Evolutionary Rate Heterogeneity of Core and Attachment Proteins in Yeast Protein Complexes

Sandip Chakraborty and Tapash Chandra Ghosh*

Bioinformatics Centre, Bose Institute, Kolkata, West Bengal, India

*Corresponding author: E-mail: tapash@jcbose.ac.in.

Accepted: June 23, 2013

Abstract

In general, proteins do not work alone; they form macromolecular complexes to play fundamental roles in diverse cellular functions. On the basis of their iterative clustering procedure and frequency of occurrence in the macromolecular complexes, the protein subunits have been categorized as core and attachment. Core protein subunits are the main functional elements, whereas attachment proteins act as modifiers or activators in protein complexes. In this article, using the current data set of yeast protein complexes, we found that core proteins are evolving at a faster rate than attachment proteins in spite of their functional importance. Interestingly, our investigation revealed that attachment proteins are present in a higher number of macromolecular complexes than core proteins. We also observed that the protein complex number (defined as the number of protein complexes in which a protein subunit belongs) has a stronger influence on gene/protein essentiality than multifunctionality. Finally, our results suggest that the observed differences in the rates of protein evolution between core and attachment proteins are due to differences in protein complex number and expression level. Moreover, we conclude that proteins which are present in higher numbers of macromolecular complexes enhance their overall expression level by increasing their transcription rate as well as translation rate, and thus the protein complex number imposes a strong selection pressure on the evolution of yeast proteome.

Key words: core protein, attachment protein, evolutionary rate, yeast, protein complex, protein complex number.

Introduction

Different proteins evolve at different rates. The central problem in molecular evolution is to comprehend the factors for the differential evolutionary rates of proteins. The origins of variation in protein evolutionary rates have remained the core issue between the selectionists and the neutralists (Fisher 1930; Kimura and Ohta 1974; Kimura 1983; Razeto-Barry et al. 2011). The neutral theory of molecular evolution accentuates that most of the nucleotide substitutions in a gene are due to the random fixation of neutral mutations (Kimura 1983), which is in accordance with the proposal that functionally important genes should evolve slower than less important genes (Kimura and Ohta 1974). According to this proposal, it has been demonstrated that essential genes evolve slower than the genes which are dispensable (Jordan et al. 2002). Moreover, proteins those interact with a large number of partners evolve at a slower rate compared with the proteins with fewer number of interacting partners (Hirsh and Fraser 2001; Jordan et al. 2002; Fraser et al. 2003; Hahn and Kern 2005; Lemos et al. 2005; Chakraborty et al. 2011; Alvarez-Ponce and Fares 2012). However, these results have been questioned by some researchers (Batada et al. 2006). On the other hand, it has been demonstrated that the gene expression level is an important constraint in protein evolution (Subramanian and Kumar 2004; Drummond et al. 2005, 2006; Drummond and Wilke 2008); highly expressed proteins are more conserved than lowly expressed proteins. This phenomenon has been attributed to natural selection (Popescu et al. 2006). Moreover, it was shown that about half of the variation in protein evolutionary rates can be explained by expression level (Drummond et al. 2005).

Proteins do not carry out their functions alone, they often act by participating in macromolecular complexes and play different functional roles (Qiu and Noble 2008; Chakraborty et al. 2010). Additionally, it has been found that the multiprotein complexes are enriched for essential genes (Semple et al. 2008), and proteins those are present in several protein complex assemblies have a tendency to be more essential than proteins that are present in fewer protein complex assemblies (Pereira-Leal et al. 2006). In contrast, Pache et al. (2009) reported that there is no such significant correlation between

gene essentiality and number of protein complex assemblies in which a protein is present. In previous studies, it has been demonstrated that complex-forming proteins evolve more slowly than the noncomplex forming proteins (Teichmann 2002), and proteins that are associated with a large number of complexes are more evolutionary conserved than those proteins which are associated with fewer numbers of complexes (Chakraborty et al. 2010; Das et al. 2013). Moreover, it has been reported that there exists a significant positive correlation between expression level and protein complex number (i.e., the number of protein complex assemblies in which the protein present) in yeast (Chakraborty et al. 2010), and thus it has been proposed that proteins associated with a large number of complexes (i.e., higher protein complex number) need to be produced in greater quantities and hence have higher expression level.

Additional studies on the protein complex assembly tell us about the architecture and the function of different protein subunits in it (Gavin et al. 2006). From the perspective of topology of complex assemblies, protein subunits are categorized mainly in two different groups, that is, core and attachment (Gavin et al. 2006; Pang et al. 2008). Core proteins are always present in all isoforms and execute the main functions (Gavin et al. 2006), whereas attachment proteins are present only in some of the isoforms and act as modifiers of the complex's function (Dezso et al. 2003). Analysis on the subunits of protein complexes revealed that the biochemical role of core proteins is essential for the complex assembly and thus they are irreplaceable (Dezso et al. 2003); although the protein abundance of core subunits in cell was found to be at a lower level than attachment subunits (Pang et al. 2008). However, two recent studies (Semple et al. 2008; Pache et al. 2009) showed that core and attachment subunits are equally important for the complex machinery to function.

In this work, we used the yeast protein complex data set (Gavin et al. 2006) for the classification pertaining to core and attachment proteins and analyzed their evolutionary rates to address the influences of gene dispensability, protein multifunctionality, protein connectivity, protein complex number, and gene expression level on protein evolutionary rates.

We report in this communication that core and attachment proteins are equally essential for survival. We also observed that core proteins evolve faster than attachment proteins in spite of their higher multifunctionality, which is incompatible with the proposal of the neutral theory of molecular evolution. Finally, we demonstrated that proteins which are present in higher numbers of macromolecular complexes increase their transcription rate (number of mRNA molecules per hour) as well as translation rate (number of proteins molecules per second). Moreover, we observed that attachment proteins have a higher protein complex number than core proteins and subsequently increases the expression level of attachment proteins. Thus, the protein complex number imposes a strong selection pressure on attachment proteins and hence can be attuned with the natural selection of protein evolution (Popescu et al. 2006).

Materials and Methods

Protein-Complex Information

The yeast protein complex data of Gavin et al. (2006) was collected from the supplementary data set of Pache et al. (Additional file 3: Gavin complexes: http://www.biomedcentral.com/1752-0509/3/74/additional, last accessed July 15, 2013) (Pache et al. 2009). There were 491 complexes with 1,487 unique protein subunits. Gavin et al. (2006) used the affinity purification technique coupled with mass spectrometry to purify the subunits of protein complexes. After repeated purification of several yeast protein complexes, they found that the same protein complexes could contain different protein subunits and termed them as "complex isoform." Proteins within each complex isoform were classified into two major classes by their clustering property, that is, core and attachment protein. In this data set, we found 357 and 339 proteins as core and attachment, respectively; however, there are 791 proteins that act as core and attachment protein subunit in the complex isoforms (supplementary table S1, Supplementary Material online). The protein complex number is defined as the number of protein complex assemblies in which a protein is present (Chakraborty et al. 2010). Recently, Benschop et al. (2010) identified 518 protein complexes, out of which 409 were determined by forward–backward module detection (i.e., FBMD). They (Benschop et al. 2010) also determined 350 consensus protein complexes by integrating their data with Pu et al. (2007) and Hart et al. (2007). This consensus data set is more accurate than other predicted data set (Benschop et al. 2010), but there is no information about core and attachment subunits in their data set. Thus, we used these consensus data sets only to compute the protein complex number.

Protein Multifunctionality

The number of unique biological processes (i.e., multifunction) (Salathé et al. 2006; Podder et al. 2009) of a protein was calculated from the Gene Ontology (GO) project (Camon et al. 2004); for each yeast protein, the corresponding child term of GO biological processes (GO-BP) was retrieved from Ensembl BioMart (version 69) (Flicek et al. 2011). Alternatively, we also used the number of biological pathways as a measure of multifunctionality. To calculate the unique pathway number, we used the MIPS-CYGD (Munich Information centre for Protein Sequences-Comprehensive Yeast Genome Database) (http://mips.helmholtz-muenchen.de/genre/proj/yeast/Search/Catalogs/searchCatfirstFun.html, last accessed July 15, 2013) (Guldener et al. 2005; Mewes et al. 2011) and the KEGG database (Kanehisa et al. 2012) separately.

Essential Genes and Calculation of Essentiality

To obtain a data set of essential genes of yeast, we downloaded the essential genes list from the OGEE database (http://ogeedb.embl.de/#summary, last accessed July 15, 2013) (Chen et al. 2012). Moreover, we used the fitness quantitative data of Steinmetz et al. (2002) to calculate the essentiality for each gene. The fitness data were downloaded from Pache et al. (2009) (Additional file 3: Quantitative fitness data: http://www.biomedcentral.com/1752-0509/3/74/additional, last accessed July 15, 2013) and considered only the growth in the YPD medium (1% Bacto-peptone, 2% yeast extract, and 2% glucose). The normalized essentiality (E_i) was calculated by

$$E_i = \left(1 - \frac{f_i}{f_{\max}}\right)$$

where f_i represents the fitness values for the deletion of the ith gene and $f_{\max}$ represents the maximum fitness value. Therefore, the gene essentiality ranges from 0 to 1, with an essentiality of 0 denoting that a gene has no measurable effect on fitness, whereas 1 denotes the gene is essential for survival.

Calculation of Evolutionary Rate

We calculated the average d*N*/d*S* ratio as a measure of evolutionary rate, which is the ratio of the number of nonsynonymous substitutions per nonsynonymous site (d*N*) to the number of synonymous substitutions per synonymous site (d*S*). To calculate the d*N*/d*S* ratio, we compared *S. cerevisiae* sequences with its orthologous sequences in *S. paradoxus*, *S. bayanus,* and *S. mikatae*, as they are sibling species (Naumov et al. 1992). We obtained whole coding sequences of four yeast strains from the Saccharomyces Genome Database (Cherry et al. 2012) (for *S. cerevisiae:* http://downloads.yeastgenome.org/sequence/S288C_reference/orf_dna/orf_coding.fasta.gz [last accessed July 15, 2013]; *S. paradoxus*: http://downloads.yeastgenome.org/sequence/fungi/S_paradoxus/archive/MIT/orf_dna/orf_genomic.fasta.gz [last accessed July 15, 2013]; *S. bayanus*: http://downloads.yeastgenome.org/sequence/fungi/S_bayanus/archive/MIT/orf_dna/orf_genomic.fasta.gz [last accessed July 15, 2013]; and *S. mikatae*: http://downloads.yeastgenome.org/sequence/fungi/S_mikatae/archive/MIT/orf_dna/orf_genomic.fasta.gz [last accessed July 15, 2013]) to calculate the evolutionary rates (d*N*/d*S*). Then, we translate these sequences into protein sequence using the EMBOSS Transeq program (http://www.ebi.ac.uk/Tools/st/emboss_transeq/, last accessed July 15, 2013). By using the National Center for Biotechnology Information BlastP program (version 2.2.17) (Altschul et al. 1997), we identified the orthologous proteins of *S. cerevisiae* in *S. paradoxus*, *S. baynus,* and *S. mikatate*. We used the expectation value 1.0×10^{-5} as a cutoff and maintained at least 75% sequence similarity between the two sequences with a minimum alignment overlap 80%. The gaps allowed in the alignment were less than 3%. We also verified our results with the results of Kellis et al. (2003). Then the pair-wise alignment was performed using the ClustalW (version 2.0) (Larkin et al. 2007) for each set of orthologs gene pair. d*N*/d*S* values were calculated by Yang and Nielsen method (Yang and Nielsen 2000) using the PAML package (version 4) (Yang 2007). We took the average values of d*N*/d*S* for each *S. cerevisiae* ORF where at least one orthologous pair was present (supplementary table S2, Supplementary Material online).

Protein–Protein Interactions Data

We collected the protein–protein interactions (PPIs) information from the DIP (Database of Interacting Proteins: http://dip.doe-mbi.ucla.edu/, last accessed July 15, 2013) (Salwinski et al. 2004), MINT (Molecular INTeraction Database: http://mint.bio.uniroma2.it/mint/, last accessed July 15, 2013) (Licata et al. 2012), and IntAct (http://www.ebi.ac.uk/intact, last accessed July 15, 2013) (Kerrien et al. 2012). In those databases, the PPIs were documented experimentally by genome wide two-hybrid screen, immune precipitation, affinity binding, antibody blockage, and so on. To collect high-throughput PPIs data from the DIP database, we used the CORE data set of *S. cerevisiae* (baker's yeast) (Scere20120228CR). In the CORE data set, the PPIs were identified by high-throughput methods and small-scale experiments, thus the data in the CORE are highly reliable (Deane et al. 2002). However, there is no predefined high confidence data in MINT and IntAct, thus we used the confidence score to collect the high-throughput data from MINT and IntAct. We calculated the average confidence score for each data set and then used this average value as a cutoff to identify the high-throughput PPIs (for MINT, the cutoff score is 0.30, and for IntAct, the cut off score is 0.40). After merging the three databases (DIP, MINT, and IntAct) and removing the redundant and self-interactions, we obtained 3,580 individual proteins and 12,624 binary interactions (fig. 1) (supplementary table S3, Supplementary Material online).

Protein Expression, Transcription Rate, and Translation Rate

The gene expression (number of mRNAs molecules per cell) and transcription rate (number of mRNAs molecules per hour) were collected from genome wide expression analysis of Holstege et al. (1998) (http://web.wi.mit.edu/young/pub/data/orf_transcriptome.txt, last accessed July 15, 2013). We also used a more up-to-date (Miller et al. 2011) expression data set for validation purposes. Dynamic transcriptome analysis was used to realistically monitor the mRNA metabolism in yeast (Miller et al. 2011). We downloaded the translation rate (number of proteins molecules per seconds) of yeast proteins from the data set of Arava et al. (2003).

Software

We used SPSS (version 13.0) (Nie et al. 1970) and Tanagra (version 1.4.36) (Rokotomalala 2005) for all statistical

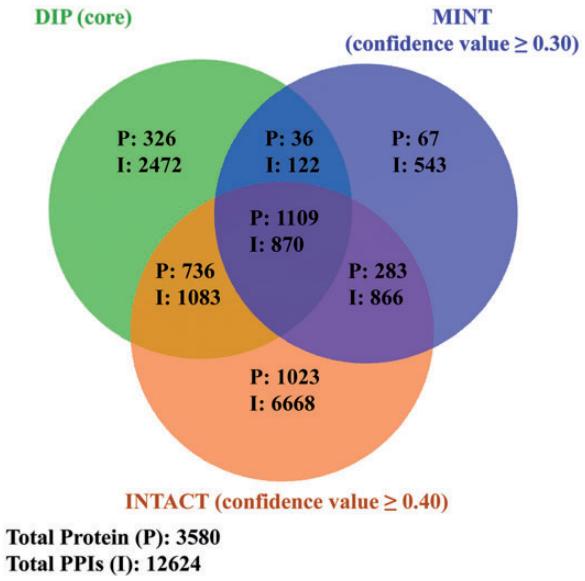

Fig. 1.—Venn diagram of PPI information collected from the DIP, MINT, and IntAct. Here protein is abbreviated by "P" and PPI is abbreviated by "I." The "light green" color indicates the DIP data set domain, "light blue" color indicates the MINT data set domain, and "light saffron" color indicates the IntAct data set domain.

calculations. In all statistical analysis, we used 95% level of confidence as a measure of significance. The network statistic (Degree) was calculated using the Pajek software package (Batagelj and Mrvar 2004).

Results

Evolutionary Rates, Multifunctionality, and Essentiality of Core and Attachment Proteins

Yeast protein complexes were taken from the supplementary data set of Gavin et al. (2006), which comprised 1,487 proteins belonging to 491 protein complexes. In their work, Gavin et al. (2006) identified 357 and 339 subunits as core and attachment proteins, respectively. The remaining 791 subunits are present in both core and attachment proteins. We have not considered those 791 proteins for evolutionary rate analyses due to their presence in both core and attachment proteins. Estimation of evolutionary rates in core and attachment proteins shows that attachment proteins are more conserved than core proteins (average evolutionary rates of core proteins: 0.0901 (±0.0031) ($N = 348$), attachment proteins: 0.0772 (±0.0040) ($N = 325$); Mann–Whitney U test, $P = 2.5 \times 10^{-5}$) (fig. 2).

The function of a protein is one of the most important parameters that influence protein evolution, and it has been reported that multifunctional genes evolve slowly (Wilson et al. 1977; Podder et al. 2009). Moreover, core proteins are always present in all different subsets of isoforms of a complex assembly, and thus they have been reported as the functional units of that complex assembly (Dezso et al. 2003; Gavin et al. 2006). Therefore, the role of each core protein is crucial for the functional integrity of a protein complex (Dezso et al. 2003). Hence, we measured the multifunctionality of core and attachment proteins by counting unique GO-BP terms, which are widely used to calculate the multifunctionality of a protein (Camon et al. 2004; Pál et al. 2006; Salathé et al. 2006; Podder et al. 2009; Su et al. 2010; Flicek et al. 2011; Yang and Gaut 2011). We observed that core proteins are involved in a higher number of biological processes than attachment proteins (table 1). Nonetheless, the GO terms, widely used for functional characterization but several drawbacks have been reported to be associated with the GO terms (Dolan et al. 2005; Park et al. 2011). In particular, GO annotation is systematically redundant and the biological domain is inconsistent (Park et al. 2011). Therefore, we also used the MIPS-CYGD (Guldener et al. 2005; Makino and Gojobori 2006; Chakraborty et al. 2010; Mewes et al. 2011) and the KEGG (Kanehisa et al. 2012) databases, and we counted the number of pathways in which a protein takes part. We found similar trends, that is, core proteins are involved in a higher number of pathways/functions than attachment proteins (table 1). However, we found that core proteins evolve faster than the attachment proteins in spite of their higher multifunctionality. Thus the prevailing idea that, "multifunctional proteins evolve at a slower rate" (Wilson et al. 1977; Podder et al. 2009) does not seem to be compatible in explaining the evolutionary rate differences between core and attachment proteins.

It has been reported that both gene essentiality and protein multifunctionality are negatively associated with protein evolutionary rates (Jordan et al. 2002; Podder et al. 2009). Therefore, it could be expected that essential genes are multifunctional. Indeed, we found a positive association between gene essentiality and multifunctionality in our data set (Spearman's $\rho_{\text{multifunctionality vs. essentiality}} = 0.1689$, $P = 1.0 \times 10^{-6}$), that is, the proportion of essential genes is higher in highly multifunctional proteins. To validate the above results, we further divided our data set into two groups by using the mean value of GO-multifunctionality as a cutoff. Genes with a multifunctionality higher than that of the corresponding mean value were considered as highly functional (HF) ($N = 436$) and those with a lower one, as low-functional (LF) ($N = 260$) groups. We found a significantly (two-sided Fisher's exact test, $P = 1.1 \times 10^{-3}$) higher proportion of essential genes in the HF group (41.28%) compared with the LF group (28.85%); indicating that HF proteins are more essential than the low functional proteins. Hence, one would expect a higher proportion of essential genes in core proteins than attachment proteins. Surprisingly, we observed a similar proportion of essential genes in core and attachment proteins (two-sided Fisher's exact test, Core: 38.66%, Attachments: 34.51%, $P = 2.7 \times 10^{-1}$). Moreover, we found that core and attachment proteins exhibit similar essentiality (average

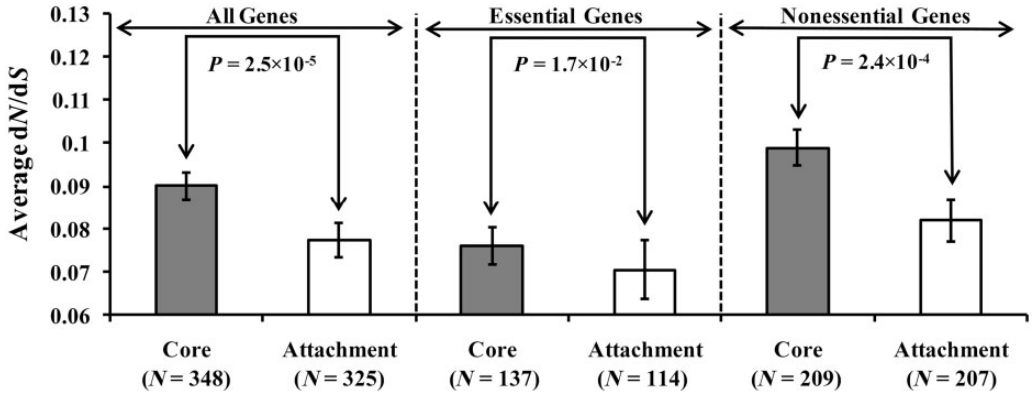

Fig. 2.—Average evolutionary rates (d*N*/d*S*) of core and attachment proteins in yeast. The statistical comparison performed by two-tailed Mann–Whitney *U* test.

Table 1
Average Multifunctionality and Pathways of Core and Attachment Proteins

Database Used	Core	Attachment	Significance Level (*P*)
KEGG	1.8992 (±0.1209) (*N* = 129)	1.6994 (±0.0939) (*N* = 173)	3.3×10^{-2}
MIPS	2.6627 (±0.0873) (*N* = 335)	2.4174 (±0.0801) (*N* = 321)	4.4×10^{-2}
GO-BP	4.2068 (±0.1934) (*N* = 266)	3.3746 (±0.1814) (*N* = 283)	2.1×10^{-5}

essentiality of core proteins: 0.1496 (±0.0091) (*N* = 214), attachment proteins: 0.1685 (±0.0086) (*N* = 213); Mann–Whitney *U* test, $P = 5.7 \times 10^{-2}$). The neutral theory of molecular evolution postulates that functionally important genes evolve more slowly than the less important genes (Hirsh and Fraser 2001; Jordan et al. 2002; Liao et al. 2006). In our data set when we separated the genes into essential and nonessential genes, we also found that essential genes evolve slower than nonessential genes (average evolutionary rate of essential genes: 0.0735 (±0.0039) (*N* = 251), nonessential genes: 0.0904 (±0.0032) (*N* = 416); Mann–Whitney *U* test, $P = 8.5 \times 10^{-5}$). Moreover, when we separated essential and nonessential genes into core and attachment proteins, we found that the evolutionary rates of attachment proteins are significantly lower than core proteins in both the gene pools (fig. 2). Therefore, the observed evolutionary rate differences between core and attachment proteins are independent of gene essentiality. Previously, Hurst and Smith (1999) demonstrated that gene essentiality is a very weak correlate of the rates of protein evolution. They observed that the "non-immune non-essential genes" ("non-immune" genes refers to those genes that do not belong to the immune system) evolve at a similar rate to the "essential genes." Even in this study we also observed no significant differences in evolutionary rates between "essential core proteins" and "non-essential attachment proteins" (average evolutionary rates of essential core proteins: 0.0760 (±0.0042) (*N* = 137), nonessential attachment proteins: 0.0818 (±0.0050) (*N* = 207); Mann–Whitney *U* test, $P = 1.0 \times 10^{-1}$). These results may indicate that there are some other constraints which can explain the variation in rates of protein evolution better than the gene essentiality or multifunctionality.

Protein Connectivity and Evolutionary Rates of Core and Attachment Proteins

In general, a biological system in a cell can be considered as a complex network, and the PPIs are one such network that acts as the source of many biological functions (Mete et al. 2008). Proteins with many interaction partners are expected to be multifunctional. We analyzed the correlation between the number of interacting partners of core/attachment proteins (whose GO-BP annotations and connectivity data were available) and its multifunctionality. As expected, we found a significant positive correlation (Spearman's $\rho_{\text{GO-BP vs. connectivity}} = 0.3005$, $P = 1.0 \times 10^{-6}$) between GO-BP and connectivity in the PPI network. Previously, it has been reported that the protein connectivity in a PPI network is negatively correlated with the evolutionary rates (Hirsh and Fraser 2001; Fraser et al. 2002, 2003; Jordan et al. 2002; Hahn and Kern 2005; Lemos et al. 2005; Chakraborty et al. 2010, 2011; Alvarez-Ponce and Fares 2012). Similar trend has been also observed in our core/attachment data set (Spearman's $\rho_{\text{connectivity vs. } dN/dS} = -0.1907$, $P = 1.0 \times 10^{-6}$). The above observations motivated us to investigate if there exists any variation of connectivity in core and attachment proteins or not. Surprisingly, we did not find any significant difference (average connectivity of core proteins: 8.0364 (±0.3565) (*N* = 357), attachment proteins: 13.8260 (±2.3445) (*N* = 339); Mann–Whitney *U* test, $P = 5.4 \times 10^{-1}$) in the connectivity between core and attachment proteins. This result indicates that the observed

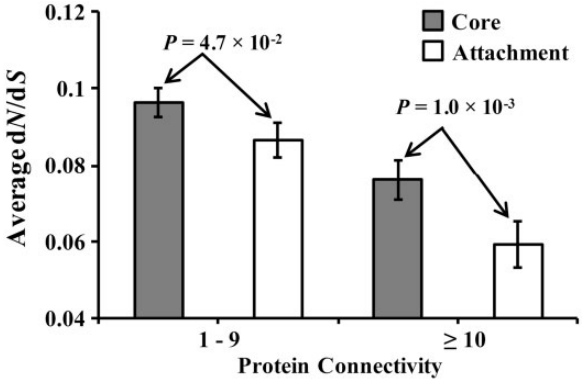

FIG. 3.—Average evolutionary rates (d*N*/d*S*) of core and attachment proteins in two bines. The statistical comparison performed by two-tailed Mann–Whitney *U* test.

difference in protein evolution between core and attachment proteins is independent of protein connectivity. To confirm the above observation, we binned core and attachment proteins into two groups according to their connectivity in the PPI network and analyzed their evolutionary rates. Interestingly, we found that core proteins are still evolving at a higher rate than attachment proteins in same PPI bin (fig. 3). Thus we can say that the observed difference in evolutionary rates between core and attachment proteins may be steered by factors other than the protein connectivity.

Protein Complex Number and Evolutionary Rates of Core and Attachment Proteins

Earlier, it has been demonstrated that protein complex number significantly modulates the rates of protein evolution when compared with protein connectivity in PPI networks (Chakraborty et al. 2010; Das et al. 2013). Thus, we calculated protein complex number in both core and attachment proteins. We observed that protein complex number is significantly higher in attachment proteins than that in core proteins (fig. 4*a*). Furthermore, we reconfirmed this result by using the data set of protein complexes developed by Benschop et al. (2010) and found similar trends (fig. 4*b*). We also found a significant negative correlation between protein complex number and d*N*/d*S* (table 2) which is in agreement with our previously published results (Chakraborty et al. 2010). Therefore, the protein complex number could be the cause of evolutionary rate differences between core and attachment proteins. Interestingly, we have found considerable differences in the distribution of protein complex numbers between core and attachment proteins, ranging from 1 to 3 for core proteins and from 1 to 22 for attachment proteins by considering the data set of Gavin et al. (2006). To verify the effect of protein complex number in controlling the evolutionary rate differences between core and attachment proteins, we divided the whole data of attachment proteins into two groups; attachment proteins with a protein

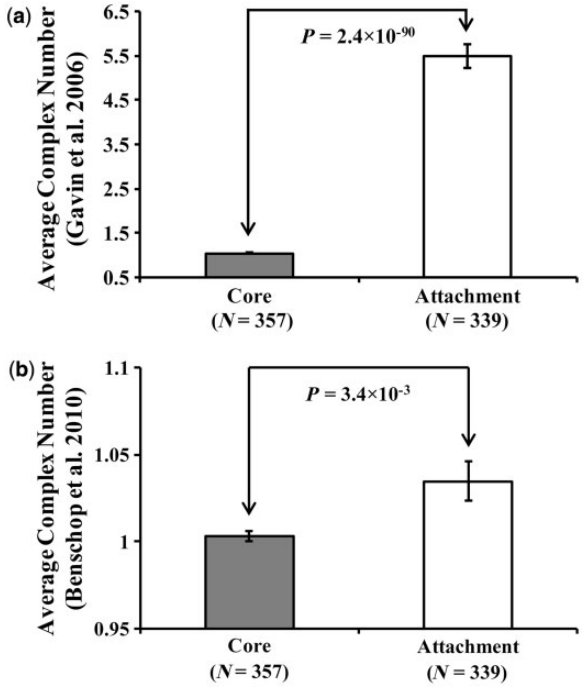

FIG. 4.—(*a*) Average protein complex number (protein complex number calculated by using the Gavin et al. [2006] data set) of core and attachment proteins in yeast. The statistical comparison performed by two-tailed Mann–Whitney *U* test. (*b*) Average protein complex number (protein complex number calculated by using the Benschop et al. [2010] data set) of core and attachment proteins in yeast. The statistical comparison performed by two-tailed Mann–Whitney *U* test.

complex number from 1 to 3, named as Attach-I, and the rest ones as Attach-II (protein complex number ranges from 4 to 22). Comparing the protein evolutionary rates between core and Attach-I proteins, we found marginally significant (Mann–Whitney *U* test, $P = 5.3 \times 10^{-2}$) differences between them (fig. 5). At this point, it should be mentioned that there is only one core protein which has complex number 3, and when this core protein was excluded from the data set and the rests were compared with attachment proteins having complex numbers 1and 2, we found no significant difference in their evolutionary rates (average evolutionary rates of core proteins: 0.0899 (±0.0031) ($N = 347$), attachment proteins: 0.0811 (±0.0057) ($N = 114$); Mann–Whitney *U* test, $P = 8.6 \times 10^{-2}$). However, the evolutionary rate differences were significant between core and the Attach-II proteins, as also between Attach-I and Attach-II proteins (fig. 5). These results indicate that the differences in evolutionary rates between core and attachment proteins disappear when the complex number is factored out.

Protein Complex Number and Expression Level

A large number of studies using organisms ranging from bacteria to mammals (Akashi 2001; Subramanian and

Table 2

Spearman's Correlation Analysis of Protein Complex Number (Taking Core and Attachment Proteins from Gavin et al. [2006]) versus Evolutionary Rates (d*N*/d*S*), Essentiality, Connectivity, Expression Level, Transcription Rate, and Translation Rate

	d*N*/d*S*	Essentiality	Connectivity	Expression Level (Holstege et al. 1998)	Expression Level (Miller et al. 2011)	Transcription Rate (Holstege et al. 1998)	Transcription Rate (Miller et al. 2011)	Translation Rate (Arava et al. 2003)
Protein complex number (Gavin et al. 2006)	$\rho = -0.1572$ $P = 4.2 \times 10^{-5}$ $N = 673$	$\rho = 0.1659$ $P = 5.7 \times 10^{-4}$ $N = 428$	$\rho = -0.0284$ $P = 4.5 \times 10^{-1}$ $N = 696$	$\rho = 0.3987$ $P = 1.0 \times 10^{-6}$ $N = 687$	$\rho = 0.3990$ $P = 1.0 \times 10^{-6}$ $N = 657$	$\rho = 0.4032$ $P = 1.0 \times 10^{-6}$ $N = 649$	$\rho = 0.3977$ $P = 1.0 \times 10^{-6}$ $N = 649$	$\rho = 0.3239$ $P = 1.0 \times 10^{-6}$ $N = 659$

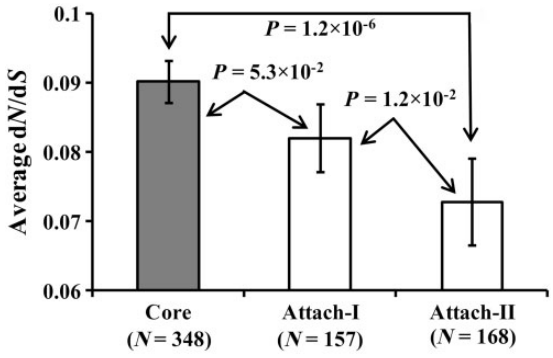

FIG. 5.—Average evolutionary rates (d*N*/d*S*) of Core, Attach-I, and Attach-II proteins in yeast. The pair wise comparison performed by two-tailed Mann–Whitney *U* test.

Kumar 2004; Drummond et al. 2006) demonstrated that gene expression levels are strongly correlated with protein evolutionary rates, and it has been hypothesized that highly expressed genes are under strong selection pressure to reduce mistranslation-induced protein misfolding (Drummond et al. 2005; Drummond and Wilke 2008). In our previous study, we found that protein complex number is one of the important constraints in guiding protein evolutionary rates and also observed that proteins those are associated with a large number of complexes (higher protein complex number) have higher expression levels (Chakraborty et al. 2010). Once we determined the average expression level of core and attachment proteins, we were able to show that attachment proteins have a higher average expression level compared with that of core proteins (fig. 6*a* and *b*), and a strong significant positive correlation exists between protein complex number and expression level (table 2). Moreover, we found that complex number and the expression level independently affect evolutionary rates (multivariate regression analysis: $\beta_{\text{expression level}} = -0.0601$, $P = 8.1 \times 10^{-5}$; $\beta_{\text{complex number}} = -0.1012$, $P = 3.2 \times 10^{-11}$), which is consistent with our previous study (Chakraborty et al. 2010). According to Drummond et al. (2005), the expression level of a protein influences the rates of protein evolution through the translation events. Furthermore, we know that protein expression level can be increased by three possible steps by 1) increasing transcription rate, 2) increasing translational rate, or 3) increasing both transcription and translation rate. Here, we analyzed the transcription rate and translation rate in the whole protein complex data set and observed that the transcription rate and translation rate increases with the increment of protein complex number (table 2). We also found that attachment proteins showed higher values of average transcription and translation rates than core proteins (for Holstege data set: average transcription rate of core proteins: 3.8349 (±0.4394) ($N = 335$), attachment proteins: 20.9643 (±2.0421) ($N = 322$); Mann–Whitney *U* test, $P = 9.6 \times 10^{-16}$; for Miller data set: average transcription rate of core proteins: 22.1872 (±0.9876) ($N = 329$), attachment proteins: 51.9403 (±3.0335) ($N = 320$); Mann–Whitney *U* test, $P = 6.9 \times 10^{-16}$; and for Arava data set: average translation rate of core proteins: 0.1893 (±0.0296) ($N = 331$), attachment proteins: 0.9446 (±0.1325) ($N = 328$); Mann–Whitney *U* test, $P = 5.8 \times 10^{-9}$). Therefore, it is clear that proteins with higher complex number increase their transcription and translation rates, and this increased transcription and translation rates enhance the overall protein expression levels. Finally, the elevated expression levels decrease the evolutionary rates (Spearman's $\rho_{\text{evolutionary rate vs. expression level\_Holstage}} = -0.4894$, $P = 1.0 \times 10^{-6}$; Spearman's $\rho_{\text{evolutionary rate vs. expression level\_Miller}} = -0.4857$, $P = 1.0 \times 10^{-6}$) due to avoidance of mistranslational-induced protein misfolding (Drummond et al. 2005, 2006; Drummond and Wilke 2008). Thus, when we controlled for expression levels, we did not obtain any significant difference in the rates of protein evolution between core and attachment proteins (fig. 7).

Discussion

It has been reported that complex-forming proteins evolve slower than the noncomplex forming proteins (Teichmann 2002), and we have also observed the same in yeast (Chakraborty et al. 2010). Moreover, within the complex-forming proteins, the subunits those are associated with a large number of protein complexes evolve slower than the proteins associated with a lower number of protein complexes

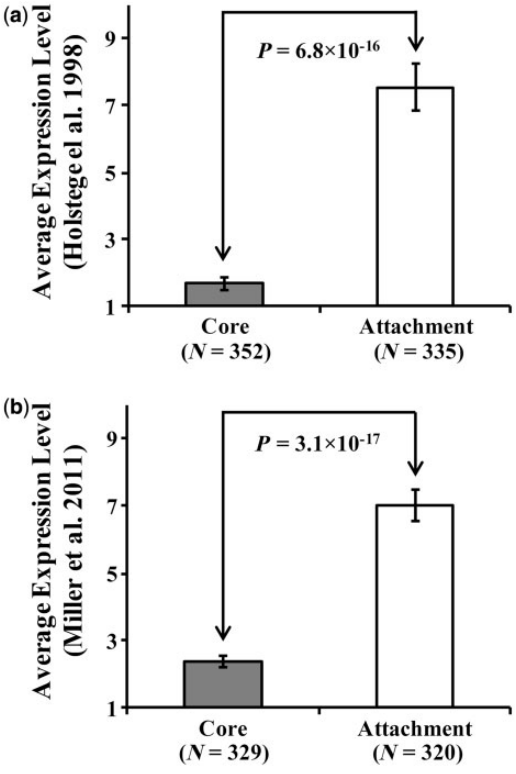

FIG. 6.—(*a*) Average expression level (using the Holstege et al. [1998] data set) of core and attachment proteins in yeast. The statistical comparison performed by two-tailed Mann–Whitney *U* test. (*b*) Average expression level (using the Miller et al. [2011] data set) of core and attachment proteins in yeast. The statistical comparison performed by two-tailed Mann–Whitney *U* test.

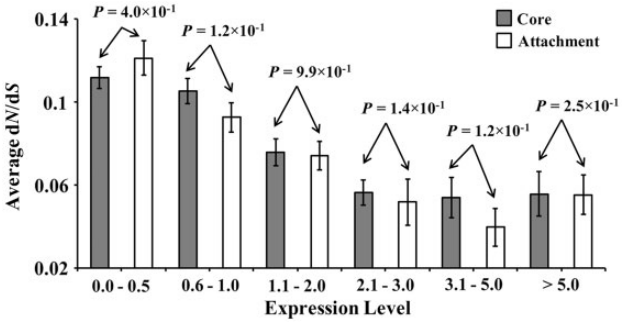

FIG. 7.—Average evolutionary rates (d*N*/d*S*) of core and attachment proteins in different bins. The statistical comparison performed by two-tailed Mann–Whitney *U* test.

(Chakraborty et al. 2010; Das et al. 2013). The protein complex number showed a strong positive correlation with gene expression levels and negative correlation with evolutionary rates, suggesting that protein complex number is a significant factor modulating protein evolutionary rates. In this work, we utilized the yeast protein complex data set (Gavin et al. 2006) for the classification pertaining to core and attachment proteins and analyzed their evolutionary rates to address the influence of gene dispensability, protein multifunctionality, protein connectivity, protein complex number, and gene expression level on protein evolutionary rates.

Analysis of complex forming versus noncomplex proteins demonstrates that proteins in complexes have higher essentiality than the noncomplex proteins (average essentiality of complex proteins: 0.1665 (±0.0045) ($N = 867$), noncomplex proteins: 0.1002 (±0.0014) ($N = 3{,}348$); Mann–Whitney *U* test, $P = 1.1 \times 10^{-65}$). Within the complex-forming proteins, core proteins were reported to be crucial for the function of the protein complex assembly (Dezso et al. 2003; Gavin et al. 2006), indicating that core proteins contribute more toward fitness of an organism. Moreover, it has been reported that essential genes are associated with multifunctional features of genes (Liao et al. 2006). Therefore, we measured multifunctionality of core and attachment proteins and found that core proteins are associated with a higher number of biological processes compared with attachment proteins. However, in our data set, we did not observe any significant difference in the essentiality between core and attachment proteins. These results indicate that the essentiality of a given gene or protein does not depend only on its multifunctionality, perhaps there are some other parameters which influence the gene/protein essentiality. In this study, we observed that attachments proteins are present in more complex assemblies than core proteins, which may increase their essentiality since they modify or enhance the activity of a large number of complex assemblies. Therefore, one can reasonably assume that protein essentiality and protein complex number are interrelated to each other. To verify our hypothesis, we determined the correlation between gene essentiality and protein complex number in our data set, and indeed, we found a strong positive association between them (Spearman's $\rho_{\text{protein complex number vs. essentiality}} = 0.2681$, $P = 1.0 \times 10^{-6}$), which contradicts the results of Pache et al. (2009) that the essentiality does not depend on the protein complex number. Thus both multifunctionality and protein complex number influence the gene/protein essentiality. We have obtained a similar proportion of essential proteins in both core and attachment proteins suggesting that core proteins execute higher multifunctionality, whereas attachment proteins have a higher protein complex number. These results are consistent with the recent study on fitness effect and protein complex that core and attachment proteins are equally important for the function of an organism (Pache et al. 2009). Evolutionary rate differences between core and attachment proteins can be interpreted in the framework of neutralist/selectionist controversy. Core proteins evolve faster than attachment proteins, although core proteins are more multifunctional than attachment proteins, and these results are not compatible with the neutral theory of molecular evolution. However, attachment proteins having higher expression levels should have higher impact on essentiality, and thus the slowly evolving nature of

attachment proteins than core proteins can apparently be accommodated in neutral theory. Moreover, we demonstrated that core proteins have higher multifunctionality and also have higher impact on gene/protein essentiality. Indeed, we found no significant difference in gene essentiality between core and attachment proteins. However, we observed significant evolutionary rate differences between core and attachment proteins, which are inconsistent with the neutral theory of molecular evolution. Recently, Razeto-Barry et al. (2011) predicted that in a selective scenario, more multifunctional proteins should evolve faster than the less multifunctional proteins when there is no difference in the size of fitness effects. They also conclude that a higher number of mutations have been fixed in more multifunctional proteins by positive selection. Our results confirm the prediction of Razeto-Barry et al. (2011).

In this study, we also observed that protein complex number has emerged as an important parameter explaining the differences in evolutionary rates between core and attachment proteins. However, attachment proteins are less multifunctional, but they participate in a high number of protein complexes and thus their transcription rates and translation rates increase. The increased transcription rate and translation rate in turn enhances the overall protein expression levels and the expression level imposes a strong selection pressure on attachment proteins. Previously, Drummond et al. (2005) also demonstrated that the majority of the variation in protein evolution is mainly guided by the expression level. This is further confirmed by our study, when we control the expression level or complex number, the difference in evolutionary rates between core and attachment proteins disappears. Hence, from our analysis, we showed that the rates of protein evolution between core and attachment proteins are mainly guided by protein complex number and expression level. To find out the relative influence in evolutionary rates, we performed a multivariate regression analysis and found that both expression level and protein complex number independently control the evolutionary rate but essentiality did not show any significant contribution.

Supplementary Material

Supplementary tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Department of Biotechnology, Government of India and Department of Science and Technology, Government of India (Sanction No. BT/PR692/BID/7/369/2011). The authors are thankful to Professor Giorgio Bernardi, Department of Biology, University Rome 3, Viale Marconi 446, Rome 00146, Italy, and Dr Bratati Kahali, Department of Internal Medicine, Division of Gastroenterology and Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109 for critical reading of the manuscript. They are also thankful to the two anonymous reviewers for their valuable comments to improve the manuscript.

Literature Cited

Akashi H. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev. 11:660–666.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Alvarez-Ponce D, Fares MA. 2012. Evolutionary rate and duplicability in the arabidopsis thaliana protein-protein interaction network. Genome Biol Evol. 4:1263–1274.

Arava Y, et al. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A. 100:3889–3894.

Batada NN, Hurst LD, Tyers M. 2006. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol. 2:748–756.

Batagelj V, Mrvar A. 2004. Pajek—analysis and visualization of large networks. In: Jünger M, Mutzel P, editors. Graph drawing software. Berlin Heidelberg (Germany): Springer. p. 77–103.

Benschop JJ, et al. 2010. A consensus of core protein complex compositions for *Saccharomyces cerevisiae*. Mol Cell. 38:916–928.

Camon E, et al. 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 32:D262–D266.

Chakraborty S, Kahali B, Ghosh TC. 2010. Protein complex forming ability is favored over the features of interacting partners in determining the evolutionary rates of proteins in the yeast protein-protein interaction networks. BMC Syst Biol. 4:155.

Chakraborty S, et al. 2011. Insights into eukaryotic interacting protein evolution. In: Pontarotti P, editor. Evolutionary biology: concepts, biodiversity, macroevolution and genome evolution. Berlin Heidelberg (Germany): Springer. p. 51–70.

Chen W-H, Minguez P, Lercher MJ, Bork P. 2012. OGEE: an online gene essentiality database. Nucleic Acids Res. 40:901–906.

Cherry JM, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 40:D700–D705.

Das J, Chakraborty S, Podder S, Ghosh TC. 2013. Complex-forming proteins escape the robust regulations of miRNA in human. FEBS Lett. 587:2284–2287.

Deane CM, Salwinski L, Xenarios I, Eisenberg D. 2002. Protein interactions—two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics. 1:349–356.

Dezso Z, Oltvai ZN, Barabási AL. 2003. Bioinformatics analysis of experimentally determined protein complexes in the yeast *Saccharomyces cerevisiae*. Genome Res. 13:2450–2454.

Dolan ME, Ni L, Camon E, Blake JA. 2005. A procedure for assessing GO annotation consistency. Bioinformatics 21:I136–I143.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102: 14338–14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327–337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: Clarendon Press.

Flicek P, et al. 2011. Ensembl 2011. Nucleic Acids Res. 39:D800–D806.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296: 750–752.
Fraser HB, Wall DP, Hirsh AE. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol. 3:11.
Gavin AC, et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. Nature 440:631–636.
Guldener U, et al. 2005. CYGD: the comprehensive Yeast Genome Database. Nucleic Acids Research 33:D364–D368.
Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 22:803–806.
Hart GT, Lee I, Marcotte ER. 2007. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics 8:236.
Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411:1046–1049.
Holstege FCP, et al. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–728.
Hurst LD, Smith NGC. 1999. Do essential genes evolve slowly? Curr Biol. 9: 747–750.
Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12:962–968.
Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40:D109–D114.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254.
Kerrien S, et al. 2012. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 40:D841–D846.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Kimura M, Ohta T. 1974. Some principles governing molecular evolution. Proc Natl Acad Sci U S A. 71:2848–2852.
Larkin MA, et al. 2007. Clustal W and clustal X version 2.0. Bioinformatics 23:2947–2948.
Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 22:1345–1354.
Liao B-Y, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol. 23:2072–2080.
Licata L, et al. 2012. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 40:D857–D861.
Makino T, Gojobori T. 2006. The evolutionary rate of a protein is influenced by features of the interacting partners. Mol Biol Evol. 23: 784–789.
Mete M, Tang F, Xu X, Yuruk N. 2008. A structural approach for finding functional modules from large biological networks. BMC Bioinformatics 9:S19.
Mewes HW, et al. 2011. MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res. 39:D220–D224.
Miller C, et al. 2011. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol Syst Biol. 7:458.
Naumov G, Naumova ES, Lantto RA, Louis EJ, Korhola M. 1992. Genetic homology between *Saccharomyces cerevisiae* and its sibling species *Saccharomyces paradoxus* and *Saccharomyces bayanus*—electrophoretic karyotypes. Yeast 8:599–612.
Nie NH, Bent DH, Hull CH. 1970. SPSS: statistical package for the social sciences. NewYork: McGraw-Hill.
Pache RA, Babu MM, Aloy P. 2009. Exploiting gene deletion fitness effects in yeast to understand the modular architecture of protein complexes under different growth conditions. BMC Syst Biol. 3:74.
Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.
Pang CNI, Krycer JR, Lek A, Wilkins MR. 2008. Are protein complexes made of cores, modules and attachments? Proteomics 8: 425–434.
Park YR, Kim J, Lee HW, Yoon YJ, Kim JH. 2011. GOChase-II: correcting semantic inconsistencies from gene ontology-based annotations for gene products. BMC Bioinformatics 12:S40.
Pereira-Leal JB, Levy ED, Teichmann SA. 2006. The origins and evolution of functional modules: lessons from protein complexes. Philos Trans R Soc Lond B Biol Sci. 361:507–517.
Podder S, Mukhopadhyay P, Ghosh TC. 2009. Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. Gene 439:11–24.
Popescu CE, Borza T, Bielawski JP, Lee RW. 2006. Evolutionary rates and expression level in *Chlamydomonas*. Genetics 172:1567–1576.
Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ. 2007. Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*. Proteomics 7:944–960.
Qiu J, Noble WS. 2008. Predicting co-complexed protein pairs from heterogeneous data. PLoS Comput Biol. 4:e1000054.
Razeto-Barry P, Diaz J, Cotoras D, Vasquez RA. 2011. Molecular evolution, mutation size and gene pleiotropy: a geometric reexamination. Genetics 187:877–885.
Rokotomalala R. 2005. TANAGRA: a free software for research and academic purposes. Advances in grid computing. EGC 2005. In: Proceedings of European Grid Conference; 2005 Feb 14–16; Amsterdam, The Netherlands. Berlin (Germany): Springer. Vol. 2, p. 697–702.
Salathé M, Ackermann M, Bonhoeffer S. 2006. The effect of multifunctionality on the rate of evolution in yeast. Mol Biol Evol. 23:721–722.
Salwinski L, et al. 2004. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 32:D449–D451.
Semple JI, Vavouri T, Lehner B. 2008. A simple principle concerning the robustness of protein complex activity to changes in gene expression. BMC Syst Biol. 2:1.
Steinmetz LM, et al. 2002. Systematic screen for human disease genes in yeast. Nat Genet. 31:400–404.
Su Z, Zeng Y, Gu X. 2010. A preliminary analysis of gene pleiotropy estimated from protein sequences. J Exp Zool B Mol Dev Evol. 314: 115–122.
Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373–381.
Teichmann SA. 2002. The constraints protein-protein interactions place on sequence divergence. J Mol Biol. 324:399–407.
Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. Annu Rev Biochem. 46:573–639.
Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol Biol Evol. 28:2359–2369.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang ZH, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.

Associate editor: Takashi Gojobori