A Model of Large-Scale Proteome Evolution

Ricard V. Solé
D. Eric Smith
Romualdo Pastor-Satorras
Thomas Kepler

SFI WORKING PAPER: 2001-08-041

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu

SANTA FE INSTITUTE

A model of large-scale proteome evolution

Ricard V. Solé[1,2,3], Romualdo Pastor-Satorras[1], Eric D. Smith[2], and Thomas Kepler[2]
[1] *Complex Systems Research Group, FEN*
Universitat Politècnica de Catalunya, Campus Nord B4, 08034 Barcelona, Spain
[2]*Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA*
[3]*NASA-associated Astrobiology Institute, INTA/CSIC, Carretera del Ajalvir Km 4, Madrid*

The next step in the understanding of the genome organization, after the determination of complete sequences, involves proteomics. The proteome includes the whole set of protein-protein interactions, and two recent independent studies have shown that its topology displays a number of surprising features shared by other complex networks, both natural and artificial. In order to understand the origins of this topology and its evolutionary implications, we present a simple model of proteome evolution that is able to reproduce many of the observed statistical regularities reported from the analysis of the yeast proteome. Our results suggest that the observed patterns can be explained by a process of gene duplication and diversification that would evolve proteome networks under a selection pressure, favoring robustness against failure of its individual components.

I. INTRODUCTION

The genome is one of the most fascinating examples of the importance of emergence from network interactions. The recent sequencing of the human genome (Lander et al., 2001; Venter et al., 2001) revealed some unexpected features and confirmed that *"the sequence is only the first level of understanding of the genome"* (Venter et al., 2001). The next fundamental step beyond the determination of the genome sequence involves the study of the properties of the proteins the genes encode, as well as their interactions (Fields, 2001). Protein interactions play a key role at many different levels and its failure can lead to cell malfunction or even apoptosis, in some cases triggering neoplastic transformation. This is the case, for example, of the feedback loop between two well-known proteins, MDM2 and p53: in some types of cancers, amplification of the first (an oncoprotein) leads to the inactivation of p53, a tumor-suppressor gene that is central in the control of the cell cycle and death (Wu et al., 1993).

Understanding the specific details of protein-protein interactions is an essential part of our understanding of the proteome, but a complementary approach is provided by the observation that network-like effects play also a key role. Using again p53 as an example, this gene is actually involved in a large number of interaction pathways dealing with cell signaling, the maintenance of genetic stability, or the induction of cellular differentiation (Vogelstein et al., 2000). The failure in p53, as when a highly connected node in the Internet breaks (Albert et al., 2000), has severe consequences.

Additional insight is provided by the observation that in many cases the total suppression of a given gene in a given organism leads to a small phenotypic effect or even no effect at all (Ross-Macdonald et al., 1999). These observations support the idea that, although some genes might play a key role and their suppression is lethal, many others can be replaced in their function by some redundancy implicit in the network of interacting proteins.

Protein-protein interaction maps have been studied, at different levels, in a variety of organisms including viruses (Bartel et al., 1996; McCraith et al., 2000; Flajolet et al., 2000), prokaryotes (Rain et al., 2001), yeast (Ito et al., 2000), and multicellular organisms such as *C. elegans* (Walhout et al., 2000). Most previous studies have used the so-called two-hybrid assay (Fromont-Racine et al., 1997) based on the properties of site-specific transcriptional activators. Although differences exist between different two-hybrid projects (Hazbum and Fields, 2001) the statistical patterns used in our study are robust.

Recent studies have revealed a surprising result: the protein-protein interaction networks in the yeast *Saccharomyces cerevisiae* share some universal features with other complex networks (Strogatz 2001). These studies actually offer the first global view of the proteome map. These are very heterogeneous networks: The probability $P(k)$ that a given protein interacts with other k proteins is given by a power law, i.e. $P(k) \sim k^{-\gamma}$ with $\gamma \approx 2.5$ (see figure 1). Additionally, they also display the so-called small-world (SW) effect: they are highly clustered (i.e. each node has a well-defined neighborhood of "close" nodes) but the minimum distance between any two randomly chosen nodes in the graph is short, a characteristic feature of random graphs (Watts and Strogatz, 1998). Finally, as pointed in Wagner's study, the average connectivity $\bar{K}$ of the proteome is close to the threshold $\bar{K}_c = 2$, above which cycles of all length begin to emerge in a random network (Bollobás, 1985; Kauffman, 1993). Specifically, there is a critical value $\bar{K}_p = 1$ such that for $\bar{K} < 1$ the graph is essentially disconnected but above it, $\bar{K} > 1$ a *giant cluster* (the giant component) Ω_∞ spans the entire graph (assuming large size). For even larger connectivities, $\bar{K} > 2$, cycles (closed paths of links between nodes) of all sizes appear. Such a small connectivity in the yeast network, close to the threshold $\bar{K}_c = 2$, might indicate a minimization of the number of (costly) links among proteins, while preserving the stability of the whole system. The presence of cycles of all lengths at $\bar{K} \approx 2$ might be related to a favoring of the

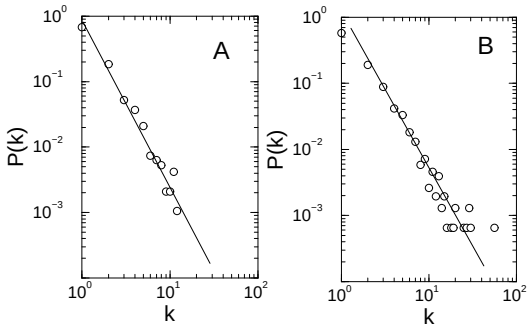

FIG. 1. Degree distributions for two different data sets from the Yeast proteome (A: Wagner, 2001; B: Jeong et al., 2001). Both distributions display scaling behavior in their degree distribution $P(k)$, i.e. $P(k) \sim k^{-\gamma}$, a sharp cut-off for large k and very small average connectivities: $\bar{K}_A = 1.83$ (total graph) and $\bar{K}_B = 2.3$ (giant component), respectively. The slopes are $\gamma_A \approx 2.5 \pm 0.15$ and $\gamma_B \approx 2.4 \pm 0.21$.

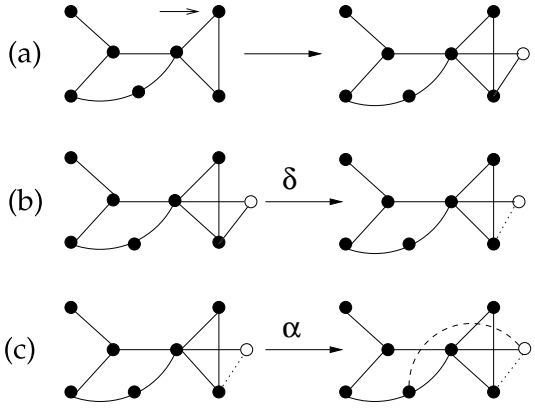

FIG. 2. Growing network by duplication of nodes. First (a) duplication occurs after randomly selecting a node (arrow). The links from the newly created node (white) now can experience deletion (b) and new links can be created (c); these events occur with probabilities δ and α, respectively.

modularity in protein interactions.

As shown in previous studies (Albert et al., 2000) this type of networks is extremely robust against random node removal but also very fragile when removal is performed selectively on the most connected nodes. SW networks appear to be present in a wide range of systems, including artificial ones (Albert et al., 2000; Amaral et al., 2000; Ferrer et al., 2001) and also in neural networks (Watts and Strogatz, 1998), metabolic pathways (Jeong et al., 2000, Fell and Wagner, 2000), even in human language organization (Ferrer and Solé, 2001). The implications of these topologies are enormous also for our understanding of epidemics (Pastor-Satorras and Vespignani, 2001; Lloyd and May, 2001).

The previous observations can be summarized as follows: (1) the proteome map is a sparse graph, with a small average number of links per protein. This observation is also consistent with the study of the global organization of the *E. coli* gene network from available information on transcriptional regulation (Thieffry et al., 1998); (2) it exhibits a SW pattern, different from the properties displayed by purely random (Poissonian) graphs and (3) the degree distribution of links is a power law with a well-defined cut-off. This would have adaptive significance as a source of robustness against mutations.

In this paper we present a model of proteome evolution aimed at capturing the main properties exhibited by protein networks. The basic ingredients of the model are gene duplication plus re-wiring of the protein iterations, two elements known to be the essential driving forces in genome evolution (Ohno, 1970). The model does not include functionality or dynamics of the proteins involved, but it is a topological-based approximation to the overall features of the proteome graph and intends to capture some of the (possibly) generic features of (real) proteome evolution.

During the completion of this work we became aware of a paper by Vázquez et al., (Vázquez et al., 2001), in which a related model of proteome evolution, showing multifractal connectivity properties, is described and analyzed.

II. PROTEOME GROWTH MODEL

Here we restrict our rules to single-gene duplications, which occur in most cases due to unequal crossover (Ohno, 1970). Multiple duplications should be considered in future extensions of these models: molecular evidence shows that even whole-genome duplications have actually occurred in *S. cerevisiae* (Wolfe and Shields, 1997; see also Wagner, 1994). Re-wiring has also been used in dynamical models of the evolution of robustness in complex organisms (Bornholdt and Sneppen, 2000).

The proteome graph at any given step t (i.e. after t duplications) will be indicated as $\Omega_p(t)$. The rules of the model, summarized in figure 2, are implemented as follows. Each time step: (a) one node in the graph is randomly chosen and duplicated; (b) the links emerging from the new generated node are removed with probability δ; (c) finally, new links (not previously present) can be created between the new node and all the rest of the nodes with probability α. Step (a) implements gene duplication, in which both the original and the replicated proteins retain the same structural properties and, consequently, the same set of interactions. The rewiring steps (b) and (c) implement the possible mutations of the replicated gene, which translate into the deletion and addition of interactions, with different probabilities.

Since we have two free parameters, we should first constrain their possible values by using the available empirical data. As a first step, we can estimate the asymptotic average connectivity exhibited by the model in a mean-field approximation (see also Vázquez et al., 2001). Let

us indicate by $\bar{K}$ the average connectivity of the system. It is not difficult to see that, when the network is composed by N nodes, the increase in the average connectivity after one iteration step of the model is proportional to

$$\Delta\bar{K} = \bar{K} - 2\delta\bar{K} + 2\alpha(N - \bar{K}). \tag{1}$$

The first term accounts for the duplication of one node, the second represents the average elimination of $\delta\bar{K}$ links emanationg from the new node, and the last term represents the addition of $\alpha(N-\bar{K})$ new connections pointing to the new node. If the model has to show a steady state with constant average connectivity, we must impose the condition, valid for large N,

$$\bar{K} - 2\delta\bar{K} + 2\alpha(N - \bar{K}) = 0, \tag{2}$$

which yields

$$\bar{K} = \frac{2\alpha N}{2\alpha + 2\delta - 1}. \tag{3}$$

In order to have a finite K, we must impose the condition $\alpha = \beta/N$, where β is a constant. That is, the rate of addition of new links (the establishment of new viable interactions between proteins) is inversely proportional to the network size, and thus much smaller than the deletion rate δ, in agreement with the rates observed in (Wagner, 2001). In this case, for large N, we get

$$\bar{K} = \frac{2\beta}{2\delta - 1}. \tag{4}$$

The previous expression imposes the boundary condition $\delta > 1/2$, necessary in order to obtain a well-defined limiting average connectivity. Eq. (4), together with the experimental estimates of $\bar{K} \sim 1.9 - 2.3$, allows to set a first restriction to the parameters β and δ. Imposing $\bar{K} = 2$, i.e., an average connectivity just on the onset of cycles of all lengths in a random graph, , we are led to the relation

$$\beta = 2\delta - 1. \tag{5}$$

Moreover, estimations of addition and deletion rates α and δ from yeast (Wagner, 2001) give a ratio $\alpha/\delta \leq 10^{-3}$. For proteomes of size $N \sim 10^3$, as in the case of the yeast, this leads to $\beta/\delta \leq 10^{-3}N \sim 1$. Using the safe approximation $\beta/\delta = 0.1$, together with the constraint (4), we obtain the values

$$\delta = 0.53, \qquad \beta = 0.06, \tag{6}$$

which will be used thought the rest of the paper. Similar values of β and δ, consistent with both constraints, were also checked, providing compatible results.

Simulations of the model start form a connected ring of $N_0 = 5$ nodes, and proceed by iterating the rules until the desired network size is achieved.

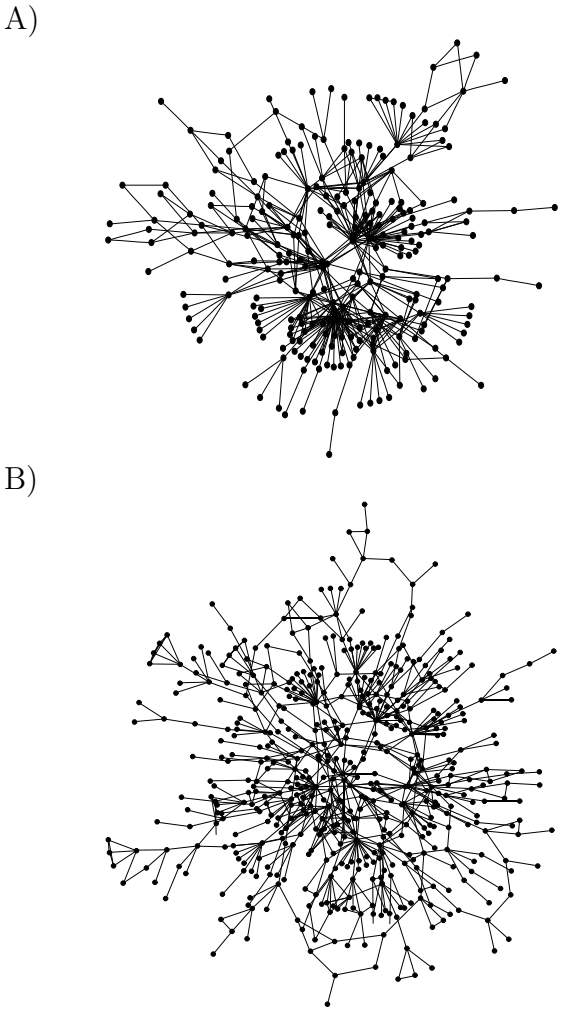

FIG. 3. A) An example of a small proteome interaction map (giant component, Ω_∞) generated by the model with $N = 10^3$, $\delta = 0.53$, and $\beta = 0.06$. B) Real yeast proteome map obtained from the MIPS database (Mewes et al. 1999). We can observe the close similitude between the real map and the output of the model.

III. RESULTS

Computer simulations of the proposed model reproduce many of the regularities observed in the real proteome data. As an example of the output of the model, in figure 3A we show an example of the giant component Ω_∞ of a realization of the model with $N = 10^3$ nodes. This figure clearly resembles the giant component of real yeast networks, as we can see comparing with figure 3B*, and we can appreciate the presence of a few highly connected hubs plus many nodes with a relatively small num-

*Figure kindly provided by W. Basalaj (see http://www.cl.cam.uk/ wb204/GD99/#Mewes).

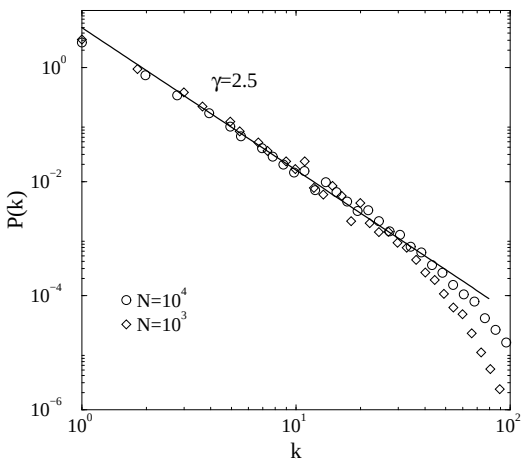

FIG. 4. Degree distribution $P(k)$ for the model, averaged over 10^4 networks of size $N = 10^3$ and 10^3 networks of size $N = 10^4$. The distributions show a characteristic power law behavior, with exponent $\gamma = 2.5 \pm 0.1$.

ber of connections. The size of the giant component for $N = 10^3$, averaged of 10^4 networks, is $|\Omega_\infty| = 360 \pm 100$, in good agreement with Wagner's data $|\Omega_\infty^W| = 466$ for a yeast with a similar total number of proteins (the high variance in our result is due to the large fluctuations in the model for such small network size N). On the other hand, in figure 4 we plot the connectivity $P(k)$ obtained for networks of size $N = 10^3$ and $N = 10^4$. In this figure we observe that the resulting connectivity distribution can be fitted to a power law for over an order of magnitude, with an scaling exponent $\gamma = 2.5 \pm 0.1$, in good agreement with the measurements reported by (Wagner, 2001) and (Jeong et al,, 2001).

An additional observation from Wagner's study of the yeast proteome is the presence of SW properties. We have found also similar topological features in our model, using the considered set of parameters. The proteome graph is defined by a pair $\Omega_p = (W_p, E_p)$, where $W_p = \{p_i\}, (i = 1, ..., N)$ is the set of N proteins and $E_p = \{\{p_i, p_j\}\}$ is the set of edges/connections between proteins. The *adjacency matrix* ξ_{ij} indicates that an interaction exists between proteins $p_i, p_j \in \Omega_p$ ($\xi_{ij} = 1$) or that the interaction is absent ($\xi_{ij} = 0$). Two connected proteins are thus called *adjacent* and the *degree* of a given protein is the number of edges that connect it with other proteins.

The SW pattern can be detected from the analysis of two basic statistical quantities: the *clustering coefficient* C_v and the *average path length* L. Let us consider the adjacency matrix and indicate by $\Gamma_i = \{p_i \,|\, \xi_{ij} = 1\}$ the set of nearest neighbors of a protein $p_i \in W_p$. The clustering coefficient for this protein is defined as the number of connections between the proteins $p_j \in \Gamma_i$ (Watts and Strogatz, 1998). Denoting

	Yeast proteome	Network model	Random network
$\bar{K}$	1.83	2.0 ± 0.1	2
$\bar{K}^g$	2.3	5.0 ± 0.5	2.0 ± 0.1
γ	2.5	2.5 ± 0.1	—
γ^g	2.5	2.5 ± 0.2	—
$\lvert\Omega_\infty\rvert$	466	360 ± 100	590 ± 73
C_v^g	2.2×10^{-2}	1.0×10^{-2}	3.1×10^{-3}
L^g	7.14	4.8 ± 0.1	7.4 ± 0.1

TABLE I. Comparison between the observed regularities in the yeast proteome (Wagner, 2001), the model predictions with $N = 10^3$, $\delta = 0.53$ and $\beta = 0.06$, and a random network with the same size and average connectivity as the model. The quantities X represent averages over the whole graph; X^g represent averages over the giant component.

$$\mathcal{L}_i = \sum_{j=1}^{N} \xi_{ij} \left[\sum_{k \in \Gamma_i} \xi_{jk} \right], \tag{7}$$

we define the clustering coefficient of the i-th protein as

$$c_v(i) = \frac{2\mathcal{L}_i}{k_i(k_i - 1))}, \tag{8}$$

where k_i is the connectivity of the i-th protein. The clustering coefficient is defined as the average of $c_v(i)$ over all the proteins,

$$C_v = \frac{1}{N} \sum_{i=1}^{N} c_v(i), \tag{9}$$

and it provides a measure of the average fraction of pairs of neighbors of a node that are also neighbors of each other.

The average path length L is defined as follows: Given two proteins $p_i, p_j \in W_p$, let $L_{min}(i, j)$ be the minimum path length connecting these two proteins in Ω_p. The average path length L will be:

$$L = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} L_{min}(i, j) \tag{10}$$

Random graphs, where nodes are randomly connected with a given probability p (Bollobás, 1985), have a clustering coefficient inversely proportional to the network size, $C_v^{\rm rand} \approx \bar{K}/N$, and an average path length proportional to the logarithm of the network size, $L^{\rm rand} \approx \log N / \log \bar{K}$. At the other extreme, regular lattices with only nearest-neighbor connections among units are typically clustered and exhibit long average paths. Graphs with SW structure are characterized by a high clustering with $C_v \gg C_v^{rand}$, while possessing an average path comparable with a random graph with the same connectivity and number of nodes.

In Table I we report the values of $\bar{K}$, γ, $|\Omega_\infty|$, C_v, and L for our model, compared with the values reported for

the yeast *S. cerevisiae* (Jeong et al., 2001; Wagner, 2001), and the values corresponding to a random graph with size and connectivity comparable with both the model and the real data. Except the average connectivity of the giant component, which is slightly larger for the model, all the magnitudes for the model compare quite well with the values measured for the yeast. On the other hand, the values obtained for a random graph support the conjecture of the SW properties of the protein network put forward in (Wagner, 2001).

IV. DISCUSSION

The analysis of complex biological networks in terms of random graphs is not new. Early work suggested that the understanding of some general principles of genome organization might be the result of emergent properties within random networks of interacting units (Kauffman, 1962; 1993). An important difference emerges, however, from the new results about highly heterogeneous networks: the topological organization of metabolic and protein graphs is very different from the one expected under totally random wiring and as a result of their heterogeneity, new qualitative phenomena emerge (such as the robustness against mutation). This supports the view that cellular functions are carried out by networks made up by many species of interacting molecules and that networks of interactions might be at least as important than the units themselves (Hartwell et al., 1999; Solé et al., 2000). It is worth mentioning that $\bar{K} \approx 2$ provides (in random graphs) a high diversity of cycles of all lengths and thus a potential source of high redundancy at a low cost in terms of wiring.

Our study has shown that the macroscopic features exhibited by the proteome are also present in our simple model. This is surprising, since it is obvious that different proteins and protein interactions play different roles and operate under very different time scales and our model lacks such specific properties, dynamics or explicit functionality. Using estimated rates of addition and deletion of protein interactions as well as the average connectivity of the yeast proteome, we accurately reproduce the available statistical regularities exhibited by the real proteome.

These results suggest that the global organization of protein interaction maps can be explained by means of a simple process of gene duplication plus diversification. These are indeed the mechanisms known to be operating in genome evolution (although the magnitude of the duplication event can be different). However, most of the classic literature within this area deal with the phylogenetic consequences of duplication and do not consider the underlying dynamics of interactions between genes. We can see, however, that the final topology has nontrivial consequences: this type of scale-free network will display an extraordinary robustness against random removal of nodes (Albert et al., 2000) and thus it can have a selective role. But an open question arises: is the scale-free organization a consequence of the pattern of duplication plus rewiring and thus we have "robustness for free"? The alternative is of course a fine-tuning of the process in which selection for robustness has been obtained by accepting or rejecting single changes. Further model approximations and molecular data might provide answers to these fundamental questions.

Acknowledgments

The authors thank J. Mittenthal, R. Ferrer, J. Montoya, S. Kauffman and A. Wuensche for useful discussions. This work has been supported by a grant PB97-0693 and by the Santa Fe Institute (RVS).

V. REFERENCES

1. Albert, R., Jeong, H. and Barabási, A-L. (2000) Error and attack tolerance of complex networks. Nature **406**, 378-382.

2. Amaral, L. A. N., Scala, A., Barthélemy, M. and Stanley, H. E. (2000) Classes of behavior of small-world networks. Proc. Nat. Acad. Sc. USA **97**, 11149-11152.

3. Bartel, P. L., Roecklein, J. A., SenGupta, D. and Fields, S. A. (1996) A protein linkage map of *Escherichia coli* bacteriohage T7. Nature Genet. **12**, 72-77.

4. Bollobás, B. (1985) *Random Graphs.* (Academic Press, London)

5. Bornholdt, S. and Sneppen, K. (2000) Robustness as an evolutionary principle. Proc. Roy. Soc. Lond. B**267**, 2281-2286.

6. Fell, D. and Wagner, A. (2000) The small world of metabolism. Nature Biotech. **18**, 1121.

7. Ferrer Cancho, R., Janssen, C. and Solé, R. V. (2001) The topology of technology graphs: small-world pattern in electronic circuits. Phys. Rev. E (in press).

8. Ferrer Cancho, R. and Solé, R. V. (2001) The small-world of human language. Procs. Roy. Soc. London B (in press).

9. Fields, S. (2001) Proteomics in genomeland. Science **409**, 861-921.

10. Flajolet, M. et al. (2000) A genomic approach to the hepatitis C virus generates a protein interaction map. Gene **242**, 369-379.

11. Fromont-Racine, M., Rain, J. C. and Legrain, P. (1997) Towards a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nature Genet. **16**, 277-282.

12. Hartwell, L. H., Hopfield, J. J., Leibler, S. and Murray, A. W. (1999) From molecular to modular cell biology. Nature **402**, c47-52.

13. Hazbun, T. R. and Fields, S. (2001) Networking proteins in yeast. Proc. Natl. Acad. Sci. USA **98**, 4277-4278.

14. Ito, T. et al. (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl. Acad. Sci. USA **97**, 1143-1147.

15. Jeong, H., Mason, S., Barabási, A. L. and Oltvai, Z. N. (2001) Lethality and centrality in protein networks Nature **411**, 41 (2001).

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

17. Kauffman, S. A. (1962) Metabolic stability and epigenesis in randomly connected nets. J. Theor. Biol. **22**, 437-467.

18. Kauffman, S. A. (1993) *Origins of Order*, Oxford, New York.

19. Lander, E. S. et al. (2001) Initial sequencing and analysis of the human genome. Nature **409**, 861-921.

20. Lloyd, A. and May, R. M. (2001) How viruses spread among computers and people. Science **292**, 1316-1317.

21. McCraith, S., Holtzman, T., Moss, B. and Fields, S. (2000) Genome-wide analysis of vaccinia virus protein-protein interactions. Proc. Natl. Acad. Sci. USA **97**, 4879-4884.

22. Mewes, H. W. et al. (1999) MIPS: a database for genomes and protein sequences. Nucleic Acids Res. **27**, 44-48.

23. Ohno, S. (1970) *Evolution by gene duplication.* Allen and Unwin, London.

24. Pastor-Satorras, R. and Vespignani, A. (2001) Epidemic spreading in scale-free networks. Phys. Rev. Lett. **86**, 066117.

25. Rain, J. -C. et al. (2001) The protein-protein interaction map of *Helicobacter pylori.* Nature **409**, 211-215.

26. Ross-Macdonald, P. et al. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature **402**, 413-418.

27. Solé, R. V., Salazar-Ciudad, I. and Newman, S. A. (2000) Gene network dynamics and the evolution of development. Trends Ecol. Evol. **15**, 479-480.

28. Strogatz, S. (2001) Exploring complex networks. Nature **410**, 268-276.

29. Thieffry, D., Huerta, A. M., Pérez-Rueda, E. and Collado-Vives, J. (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli.* BioEssays **20**, 433-440.

30. Vázquez, A,. Flammini, A., Maritan, A., and Vespigani, A. (2001) Modelling of protein interaction networks. e-print cond-mat/0108043.

31. Venter, J. C. et al. (2001) The sequence of the human genome. Science **291**, 1305.

32. Vogelstein, B., Lane, D. and Levine, A. J. (2000) Surfing the p53 network. Nature **408**, 307-310.

33. Wagner, A. (1994) Evolution of gene networks by gene duplications: A mathematical model and its implications on genome organization. Proc. Natl. Acad. Sci. USA **91**, 4387-4391.

34. Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. J. Molec. Evol. (in press).

35. Walhout, A. J. M. et al. (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. Science **287**, 116-122.

36. Watts, D. J. and Strogatz, S. H. (1998) Collective Dynamics in 'small-world' networks, *Nature (Lond.)* **393**, 440-442

37. Wolfe, K. H. and Shields, D. C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**, 708-713.

38. Wu, X., Bayle, J. H., Olson, D. and Levine, A. J. (1993) Genes Dev. **7**, 1126.