nature genetics

Evolutionary conservation of motif constituents in the yeast protein interaction network

S Wuchty[1], Z N Oltvai[2] & A-L Barabási[1]

Understanding why some cellular components are conserved across species but others evolve rapidly is a key question of modern biology[1–3]. Here we show that in *Saccharomyces cerevisiae,* proteins organized in cohesive patterns of interactions are conserved to a substantially higher degree than those that do not participate in such motifs. We find that the conservation of proteins in distinct topological motifs correlates with the interconnectedness and function of that motif and also depends on the structure of the overall interactome topology. These findings indicate that motifs may represent evolutionary conserved topological units of cellular networks molded in accordance with the specific biological function in which they participate.

Many biological functions are carried out by the integrated activity of highly interacting cellular components, referred to as functional modules[4,5]. Motifs, considered to be topologically distinct interaction patterns within complex networks, may represent the simplest building blocks of such modules[6,7]. Owing to their small size, motifs can be explicitly identified and enumerated in various cellular networks[6–8], but their biological importance, if any, has not yet been determined. A well known signature of the conservation of specific cellular functions is the evolutionary retention of orthologous proteins that are responsible for selected functions. Therefore, the tendency to conserve evolutionarily the protein components of topologically distinct motifs could be indicative of their importance and involvement in specific biological functions.

To test the correlation between a protein's evolutionary rate and the structure of the motif it is embedded in, we first identified all two-, three- and four-node motifs and some five-node motifs in the protein interaction network of *S. cerevisiae* using the DIP protein interaction database[9]. Although the quality of results from two-hybrid studies, which supply the core of the data, is debated[10], the manually curated DIP database represents our current best approximation for yeast protein interactions and provides sufficient data for their unambiguous statistical analyses (see **Supplementary Note** online). The network of 3,183 interacting yeast proteins encodes 10^3–10^6 copies of the specific motif types (**Table 1**).

If there is evolutionary pressure to maintain specific motifs, their components should be evolutionarily conserved and have identifiable orthologs in other organisms. To test this hypothesis, we studied the conservation of 678 *S. cerevisiae* proteins with an ortholog in each of five higher eukaryotes (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*) deposited in the InParanoid database[11]. We found substantially different conservation rates for proteins in the different motifs: less than 5% of the

Table 1 Evolutionary conservation of motif constituents

#	Motifs	Number of yeast motifs	Natural conservation rate	Random conservation rate	Conservation ratio
1		9,266	13.67%	4.63%	2.94
2		167,304	4.99%	0.81%	6.15
3		3,846	20.51%	1.01%	20.28
4		3,649,591	0.73%	0.12%	5.87
5		1,763,891	2.64%	0.18%	14.67
6		9,646	6.71%	0.17%	40.44
7		164,075	7.67%	0.17%	45.56
8		12,423	18.68%	0.12%	157.89
9		2,339	32.53%	0.08%	422.78
10		25,749	14.77%	0.05%	279.71
11		1,433	47.24%	0.02%	2,256.67

The third column gives the number of motifs of a given kind found in the yeast protein interaction network of 3,183 proteins, which we obtained by counting all subgraphs of two-node to five-node motifs (from the set of 28 five-node motifs, we show only two, #10 and #11). We identified 678 proteins that have an ortholog in each of the five higher eukaryotes that we studied and identified all motifs for which each component belongs to this evolutionary conserved protein subset. The natural conservation rate indicates the fraction of the original yeast motifs that is evolutionarily fully conserved, meaning that each of their protein components belongs to the 678 orthologs of the list. For example, we find that 47% of the 1,433 fully connected pentagons (#11) found in yeast have each of their five proteins conserved in each of the five higher eukaryotes. If the topology of motifs does not interfere with the conservation rate of its constituting proteins, a random ortholog distribution should give the same conservation rate for specific motifs as seen in the natural sample. The random conservation rate therefore represents the fraction of motifs that is fully conserved for the random ortholog distribution. The last column gives the ratio between the natural and the random conservation ratios, indicating that all motifs are highly conserved, some (for example, #11) having a natural conservation rate 2,256 times higher than expected in the absence of correlations between protein conservation rate and the topology of a given motif.

[1]Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA. [2]Department of Pathology, Northwestern University, Chicago, Illinois 60611, USA. Correspondence should be addressed to A.-L.B. (alb@nd.edu) or Z.N.O. (zno008@northwestern.edu).

Published online 14 September 2003; doi:10.1038/ng1242

linear three-node motifs (#2) were completely maintained (meaning that all three component proteins had an ortholog), whereas 47% of the fully connected pentagons (#11) were completely conserved across each of the other five eukaryotes (**Table 1**).

These results indicate that the orthologs are not randomly distributed in the yeast protein interaction network but are the building blocks of cohesive motifs, which tend to be evolutionary conserved. We need, however, to test the validity of this finding against a random set of orthologs. If the same number of orthologs were randomly placed on the yeast protein interaction network, with no correlation between the network topology and the ortholog position, the motif conservation described above should disappear. In fact, under such random ortholog distribution, the conservation of motifs showed a trend opposite to that observed for the original system: the larger the motif, the smaller was the likelihood that each of its components was conserved (**Table 1**). For example, 4.6% of the randomized two-node motifs (#1) were retained with randomized orthologs, but only 1.01% of the triangle motifs (#3), 0.08% of the fully connected square motifs (#9) and 0.02% of the fully connected pentagon motifs (#11) were retained with randomized orthologs.

The influence of the global network topology on the retention rate of specific local motifs is best quantified by calculating the ratio between the real and the random conservation rates. We found that this conservation ratio for each motif was greater than one and increased considerably for larger motifs. For the two-node motifs (#2), the conservation ratio was 2.94, whereas for the larger fully connected motifs, such as the triangle (#3) and square motifs (#9), it was 20.28 and 422.78, respectively. Moreover, the conservation rate of proteins participating in fully connected pentagon motifs (#11) was 2,256 times higher than would be expected if the network topology did not influence the natural placement of orthologs (**Table 1**).

We also observed that larger motifs tended to be conserved as a whole, each of their components having an ortholog. For example, less than 1% of the fully connected pentagon motifs disappeared completely, so that none of their protein components were conserved in other eukaryotes, and less than 2% of such pentagons had only one conserved protein (**Fig. 1b**). In contrast, for 69% of the fully connected pentagons, each of the subunits had an ortholog in humans. We observed a similar trend toward complete conservation of larger motifs for each of the five higher eukaryotes (**Fig. 1a**,**b**). In general, as the number of nodes in a motif and number of links among its constituents increased, the evolutionary retention of the constituent proteins was more complete. In particular, we observed a clear correlation between the conservation rate and the degree of saturation of the motif. Of the four-node motifs, the more intraconnected ones (#8 and #9; **Table 1**) had a much higher conservation rate than their less intraconnected counterparts (#4, #5, #6 and #7; **Table 1**). Overall, these exceptionally high conservation rates strongly suggest that participation in motifs substantially influences the evolutionary conservation of the specific components.

To examine the relationship between the local interconnectedness of the network and the retention rates of the protein components, we also measured the correlation between the clustering coefficient and the conservation rates of the interacting proteins (**Fig. 1c**). The clustering coefficient is high ($C_i = 1$) in a highly cohesive region of the network if all neighbors of a protein *i* have links to each other and is small ($C_i = 0$) if the network is locally sparse[12,13]. We found that from 65% ($C = 0$) to 84% ($C = 1$) of neighbors of a human ortholog were also human orthologs (**Fig. 1c**) and that the conservation rate increased with the neighborhood's cohesiveness. In contrast, the conserved fraction of the nonorthologous protein's neighborhood was markedly smaller, from 40% ($C = 0$) to 20% ($C = 1$; **Fig. 1c**). Therefore, groups of proteins forming a highly interlinked cluster tend to be conserved (or nonconserved) in a cohesive group if they represent an evolutionary conserved (or nonconserved) functional module.

Motifs and the clustering coefficient probe the network's small-scale properties, addressing the influence of a protein's immediate neighbors on its conservation rate. But the proposed hierarchical modularity of metabolic[13] and protein interaction networks[14] suggests that highly interconnected motifs may combine into larger, less cohesive modules. To examine if the observed correlations between the conservation rate and the network topology are relevant beyond the protein's immediate vicinity, we identified all proteins, starting from ortholog *i*,

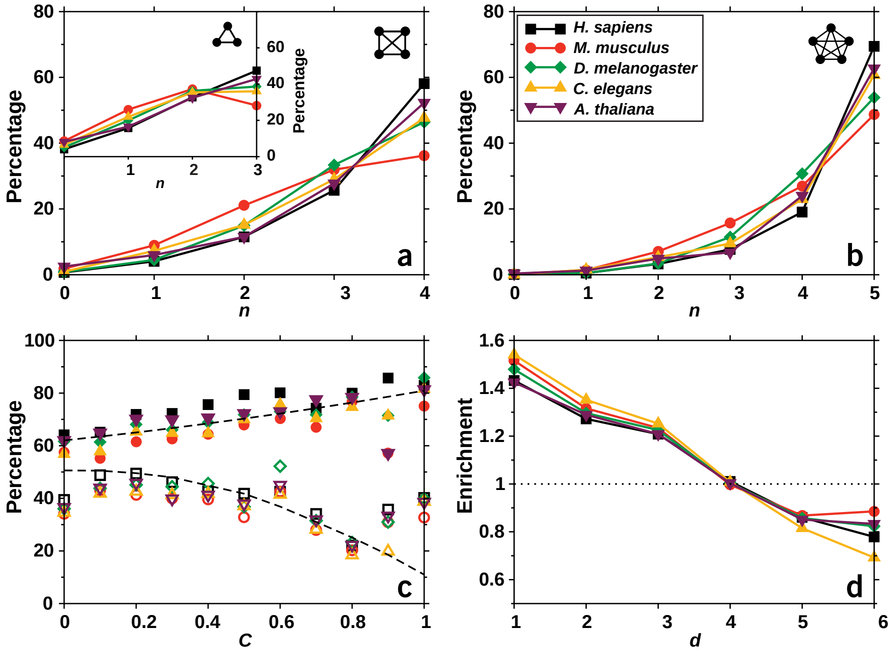
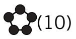
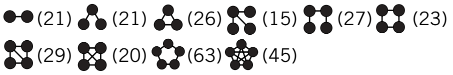
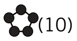
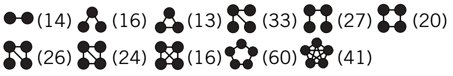
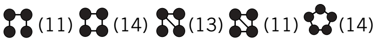
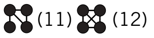
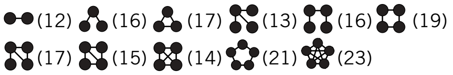
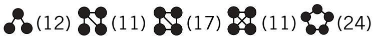


Figure 1 Relationship between the topology of a protein interaction network and the evolutionary conservation of individual proteins. (**a**,**b**) Detailed conservation rates of fully connected three-node (inset in **a**), four-node (**a**) and five-node (**b**) motifs. (**b**) In humans, less than 1% of the 1,433 pentagon motifs found originally in yeast have fully disappeared (meaning that none of their components have an ortholog), and only 1.5% of motifs have a single ortholog component ($n = 1$), whereas for more than 69% of the motifs, each of the five proteins have been conserved ($n = 5$). The five curves correspond to the five studied eukaryotes, and the key in **b** identifies the corresponding symbols and colors used throughout. (**c**) The conserved fraction of the immediate neighbors of an orthologous protein *i* (filled symbols) correlates positively with the node's clustering coefficient *C*. Open symbols show the fraction of orthologs in the vicinity of a nonorthologous protein, which correlates negatively with *C*. (**d**) The enrichment, defined as the ratio between the percentages of orthologous proteins at distance *d* from an ortholog in the natural and the random orthologous sets, indicates decreasing overrepresentation of orthologs with increasing distance.

that are $d = 1, 2, 3...$ links from i, where d represents the shortest distance between i and a target protein measured along the network links. We separately determined the fraction of orthologous proteins at distance d for both the natural and the random ortholog distributions. The ratios in the natural and random fractions of orthologous proteins (**Fig. 1d**) indicated a considerable enrichment for orthologs at distances $d = 1$, 2 or 3, which disappears for $d > 3$. Proteins that interact directly with an ortholog at $d = 1$ had a 50% higher (or more) chance of conservation than would be expected for a random ortholog distribution, and those at $d = 2$ had a 25–35% higher rate of conservation. We also observed enrichment (20–25%) for those at $d = 3$, indicating that the extended vicinity of an orthologous protein is enriched with orthologs, thus supporting the extension of conservation to larger modules as well.

To examine if the specific function of the yeast proteins within motifs affects their rate of evolutionary conservation, we assigned each motif to the functional class to which its protein components belong, using the classification of the MIPS database[15]. Larger motifs had a notable functional homogeneity. For 95% of those fully connected yeast pentagon motifs (#11) whose proteins had an ortholog in each of the five higher eukaryotes, all components shared at least one common functional class. In contrast, only 10% of the two-node motifs (#1) were functionally homogenous. We identified the type and number of evolutionary fully conserved motifs of each functional class in *S. cerevisiae*, limiting our study to those proteins that had an ortholog in humans. The ratio of the number of motifs identified for the natural and random ortholog distributions indicated substantial functional class–dependent differences in the evolutionary conservation of motifs (**Table 2**). For three functional classes (subcellular localization, protein fate and transcription), each of the 11 studied motifs were considerably overrepresented. In contrast, a few functional classes, such as transport facilitation, regulation and cellular transport, had only one or two characteristic motifs, and others had none. These results indicate that the different functions not only are associated with characteristic topological motifs but also conserve these motifs at different rates during evolution.

Table 2 Overrepresentation of human orthologous motifs in various functional classes of yeast proteins

Functional class	Overrepresented motifs
Transport facilitation	 (10)
Subcellular localization	 (21) (21) (26) (15) (27) (23) (29) (20) (63) (45)
Regulation	 (10)
Protein fate	 (14) (16) (13) (33) (27) (20) (26) (24) (16) (60) (41)
Cell cycle	 (11) (14) (13) (11) (14)
Cellular transport	 (11) (12)
Transcription	 (12) (16) (17) (13) (16) (19) (17) (15) (14) (21) (23)
Protein synthesis	 (12) (11) (17) (11) (24)

We determined the number of motifs for the subnetworks defined by proteins belonging to a specific functional class, as well as the number of these motifs (μ_h) that are fully conserved in humans. Finally, for 100 randomized human orthologous sets we determined the average number of motifs (μ_r) in the random ortholog samples and the standard deviation (σ_r) for each motif. The table lists all motifs that are overrepresented by a factor of at least ten compared with a random configuration ($Z > 10$), with the specific Z values shown next to the motifs. We did not find overrepresented motifs for the classes of transposable elements, energy, cellular fate, cellular communication, cellular rescue, cellular organization, metabolism, protein activity, protein binding and proteins that are not yet classified or that are classified unclearly. If all proteins of a given motif simultaneously belong to more than one functional class, the motif will also appear in multiple functional classes.

Motifs may represent various types of protein interactions, and the fully connected motifs (#9 and #11), as expected, tend to identify protein complexes. Smaller complexes in which each of the proteins interacts with all others should appear as fully connected n-node motifs in the protein interaction network. In larger protein complexes, however, not all proteins have direct interactions with each other, and thus motifs are expected to capture only some local, physically interacting components of the whole complex. For example, proteins found in the fully connected pentagons contained components of known yeast proteasome complexes RPN (rpn1, rpn2, rpn3, rpn4, rpn6, rpn7, rpn9, rpnA and rpnC), PSA (psa1, psa2, psa3, psa4, psa6 and psa7), PSB (psb2, psb3, psb4, psb5 and psb6) and PRS (psa4, psa6, psa7, psa8 and psaA). These complexes interact extensively with each other as well as with seven other proteins (sug2, mpr1, ra23, ubp6, pyrg, p2a2 and psda) that are not known to be part of the specific complexes. A separate cluster of proteins, in contrast, did not represent a protein complex but consisted of an interlinked collection of nucleolar (nop2, nop4 and nog1), kinase (kc21) and RNA helicase (mak5 and has1) proteins and four proteins with unknown function (ymt9, ytm1, yo26 and yev6). The large number of interactions with these uncharacterized proteins may indicate functional relatedness, suggesting that a combination of evolutionary retention and dense interactions, as selected by the specific motifs, could be used to predict *in silico* the functional role of the unknown protein components. But the mere existence of protein complexes cannot explain the observed trends towards higher conservation rates of the highly connected motifs. In fact, the basic conservation trends were not altered after we removed proteins that are part of known complexes, although the actual conservation ratios did change. Similarly, although the protein interaction and ortholog databases are incomplete and contain numerous false positives, an error analysis confirmed that our main findings and conclusions are not affected by such data inconsistencies, indicating the robustness of the observed evolutionary trends (see **Supplementary Note** online).

Further studies on the evolutionary conservation of topological modules and motifs would benefit from the simultaneous study of the retention rate of both nodes (proteins) and the links (interactions) among them. Because protein interactions are available systematically for only *S. cerevisiae* among all eukaryotes, our study is limited to the orthologous retention of the protein components of selected motifs. The high retention rates of many of the constituents of highly connected motifs (**Table 1**) strongly suggest that interactions between the proteins of these motifs may be preserved in other organisms, a hypothesis that could be confirmed once protein interaction databases are established for other eukaryotic species.

Previous results suggest that the evolutionary rate of a protein correlates with the protein's essentiality and individual fitness[16–18] and its level of interactions with other proteins[19], but the quantitative correlations supporting some of these hypotheses have occasionally been questioned[16,20,21]. As these hypotheses aim to relate the properties of cellular components to their evolutionary rate, the contradictory nature of some of these conclusions might have biological origins. Natural selection is expected to preserve components only to the degree that they contribute to conserved cellular functions. A given biological function can rarely be assigned to a single protein, gene or metabolite, however, but rather emerges from the interaction of many separate components forming distinct functional modules[4,5,22]. Thus, the identified motif conservation may represent the network equivalent of domain and residue conservation in protein sequences. Our results indicate that understanding the evolutionary rate of single proteins must address the need to preserve evolutionarily the specific functional modules and the topologic features of the network in which their respective proteins are embedded. In agreement with this hypothesis, we found that the conservation rate of motif constituents was tens to thousands of times higher, an enhancement that is clearly unparalleled in measurements focusing on the evolutionary rate of single components.

METHODS

Databases. For a list of experimentally detected protein-protein interactions in *S. cerevisiae*, we used the manually curated DIP database[9] (as of March 2003), which contains 3,183 proteins with 9,463 interactions. We assigned to each protein its known functional classification according to the MIPS database[15], which compiles genetic, biochemical and cell biological knowledge of yeast genes and proteins extracted from the literature. If a protein belonged to more than one functional class, its corresponding motif was assigned to both groups.

Motif identification. Similar to the method of Milo *et al.*[7] for detecting all *n*-node subgraphs, our algorithm scans all rows of the adjacency matrix *M*. For each non-zero element (i,j) representing a link, it scans through all neighbors of (i,j), M_{ik}, M_{ki}, M_{jk} and $M_{kj} = 1$. This is done recursively for all other elements (i,k),(k,i),(k,j) and (j,k) until a specific *n*-node subgraph is detected. The detected subgraphs are then compared to the subgraphs found in previous steps and eliminated if they are already in the database. In contrast to ref. 7, where motifs were defined as overrepresented subgraphs, here we used the terms motifs and subgraphs interchangeably.

Assigning orthologs. The InParanoid database[11] provides orthologous sequence cluster information between organism pairs of *S. cerevisiae* and *H. sapiens*, *D. melanogaster*, *M. musculus*, *C. elegans* and *A. thaliana*. For our study, we chose only the core orthologous sequence pair of each cluster, providing a bootstrap value of 100%. Each yeast protein that is engaged in orthologous core pairs in a specific eukaryote was labeled accordingly. Therefore, 2,174 proteins were labeled to have orthologs in *H. sapiens*, 2,093 in *A. thaliana*, 1,696 in *C. elegans*, 1,674 in *M. musculus* and 1,958 in *D. melanogaster*. We used this detailed ortholog information to calculate the results depicted in **Figure 1**. For the data presented in **Tables 1** and **2**, we identified 678 yeast proteins with an ortholog in each of the five higher organisms, representing the cross-section of the orthologous sets derived for the five organisms.

Random ortholog distribution. As a negative control set, we selected 678 proteins randomly on the yeast protein interaction network, assigned them as random orthologs and determined again the number of specific yeast motifs that were fully conserved (meaning each of their components belonged to the random ortholog set). The random conservation rate of a motif with *n* proteins is well approximated by p^n, where *p* is the probability that a protein has an ortholog across all five higher eukaryotes, given by $P = 678/3{,}128 = 0.216$. Indeed, p^n gives 4.6%, 1.01%, 0.22% and 0.047% for the two-, three-, four- and five-node motifs, respectively, in agreement with the numbers shown in **Table 1** for the random conservation rate.

Enrichment of orthologous proteins. We identified all proteins at distance *d* from an orthologous protein *i* and denoted their number as $N(d)$. For example, $N(1)$ is the number of proteins directly interacting with protein *i*. Of the $N(d)$ proteins, we also identified the number $n(d)$ that had an ortholog in a reference eukaryote. The ratio $r(d) = n(d)/N(d)$ represents the fraction of orthologs at distance *d* from protein *i*. If the orthologs are randomly placed on the network, this ratio should be independent from *d* and have the value $r = n/N$, where *n* is the total number of yeast orthologs in the reference organism and *N* is the total number of proteins in the network. The ratio $E(d) = r(d)/r$ gives the orthologous enrichment, which is equal to 1 for any *d* if there is no clustering of orthologs in the network. $r(d) >> 1$ implies that orthologs are overrepresented among proteins at distance *d* from *i*, which is a signature of clustering. To decrease the noise level in **Figure 1d**, we averaged $r(d)$ values over all yeast orthologs chosen as *i*.

Functional classes. We determined the number of motifs (μ_{h}) for the subnetworks defined by yeast proteins as belonging to a specific functional class and found to be fully conserved in humans. To identify overrepresented motifs in each functional class, we determined the average number of each motif (μ_{r}) and the respective standard deviation (σ_{r}) using 100 random human ortholog sets. The parameter $Z = (\mu_{\mathrm{h}} - \mu_{\mathrm{r}})/\sigma_{\mathrm{r}}$ offers a quantitative measure of the degree to which a motif is overrepresented in a specific functional class: $Z >> 1$ implies that there were substantially more motifs in that class than a random distribution of ortholog placement could support. For functional classification we used the MIPS database[15], which classifies yeast proteins in 17 distinct functional classes. This coarser classification offers better statistics for most classes.

Clustering. To characterize the degree of clustering in the network (**Fig. 1c**), we used the clustering coefficient, defined as $C_i = 2n_i/k_i(k_i - 1)$, where n_i is the number of direct links between the k_i neighbors of protein *i* (ref. 12). The clustering coefficient is 1 if all neighbors of node *i* are connected to each other and 0 if none of the neighbors are linked to each other.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

Research at the University of Notre Dame and Northwestern University was supported by grants from the US National Institutes of Health (National Institute of General Medical Sciences) and the Department of Energy Genomes to Life Program. Research at the University of Notre Dame was also supported by the National Science Foundation.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 5 June; accepted 26 August 2003
Published online at http://www.nature.com/naturegenetics/

1. Hasty, J., McMillen, D. & Collins, J.J. Engineered gene circuits. *Nature* **420**, 224–230 (2002).
2. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
3. Rao, C.V., Wolf, D.M. & Arkin, A.P. Control, exploitation and tolerance of intracellular noise. *Nature* **420**, 231–237 (2002).
4. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
5. Oltvai, Z.N. & Barabási, A.-L. Life's complexity pyramid. *Science* **298**, 763–764 (2002).
6. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68 (2002).
7. Milo, R., Shen-Orr, S.S., Itzkovitz, S., Kashtan, N. & Alon, U. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
8. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
9. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
10. Von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein protein interactions. *Nature* **417**, 399–403 (2002).
11. Remm, M., Storm, C.E.V. & Sonnhammer, E.L. Automatic clustering of orthologs and inparalogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
12. Watts, D.J. & Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
13. Ravasz, E., Somera, A.L., Mongru, D. A., Oltvai, Z.N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
14. Rives, A.W. & Galitski, T. Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA* **100**, 1128–1133 (2003).
15. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).
16. Hurst, L.D. & Smith, N.G. Do essential genes evolve slowly? *Curr. Biol.* **9**, 747–750 (1999).
17. Hirsh, A.E. & Fraser, H.B. Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).
18. Hirsh, A.E. & Fraser, H.B. Genomic function (communication arising): rate of evolution and gene dispensability. *Nature* **421**, 497–498 (2003).
19. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. & Feldman, M.W. Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
20. Pal, C., Papp, B. & Hurst, L.D. Genomic function (communication arising): rate of evolution and gene dispensability. *Nature* **421**, 496–497 (2003).
21. Jordan, I.K., Rogozin, I.B., Wolf, Y.I. & Koonin, E.V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962–968 (2002).
22. Snel, B., Bork, P. & Huynen, M.A. The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci. USA* **99**, 5890–5895 (2002).