Topological and causal structure of the yeast transcriptional regulatory network

Nabil Guelzim[1,2], Samuele Bottani[3], Paul Bourgine[2] & François Képès[1]

Published online: 22 April 2002, DOI: 10.1038/ng873

Interpretation of high-throughput biological data requires a knowledge of the design principles underlying the networks that sustain cellular functions. Of particular importance is the genetic network, a set of genes that interact through directed transcriptional regulation. Genes that exert a regulatory role encode dedicated transcription factors (hereafter referred to as regulating proteins) that can bind to specific DNA control regions of regulated genes to activate or inhibit their transcription. Regulated genes may themselves act in a regulatory manner, in which case they participate in a causal pathway. Looping pathways form feedback circuits. Because a gene can have several connections, circuits and pathways may crosslink and thus represent connected components. We have created a graph of 909 genetically or biochemically established interactions among 491 yeast genes. The number of regulating proteins per regulated gene has a narrow distribution with an exponential decay. The number of regulated genes per regulating protein has a broader distribution with a decay resembling a power law. Assuming in computer-generated graphs that gene connections fulfill these distributions but are otherwise random, the local clustering of connections and the number of short feedback circuits are largely underestimated. This deviation from randomness probably reflects functional constraints that include biosynthetic cost, response delay and differentiative and homeostatic regulation.

In integrating genome-wide data on transcript abundance[1] into a dynamic view of gene networks, recent studies have focused on abstracting the principles that underlie the architecture and causal interplay of these networks. At present, the yeast *Saccharomyces cerevisiae* is the most suitable eukaryotic organism for achieving this goal, as much information about its transcriptional regulations has been accumulated[2,3]. Of roughly 6,000 yeast genes, 124 have been shown through genetic and biochemical experiments to encode regulating proteins that can influence the expression of specific genes[2]. These data were obtained from a previous review[2] and were validated and updated, until July 2001, by manual inspection of the websites of MIPS, SwissProt, Yeast Protein Database, *S. cerevisiae* Promoter Database and the *Saccharomyces* Genome Database (see Web Note A online). The elements of the general transcription initiation machinery were excluded from this study, although some have differential roles in transcription of large subsets of genes[3]. Some of the 124 regulatory genes transcriptionally control a set of 367 non-regulatory genes (Fig. 1) through 837 connections (see Web Table A online). Of the 124 regulatory genes, 52 interact with themselves or with other regulatory genes through 72 additional links (see Web Table A online). A transcriptional regulatory network can thus be represented as a graph where vertices are genes and directed edges denote activating or repressing effects on transcription. The graph of these 52 'interregulatory' genes comprises mainly several small disconnected components (Fig. 1).

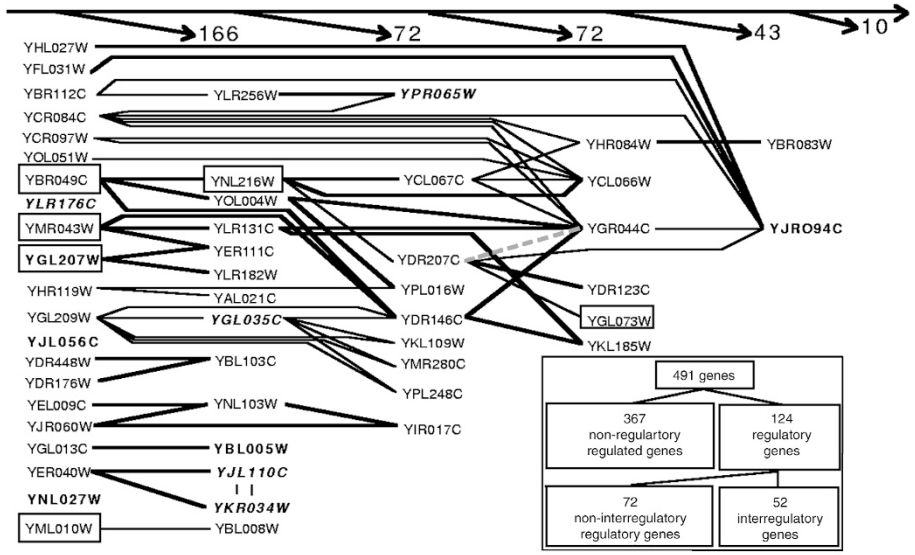

Fig. 1 Data set on the experimentally established genetic interactions in yeast. The graph shows causal relations among the 52 interregulatory genes. To indicate downstream causality (top arrow), genes lacking a known regulator other than mutual or self-regulation are listed in the left column. The other genes are then placed in the leftmost column so that all of their regulators locate to their left. Following the same rule, above are shown the numbers of nonregulatory genes (names omitted) regulated by genes from each column. From left to right, downstream causation emphasizes the consequences of altering a gene's activity for other genes. From right to left, upstream causation reveals the sources of a gene's perturbation. Bold type indicates self-activation, bold italics indicates self-inhibition and borders indicate essential genes. Thick lines represent activation, thin lines represent inhibition and the dashed gray line represents dual regulation.

[1]ATelier de Génomique Cognitive, Centre National de la Recherche Scientifique ESA 8071, genopole®, 523 Terrasses de l'Agora, 91000 Evry, France. [2]Centre de Recherche en Épistémologie Appliquée, École Polytechnique, Paris, France. [3]Groupe de Modélisation Physique Interfaces Biologie, UFR de Physique, Université Paris 7, Paris, France. Correspondence should be addressed to F.K. (e-mail: Francois.Kepes@genopole.cnrs.fr).

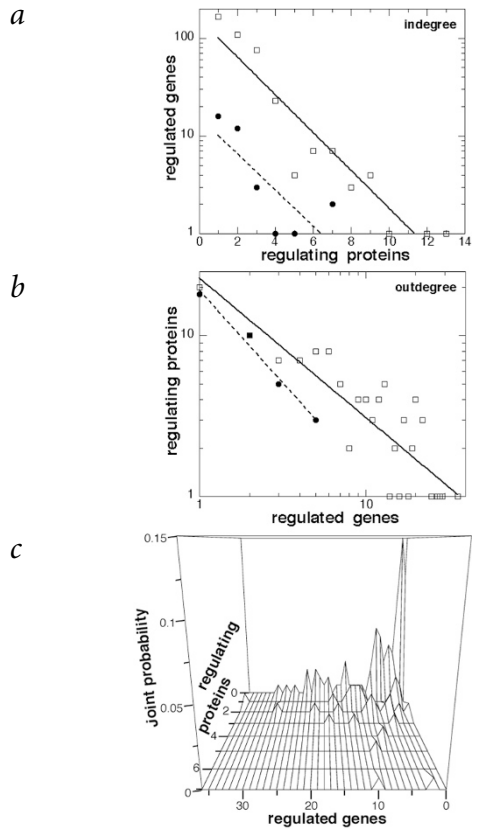

Fig. 2 Connectivity of the yeast genetic network. ***a***, Arriving connectivity distribution (semi-log plot). The number of regulating proteins per regulated gene follows an exponential distribution (least-square method) both for all 402 regulated genes (367 nonregulatory and 35 interregulatory genes—17 interregulatory genes are excluded because they lack a known regulator; 909 connections; open squares, full line; p_k=157e$^{-0.45k}$; R=0.99) and for the subset of 35 interregulatory genes (72 connections; filled circles, broken line; p_k=15e$^{-0.43k}$; R=0.94). ***b***, Departing connectivity distribution (log/log plot). The number of regulated genes per regulating protein better fits a power-law distribution both for all 124 regulating proteins (909 connections; open squares, full line; P_k=23$k^{-0.87}$; R=0.95) and for its subset of 37 regulating proteins that control regulatory genes (72 connections; filled circles, broken line; P_k=19$k^{-1.14}$; R=0.99). Opposite views (*a* versus *b*; exponential departing and power-law arriving connectivities) would give lower correlation coefficients (R=0.88, 0.91, 0.83 and 0.98, respectively) and very different slopes for global versus interregulatory genes. Null values were discarded. ***c***, Joint distributions. The probability that a randomly chosen gene has *in* arriving and *out* departing connections is distributed on this linear plot as a function of *in* (regulating proteins) and *out* (regulated genes).

Most networks fall into two major categories on the basis of their connectivity distribution, p_k, which represents the probability that a vertex in the network is connected to k other vertices. One category of networks is characterized by a p_k that peaks at an average k_{mean} and decays exponentially for large k[4,5]. In these exponential networks, most vertices have approximately the same number of links. By contrast, metabolic pathways[6–8] belong to a category of nonhomogeneous networks, where p_k decays as a power law. As the connections are inherently oriented in a transcriptional regulatory network, we separately analyzed the number of regulating proteins per regulated gene (arriving connectivity) and the number of regulated genes per regulating protein (departing connectivity), to determine whether they were best described by the exponential or power-law models.

The arriving connectivity of the yeast network has an exponential distribution, with 93% of the genes being regulated by 1–4 regulating proteins (Fig. 2*a*). The probability p_k that a given target gene is regulated by k regulating proteins decreases roughly as $Ce^{-\beta k}$ (C is a constant), with β~0.45 for both the total set of regulated genes and its interregulatory subset. The available data for *Escherichia coli*[9] are compatible with an exponential distribution of arriving connections, with β~1.2; this higher β coefficient means that fewer targets have many regulators. This coefficient thus reflects the molecular limits on the number of regulating proteins that can combinatorially exert an effect on the target gene expression. Consequently, lower coefficients are predicted for multicellular organisms with a more sophisticated genetic regulatory machinery.

The departing connectivity of the yeast network does not seem to be distributed according to an exponential law (Fig. 2*b*). It fits better a power law, although there are insufficient data to rule out other possibilities. The probability p_k that a given regulating protein regulates k target genes decreases as approximately $Ck^{-\gamma}$, with γ~1 for both the global set of 124 regulatory proteins and its interregulatory subset. For *E. coli* as well, γ~1 (our best fit computed from ref. 9; see also refs 8,10). Because γ~1, the number of departing connections (kp_k~kCk^{-1}=C) is distributed almost equally over k, unlike the connections present in metabolic networks (γ~1.5–3)[6–8]. Thus, bacterial and fungal genetic networks are free of a characteristic scale with respect to the distributions of both regulating proteins and departing connections.

The differing distribution laws for arriving and departing connectivities suggests that there is a correlation between them. A joint distribution (Fig. 2*c*) shows that genes with few regulators also tend to have few targets. Because there are many such genes, inactivating a gene selected at random has a low probability of altering the pathway structure of other genes. In contrast, inactivating one of the few highly connected genes would greatly decrease the communication between the remaining genes[11] and could be lethal. Of 124 regulatory genes, 10 are essential, including 6 interregulatory genes that tend to be located upstream in the causal graph (Fig. 1). Indeed, their overall influence (direct and indirect targets) is twice as big on average as that of nonessential genes.

To evaluate the generality of the predicted topology, two things must be determined: (i) to what extent the present compilation differs from a complete yeast data set and (ii) whether the observed global topology is likely to hold true as more data accumulate. On the basis of sequence homology, at most, 77 additional yeast genes encode putative regulating proteins (see Web Note A online); however, recent work has investigated the genome-wide locations of 12 DNA-binding proteins, using chromatin immunoprecipitation and microarrays[12–15]. Depending on the laboratory, the number of targets thus obtained is on average 3.5-fold[12,15] and 26-fold[13,14] greater than the number found here for the same regulators (see Web Table A online). Although the exact number of targets depends on a somewhat arbitrary threshold, it is already clear that this new method has the potential to reveal many unsuspected links[12–15]. It is therefore essential to re-evaluate the topology of the yeast network once a sufficient set of regulatory genes has been studied with this genome-wide approach and universal threshold definitions. Moreover, theoretical considerations, consistent with the comparison of a subset to the whole set (Fig. 2*a*,*b*), suggest a way in which future data may affect the described network structure. If departing connectivity is free of a characteristic scale, future data should presumably not alter the power-law parameters. If arriving connectivity is shaped by the sophistication of the regulatory machinery, additional data would probably increase C while maintaining β.

Table 1 • Structure of the yeast transcriptional regulatory network

	Connectivity distributions			
	(a) Actual data	(b) Empirical	(c) Expo/Power	(d) Poisson
Giant component %[a]	0	0	0	77
$\langle in_1 \rangle$ (=$\langle out_1 \rangle$)[b]	1.9	1.9	2.5	1.9
$\langle in_2 \rangle$ (=$\langle out_2 \rangle$)[c]	1.0	1.3	1.4	3.7
$\langle in_\infty \rangle$ (=$\langle out_\infty \rangle$)[d]	4.0	7.2	7.0	NA[h]
NΔ[e]	135	29	26	7
SC_{in}[f]	0.050	0.010	0.010	0.004
SC_{out}[f]	0.010	0.002	0.002	0.004
FC_1[g]	10	0.7	0.6	2
FC_∞[g]	11	2.2	1.4	NA[h]

[a]Percentage of genes in the largest connected component (with at least one oriented path between any couple of vertices). [b]Average number of regulating proteins (regulated genes) one step away. [c]Idem two steps away. [d]Idem at all distances (average component size). [e]Number of oriented triangular interactions. [f]SC_{in} (SC_{out}), upstream (downstream) semi-clustering coefficient. [g]FC_1 (FC_∞), number of feedback circuits comprising one (any number of) gene. [h]Not directly applicable, owing to the presence of a giant component in (d). The features of the actual genetic network (a) were compared with those of a directed random graph with an equal number of vertices and edges (see Web Note B online). Connectivity distributions were as empirically observed (b), or followed the exponential (power, respectively) law for arriving (departing) regulations that had been determined from Fig. 2 (c), or followed a Poisson law (d).

To assess how accurately various models represent the biological situation, the actual yeast genetic network (a) was compared with directed random graphs modeled under three assumptions (see Web Note B online and Fig. 2): the connectivity distributions conform with (b) the empirical data (c) the laws deduced in Fig. 2 and (d) a Poisson law. A uniformly distributed connectivity (d) favors the emergence of a connected component that comprises the majority of the genes (Table 1), which is not observed. By contrast, both random graphs with constrained connectivity distributions (b or c; Fig. 2) reasonably approximate the average number of neighbors one or two steps away. At a more refined grain, however, they are no longer acceptable approximations. The local attribution of a few edges per vertex in a sparse graph is an important parameter that affects the network dynamics. It could be uniform, as in random graphs[4], or highly clustered, as in small worlds[5]; extreme local clustering would result in global fragmentation, unlike small worlds, which still retain large connected components. Global fragmentation is observed (Fig. 1), beyond that expected from the empirical data or the deduced laws (b or c; Table 1). A clustering coefficient has been proposed to quantify the propensity of the links reaching an individual to involve him or her in local social interactions within 'cliques'[5]. Because genetic networks are directed, we introduce the notion of upstream or downstream 'semi-cliquishness' (see Web Note B online). The corresponding semi-clustering coefficients are approximately fivefold higher than those expected for the yeast network in a constrained random graph (Table 1). Along the same lines, the total number of observed feedback circuits is fivefold higher than that predicted by (b) or (c), and 14-fold higher for single-gene circuits (Table 1).

These circuits are crucial to the dynamics of the system. Positive circuits comprise an even number of inhibitory interactions and contribute to multistationarity, whereas negative circuits comprise an odd number of inhibitory interactions and contribute to homeostasis[16]. In this view, higher organisms are expected to rely more heavily than lower ones on positive circuits, particularly to achieve cellular differentiation, with each cell type corresponding to one of several stationary states. We observed five negative and six positive circuits in 52 yeast interregulatory genes (Fig. 1). As expected, this is in marked contrast to the genes of *E. coli*, where 45 circuits (39 negative, 3 positive, 3 dual) were observed for 55 interregulatory genes[9]. Yeast positive circuits control switching processes, such as those leading to pseudohyphal growth (*YJL110C*/*YKR034W*, controlled by *YER040W*)[17], sporulation (*YJR094C*)[18] or multiple-drug resistance (*YBL005W*)[19]. Negative circuits are constituted by (self-) inhibitors that finely control responses to the absence of glucose (*YGL035C*)[20], DNA damage (*YLR176C*)[21] or oxygen (*YPR065W*)[22].

As a whole, the yeast transcriptional regulatory network combines a small maximal diameter, an elevated local semi-clustering, a high number of feedback circuits and a global fragmentation. This departure from a random distribution must reflect functional constraints. Indeed, each small connected piece implements a biological function, and the global fragmentation may serve to limit inter-functional crosstalk at the transcriptional level. The elevated clustering and feedback content probably implement differentiative and homeostatic requirements. Single-gene feedback circuits are predominant (this study and ref. 9) and may have been selected through evolution for several reasons: (i) they decrease the biosynthetic cost (roughly proportional to the amount of transcripts and proteins to be produced), (ii) together with the small diameter, they reduce the response delay (often a consequence of macromolecular synthesis) and (iii) they stabilize the fluctuations of expression of the involved genes[23]. Similar laws seem to govern the local and global network topologies in eukaryotes and prokaryotes, notwithstanding the circuit sign. When prior knowledge of the specific transcriptional connections is limited, these laws may prove general enough to facilitate the integration of transcriptomic data into dynamic models of genetic networks.

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

Acknowledgments

We thank M.-H. Mucchielli for help with the statistical analysis and M. Gromov, V. Norris and B. Prum for critically reading the manuscript. This work was supported by funding from CNRS and Conseil Régional d´Ile de France.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 6 December 2001; accepted 18 March 2002.

1. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
2. Svetlov, V.V. & Cooper, T.G. Review: compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast* **11**, 1439–1484 (1995).
3. Holstege, F.C.P. *et al*. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
4. Erdös, P. & Rényi, A. On random graphs. *Publicationes Mathematicae* **6**, 290–297 (1959).
5. Watts, D.J. & Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
6. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
7. Fell, D.A. & Wagner, A. The small world of metabolism. *Nature Biotech.* **18**, 1121–1122 (2000).
8. Raine, D.J. & Norris, V. Network structure of metabolic pathways. *Interjournal Complex System* #361 (International Conference on Complex Systems, Nashua, New Hampshire, 21–26 May 2000).
9. Thieffry, D., Huerta, A.M., Pérez-Rueda, A. & Collado-Vides, J. From specific gene regulation to genomic networks: a global analysis to transcriptional regulation in *Escherichia coli*. *Bioessays* **20**, 433–440 (1998).
10. Karp, P.D. Pathway databases: a case study in computational symbolic theories. *Science* **293**, 2040–2044 (2001).

11. Albert, R., Jeong, H. & Barabasi, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
12. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
13. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
14. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genet.* **28**, 327–334 (2001).
15. Simon, I. *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697–708 (2001).
16. Thomas, R. & D'Ari, R. *Biological Feedback* (CRC, Boca Raton, 1990).
17. Lorenz, M.C. & Heitman, J. The MEP2 ammonium permease regulates pseudohyphal differentiation in *Saccharomyces cerevisiae. EMBO J.* **17**, 1236–1247 (1998).
18. Vershon, A.K. & Pierce, M. Transcriptional regulation of meiosis in yeast. *Curr. Opin. Cell Biol.* **12**, 334–339 (2000).
19. Rogers, B. *et al.* The pleiotropic drug ABC transporters from *Saccharomyces cerevisiae. J. Mol. Microbiol. Biotechnol.* **3**, 207–214 (2001).
20. Carlson, M. Glucose repression in yeast. *Curr. Opin. Microbiol.* **2**, 202–207 (1999).
21. Huang, M., Zhou, Z. & Elledge, S.J. The DNA replication and damage checkpoint pathways induce transcription by inhibition of the Crt1 repressor. *Cell* **94**, 595–605 (1998).
22. Zhang, L. & Hach, A. Molecular mechanisms of heme signaling in yeast: the transcriptional activator Hap1 serves as the key mediator. *Cell. Mol. Life. Sci.* **56**, 415–426 (1999).
23. Becsksei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).