Supplementary material

The coordinated evolution of yeast proteins is constrained by functional modularity

Yiwen Chen and Nikolay V. Dokholyan

Department of Physics and Astronomy and Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
Corresponding author: Dokholyan, N.V. (dokh@med.unc.edu).

This supplementary material accompanies the article by Chen *et al.* published in the August issue of *Trends in Genetics*.

Material and methods

Evolutionary rates estimation

We used the data set of the evolutionary rates of yeast proteins from a recent study by Hirsh *et al.* [1] The evolutionary rate d_i of a protein i is defined as the ratio of non-synonymous (dN) substitution per site to synonymous substitution (dS) per site of this protein [2]. The evolutionary rates of 3036 open reading frames were calculated by Hirsh *et al.* [1], based on four complete genomes of *Saccharomyces* species (*Saccharomyces paradoxus*, *Saccharomyces cerevisiae*, *Saccharomyces mikatae* and *Saccharomyces bayanus*) [3], with corrections for non-neutral dS [1].

Measure of co-evolution strength of expression level

We used codon adaptation index (CAI) [4] of each gene as a proxy for its expression level [2,5–8], kindly provided by Aaron. E. Hirsh [1]. The CAI value is defined as following:

$$\mathrm{CAI} = \left(\prod_{k=1}^{L} w_k\right)^{\frac{1}{L}},$$

where L is the total number of codons in a given gene and w_k is the ratio of the frequency of the k^{th} codon over the frequency of the optimal codon for the same amino acid. Both frequencies are calculated from a set of highly expressed genes in the organism being studied [4]. For a given pair of orthologous gene sets, we use the Pearson correlation coefficient between the CAI values in the corresponding sets (four CAI values in each set) as a measure of the co-evolution strength of expression level between this pair [5]. We calculated all pairwise CAI correlation coefficients within and between modules.

Characterization of functional modules

The formation of functional modules by proteins is achieved by physical or functional interactions between proteins. Here, we define interactions that result in proteins association as physical interactions and those between two proteins having no direct physical interactions as functional interactions. For example, a repressor protein has a functional interaction with a protein that is encoded in the gene it is repressing. Various functional genomics techniques such as microarray [9], yeast two-hybrid [10,11], synthetic lethal screens [12] and affinity purification coupled with mass spectrometry [13,14] are available for characterizing either physical or functional interactions between proteins. The interactions characterized in these experiments feature functional associations between proteins. However, each of these techniques captures different aspects of functional associations. Methods that rely solely on the experimental data obtained from one technique can give rise to systematic errors in the characterization of functional modules. For example, some modules formed through functional interactions can be missed in a method of module characterization solely based on the physical protein–protein interaction data

[10,11] because this data does not have explicit information on functional interactions. In addition, different techniques vary considerably in accuracy and precision, and functional associations established in one experiment can have different confidence levels from those established in other types of experiments. Therefore, to characterize functional modules comprehensively and reliably, it is essential to integrate the information obtained from different functional genomics experiments and use well-curated annotation references to test the validity of the characterization.

Lee *et. al.* [15] systematically characterized functional modules in *S. cerevisiae* by integrating the information obtained from different functional genomics experiments [9,10,12–14], comparative genomics analysis and literature mining. They showed the validity of the characterization by comparing the characterized modules with well-curated annotation references. Here, we use the functional modules characterized by Lee *et al.* [15] containing 3285 genes in *S. cerevisiae.* The data set of the characterized functional modules and a detailed description of the methodologies can be found in Ref. [15] and the supplementary materials therein.

Identification of duplicate genes

The orthologous and paralogous genes in 43 genomes of bacteria, archaea and *S. cerevisiae* are cataloged as clusters of orthologous groups (COG) [16] at NCBI COG database. Here, we used the gene family data of *S. cerevisiae* to identify all duplicate genes in the same module.

Controlling for alternative factors

Proteins with no direct physical interactions

Because proteins with physical interactions have been previously shown to co-evolve in both protein sequence and expression, it is important to control for the possibility that the observed co-evolution within modules is a simple consequence of co-evolution between physically interacting proteins. To estimate how much proteins with no direct physical interactions account for the observed similarity of evolutionary rates, we excluded protein pairs in the same module (~13% of total pairs) that have experimentally supported physical interactions from our analysis [10,11,13,14,17,18]. We found that the differences in evolutionary rates between the remaining pairs in the same module are almost indistinguishable from the original set that contains all pairs (KS test, P >0.6), indicating that the observed pattern for all pairs in the same module is not simply a result of physically interacting proteins. Interestingly, pairs with physical interactions show greater similarity in evolutionary rates than those with no physical interaction (Figure S1a). A similar analysis was performed for co-evolution of expression levels when physically interacting protein pairs in the same module are excluded (Figure S1b).

Expression levels

Recent studies showed that the expression level of a protein has a significant effect on its evolutionary rate [7,8,19]. Thus, it is important to control for the possibility that the greater similarity of the evolutionary rates observed within the same modules is simply due to the greater similarity in expression levels.

We calculate Kendall's partial tau-values [20], a nonparametric measure of partial correlation between the pairwise evolutionary rate difference and the following two variables (while controlling for the effect of each variable; see main text). As before, we used CAI of each gene as a proxy for its expression level [2,5–8]. We found that the evolutionary rate difference exhibits significantly greater correlation with the category variable (τ = 0.17, P <10^{-106}) that reflects functional modularity than with the expression level difference (τ = 0.07, P <10^{-19}), indicating that the expression level contributes much less significantly to the observed patterns.

Duplicate genes

The duplicate gene pairs represent only a small fraction of all pairs in the same module. Therefore, the exclusion of those pairs has almost no effect on the distribution of similarity in evolutionary rates (KS test, P~1).

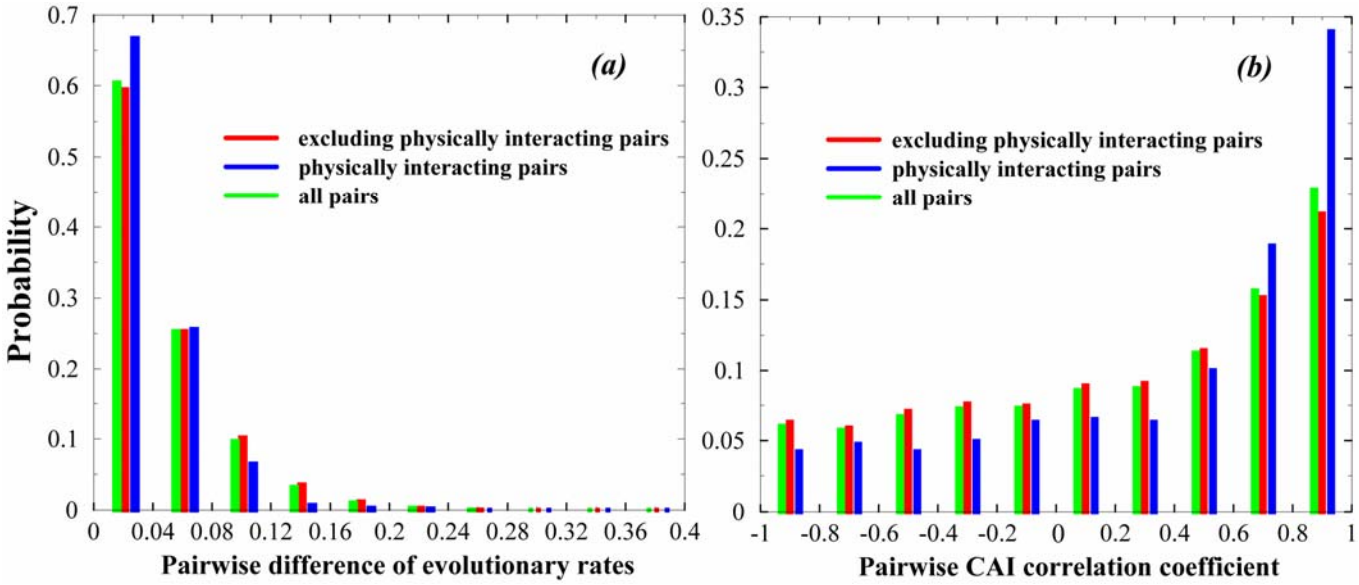

Figure S1. The histograms of **(a)** pairwise differences in evolutionary rates and **(b)** pairwise CAI correlation coefficients between proteins in the same functional module. Three distributions are plotted: all pairs (green), pairs with experimentally-supported interactions (blue) and the pairs with no experimentally supported interactions (red). The histograms in (a) are plotted in the range between 0 and 0.4 with the bin size of 0.04. The histograms in (b) are plotted in the range between –1 and 1 with the bin size of 0.2.

References

1. Hirsh, A. E., *et al.* (2005) Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol. Biol. Evol.* 22, 174-177
2. Wall, D. P. *et al.* (2005) Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. U.S.A* 102, 5483-5488
3. Kellis, M., *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241-254
4. Sharp, P. M. and Li, W. H. (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* 15, 1281-1295
5. Fraser, H. B., *et al.* (2004) Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad .Sci. U.S.A* 101, 9033-9038
6. Pal, C., *et al.* (2003) Rate of evolution and gene dispensability. *Nature* 421, 496-497
7. Drummond, D. A., *et al.* (2005) Why highly expressed proteins evolve slowly. *Proc. Natl. Acad .Sci. U.S.A* 102, 14338-14343
8. Pal, C., Papp, B. and Hurst, L. D. (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158, 927-931
9. Gollub, J. *et al.* (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Research* 31, 94-96
10. Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623-627
11. Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A* 98, 4569-4574
12. Tong, A. H. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294, 2364-2368
13. Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180-183
14. Gavin, A. C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141-147

15. Lee, I., Date, S. V., Adai, A. T., & Marcotte, E. M. (2004) A probabilistic functional network of yeast genes. *Science* 306, 1555-1558
16. Tatusov, R. L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC. Bioinformatics.* 4, 41
17. Tong, A. H. *et al.* (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295, 321-324
18. Xenarios, I. *et al.* (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303-305
19. Rocha, E. P. C. and Danchin, A. (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular Biology and Evolution* 21, 108-116
20. Gibbons, J. D. (1993) *Nonparametric Measures of Association*. Sage Publications