Functional Evolution of the Yeast Protein Interaction Network

Victor Kunin,[1] *José B. Pereira-Leal,*[1] *and Christos A. Ouzounis*

Computational Genomics Group, The European Bioinformatics Institute EMBL Cambridge Outstation, Cambridge, UK

Protein interactions are central to most biological processes. We investigated the dynamics of emergence of the protein interaction network of *Saccharomyces cerevisiae* by mapping origins of proteins on an evolutionary tree. We demonstrate that evolutionary periods are characterized by distinct connectivity levels of the emerging proteins. We found that the most-connected group of proteins dates to the eukaryotic radiation, and the more ancient group of pre-eukaryotic proteins is less connected. We show that functional classes have different average connectivity levels and that the time of emergence of these functional classes parallels the observed connectivity variation in evolution. We take these findings as evidence that the evolution of function might be the reason for the differences in connectivity throughout evolutionary time. We propose that the understanding of the mechanisms that generate the scale-free protein interaction network, and possibly other biological networks, requires consideration of protein function.

Introduction

Protein-protein interactions are intrinsic to the vast majority of cellular processes. They form complex networks where individual molecular components act concertedly to perform all the multitude of cellular functions. The functional properties of the network transcend the individual properties of each of its molecular components. A wealth of protein interaction data is becoming available from a variety of sources. These sources include compilations of previously identified groups of interacting molecules from the biomedical literature (Xenarios et al. 2002; Bader, Betel, and Hogue 2003), high-throughput methods such as the yeast two-hybrid system (Uetz et al. 2000; Ito et al. 2001), and the computational prediction of protein interactions via genome context (Enright et al. 1999; Marcotte et al. 1999). These data are particularly rich for the model organism *Saccharomyces cerevisiae* (budding yeast) (von Mering et al. 2002).

Analysis of this information has revealed that the protein interaction network of the budding yeast is a small-world network (Jeong et al. 2001; Maslov and Sneppen 2002), which is characterized by small average path length between nodes. It also follows a power-law distribution of connectivity, indicating a scale-free topology (Jeong et al. 2001; Wagner 2001). These defining properties are observed in some, but not necessarily all, organized networks such as the metabolic network (Jeong et al. 2000), the protein similarity network (Harrison and Gerstein 2002; Enright, Kunin, and Ouzounis 2003), ecological networks (Dunne, Williams, and Martinez 2002), and the Worldwide Web (Willinger et al. 2002). The scale-free topology suggests an explanation for the robustness of the network, as removal of most nodes has little or no detectable effect, and only the removal of the most central (connected) nodes will cause the network to collapse (Albert, Jeong, and Barabasi 2000). This was elegantly demonstrated for the yeast-protein interaction network by Jeong et al. (2001), who observed that centrality in the network correlates positively with the probability of lethality of a node. However, a protein may be vital without being highly connected in the protein-protein interaction network. It may in stead participate in other types of interactions that are not captured in these data sets; for example, protein-DNA interactions in the transcriptional network or protein-metabolite interactions in the metabolic network.

A fundamental problem in the study of biological networks is the understanding of how the scale-free topology emerges. Barabasi and Albert (1999) proposed that a simple preferential attachment rule is sufficient for the emergence of this topology. The principle is that new nodes will preferentially bind the most-connected existing nodes—also described as "the richer gets richer" principle (Barabasi and Albert 1999). A mathematical model of the growth of networks based on this principle produces scale-free topologies with topological parameters comparable to those of real-world networks (Barabasi and Albert 1999). This model predicts that the degree of all nodes evolves the same way, according to the equation $k_i(t) = m(t/t_i)^{\beta}$ with $\beta = ½$ where k_i is the connectivity degree of node i, and m, the initial connectivity of all nodes, are assumed to be identical, and t represents time. All nodes increase their connectivity with time, following a power-law dependence with the same dynamic exponent β. This preferential attachment model then predicts that older nodes should display higher connectivity values, whereas more recent nodes should be the least connected (Albert and Barabasi 2002).

In contrast with this prediction are anecdotal examples of proteins considered to be of very old origin that display very low connectivity levels. For example, triose-phosphate isomerase from *S. cerevisiae* (YDR050C) has only one interaction documented with a hypothetical protein (YNLI27W) (Xenarios et al. 2002; Gavin et al. 2002). It is an essential enzyme in glycolysis, with wide phylogenetic distribution that suggests its ancient origin. A contrasting example is the *Saccharomyces*-specific, hence, of presumably recent origin, regulatory protein MSN3 (YOR047C). This protein is reported to be involved in transcriptional regulation and has 32 reported interacting partners (Xenarios et al. 2002). Its interaction partners include transcription factors such as YGL237C, YMR236W, YMR280C, YNL314W, YNRO52C, and

Key words: network, protein interaction, functional evolution, *Saccharomyces cerevisiae*.

E-mail: kunin@ebi.ac.uk.

[1] These authors contributed equally to this work.

Mol. Biol. Evol. 21(7):1171–1176. 2004
doi:10.1093/molbev/msh085
Advance Access publication April 7, 2004

YKLO38W (Xenarios et al. 2002). This led us to question whether these are isolated cases or whether the general preferential attachment model fails to capture the complexity of the protein interaction network.

In this study, we trace the origin of each protein interaction network. We use this information to test the predictions of the preferential attachment model and fail to observe the expected correlation between the number of interacting partners and the age of the protein. We propose that this is because of the functional heterogeneity of the protein interaction network as protein function determines types of binding partners, the degree of connectivity, and the time of emergence in the network.

Methods

We derived a protein interaction network *Saccharomyces cerevisiae* from the Database of Interacting Proteins (DIP), release January 2003 (Xenarios et al. 2002). This network consists of 4,715 proteins and 15,114 associations and includes interactions obtained from small-scale and large-scale studies. These data were used to calculate the connectivity (k) of each node.

To find the age of the proteins in the data set, we used the GeneTrace algorithm (Kunin and Ouzounis 2003*a*) with default parameters (Kunin and Ouzounis 2003*b*). This algorithm deduces the most likely history of a protein family, including the timing of the origin of the family, given a phylogenetic profile of protein family and species tree. The protein families and their phylogenetic profiles were derived from the TRIBES database (Enright, Kunin, and Ouzounis 2003) comprising protein families from 83 complete genomes. These families were also used to construct a gene content–based phylogenetic tree (Snel, Bork, and Huynen 1999). The tree was subsequently manually rooted on the node joining the tree domains of life (Eukaryota, Bacteria, and Archaea), and resolution of the eukaryotic part of the tree was edited to capture major evolutionary events (fig. 1).

The use of protein families as opposed to groups of orthologous proteins has the advantage of the large coverage of the TRIBES database. Also, defining orthologous groups genes across vast phylogenetic distances, such as across domains of life, is a difficult problem because orthology is defined not only in terms of sequence similarity but also in terms of evolutionary relationships, which are not always known.

In contrast, the relative lucidity in defining protein families is of a great advantage. However, protein families with large numbers of paralogs may contain members with different functions and degrees of connectivity, masking the genuine signal and producing noise. Nevertheless, as distribution of protein family sizes in yeast follows a power-law (Qian, Luscombe, and Gerstein 2001), a very small number of protein families have a large number of paralogs and, thus, this effect is actually negligible (Enright, Kunin, and Ouzounis 2003).

Functional classifications of proteins were derived from the GeneQuiz automatic protein annotation tool (Andrade et al. 1999). Functional classes describing various metabolic enzymes ("amino acid biosynthesis," "biosynthesis of cofactors, prosthetic groups and carriers," "cell intermediary metabolism," "energy metabolism," "fatty acid and phospholipids metabolism," "purines, pyrimidines, nucleosides, and nucleotides") were unified to a single-class "metabolism."

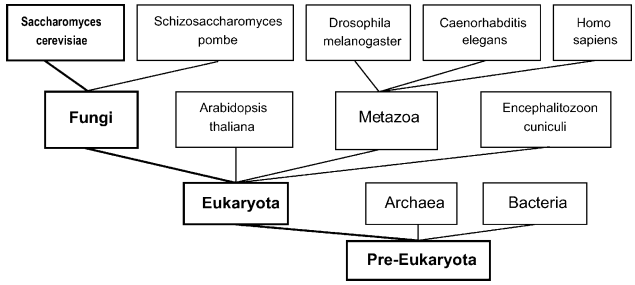

FIG. 1.—Schematic representation of part of the tree used in this study. The path with evolutionary time points leading to *S. cerevisiae* is highlighted in bold. Note that because of the structure of the tree (three domains of life diverging from the last common ancestor) the pre-Eukaryota group does not represent proteins present in the last common ancestor, but rather proteins that evolved before the appearance of Eukaryota.

Results and Discussion

Connectivity Versus Time of Origin

For each protein in the data set we found the timing of the most likely origin of its corresponding family using GeneTrace (Kunin and Ouzounis 2003*a*). Based on the analysis of a phylogenetic profile, this method detects the time of origin of a protein family on a species tree (see *Methods*). In this study, the times of origin represent nodes on the evolutionary tree on the path leading to *Saccharomyces* (fig. 1). The most ancient timepoint groups proteins of pre-eukaryotic origin. These proteins are assumed to belong to families that emerged before the fusion of Bacteria-like and Archaea-like organisms into primordial Eukaryota (Golding and Gupta 1995; Anderson et al. 1998). The second group includes proteins that appeared before radiation of Eukaryota to Viridiplantae, Metazoa, and Protista (fig. 1). The third group represents proteins evolving before the separation of baker's yeast and fission yeast. Finally, the last group contains proteins found solely in *Saccharomyces*.

We used these data to determine how the average protein connectivity correlates with the estimated time of origin (fig. 2*A*). On average, the Eukaryota group contains the most-connected proteins (fig. 2*A*). The more recent Fungi and *Saccharomyces* group, display reduced levels of connectivity. Surprisingly, the proteins of oldest origin do not display the highest connectivity. This in is contrast with the prediction of the preferential attachment model that oldest proteins are expected to display the highest levels of connectivity (Albert and Barabasi 2002).

The average connectivity may be misleading because of the power-law distributions of connectivity observed in each group (data not shown). Thus, we divided the proteins in the yeast-protein interaction network into four

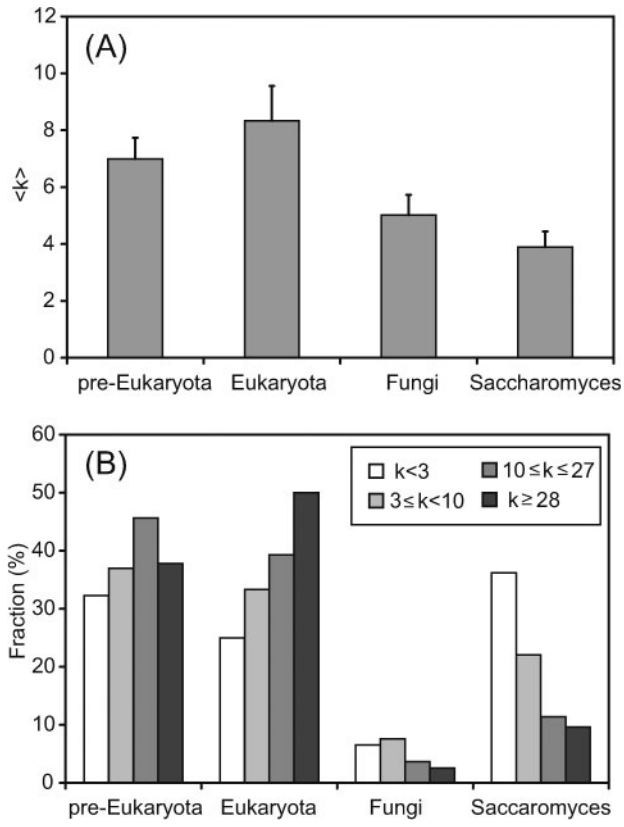

FIG. 2.—(*A*) Average connectivity <k> of proteins at evolutionary timepoints. Error was estimated from 100 random samplings of the data set, comprising 100 proteins per set. (*B*) Evolutionary origins of proteins with various degrees of connectivity (k).

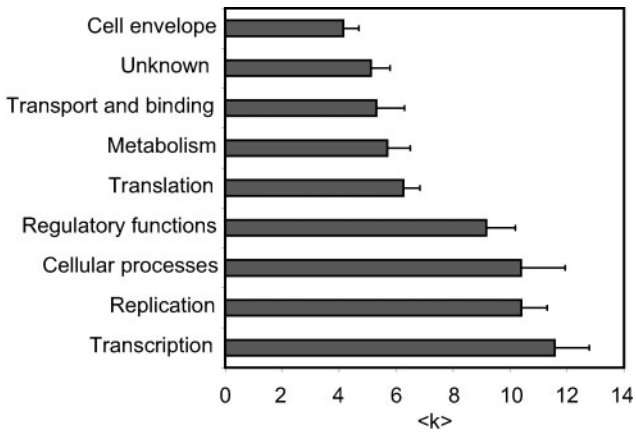

FIG. 3.—Average connectivity levels <k> of functional classes. Error was estimated from 100 random samplings of each functional class with 50 proteins each.

groups according to the number of their interacting partners, in the form of exponentially increasing bins. We then asked, what is the contribution of each of these connectivity groups at each evolutionary timepoint (fig. 2*B*)? Highly connected proteins are the most underrepresented in the *Saccharomyces*-specific data set, consistent with the lowest average connectivity shown in figure 2*A*. Also consistent with the trend observed on average connectivity, the majority of the most highly connected proteins are found to emerge during the eukaryotic radiation. This might reflect the emergence of proteins playing a central role in eukaryotic cell organization. This is illustrated by multiple examples of highly connected proteins central to eukaryotic cell organization, such as cytoskeleton proteins (e.g., actin), the complex multicomponent structure of the transcription apparatus, and highly connected nuclear pore proteins. All these types of proteins are found in most known Eukaryota but not in prokaryotes.

Interestingly, very few proteins show evidence of appearing in the time period between the separation of Eukaryotes and the two yeasts. The proteins evolving at this stage constitute less than 10% of total. However, this does not necessarily imply that few protein families emerged in this period. Tree-based phylogenetic analysis suggests that yeasts might have evolved from multicellular Fungi, acquiring secondary unicellularity (Hedges 2002). Thus, proteins found in multicellular organisms, with functions such as cell-cell interaction, may be lost in yeasts. The current data coming from yeast genomes may thus not provide the full insight into this period.

In summary, although we observed that the most recent proteins tend to be of lower connectivity, we failed to detect the steady increase of connectivity with the protein age.

Connectivity Versus Function

The preferential attachment model assumes that all nodes have the same properties; that is, that all nodes are equal in everything but their connectivity. This way, any node can bind any other node. This assumption limits the scope of this model, as recognized by its authors (Albert and Barabasi 2002). Furthermore, it contrasts with the observation that proteins might preferentially bind within their functional class (von Mering et al. 2002). This led us to hypothesize that the observed distribution of connectivity levels over evolutionary times reflects appearance of functional aspects characterizing each evolutionary stage.

To test this hypothesis, we used GeneQuiz functional classifications (Andrade et al. 1999). GeneQuiz is an automated genome annotation system that performs large-scale functional analysis of protein sequences. Although the use of manual classification systems such as GO (Ashburner et al. 2000) would be more desirable, we find that these functional classes are not evenly distributed and have markedly different sizes in our data set, which complicates the analysis. In contrast, GeneQuiz automatically assigns each protein to a single functional class (e.g., "Transcription" or "Regulatory functions"). The functional classes are distributed relatively evenly between *S. cerevisiae* proteins and are more amenable to this analysis.

We observe that functional classes display different average connectivity levels (fig. 3). Proteins involved in the yeast cell envelope appear as the least connected,

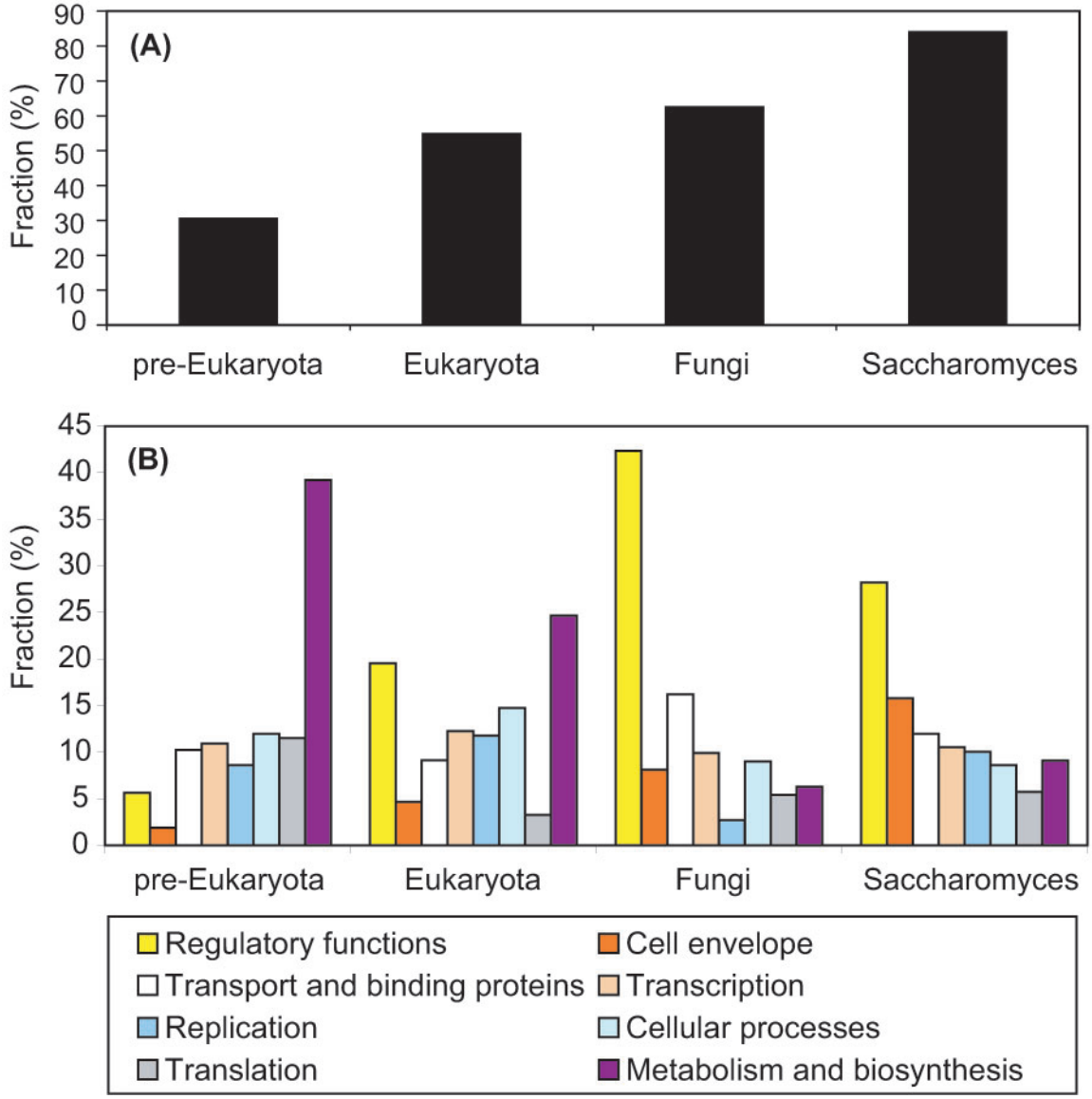

FIG. 4.—(*A*) Proportion of proteins of "Unknown" function arising at evolutionary timepoints. (*B*) Distribution of functional classes arising at evolutionary timepoints. The relative frequency of each functional class is calculated independently for each timepoint.

followed by proteins involved in transport and binding and metabolism. At the other extreme, proteins involved in transcription, replication, cellular processes, and regulatory functions have, on average, almost twice as many binding partners. Interestingly, proteins of unknown function ("Unknown" functional class) are close to the lower end of connectivity.

Function Versus Time of Origin

As distinct connectivity levels associate with functional classes, we expect that those functions associated with higher connectivity levels emerge during the eukaryotic radiation. Conversely, for those functions that display lower connectivity levels, we expect appearance at the speciation stage. Figure 4 shows the distribution of the proteins in each functional class in the evolutionary path leading to *Saccharomyces*. It is striking how well the age of protein family correlates with the knowledge about the function of its members (fig. 4*A*). Whereas only 31% of pre-eukaryotic proteins belong to the "Unknown" functional class, the proportion grows as the origin of the protein family becomes more recent, and reaches 83% for *Saccharomyces*-specific proteins. This observation is a logical confirmation of our method as more phylogenetically extended protein families are also more likely to be better characterized.

Variation in the proportion of unknown functional class over the evolutionary periods masks trends from other functional classes. Several trends emerge in the

dynamics of their appearance, once the unknown functional class is removed (fig. 4*B*). The most dominating group of proteins in the pre-eukaryotic era are metabolic enzymes, characterized by low connectivity levels (fig. 3). This is consistent with previous reports that metabolism is one of the most conserved functional groups and known to appear very early in evolution (Peregrin-Alvarez, Tsoka, and Ouzouni 2003). Translation is another functional class appearing mostly in the pre-eukaryotic era and having only very low addition levels at the later stages. This is consistent with the view that translation is one of the most ancient processes in the cell. (Kunin 2000; Ouzounis and Kyrpides 1996).

In contrast, the dominating functional class in Eukaryotes is "Regulatory Functions." This functional class includes proteins involved in genetic, transcriptional, and posttranslational regulation. Proteins involved with cell wall biogenesis, which display low connectivity levels, have a clear trend of later appearance. Indeed, the chitin-based cell wall of fungi differs from prokaryotic-types and plant-types of cell wall (Klis et al. 2002; Martin Bhatt, and Baumann 2001). Other functional classes, such as transcription, transport and binding, and cellular functions maintain a relatively constant rate of appearance in the timepoints analyzed.

Overall, there is a clear distinction in the functional nature of protein families emerging at different stages of evolution, at least for the four distinct functional classes discussed above. Functional classes that display high average connectivity predominate in the time periods when highly connected proteins emerge, and lowly connected functional classes predominate in periods when lowly connected proteins emerge. It is unclear if the lack of variation observed in the remaining classes represents a true biological phenomenon or a limitation of resolution.

Conclusion

In conclusion, we have shown that each evolutionary period gives rise to distinct connectivity patterns. We have demonstrated that this is, in part, because of the emergence of different functions in each evolutionary timepoint, each characterized by a distinct level of connectivity. Considering that functional classes are constrained in their choice of binding partners (von Mering et al. 2002), it follows that preferential attachment in the protein interaction network operates within functional constrains.

The preferential attachment model aims to capture a general mechanism of network evolution capable of producing the observed systemic properties, namely the, scale-free topology. However, the mechanism by which preferential attachment operates is likely to be system-specific (Albert and Barabasi 2002). Our results suggest that function represents a constraint to the preferential attachment in the evolution of biological networks and that models of proteome evolution will have to take this into account.

Acknowledgments

J.B.P.-L. acknowledges support from the Fundação para a Ciência e Technologia–Portugal. C.A.O. acknowledges support from the UK Medical Research Council and IBM Research. Additional support was provided by the European Molecular Biology Laboratory. We thank Dr. Richard Coulson for critical reading of this manuscript.

Literature Cited

Albert, R., and A. L. Barabasi. 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. **74**:47–87.

Albert, R., H. Jeong, and A. L. Barabasi. 2000. Error and attack tolerance of complex networks. Nature **406**:378–382.

Anderson, S. C., A. Zomorodipour, J. O. Andersgon, T. Sicheritz-Ponten, U. C. Aismark, R. M. Podowski, A. K. Naslund, A. S. Eriksson, R. H. Winkler, and C. G. Kurland. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature **396**:133–140.

Andrade, M. A., N. P. Brown, C. Leroy et al. (11 co-authors). 1999. Automated genome sequence analysis and annotation. Bioinformatics **15**:391–412.

Ashburner, M., C. A. Ball, J. A. Blake et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**:25–29.

Bader, C. D., D. Betel, and C. W. Hogue. 2003. BIND: the biomolecular interaction network database. Nucleic Acids Res. **31**:248–250.

Barabasi, A. L., and R. Albert. 1999. Emergence of scaling in random networks. Science **286**:509–512.

Dunne, J. A., R. J. Williams, and N. D. Martinez. 2002. Food-web structure and network theory: the role of connectance and size. Proc. Natl. Acad. Sci. USA **99**:12917–12922.

Enright, A. J., I. Iliopolous, N. C. Kyrpides, and C. A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature **402**:86–90.

Enright, A. J., V. Kunin, and C. A. Ouzounis. 2003. Protein families and TRIBES in genome sequence space. Nucleic Acids Res. **31**:4632–4638.

Gavin, A. C., M. Bosche, R. Krause et al. (38 co-authors). 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**:141–147.

Golding, G. B., and R. S. Gupta. 1995. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. Evol. **12**:1–6.

Harrison, P. M., and M. Gerstein. 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J. Mol. Biol. **318**:1155–1174.

Hedges, S. B. 2002. The origin and evolution of model organisms. Nat. Rev. Genet. **3**:838–489.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA **98**:4569–4574.

Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. Nature **411**: 41–42.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. 2000. The large-scale organization of metabolic networks. Nature **407**:651–654.

Klis, F. M., P. Mol, K. Hellingwerf, and S. Brul. 2002. Dynamics of cell wall structure in *Saccharomyces cerevisiae*. FEMS Microbiol. Rev. **26**:239–256.

Kunin, V. 2000. A system of two polymerases: a model for the origin of life. Orig. Life Evol. Biosph. **30**:459–466.

Kunin, V., and C. A. Ouzounis. 2003*a*. GeneTRACE-reconstruction of gene content of ancestral species. Bioinformatics **19**:1412–1416.

———. 2003*b*. The balance of driving forces during genome evolution in prokaryotes. Genome Res. **13**:1589–1594.

Marcotte, F. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science **285**:751–753.

Martin C., K. Bhatt, and K. Baumann. 2001. Shaping in plant cells. Curr. Opin. Plant Biol. **4**:540–549.

Maslov, S., and K. Sneppen. 2002. Specificity and stability in topology of protein networks. Science **296**:910–933.

Ouzounis, C., and N. Kyrpides. 1996. The emergence of major cellular processes in evolution. FEBS Lett. **390**:119–123.

Peregrin-Alvarez, J. M., S. Tsoka, and C. A. Ouzounis. 2003. The phylogenetic extent of metabolic enzymes and pathways. Genome Res. **13**:422–427.

Qian, J., N. M. Luscombe, and M. Gerstein. 2001. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J. Mol. Biol. **313**:673–681.

Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gone content. Nat. Genet. **21**:108–110.

Uetz, P., L. Glot, and G. Cagney et al. (20 co-authors). 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature **403**:623–627.

von Mering, C., R. Krause, B. Snel, M. Cornell, S. C. Oliver, S. Fields, and P. Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417**:399–403.

Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol. Biol. Evol. **18**:1283–1292.

Willinger, W., R. Govindan, S. Jamin, V. Paxson, and S. Shenker. 2002. Sealing phenomena in the Internet: critically examining criticality. Proc Natl Acad Sci USA **99**(suppl 1): 2573–2580.

Xenarios, I., T. Salwinsiki, X. J. Duan, P. Higney, S. M. Kim, and D. Pisenberg. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30**:303–305.

Michele Vendruscolo, Associate Editor

Accepted December 12, 2003