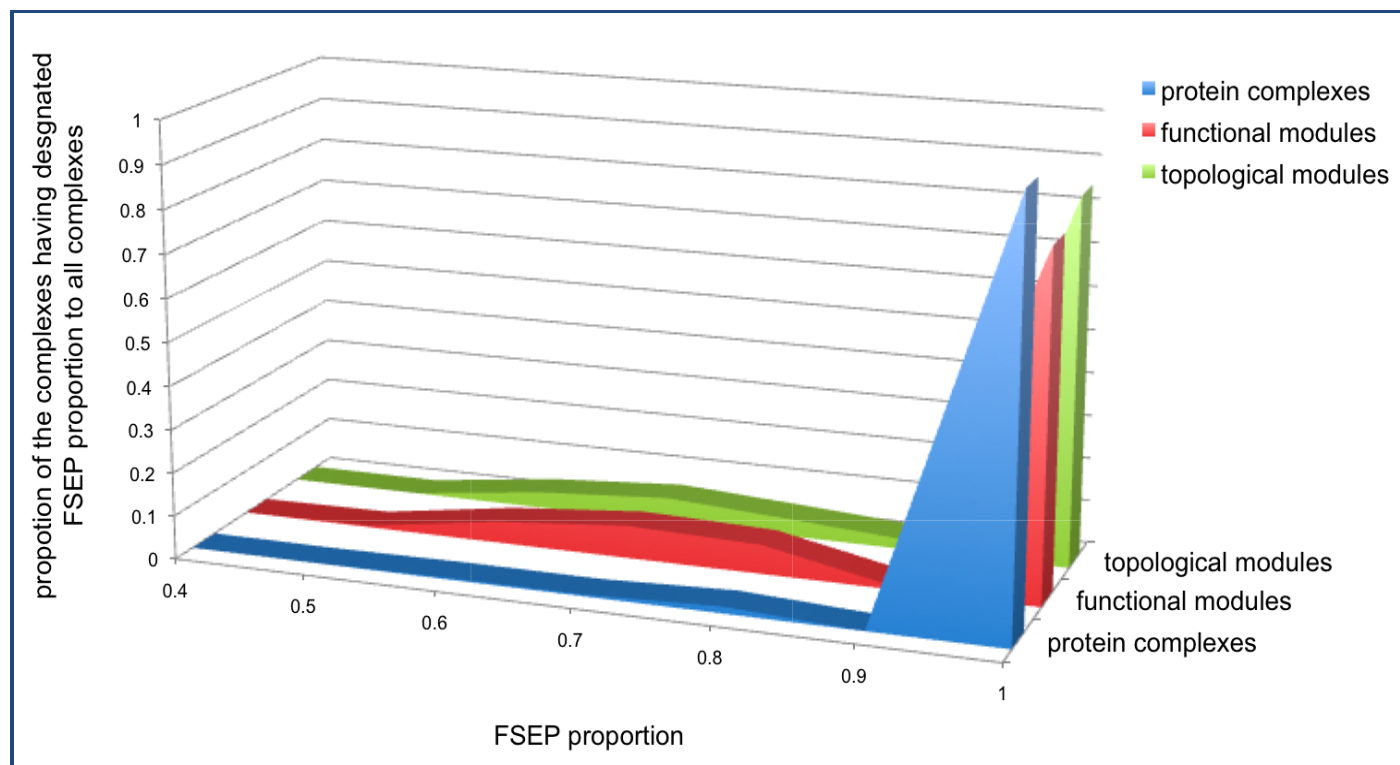

Figure 2: Modularity in the evolutionary process of protein complexes, functional modules and topological modules. Horizontal axes indicates the proportion of the FSEP proteins to all member proteins (FSEP proportion), vertical axes indicates the proportion of the complexes having the designated FESP proportion to total complexes.

Discussion:

All the hub proteins are not evolutionarily old

We first assigned yeast proteins into six evolutionary ages by constructing a phylogenetic profile. A phylogenetic profile is a profile of the presence/absence of orthologous proteins to the correspondent protein. The numbers of proteins of six evolutionary ages (bacteria, plasmodium, plant, social amoeba, fission yeast and yeast) are shown in **(Figure 1A).** We examined relationships between the evolutionary ages of yeast proteins and their connection degrees **(Figure 1B).** We found that the distribution of connection degrees of the newest age proteins (yeast age proteins) is similar to that of the oldest age proteins (bacterial age proteins). This result contradicts the preferential attachment model, in that old proteins should

Show high connection degree **[6].** Moreover, almost half of high- degree proteins are evolutionarily new. In fact, rates of high-degree proteins (>30 connection degrees) in bacterial (oldest) and yeast (newest) age are 0.0026 (14/1,530) and 0.0079 (10/1,267), respectively. To explain these results, newly emerged proteins are considered to prefer to connect not to already-existed proteins but to newly emerged proteins in the same evolutionary age. It suggests a hypothesis of modularity in the evolution of yeast protein interaction network that hub

proteins appeared and interacted with other simultaneously emerged proteins that form what we call "modules".

Protein complexes, functional modules and topological modules

What are the correspondences with "modules"? We considered that protein complexes, functional modules and topological modules correspond with "modules."We examined 1,142 protein complexes, 598 functional modules and 43 topological modules, and inferred the evolutionary processes of them by assigning each evolutionary age to member protein. We identified the firstly and secondly evolutionarily-populated (FSEP) proteins in each module defined by proteins of the top-two populated evolutionary ages. That is, we identified the firstly and secondly largest groups of member proteins in each module which appear in the same evolutionary ages. We then examined their compositions to form protein complexes, functional modules and topological modules **(Figure 2).** Our results showed that the FESP proportion is remarkably concentrated at 1. This tendency does not result from the background bias in the numbers of proteins of each evolutionary age;67.2% of complexes, 56.0% of functional modules and 45.5% of topological modules are significantly biased in their evolutionary compositions ($\chi 2$ test of goodness-of-fit, p-value<0.05).These results suggest that protein complexes, functional modules and topological

modules tends to be formed by proteins that appeared in only one or two evolutionary ages, therefore they did not appear in all six ages continuously and incrementally, but instead in only one or two evolutionary ages simultaneously.

Low evolutionary rates of proteins in modules

If the yeast protein interaction network evolved by"module", proteins in modules should be more conserved than those not in modules. We examined the differences of the evolutionary rates of proteins between them. In fact, proteins in modules showed significantly low evolutionary rates than those not in modules, in both protein complexes(Wilcoxon rank sum test, p-value < $2.2×10^{-16}$) and topological modules (p-value < 0.0027).On the other hand, proteins in functional modules did not show low evolutionary rates. To exclude the effect of low evolutionary rates of more interactors (hub proteins) [14], we conducted Wilcoxon rank sum test on module proteins that do not contain hub proteins (>30 connection degrees). Proteins in modules, even not containing hub proteins, also show significantly low evolutionary rates than those not in modules, in both protein complexes (p-value < $2.2×10^{-16}$) and topological modules (p-value < 0.004577). It shows that proteins in modules will be more conserved than those not in modules.

Modularity in the evolution of yeast protein interaction network

As described above, we proposed a hypothesis of modularity in the evolution of the yeast protein interaction network. This modular evolution hypothesis is consistent with the finding by Qin et al. that two proteins tend to interact with each other if they are in the same or similar evolutionary categories [8]. Moreover, this modular evolution model is also consistent with the finding by Fernandez that modularity (Q-measure) was increased associated with evolutionary change in the yeast protein interaction network [10]. Van Dam et al. showed the evolutionary dynamics of protein complexes are, by and large, not the result of network rewiring, but mainly due to genomic acquisition or loss of genes coding for subunits [15]. Proteins do not function by themselves and need to form modules. Modularity is a remarkable characteristic of biological network [8, 16-18], and we also saw it in the yeast protein interaction network.

Conclusion:

We here propose a hypothesis of modularity in the evolution of yeast protein interaction network based on molecular evolutionary evidence. We found that all the almost half of hub proteins are evolutionarily new. Newly emerged proteins are considered to prefer to connect not to already-existed proteins but to newly emerged proteins in the same evolutionary age. It suggests that hub proteins appeared and

interacted with other simultaneously emerged proteins that form what we call "modules". Examining the evolutionary processes of protein complexes, functional modules and topological modules, we also found that member proteins of these modules tend to appear in one or two evolutionary ages. Moreover, proteins in protein complexes and topological modules show significantly low evolutionary rates than those not in these modules. Our results suggest a hypothesis of modularity in the evolution of yeast protein interaction network as systems evolution.

Conflict of interests:

No conflict of interest was declared.

Reference:

[1] Guldener U *et al*. *Nucleic Acids Res.* 2005 **33**: D364 [PMID: 15608217]
[2] Giot L *et a*l. *Science* 2003 **5**: 1727 [PMID: 14605208]
[3] Watts D & Strogatz S, *Nature* 1998 **393**: 440 [PMID: 9623998]
[4] Barabasi AL & Albert R, *Science* 1999 **286**: 509 [PMID: 10521342]
[5] Eisenberg E & Levanon EY, *Phys Rev Letter*. 2003 **91**: 138701 [PMID: 14525344]
[6] Romualdo PS *et al. Santa Fe Institute working paper* 2002 02-02-008
[7] Qin H *et al*. *Proc Natl Acad Sci USA* 2003 **100**: 12820 [PMID: 14557537]
[8] Yook SH *et al. Proteomics*. 2003 **4**: 928 [PMID: 15048975]
[9] Barabási AL & Oltvai ZN, *Nat Rev Genet.* 2004 **5**: 101 [PMID: 14735121]
[10] Fernandez A, *PLoS Comp Biol.* 2007 **3**: e226 [PMID: 17997598]
[11] Altschul SF *et al. J Mol Biol.* 1990 **215**: 403 [PMID: 2231712]
[12] Pearson WR & Lipman DJ, *Proc Natl Acad Sci USA.* 1988 **85**: 2444 [PMID: 3162770]
[13] Hirsh AE *et al*. *Mol Biol Evol.* 2005 **22**: 174 [PMID: 15371530]
[14] Fraser HB *et al. Science*. 2006 **296**: 750 [PMID: 11976460]
[15] Van Dam TJ & Snel B, *PLoS Comput Biol.* 2008 **4**: e1000132 [PMID: 18711636]
[16] Han JD *et al*. *Nature* 2004 **430**: 88 [PMID: 15190252]
[17] Simonis N *et al*. *J Mol Biol.* 2006 **363**: 589 [PMID: 16973176]
[18] Zhao J *et al. BMC Bioinformatics* 2006 **7**: 386 [PMID: 16916470]