PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2005) **272**, 1721–1725
doi:10.1098/rspb.2005.3128
Published online 14 July 2005

Gene essentiality and the topology of protein interaction networks

Stéphane Coulomb[1], Michel Bauer[1], Denis Bernard[1] and Marie-Claude Marsolier-Kergoat[2,*]

[1]*Service de Physique Théorique, CEA/Saclay-Orme des Merisiers, 91191 Gif-sur-Yvette Cedex, France*
[2]*Service de Biochimie et de Génétique Moléculaire, CEA/Saclay, 91191 Gif-sur-Yvette Cedex, France*

The mechanistic bases for gene essentiality and for cell mutational resistance have long been disputed. The recent availability of large protein interaction databases has fuelled the analysis of protein interaction networks and several authors have proposed that gene dispensability could be strongly related to some topological parameters of these networks. However, many results were based on protein interaction data whose biases were not taken into account. In this article, we show that the essentiality of a gene in yeast is poorly related to the number of interactants (or degree) of the corresponding protein and that the physiological consequences of gene deletions are unrelated to several other properties of proteins in the interaction networks, such as the average degrees of their nearest neighbours, their clustering coefficients or their relative distances. We also found that yeast protein interaction networks lack degree correlation, i.e. a propensity for their vertices to associate according to their degrees. Gene essentiality and more generally cell resistance against mutations thus seem largely unrelated to many parameters of protein network topology.

Keywords: network; protein interaction; topology; mutational resistance; gene dispensability; yeast

1. INTRODUCTION

Living organisms are resistant against mutations. Only 18.7% of *Saccharomyces cerevisiae* genes prove essential for growth on rich glucose medium and only 15% of all viable homozygous deletion strains exhibit a slow growth phenotype under optimal conditions (Giaever *et al.* 2002). When a systematic analysis of synthetic lethal or sick (SLS) interactions (in which the combination of mutations in two genes causes cell death or reduced fitness) was performed in yeast, the average number of SLS partners per studied gene was only 28 out of about 4700 interactions tested (Tong *et al.* 2004).

Gene dispensability and overall cell resistance to mutations can be explained either by the redundancy of duplicate genes or by other kinds of functional compensations involving sequence-unrelated genes. Gu *et al.* (2003) have found a significantly higher probability of functional compensation for a duplicate gene than for a single-copy gene and a high correlation between the frequency of compensation and the sequence similarity of two duplicates. However, even if there is clear evidence that duplicate genes play a significant role in mutational robustness, they do not seem to account for the main part of it.

The accumulation of small-scale studies and the development of high-throughput technologies have recently led to the construction of large protein interaction networks whose vertices (the proteins) are connected when evidence for physical interaction between them has been found. Several authors have proposed that the topological properties of proteins in interaction networks could be strongly related to gene essentiality and cell robustness against mutations (Jeong *et al.* 2001; Maslov & Sneppen 2002; Barabasi & Oltvai 2004). However these conclusions were based on protein interaction data whose biases had been overlooked. We have therefore reassessed systematically the relationships between gene essentiality and several topological characteristics of proteins in the yeast interaction networks. We found that the dispensability of a gene is only weakly related to the number of links of the corresponding protein and seems completely unrelated to all the other topological parameters that we tested, suggesting that protein network topology has little influence on the essentiality of specific genes and more globally on cell mutational robustness.

2. METHODS

Our study deals exclusively with data related to the yeast *S. cerevisiae*. To measure the degrees of essential and non-essential genes, the data corresponding to the study by Ito *et al.* (2001) was downloaded from the web page http://genome.c.kanazawa-u.ac.jp/Y2H/. A list of essential genes was obtained from the SGD (*Saccharomyces* Genome Database) webpage (http://db.yeastgenome.org/cgi-bin/SGD/phenotype/phenotype.pl?phenotype=inviable). To evaluate the average numbers of articles dealing with essential and non-essential genes, we considered all the open reading frames (ORFs) that were classified as genuine by the study of Ghaemmaghami *et al.* (2003). The number of SGD curated references was retrieved automatically from SGD web site at http://www.yeastgenome.org/.

To measure the degrees of SLS partners, the average degrees of nearest neighbours, the clustering coefficients and the distances between SLS partners, we have mostly used two

[*] Author for correspondence (marie-claude.marsolier-kergoat@cea.fr).

Received 20 December 2004
Accepted 7 April 2005

databases, which we considered the most comprehensive and reliable: the database of interacting proteins (DIP) core (Xenarios *et al.* 2000) and the 'affinity purification' databases. The DIP core database records data derived from both small-scale and large-scale experiments that has been validated by different criteria such as its reproducibility or the occurrence of the interaction between paralogous proteins in different species (Deane *et al.* 2002). The DIP core database was downloaded in May 2004 from the web page http://dip.doe-mbi.ucla.edu/dip/Download.cgi. We constructed the affinity purification database from the GRID database (http://biodata.mshri.on.ca:80/yeast_grid/servlet/SearchPage) in May 2004 by selecting all the protein interactions that have been found either by affinity chromatography or affinity precipitation: the large majority of the data (about 90%) corresponds to the results of Gavin *et al.* (2002) and Ho *et al.* (2002) and some data had also been collected from the small-scale studies' results recorded in the MIPS database (http://mips.gsf.de/genre/proj/yeast/index.jsp). Whereas the two-hybrid system is assumed to mostly reveal binary interactions, the affinity methods also yield complex composition data. In the case where proteins B and C are retrieved after affinity purification of protein A, the databases we used record the A/B and A/C interactions whereas A, B and C could be part of the same complex, B could interact with C and the A/C interaction could even be mediated by B. Thus the data produced from affinity purification techniques are not subject to the same biases as the two-hybrid system, but certainly have their own limits. The purpose of this double choice (DIP core and affinity purification databases) was to use large databases with *a priori* different biases.

The data from analyses of Tong *et al.* (2004) was collected from the GRID database (http://biodata.mshri.on.ca:80/yeast_grid/servlet/SearchPage).

3. RESULTS

(a) *The correlation between the essentiality of a gene and its degree in interaction networks*

Protein interaction networks exhibit a large degree distribution, which can be approximately fitted to a power law (Jeong *et al.* 2001): the probability $P(k)$ that a protein interacts with k other proteins roughly decreases like $k^{-\gamma}$. These networks consist in a majority of low-connected vertices and a minority of highly connected vertices or hubs. For convenience, we will consider the corresponding interaction networks linking genes whose protein products interact physically and we will refer to the topological parameters of these gene interaction networks. Several authors have found that the fraction of essential genes is 3 to 5 times higher among highly connected genes than among low-connected genes and have proposed that the phenotypic consequences of a gene deletion in yeast are affected to a large extent by the gene degree in the interaction networks (Jeong *et al.* 2001; Wuchty 2004; Yu *et al.* 2004). However these analyses used data derived from Uetz *et al.* (2000) and/or from the DIP database (Xenarios *et al.* 2000), which present some biases. The study by Uetz *et al.* suffers from a specific bias in that it mixes results from individual analyses of 192 genes (which remain unidentified) that have yielded 281 interacting protein pairs and from high-throughput screening procedures involving almost 6000 proteins that resulted in 692 interacting protein pairs. The DIP database also contains some biases regarding the number of interactants of essential and non-essential genes because it records data derived from small-scale experiments. We have found that essential genes are the objects of more articles than non-essential ones as they have on average 1.75 times more SGD curated references than non-essential genes (the SGD web site performs systematic searches through all PubMed literature for all papers mentioning yeast genes). The fact that essential genes are more intensely studied than non-essential ones is likely to increase the number of their known interactants.

In order to avoid the biases of databases recording the results of small-scale studies, we determined to use only interaction data derived from large-scale analyses. Besides the study by Uetz *et al.* (2000), three large-scale analyses have been conducted: Ito *et al.* (2001) have tested all possible interactions between about 6000 yeast proteins and have defined a data core containing only interactions that have been observed more than thrice, while Gavin *et al.* (2002) and Ho *et al.* (2002) have analysed the proteins found associated to respectively 493 and 589 tagged proteins. The study by Ito *et al.* is the only one that can be considered as unbiased regarding the proteins that were analysed since all yeast proteins were tested. In contrast, Ho *et al.* and Gavin *et al.* analysed a small fraction of yeast proteins and these proteins were not chosen randomly. We therefore focused on the data core of the Ito *et al.* analysis and we found that the fractions of essential genes present among the 10% less-connected and the 10% most-connected genes are 0.24 and 0.27, respectively. The average degrees of essential and non-essential genes are 2.2 and 1.8, respectively, in the Ito *et al.* study, which means that essential genes only have on average 1.2 times more links than non-essential ones. If the highest degrees of the genes present in the Ito *et al.* data are artificially attributed to the essential genes of this database, this ratio amounts to 3.8.

As another test of the correlation between the topological positions of genes and the physiological consequences of their deletions, we calculated the average degree of SLS partners. We used the study by Tong *et al.* (2004) which describes the systematic screenings for SLS partners of 132 genes, termed query genes. The degrees of the SLS genes in the interaction networks were estimated using different databases (the DIP core (Xenarios *et al.* 2000) and the affinity purification databases) as it was difficult to evaluate what could be their respective biases for this measure (see §2). If the topological positions of genes in the networks were related to the phenotypes of their deletions, SLS partners would be expected to exhibit higher degrees than the bulk of non-essential genes. However we found that the average degree of SLS partners is similar to the average degree of the whole set of non-essential genes, with 4.4 and 5.3 links for SLS genes versus 3.6 and 4.8 links for all non-essential genes in the DIP core and the affinity purification databases, respectively.

We observed that, when considering interaction results that are *a priori* unbiased for essential genes, essential genes only have slightly more interactants than non-essential ones. Similarly, we found little difference between the average numbers of links of SLS genes and of all non-essential genes. Our results suggest that the physiological consequences of gene deletions are only weakly related to gene degrees in interaction networks.

(b) *The essentiality of a gene and the average degree of its neighbours*

Another structural feature that has been suggested to influence gene essentiality is the average degree *K1* of their neighbours. Considering genes with a given degree *k*, it has been proposed that the deletion of genes with low *K1* is less deleterious than the deletion of genes with high *K1* because low *K1* should restrict the influence of the perturbations brought about by gene deletion (Maslov & Sneppen 2002). We therefore measured the average degree *K1* of nearest neighbours as a function of *k* for essential and non-essential genes, using either the affinity purification database or the DIP core database. We found no significant difference between the *K1* of essential and non-essential genes whatever their degrees *k* (figure 1 and data not shown). Hence the essentiality of a gene does not seem to be related to the average degree of its neighbours in interaction networks.

Interestingly, we also observed that *K1* is independent of *k* for both essential and non-essential genes. We confirmed this observation by measuring *K1* as a function of *k* for all genes with the affinity purification and the DIP core databases (figure 2*a*). This absence of correlation between *k* and *K1* contradicts several reports stating that yeast protein interaction networks exhibit a negative degree correlation, that is a propensity for high-degree vertices to attach to low-degree vertices (Maslov & Sneppen 2002; Newman 2002; Newman 2003). However, these studies used the full sets of data derived from systematic two-hybrid screens (Uetz *et al.* 2000; Ito *et al.* 2001), which suffer from several flaws, the most problematic one in this case being the number of false positives arising from the spurious activation of the reporter genes by proteins behaving as weak transcriptional activators (Aloy & Russell 2002). To illustrate the facility with which very few spurious hubs can bias these measures, we introduced into the experimental database affinity purification three artificial vertices randomly connected to 100, 150 and 200 vertices, mimicking the kind of artifacts that can be observed in two-hybrid experiments. As shown in figure 2*b*, the introduction of these three hubs enriches the network into low-degree vertices linked to high-degree ones and brings about a negative degree correlation comparable to that previously reported (Maslov & Sneppen 2002).

(c) *The essentiality of a gene and its clustering coefficient*

Essential and non-essential genes have also been proposed to differ in their clustering coefficients (Yu *et al.* 2004). The clustering coefficient of a vertex *i*, C_i, is defined as the ratio between n_i, the number of links that exist between the k_i neighbours of the vertex *i*, and $k_i\ (k_i-1)/2$, the number of all possible links between these k_i neighbours. The average clustering coefficients of essential and non-essential genes are almost identical for the DIP core database (0.31 and 0.29, respectively) and the data core of the Ito *et al.* study (0.09 and 0.11, respectively), but quite different with the affinity purification database (0.29 and 0.18, respectively) and the database derived from the Gavin *et al.* study (0.40 and 0.30, respectively). Since the affinity purification database and the results of the Gavin *et al.* study are probably enriched in interactions within stable protein complexes, our observations could hint at a relationship between the clustering coefficient and the essentiality of genes specifically when their protein products are part of stable complexes. However this potential connection is not valid when considering more diverse sets of interactions like those recorded in the Ito *et al.* study or in the DIP core database, and we cannot reliably associate gene essentiality with this parameter.

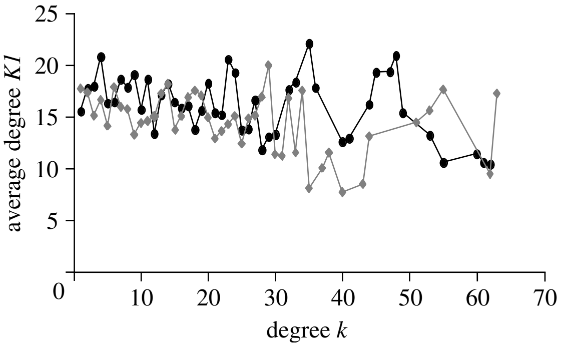

Figure 1. The average degree *K1* of nearest neighbours as a function of the degree *k* for essential (black circles) and non-essential genes (grey diamonds). The list of interactions used is defined by the affinity purification database and similar results were obtained with the DIP core database.

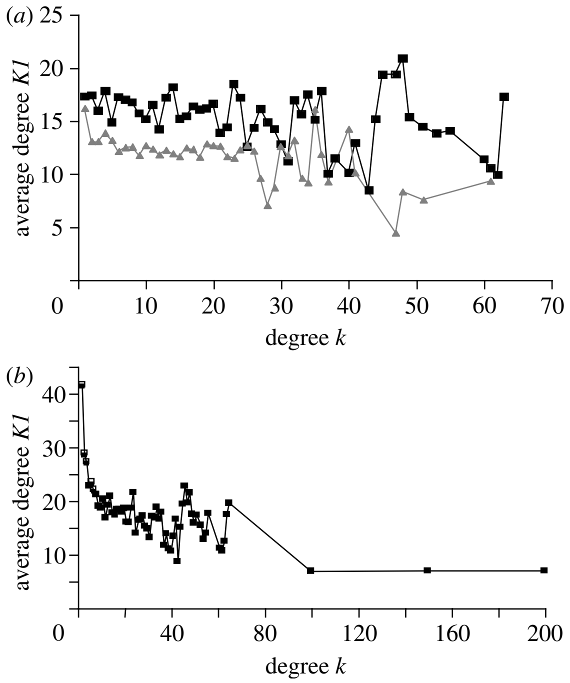

Figure 2. Absence of correlation between the degrees *k* of genes and the average degrees *K1* of their nearest neighbours. (*a*) The average degrees *K1* of the nearest neighbours of genes are plotted against their degrees *k* using either the affinity purification database (black squares) or the DIP core database (grey triangles)—the point ($k=111$; $K1=5.2$) for the DIP database is out of the range of the figure. (*b*) *K1* is plotted as a function of *k* for the affinity purification database complemented with three artificial hubs randomly connected to 100, 150 and 200 vertices. The introduction of the three spurious hubs and the subsequent measure of *K1* were repeated 100 times and the averaged results are represented here. The first point ($k=1$) has been removed as it corresponds exclusively to vertices that have not been directly affected by the introduction of the spurious hubs.

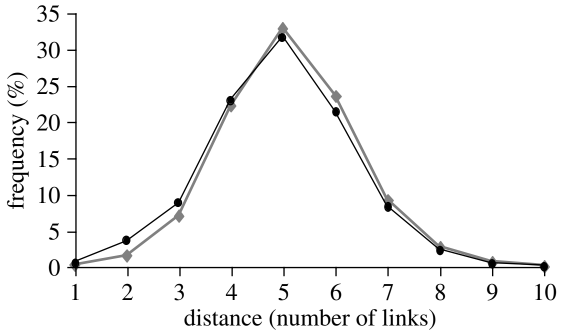

Figure 3. The distributions of the distances separating query genes from their SLS partners (black circles) and from the whole set of non-essential genes (grey diamonds). The distances represent the numbers of intermediate links in the gene interaction network defined by the DIP core database. They were measured using a breadth-first search algorithm. Only query genes whose complete deletion has been studied were included in this analysis. Similar results were obtained with the affinity purification database.

(d) *Synthetic lethality/sickness and the relative distances between genes*

Finally we tested whether the distance between genes in the interaction networks could be related to the phenotypes of double mutations. Using again the study by Tong *et al.* (2004) and either the affinity purification database or the DIP core database, we found that the average distance separating query genes from their SLS partners in the interaction networks (4.7 and 5.3 links for the affinity purification and the DIP core databases, respectively) is similar to the average distance separating query genes from the whole set of non-essential genes (4.6 and 5.6 links for the affinity purification and the DIP core databases, respectively). A closer analysis showed that the distributions of the distances between query genes and their SLS partners or the global set of non-essential genes are almost identical (figure 3 and data not shown). The synthetic lethality or sickness of two genes in yeast thus does not seem related to their relative topological positions in interaction networks.

4. DISCUSSION

We have looked systematically for correlations between the essentiality of genes and several of their topological characteristics in interaction networks. We have shown that the essentiality of genes or their synthetic lethality or sickness in yeast are only weakly related to their degrees in interaction networks and that the physiological consequences of gene deletions are unrelated to the average degrees of the genes' neighbours or to their relative distances. The difference in the clustering coefficients of essential and non-essential genes proved highly variable from one interaction database to another, which precludes any general conclusion about the relationship between the essentiality of a gene and its cliquishness. So far no topological feature of genes in the interaction networks has been found to be strongly related to their essentiality. We also showed that yeast interaction networks lack degree correlation. Our results contradict several reports (Jeong *et al.* 2001; Maslov & Sneppen 2002; Newman 2002; Newman 2003; Wuchty 2004; Yu *et al.* 2004) that argue for the existence of a negative degree correlation in yeast protein interaction networks or for a strong correlation between the essentiality of a gene and its topological characteristics in interaction networks. One of the main problems in studying interaction networks lies in the choice of the interaction databases. All interaction databases are probably biased and it appears crucial to try to understand these biases and to test several bases recording different kinds of data.

Several authors have proposed that topological characteristics of protein interaction networks like their degree distribution or degree correlation could account for cell robustness against mutations (Jeong *et al.* 2001; Maslov & Sneppen 2002; Barabasi & Oltvai 2004). However, these hypotheses are inconsistent with the observation that the essentiality of genes is either poorly related or completely unrelated to their topological parameters in the interaction networks. The fact that the structure of interaction networks has little or no effect on gene essentiality and mutational robustness is compatible with the hypotheses that the topology of interaction networks would not be submitted to evolutionary constraints and that their structural features would simply be the consequences of construction processes. Indeed, it has been shown that protein interaction networks with a large degree distribution could be generated with evolution processes involving gene duplication and addition and deletion of links, without invoking natural selection on the degree distribution itself (Wagner 2003). More recently, Amoutzias *et al.* (2004) and van Noort *et al.* (2004) have also shown that many structural characteristics of experimentally observed coexpression and interaction networks could be generated by neutralist models without the need of selection. Likewise, the fact that yeast protein interaction networks lack degree correlation is compatible with random processes of network construction, even if it excludes some models of network evolution based on asymmetric link attachment (Berg *et al.* 2004).

In conclusion, our results show that gene dispensability has little if any relationship to the structure of protein interaction networks. Further investigations will be required to unravel the molecular mechanisms of gene essentiality and of cell mutational resistance.

We would like to thank Jean-Louis Sikorav and Arach Goldar for critical comments on the manuscript, and Raphaël Guérois for the automatic retrieval of the numbers of SGD curated references.

REFERENCES

Aloy, P. & Russell, R. B. 2002 Potential artefacts in protein-interaction networks. *FEBS Lett.* **530**, 253–254.

Amoutzias, G. D., Robertson, D. L., Oliver, S. G. & Bornberg-Bauer, E. 2004 Convergent networks by single-gene duplications in higher eukaryotes. *EMBO Rep.* **5**, 274–279.

Barabasi, A. L. & Oltvai, Z. N. 2004 Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113.

Berg, J., Laessig, M. & Wagner, A. 2004 Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.* **4**, 51.

Deane, C. M., 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.

Gavin, A. C. *et al.* 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.

Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K. & Weissman, J. S. 2003 Global analysis of protein expression in yeast. *Nature* **425**, 737–741.

Giaever, G. *et al.* 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W. & Li, W. H. 2003 Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66.

Ho, Y. *et al.* 2002 Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183.

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

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

Maslov, S. & Sneppen, K. 2002 Specificity and stability in topology of protein networks. *Science* **296**, 910–913.

Newman, M. E. 2002 Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208 701.

Newman, M. E. 2003 Mixing patterns in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **67**, 026 126.

Tong, A. H. *et al.* 2004 Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813.

Uetz, P. *et al.* 2000 A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627.

van Noort, V., Snel, B. & Huynen, M. A. 2004 The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* **5**, 280–284.

Wagner, A. 2003 How the global structure of protein interaction networks evolves. *Proc. R. Soc. B* **270**, 457–466. (doi:10.1098/rspb.2002.2269.)

Wuchty, S. 2004 Evolution and topology in the yeast protein interaction network. *Genome Res.* **14**, 1310–1314.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. 2000 DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289–291.

Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. & Gerstein, M. 2004 Genomic analysis of essentiality within protein networks. *Trends Genet.* **20**, 227–231.