

the deleterious effect of A53 became more prominent, its replacement by T53 was advantageous. Nevertheless, the fixations of such mutations might not be under strong positive selection because they have a negligible effect on the reproductive success of the organism.

Compensatory changes?

Although only five FDDAMs are associated with fatal early-onset diseases, all 16 early-onset diseases reduce the fitness of the affected individuals to certain degrees. The mutations associated with these diseases are probably under strong purifying selection and it is difficult to explain why they have been fixed in mice. Even for *Sgca*, in which the P30L substitution probably occurred during mouse breeding (Figure 2), the mouse protein is apparently functional because the null mutation leads to the expected disease phenotype [13]. In other words, a founder effect cannot adequately explain why P30L did not lead to the disease in mice. Kondrashov *et al.* [9] suggested that compensatory mutations at other sites of the same or a different protein render the deleterious mutations neutral. Compensatory mutations have been experimentally demonstrated in several proteins [14,15]. In general, if (i) a mutation from amino acid A1 to A2 at position A causes a disease in human; (ii) human has amino acid B1 at position B (of the same or a different protein); and (iii) amino acid B2 at position B can compensate the deleterious A2, the compensatory hypothesis predicts that all species with A2 should also have B2. Such predictions can be tested using comparative methods when sequence data from multiple species are available. Two empirical examples have been extensively discussed in Ref. [9]. Given the pervasiveness of protein–protein interaction and possibility of genotype–environment interaction, compensatory changes might occur in different genes or even in the environment.

Concluding remarks

In summary, our analyses of 20 cases of human–mouse FDDAMs show that a small number of FDDAMs might be explained by either a founder effect in mouse breeding or reduced selection against late-onset disease phenotypes, but the majority require other explanations. The compensatory hypothesis appears to be the most likely one and identification of the compensatory changes would greatly

help understand the mechanisms behind the FDDAMs in mice and other species.

Acknowledgements

We thank Priscilla Tucker for the DNA samples of various *Mus* species and David Webb for valuable comments. This work was supported by a startup fund of the University of Michigan and an NIH grant (GM67030) to J.Z.

References

- Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
- Ahmed, S.F. *et al.* (2000) Phenotypic features, androgen receptor binding, and mutational analysis in 278 clinical cases reported as androgen insensitivity syndrome. *J. Clin. Endocrinol. Metab.* 85, 658–665
- Ferris, S.D. *et al.* (1982) Evidence from mtDNA sequences that common laboratory strains of inbred mice are descended from a single female. *Nature* 295, 163–165
- Tucker, P.K. *et al.* (1992) Geographic origin of the Y chromosomes in “old” inbred strains of mice. *Mamm. Genome* 3, 254–261
- Wade, C.M. *et al.* (2002) The mosaic structure of variation in the laboratory mouse genome. *Nature* 420, 574–578
- Guenet, J.L. and Bonhomme, F. (2003) Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet.* 19, 24–31
- Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425
- Thompson, J.D. *et al.* (1997) The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882
- Kondrashov, A.S. *et al.* (2002) Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14878–14883
- Schaner, P. *et al.* (2001) Episodic evolution of pyrin in primates: human mutations recapitulate ancestral amino acid states. *Nat. Genet.* 27, 318–321
- Ohta, T. (1995) Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* 40, 56–63
- Hamilton, W.D. (1966) The moulding of senescence by natural selection. *J. Theor. Biol.* 12, 12–45
- Duclos, F. *et al.* (1998) Progressive muscular dystrophy in alpha-sarcoglycan-deficient mice. *J. Cell Biol.* 142, 1461–1471
- Izumi, T. *et al.* (1999) Intragenic suppression of an active site mutation in the human apurinic/apyrimidinic endonuclease. *J. Mol. Biol.* 287, 47–57
- Zhang, J. and Rosenberg, H.F. (2002) Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5486–5491
- Murphy, W.J. *et al.* (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* 294, 2348–2351

0168-9525/\$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2003.10.002

Protein interaction: same network, different hubs [☆]

Robert Hoffmann and Alfonso Valencia

National Center of Biotechnology, CNB-CSIC. Cantoblanco Madrid M-28049, Spain

Recently, large-scale experiments have provided new insights into the complex protein interaction network

in yeast. However, previous analyses have shown that the number of interacting pairs that are common to different methods is extremely low and, therefore, less informative than expected. In this article, we show that comparing the connectivities of individual proteins can reveal that a common tendency between methods

* Supplementary data associated with this article can be found at doi: 10.1016/j.tig.2003.10.011

Corresponding author: Robert Hoffmann (hoffmann@cnb.uam.es).

has been missed by the pairwise comparison of interactions. We found significant correlations between experimental methods and also between various *in silico* methods. Exceptionally, a computational method, gene neighbourhood, correlates with both *in silico* and experimental approaches.

Mering [1] and Bader *et al.* [2] compared experimental and *in silico* network predictions for yeast and concluded that the agreement between individual methods was extremely low. Tandem affinity purification (TAP) and high-throughput mass spectrometric complex identification (HMS), for instance, are technically similar, but they agree only in a fraction of their interacting pairs. This low overlap of interactions between methods is usually explained by the intrinsic experimental errors and complexity of the system [3]. However, the possibility exists that different methods are simply measuring different types of interactions (e.g. transient or stable) [4]. In this article, we adopt a different perspective and compare the organization of predicted networks by individual protein connectivity. In other words, we compare the number of connections that a protein has in different experimental conditions and not the particular nature of each connection.

Although based on various technical approaches, current methods all predict a scale-free topology for the intracellular network [5,6]. In such networks, a few central hub proteins interact with many other proteins, but the majority of proteins have few interactions and are placed at the periphery of the network. The importance of hub proteins has been discussed in the context of essential functions [5] and the evolution of interaction networks [7]. We argue that it is difficult to imagine how two result sets can represent the same network, if they do not agree on the general organization of individual proteins within the network. In other words, hub proteins, should be predicted as central by all compatible methods, even if there is no agreement on all specific interactions. Therefore we propose that the connectivity of proteins is a more suitable measure for the comparison of two protein networks than the comparison of single pairwise interactions (Figure 1a). Furthermore, measuring the connectivity takes all data into account because individual proteins, and not only pairs of proteins, are considered.

The organization of protein networks

To compare our approach with previous analyses [1,2] we have tested data collected from various sources by Mering *et al.* [1]. For every protein in each method (Box 1) we have calculated the degree of interaction. Then, all methods were compared in pairs, taking into account only those proteins that had been covered by both methods. (Figure 1b). Experimental methods such as spectrometric complex identification (HMS) and tandem affinity purification (TAP) correlate significantly with the curated comparison standard

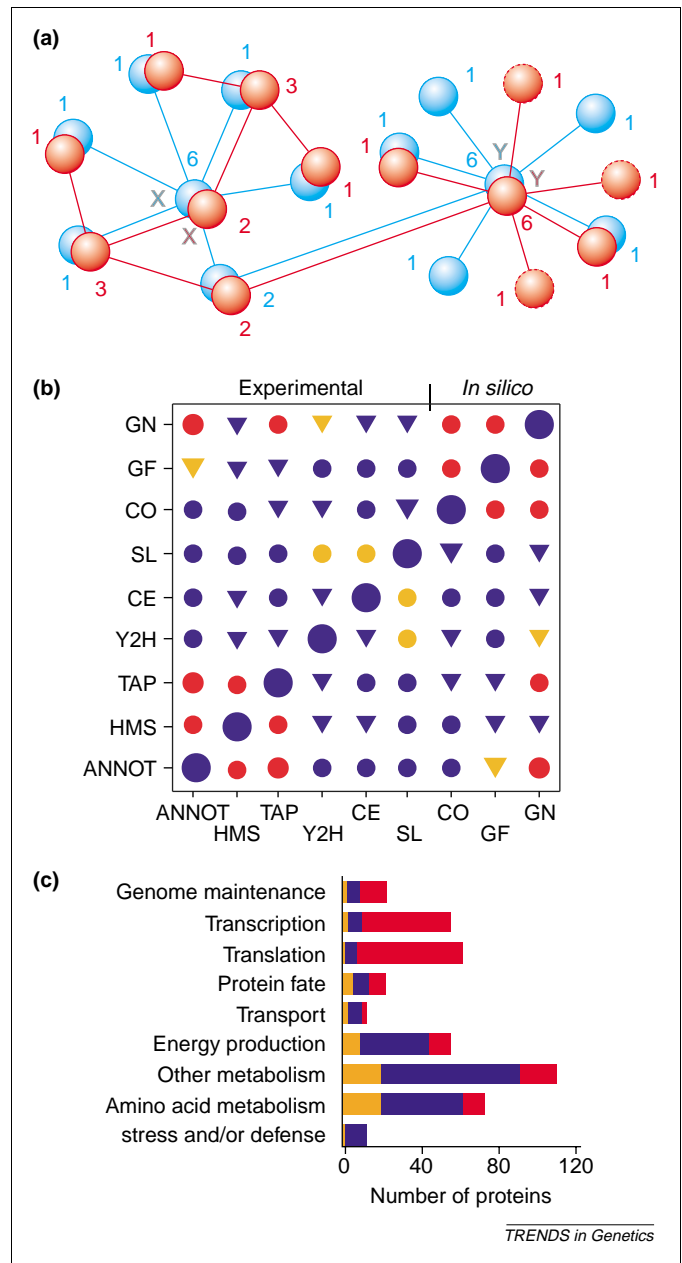


Figure 1. (a) Topological comparison of two hypothetical networks. The number of connections is the same in both networks (red and blue), however, the connectivity of individual proteins is distributed differently. For example, protein X has a central position only in the blue network. The degree of interaction of each protein is determined by its total number of connections, independently of the identity of its binding partners. Protein Y, for example, has a connectivity of six in both networks, although the individual interactions are not the same. This way, the complete data can be taken into account and the comparison of networks is not restricted to the small number of identical interactions. (b) Correlation matrix of protein connectivities detected by experimental and *in silico* methods. Every field in the matrix represents a nonparametric comparison of protein connectivities between two methods. Symbol sizes correspond to correlation coefficients (Spearman's rho). A nonparametric correlation analysis was applied, because the distribution of connectivity follows a power law and the assumption of normal distribution is not valid. Positive correlation is shown as circles, negative correlation as triangles. Significant correlation coefficients are indicated in red (P -value = 0.01) and in orange (P -value = 0.05). See supplementary information online. (c) Functional classes of proteins involved in GN correlations. Intersection of GN and TAP is shown in red, intersection of GN and CO in blue and of GN, TAP and CO in orange. The correlation between methods increases significantly when restricted to the corresponding functional categories (see supplementary information online). The correlation between methods increases significantly when restricted to the corresponding functional categories (see supplementary information online). Abbreviations: ANNOT, reference set (MIPS and YPD); CO, co-occurrence of genes; CE, correlated mRNA expression; GF, gene fusion events; GN, conserved gene neighbourhood; HMS, spectrometric complex identification; SL, synthetic lethal interactions; TAP, tandem affinity purification; Y2-H, yeast two-hybrid.

Box 1. Predicting functional interaction

Various experimental and *in silico* methods have been developed to predict the functional or physical interaction of proteins. *In silico* methods are generally based on the comparative analysis of bacterial genomes, and functional interactions in yeast are inferred by homology between yeast and bacterial proteins.

Conserved gene neighbourhood (GN)

This method predicts the functional interaction of genes on the basis of the conservation of gene clusters in bacterial genomes [9,10].

Co-occurrence of genes (CO)

The underlying assumption of this method is that genes that function together in a pathway or structural complex will have matching or similar phylogenetic profiles [11], exhibiting a similar pattern of presence and absence over genomes.

Gene fusion events (GF)

Gene fusion is based on the assumption that the functional interaction of proteins might lead to the fusion of the corresponding genes in some species [12,13].

Correlated mRNA expression (CE)

The prediction of functional interaction from mRNA expression data is based on the hypothesis that genes will have a functional relationship if their expression profiles correlate across several experiments (conditions) [1].

Synthetic lethal interactions or genetic interaction (SL)

This method is based on the assumption that two genes functionally interact if only the absence (i.e. deletion) of both genes has a lethal effect in yeast [14].

(ANNOT) and with each other. The other group is made up of the *in silico* methods gene neighbourhood (GN), co-occurrence (CO) and gene fusion (GF). At the network organization level, there is no consistency between yeast two-hybrid (Y2-H) and the other methods, although reasonable consistency between yeast two-hybrid datasets from different laboratories can be found, something that was not detected by previous assessments [8] (see supplementary material online).

Interestingly, the *in silico* gene neighbourhood method (GN), which is based on evolutionary constraints on genome organization, correlates with most other methods, thereby linking experimental and *in silico* predictions. This seems to be because of its broad coverage of proteins from diverse functional classes; the correlation between GN and TAP is caused essentially by proteins involved in transcription and translation, whereas the correlation between GN and CO is based mainly on proteins that function in metabolism and energy production (Figure 1c). For example, both GN and TAP agree on the central role of MRPL8 (large subunit of the mitochondrial ribosome),

whereas they completely disagree on the connectivity of ACACA (acetyl-Coenzyme A carboxylate α), which is involved in fatty acid biosynthesis. Further information is available from the authors at http://www.pdg.cnb.uam.es/supplement/protein_network/index.html.

Conclusion

Although the pairwise comparison of interactions revealed little coincidence between the various experimental or computational methods, we have shown that comparing connectivities of individual proteins can uncover characteristic experimental bias and common tendencies within network organizations.

Acknowledgements

We thank Michael Tress and David Juan for helpful discussion. This work was supported in part by the ORIEL and TEMPLOR EC projects.

References

- 1 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403
- 2 Bader, G.D. and Hogue, C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* 20, 991–997
- 3 Grunfelder, B. and Winzler, E.A. (2002) Treasures and traps in genome-wide data sets: case examples from yeast. *Nat. Rev. Genet.* 3, 653–661
- 4 Aloy, P. and Russell, R.B. (2002) The third dimension for protein interactions and complexes. *Trends Biochem. Sci.* 27, 633–638
- 5 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42
- 6 Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* 18, 1283–1292
- 7 Fraser, H.B. *et al.* (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752
- 8 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574
- 9 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
- 10 Huynen, M. *et al.* (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210
- 11 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288
- 12 Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86
- 13 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
- 14 Tong, A.H. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294, 2364–2368