Chinese Science Bulletin 2003 Vol. 48 No. 20 2226—2230

Analysis of correlations between protein complex and protein-protein interaction and mRNA expression

CAI Lun[1], XUE Hong[2], LU Hongchao[1], ZHAO Yi[1], ZHU Xiaopeng[2], BU Dongbo[1], LING Lunjiang[2] & CHEN Runsheng[2,1]

1. IIP Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2. Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

Correspondence should be addressed to Chen Runsheng (e-mail: crs@sun5.ibp.ac.cn)

Abstract **Protein-protein interaction is a physical interaction of two proteins in living cells. In budding yeast *Saccharomyces cerevisiae*, large-scale protein-protein interaction data have been obtained through high-throughput yeast two-hybrid systems (Y2H) and protein complex purification techniques based on mass-spectrometry. Here, we collect 11855 interactions between total 2617 proteins. Through seriate genome-wide mRNA expression data, similarity between two genes could be measured. Protein complex data can also be obtained publicly and can be translated to pair relationship that any two proteins can only exist in the same complex or not. Analysis of protein complex data, protein-protein interaction data and mRNA expression data can elucidate correlations between them. The results show that proteins that have interactions or similar expression patterns have a higher possibility to be in the same protein complex than randomized selected proteins, and proteins which have interactions and similar expression patterns are even more possible to exist in the same protein complex. The work indicates that comprehensive integration and analysis of public large-scale bioinformatical data, such as protein complex data, protein-protein interaction data and mRNA expression data, may help to uncover their relationships and common biological information underlying these data. The strategies described here may help to integrate and analyze other functional genomic and proteomic data, such as gene expression profiling, protein-localization mapping and large-scale phenotypic data, both in yeast and in other organisms.**

Keywords: protein-protein interaction, mRNA expression, protein complex, correlation, yeast.

DOI: 10.1360/03wc0123

Genomic researches have entered a post-genomic era, when new techniques and new methods are introduced to produce large-scale original biological data, such as mRNA mciroarry data[1,2], protein-protein interaction data[3—7], protein-localization mapping data[8] and large-scale phenotypic data[9]. Basic cellular physiological and biochemical principles are underlying in these large-scale data. So it is a challenge in bioinformatics that how to analyze these data[10—12]. Budding yeast *Saccharomyces cerevisiae* is the simplest model eukaryotic organism, whose genetic background was well studied. To date, large-scale protein-protein interaction data, protein complex data and genome-wide mRNA expression data have largely accumulated in yeast[13,14], while only limited such data sets have accumulated in other organisms, including prokaryotic organisms. So correlations between protein-protein interaction, protein complex and mRNA expression can be analyzed based on these large-scale data in yeast. Some studies have reported correlations between protein-protein interaction data from yeast two-hybrid systems (Y2H) and mRNA expression profiling[15]. But debates on this issue came later[16]. Recently, the reports by Ho Y et al. and Giaever G et al. were the largest contributions to dating in the effort to generate large-scale protein-protein interaction data sets[5,6]. Using these additional protein-protein interaction data, we can re-study the correlation between protein-protein interaction and gene expression more comprehensively. No report has focused on correlations of protein complex, protein-protein interaction and mRNA expression together. Here, their correlations will be analyzed and the contributions of protein-protein interaction and mRNA expression to a protein pair being in same complex will be studied.

1 Data and methods

(i) Data sources. Protein-protein interaction data, mRNA expression data and protein complex data of yeast were obtained publicly from Internet. Protein-protein interaction data were from Mering C et al.[7]. For more reliable data to be used, we selected the high- and medium-confidence summing up to 11855 interactions among 2617 proteins through the total more than 80000 interactions for later analysis. Genome-scale gene-expression data of three yeast cell lines were from Yeast Cell Cycle Analysis Project in Stanford University[17]. Three yeast's cell lines were cells of cell division cycle synchronized by factor arrest (ALPHA; 18 time points), after centrifugal elutriation (ELU; 14 time points), and with a temperature-sensitive cdc15 mutant (CDC15; 15 time points). Yeast protein complex data, which include 789 complexes, were from BIND (The Biomolecular Interaction Network Database, http://www.bind.ca/)[14], including 1573 proteins and 6.27 proteins in each complex meanly. Total 2617 proteins, which are in the protein-protein interaction network, and their corresponding expression data and complex data were selected for later analysis.

(ii) Translate complex data sets and expression data sets to pairwise formats. Protein-protein interaction data indicate relationships of each protein pairs in yeast genome, while expression data indicate seriate expression levels of each gene and complex data indicate the relationships of many proteins. For analysis together, the mRNA expression data and protein complex data were both

translated to pairwise formats.

(1) Translate gene expression data sets to pairwise formats. Gene expression data indicate that seriate expression levels and the expressions of a gene at seriate time points may be seen as its expression vector. Correlation between to genes' expressions is the similarity of their expression vectors. Pearson product-moment correlation coefficient r between genes of each pair was measured as the pairwise similarity. The formula is

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{N}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{N}(y_i - \bar{y})^2\right)}}, \quad (1)$$

where both genes have expressions of seriate N time points; and x_i and y_i indicate their expression levels in ith time point respectively; and $\bar{x}$ and $\bar{y}$ indicate their mean expression levels of N time points respectively. The value of a correlation coefficient r is from +1 to –1. And if a pair of gene co-express, it would be significantly positive, and else it would be significantly negative. If expression data missed on a time point, the gene was removed in later analysis.

(2) Translate protein complex data sets to pairwise formats. Protein complex data indicate the relationships of many proteins, and a simple strategy was applied to translating them to pairwise formats. If two proteins exist in the same complex, they are correlated (indicated as 1); and else they are uncorrelated (indicated as 0).

(iii) Correlation analyses between protein complex and protein-protein interaction and mRNA expression. After the translation above, mRNA expression data and protein complex data were indicated as pairwise formats. So they can be analyzed together with protein-protein interaction data.

Firstly, the correlation between protein-protein interaction and mRNA expression was analyzed. The value of a correlation coefficient r, from +1 to –1, was split into 20 even spans. Each protein pair of 2617 proteins was assigned to one of the 20 spans according to their values of r. Total pair number N and the interacted pair number M in each of the 20 spans were counted, and their ratio P was calculated:

$$P = 100\% \times M / N. \quad (2)$$

Secondly, the probabilities of an interacted protein pair and a randomized protein pair being in the same complex were compared. For all protein pairs of 2617 proteins, the ratio of co-complex pair number and total protein pair number was calculated. For all interacted protein pairs, the similar ratio was calculated too.

Thirdly, the correlation between protein complex and mRNA expression was analyzed. Total pair number N and the co-complex pair number M in each of the 20 correlation coefficient spans were counted, and their ratio P was calculated as eq. (2).

Lastly, for total interacted protein pairs, the correlation between protein complex and mRNA expression was analyzed. Total interacted pair number N and the co-complex interacted pair number M in each of the 20 correlation coefficient spans were counted, and their ratio P was calculated as eq. (2).

2 Results

(i) Correlation between protein-protein interaction and mRNA expression. The correlations between protein-protein interaction and three mRNA expression data sets were analyzed, i.e. mRNA expression of three lines of yeast cells of the cell cycle synchronized by factor arrest (ALPHA), after centrifugal elutriation (ELU), and with a temperature-sensitive cdc15 mutant (CDC15).

In all three cell lines, mRNA expression and protein-protein interaction have obvious correlation, and more similar two proteins' expressions are, more possibly do they interact each other (Fig. 1). Such as, in the correlation coefficient span [−0.2, −0.3], the interacted protein pair has a ratio of 1.4 ‰(ALPHA), 2.2 ‰(CDC15) and 1.1 ‰ (ELU), respectively, and 1.8 ‰meanly; and in the correlation coefficient span [0,0.1], the interacted protein pair has a ratio of 2.8 ‰(ALPHA), 3.1 ‰(CDC15) and 2.6 ‰ (ELU) respectively, and 2.8‰ meanly; while in the correlation coefficient span [0.9, 1], the interacted protein pair has a ratio of 71.8 ‰(ALPHA), 72.3 ‰(CDC15) and 46.5 ‰(ELU) respectively, and 63.5 ‰meanly.

(ii) Interacted proteins are more possible to be in the same complex. Analysis of 11855 protein-protein interactions and 789 protein complexes shows that two interacted proteins have a possibility of 23.4% being in the same complex, while randomized selected protein pair's possibility is 0.28%. The result is consistent with some basic processes of protein in living cells. Proteins in a complex must interact directly or indirectly, while interacted proteins may not be in the same complex for they may regulate each other through interactions instantaneously.

(iii) Proteins with similar expressions have more abilities to being in the same complex. In all three yeast cell lines, expression correlation coefficient of each pair of total 2617 proteins was calculated and assigned to the 20 spans of correlation coefficient. In each span, the ratio of co-complex pair number and total pair number was calculated (Fig. 2). In the correlation coefficient span [0.9, 1], the protein pair has possibility of 6.8% (ALPHA), 4.8% (CDC15) and 4.1% (ELU) being in the same complex respectively, while randomized protein pair has possibility of 0.56% (ALPHA), 0.51% (CDC15) and 0.49% (ELU) being in the same complex respectively. The result shows that, in all three cell lines, protein pairs with similar expressions (correlation coefficient significant positive) are more possible in the same complex than that with no

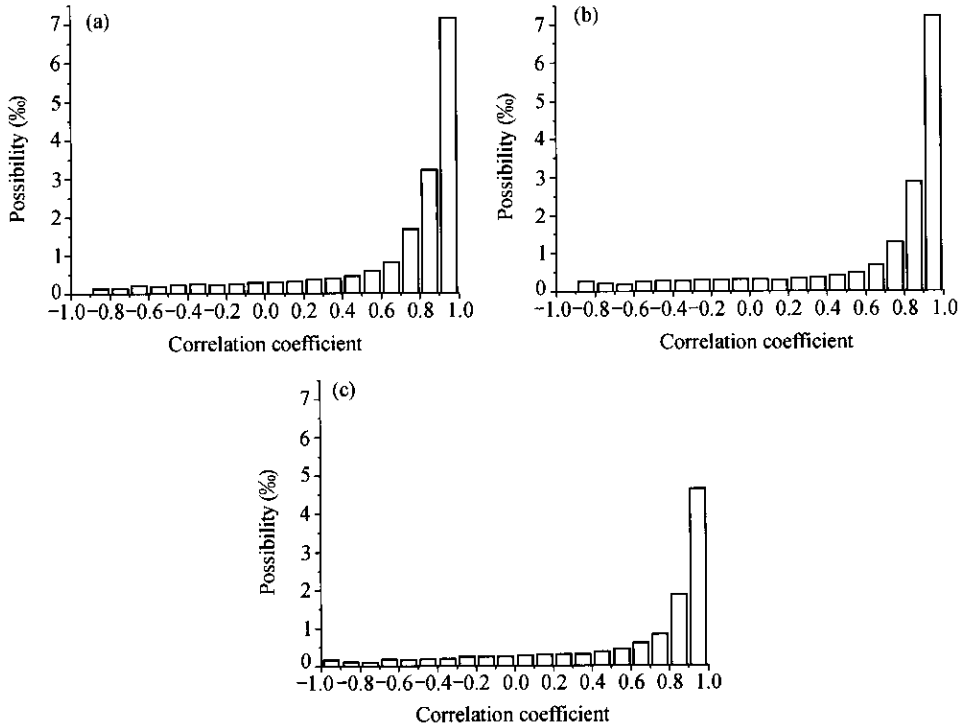

Fig. 1. Possibility (‰) of interacted protein pair (vertical axis) in each of 20 correlation coefficient spans (horizontal axis) of three yeast gene expression data sets are summarized. ALPHA: Cell cycle synchronized by factor arrest; CDC15: with a temperature-sensitive cdc15 mutant; ELU: after centrifugal elutriation.

similar expressions or random protein pairs. So, proteins in the same complex are more possible to co-express in living cells.

Are proteins in the same complex more possible to suppress the expressions of each other? Fig. 2 shows that, in the correlation coefficient span of [−1, −0.9] of all three cell cycles, the protein pair has possibility of 5.3% (ALPHA), 1.4% (CDC15) and 0.9% (ELU) to be in the same complex respectively. It seems that protein pairs with opposite expressions (correlation coefficient significant negative) are more possible in the same complex than that with no similar expressions or random protein pairs, especially to ALPHA (Fig. 2). But the analysis here cannot prove this issue for the pair numbers in this span are too little for statistic analysis. In the correlation coefficient span of [–1, –0.9] of all three cell lines, the protein pair number is 1 (ALPHA), 1 (CDC15) and 32 (ELU) respectively. To test this issue, more protein complex data and mRNA expression data are needed for more comprehensive analysis.

(iv) Correlation between protein complex and mRNA expression of interacted proteins. After analyses of the correlations between any two of protein complex, protein-protein interaction and mRNA expression, the interacted protein pairs were used to analyze the correlation between mRNA expression and protein complex. Fig. 3 shows the possibilities of interacted protein pair being in the same complex in the 20 correlation coefficient spans. The result shows that a protein pair with interaction and similar expression tends to be in the same complex. In all three cell lines, an interacted protein pair is more possible to be in the same complex if they are co-expressed in cell lines. Such as, in the correlation coefficient span [0.9, 1], the interacted protein pair has a ratio of 49% (ALPHA), 39% (CDC15) and 45% (ELU) being in the same complex respectively, and 44.7% meanly.

3 Conclusion and discussion

The analyses in this article were based on public protein complex data, protein-protein interaction data and mRNA expression data. First, the correlation between protein-protein interaction and mRNA expression was analyzed. Second, the correlation of protein-protein interaction and protein complex and the correlation of mRNA expression and protein complex were analyzed. Third, the correlations of protein-protein interaction, mRNA expression and protein complex were analyzed. The results show that there are correlations between any two of protein complex, protein-protein interaction and mRNA expression. Proteins in the same complex tend to interact each other and co-express. In the three yeast cell lines, i.e. ALPHA, CDC15 and ELU, two interacted and co-expressing proteins are more possible to be in the same

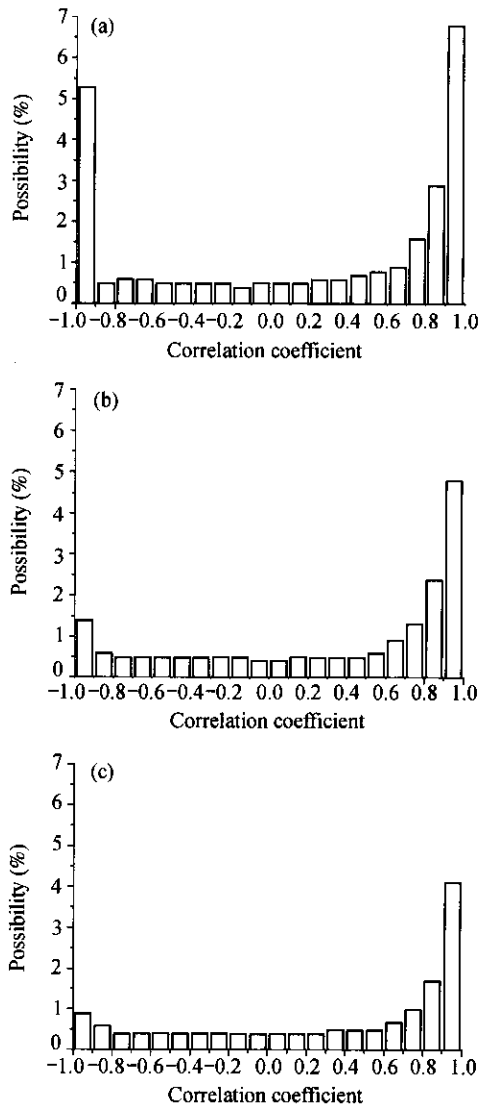

Fig. 2. Possibility (%) of a protein pair being in same protein complex (vertical axis) in each of 20 correlation coefficient spans (horizontal axis) of three yeast gene expression data sets are summarized. ALPHA: cell cycle synchronized by factor arrest; CDC15: with a temperature-sensitive cdc15 mutant; ELU: after centrifugal elutriation.

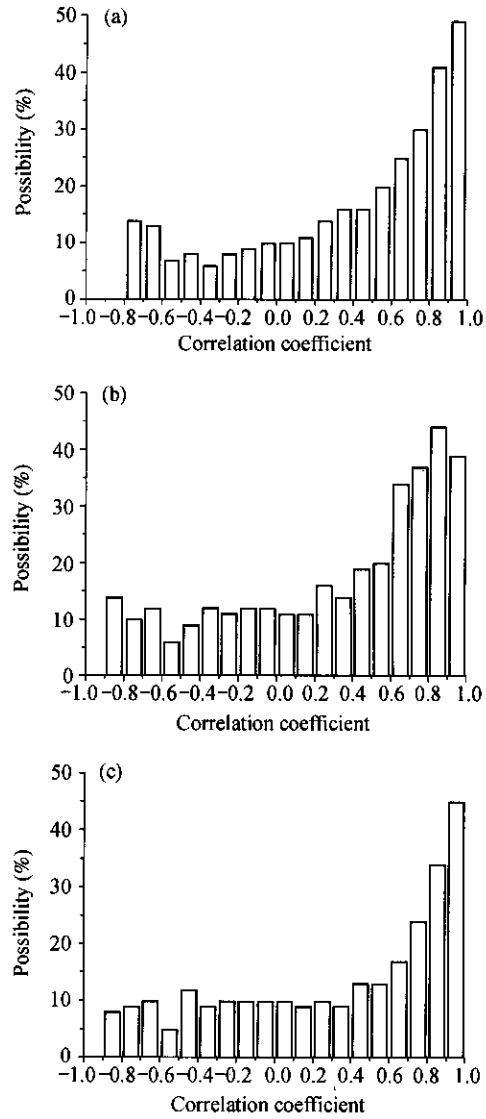

Fig. 3. Possibility (%) of an interacted protein pair being in same protein complex (vertical axis) in each of 20 correlation coefficient spans (horizontal axis) of three yeast gene expression data sets are summarized. ALPHA: Cell cycle synchronized by factor arrest; CDC15: with a temperature-sensitive cdc15 mutant; ELU: after centrifugal elutriation.

complex, and the mean possibilities are 44.7%, while randomized selected two proteins only have a possibility of 0.28% being in the same complex.

Our analysis was based on public large-scale bioinformatical data from Internet. So the related data sets must be collected firstly. Gene expression data of different experiment conditions are abundant and freely acquired. Protein-protein interaction data and protein complex data are uncompleted to date. Although there are these large-scale two kinds of data sets available in budding yeast, only part of the total more than 6000 proteins have been included[7]. Some classic pathways have not been in the protein-protein interaction network (not published), and some positive false and negative false in protein-protein interaction data should be determined[7]. Protein complex data are also uncompleted. So, based on these uncompleted data, only limited analyses could be done now. With more protein-protein interaction data and more protein complex data, the analyses would bring better results. If large-scale such data accumulated in other organisms, the strategies could be used on these organisms.

Genome-wide studies produced large-scale data of

protein functions, gene regulations and protein-protein interactions. But one of the most challenges is how to integrate and analyze these different source data together. We introduced simple and intuitional methods to deal with mRNA expression data and protein complex data, and analyzed them together with protein-protein interaction data. Clustering was widely used in mRNA expression data analysis[17]. Here, to obtain expression relationship of protein pair, Pearson product-moment correlation coefficient was introduced. Protein complex data were also been translated to the relationship of protein pair. So these three different source large-scale data sets could be analyzed together.

The correlation between protein-protein interaction and mRNA expression is uncertain to date[15,16]. Our result assists with the correlation between them. As mentioned above, the protein-protein interaction data are uncompleted and have many false positive and false negative. So, the analysis is influenced by what protein-protein data sets were used. Different data sets could result in different conclusions. Most of previous studies on this issue were based on protein-protein interactions determined, primarily, using the yeast-two hybrid system[3,4]. We used mixed protein-protein data sets predicted by low-throughput genetic, biochemical and biophysical methods in literary, high-throughput yeast tow-hybrid systems and two newly protein complex purification techniques based on mass-spectrometry[5,6]. These are the mostly comprehensive data available to date, and so the analysis would be close to the real instance.

We analyzed mRNA expression of three lines of yeast cells of the cell cycle synchronized by factor arrest (ALPHA), after centrifugal elutriation (ELU), and with a temperature-sensitive cdc15 mutant (CDC15). The results show that, in all three expression data sets, there are correlations between protein-protein interaction, protein complex and mRNA expression, which indicate the correlations are universal. But in some cases, results from the three cell lines are inconsistent. Such as, in Result (iii), the phenomenon that proteins in the same complex tend to suppress the expressions of each other needs more protein complex data for further analysis. In addition, in Fig. 3 CDC15, that the ratio of span [0.8,0.9] (44%) is larger than that of span [0.9, 1] (39%) would be a chance.

Acknowledgements This work was supported by the National High Technology Development Program of China (Grant No. 2002AA231031), the "973" Program in Biology (Grant No. 2002CB713805), the National Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KSCX2-2-07), Beijing Science and Technology Commission (Grants Nos. H010210010113 and H020220030130), and the Bioinformatics Program of Institute of Computing Technology of the Chinese Academy of Sciences (Grant No. 20016200C).

References

1. Cho, R. J., Campbell, M., Winzeler, E. et al., A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell, 1998, 2: 65—73.
2. Hughes, T. R., Marton, M. J., Jones, A. R. et al., Functional discovery via a compendium of expression profiles, Cell, 2000, 102: 109—126.
3. Uetz, P., Giot, L., Cagney, G. et al., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, Nature, 2000, 403: 623—627.
4. Ito, T., Chiba, T., Ozawa, R. et al., comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl Acad. Sci. USA, 2001, 98: 4569—4574.
5. Gavin, A. C., Bosche, M., Krause, R. et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature, 2002, 415: 141—147.
6. Ho, Y., Gruhler, A., Heilbut, A. et al., Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, Nature, 2002, 415: 180—183.
7. Mering, C., Krause, R., Snel, B. et al., Comparative assessment of large-scale data sets of protein-protein interactions, Nature, 2002, 417(6887): 399—403.
8. Kumar, A., Agarwal, S., Heyman, J. A. et al., Subcellular localization of the yeast proteome, Genes Dev., 2002, 16(6): 707—719.
9. Tong, A. H. Y., Evangelista, M., Parsons, A. B. et al., Systematic genetic analysis with ordered arrays of yeast deletion mutants, Science, 2001, 294: 2364—2368.
10. Clark, T. A., Sugnet, C. W., Manuel, A. J., Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays, Science. 2002, 296(5569): 907—910
11. Maslov, S., Sneppen, K., Specificity and stability in topology of protein networks. Science, 2002, 296: 910—913.
12. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M. et al., Evolutionary rate in the protein interaction network, Science, 2002, 296:750—752.
13. Mewes, H. W., Frishman, D., Güldener, U. et al., MIPS: a database for genomes and protein sequences, Nucleic Acids Res., 2002, 30: 31—34.
14. Bader, G. D., Donaldson, I., Wolting, C. et al., BIND--the biomolecular interaction network database, Nucleic Acids Res., 2001, 29(1): 242—245.
15. Ge, H., Liu, Z., Church, G. M. et al., Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae, Nature Genet., 2001, 29: 482—486.
16. Mrowka, R., Liebermeister, W., Holste, D., Does mapping reveal correlation between gene expression and protein-protein interaction? Nature Genet., 2003, 33(1): 15—16.
17. Eisen, M. B., Spellman, P. T., Brown, P. O. et al., Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA, 1998, 95(25): 14863—14868.

(Received June 26, 2003; accepted August 14, 2003)