Genomic analysis of regulatory network dynamics reveals large topological changes

Nicholas M. Luscombe[1]*, **M. Madan Babu**[4]*, **Haiyuan Yu**[1], **Michael Snyder**[2], **Sarah A. Teichmann**[4] **& Mark Gerstein**[1,3]

[1] *Department of Molecular Biophysics and Biochemistry,* [2] *Department of Molecular, Cellular and Developmental Biology,* [3] *Department of Computer Science, Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA*

[4] *Division of Structural Studies, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

* These authors contributed equally to this work

Network analysis has been applied widely, providing a unifying language to describe disparate systems ranging from social interactions to power grids. It has recently been used in molecular biology, but so far the resulting networks have only been analysed statically[1–8]. Here we present the dynamics of a biological network on a genomic scale, by integrating transcriptional regulatory information[9–11] and gene-expression data[12–16] for multiple conditions in *Saccharomyces cerevisiae*. We develop an approach for the statistical analysis of network dynamics, combining well-known global topological measures, local motifs and newly derived statistics. We uncover large changes in underlying network architecture that are unexpected given current viewpoints and random simulations. In response to diverse stimuli, transcription factors alter their interactions to varying degrees, thereby rewiring the network. A few transcription factors serve as permanent hubs, whereas most act transiently only during certain conditions. By studying sub-network structures, we show that environmental responses facilitate fast signal propagation (for example, with short regulatory cascades), whereas the cell cycle and sporulation direct temporal progression through multiple stages (for example, with highly inter-connected transcription factors). Indeed, to drive the latter processes forward, phase-specific transcription factors inter-regulate serially, and ubiquitously active transcription factors layer above them in a two-tiered hierarchy. We anticipate many of the concepts presented here—particularly large-scale topological changes and hub transience—will apply to other biological networks, including complex sub-systems in higher eukaryotes.

We began by assembling a static representation of known regulatory interactions from the results of genetic, biochemical and ChIP (chromatin immunoprecipitation)-chip experiments. Figure 1 illustrates the complexity of the resultant network, which contains 7,074 regulatory interactions between 142 transcription factors and 3,420 target genes (interactions can be between transcription factors and non-transcription factor targets, or two transcription factors). To get a dynamic perspective, we integrated gene-expression data for the following five conditions: cell cycle[13], sporulation[14], diauxic shift[12], DNA damage[16] and stress response[15]. From these data, we traced paths in the regulatory network that are active in each condition using a back-tracking algorithm (see Methods).

Figure 1b presents the sub-networks active under different cellular conditions, and gross changes are apparent in the distinct sections of the network that are highlighted. Recent functional genomics studies have analysed the dynamics of a few transcription factors[17,18]; however, Fig. 1 represents the first dynamic view of a genome-scale network.

Half of the targets are uniquely expressed in only one condition; in contrast, most transcription factors are used across multiple processes. The active sub-networks maintain or rewire regulatory interactions, and over half of the active interactions (1,476 of 2,476 total) are completely supplanted by new ones between conditions. Only 66 interactions are retained across four or more conditions; these comprise 'hot links'[6] that are always on (compared with the rest of the network) and mostly regulate house-keeping functions.

The large number of changing interactions makes rigorous comparison of active sub-networks impossible visually. Consequently, we introduce an approach called the 'statistical analysis of network dynamics' (SANDY) that combines: standard measures of network connectivity (involving global topological statistics[6] and local network motifs[4]), newly derived follow-on statistics, and comparisons against simulated controls to assess the significance of each observation.

Overall, our calculations divide the five condition-specific sub-networks into two categories: endogenous and exogenous (Fig. 1). This allows us to rationalize the different sub-network structures in terms of the biological requirements of each condition. Endogenous processes (cell cycle and sporulation) are multi-stage and operate with an internal transcriptional program, whereas exogenous states (diauxic shift, DNA damage and stress response) constitute binary events that react to external stimuli with a rapid turnover of expressed genes.

We begin SANDY by examining global topological measures that quantify network architecture (Fig. 1c)[6]. The view from recent studies is that these statistics are remarkably constant across many biological networks (including regulatory systems)[1,5,6,19,20]. Moreover, most of them remain invariant between randomly simulated sub-graphs of different sizes (Methods).

In fact, we show that topological measures change considerably between the endogenous and exogenous sub-networks. Furthermore, most of the observed measurements differ significantly from random expectation and are insensitive to addition of noise in the underlying network (Methods). The 'in-degree' (k_{in}) is the number of incoming edges per node (that is, the number of transcription factors regulating a target). Its average across each sub-network decreases by 20% from endogenous to exogenous conditions. (The probability, *P*, that these values originate from the same population is $<3 \times 10^{-4}$; Supplementary Information.) The 'out-degree' (k_{out}) represents the number of outgoing edges per node (that is, the number of target genes for each transcription factor). Average values double from endogenous to exogenous conditions ($P < 2 \times 10^{-3}$). The 'path length' (l) is the shortest distance between two nodes (here, it is the number of intermediate regulators between a transcription factor and a terminating target gene). Its average halves from endogenous to exogenous conditions ($P < 10^{-10}$). Finally, the 'clustering coefficient' (c) gauges the level of inter-connectivity around a node (that is, the level of transcription factor inter-regulation). Values range from 0 for totally dispersed nodes to 1 for fully connected ones. Average coefficients nearly halve from endogenous to exogenous conditions ($P < 0.01$).

In biological terms, the small in-degrees for exogenous conditions indicate that transcription factors are regulating in simpler combinations, and the large out-degrees signify that each transcription factor has greater regulatory influence by targeting more genes simultaneously. The short paths imply faster propagation of the regulatory signal. Conversely, long paths in the multi-stage, endogenous conditions suggest slower action arising from the formation of regulatory chains to control intermediate phases. Finally, high clustering coefficients in endogenous conditions signify greater inter-regulation between transcription factors. In summary, sub-networks have evolved to produce rapid, large-scale responses in exogenous states, and carefully coordinated processes in endogenous conditions (Fig. 1a).

SANDY also examines sub-networks locally by calculating the occurrence of network motifs[4], which are compact, specific patterns of inter-connection between transcription factors and targets. We show the occurrence of the most common motifs in Fig. 1c: single-input, multiple-input, and feed-forward-loop motifs (SIMs, MIMs, and FFLs). In SIMs a single transcription factor targets many genes; in MIMs multiple transcription factors co-regulate sets of genes; and in FFLs a primary transcription factor regulates a secondary one, and both target a final gene. Motifs appear at similar relative frequencies across regulatory networks of diverse organisms (although individual motifs are not conserved)[3,7], and this is also true for the randomly simulated sub-graphs. Therefore, constancy in motif usage is expected across conditions.

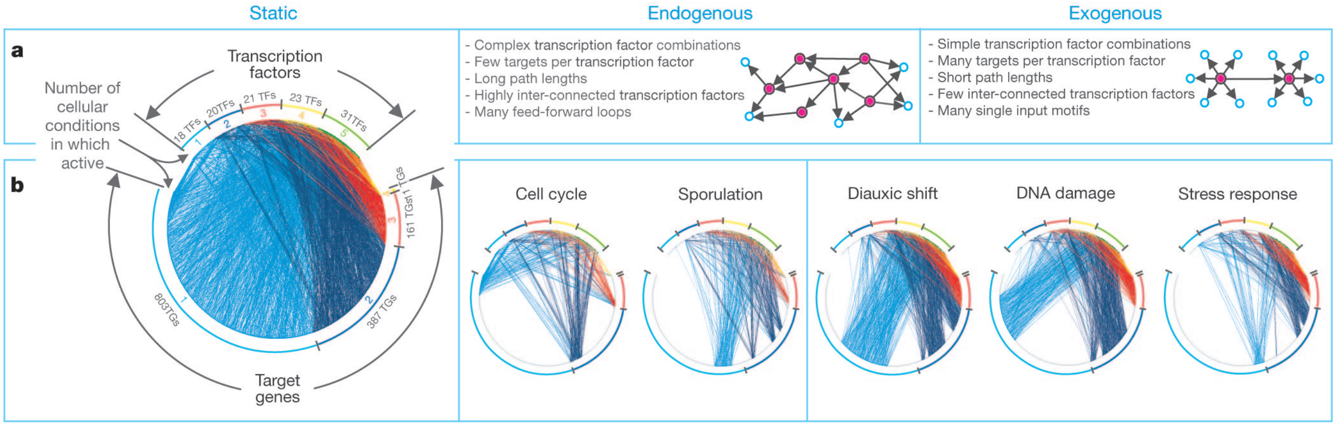
However, Fig. 1c shows that the relative occurrence of motifs varies considerably between endogenous and exogenous conditions ($P < 10^{-9}$). SIMs are favoured in exogenous sub-networks where they comprise >55% of regulatory interactions in motifs. But the frequency drops to ~35% in endogenous processes. Instead, these states favour FFLs (~44%). MIMs do not significantly change their usage.

Previous studies defined precise regulatory properties and information processing tasks for motifs[4]. SIMs and MIMs are implicated in conferring similar regulation over groups of genes, so they are ideal for directing the large-scale gene activation found in exogenous conditions. FFLs are buffers that respond only to persistent input signals. They are suited for endogenous conditions, as cells cannot initiate a new stage until the previous one has stabilized. Though used sparingly in exogenous processes they may be important in filtering spurious external stimuli.

Having quantified global and local changes with standard topological measures, we then moved to the follow-on statistics in SANDY (Fig. 2). Like many large-scale networks, the regulatory system displays scale-free characteristics (the probability P_k that a transcription factor targets k genes is proportional to $k^{-\gamma}$ for constant γ). This behaviour (maintained across all active sub-networks) signifies the presence of regulatory hubs targeting disproportionately large numbers of genes. Hubs are of general interest as they represent the most influential components of a network[6] and, accordingly, tend to be essential[21]. They are thought to target a broad spectrum of gene functions[4,11,22] and are commonly located upstream in the network[2] to expand their influence via secondary transcription factors[23]. These observations suggest that hubs would be invariant features of the network across conditions, and this expectation is supported by the random simulations that converge on similar sets of transcription factor hubs.

Figure 2a shows the observed regulatory hubs in each of the five conditions (Methods). They divide into two groups. The smaller one represents permanent hubs, which, in line with expectation, are important regardless of cellular state. They mainly comprise multifunctional transcription factors (such as Abf1) and house-keeping regulators (such as Mig1/2), and are responsible for maintaining the hot links. However, contrary to expectation, most hubs (78%) are transient; that is, they are influential in one condition, but less so in others. Exogenous conditions have fewer hubs, suggesting a more centralized command structure. (This is reflected in different γ; Supplementary Information.) About half of the transient hubs are known to be important for their respective conditions (for example, Swi4 in the cell cycle; Methods). For the remainder with sparse annotations, their transient-hub status in a particular condition considerably augments their functional annotation (for example, Sok2 in the cell cycle). These hubs may also relate to condition-dependent lethality, and this has clear implications for identifying specific drug targets.

The defining feature of transient hubs is their capacity to change interactions between conditions. We attempted to quantify this rewiring more broadly for every transcription factor in the network with the interchange index, I. This is defined so that higher values


Static
Endogenous
Exogenous
a
Transcription factors
Number of cellular conditions in which active
18 TFs
20TFs
21 TFs
23 TFs
31TFs
1
2
3
4
5
- Complex transcription factor combinations
- Few targets per transcription factor
- Long path lengths
- Highly inter-connected transcription factors
- Many feed-forward loops
- Simple transcription factor combinations
- Many targets per transcription factor
- Short path lengths
- Few inter-connected transcription factors
- Many single input motifs
b
803TGs
387 TGs
161 TGs
11 TGs
Target genes
Cell cycle
Sporulation
Diauxic shift
DNA damage
Stress response

c Size	No. of transcription factors	142	70	74	71	72	63
	No. of target genes	3,420	280	257	748	678	362
	No. of regulatory interactions	7,074	550	481	1,217	1,082	566
Topological measures	In-degree (k_{in})	2.1	2.0	1.9	1.6	1.6	1.6
	Out-degree (k_{out})	49.8	7.9	6.5	17.1	15.0	9.0
	Path length (l)	4.7	4.5	3.4	2.1	2.0	2.2
	Clustering coefficient (c)	0.11	0.15	0.14	0.09	0.09	0.08
Motifs (%)	Single input (SIM)	1,748 (37.6%)	130 (32.0%)	117 (38.9%)	438 (57.4%)	462 (55.7%)	228 (59.1%)
	Multiple input (MIM)	325 (7.0%)	96 (23.7%)	50 (16.6%)	180 (23.6%)	226 (27.3%)	78 (20.2%)
	Feed-forward loop (FFL)	2,581 (55.5%)	180 (44.3%)	134 (44.5%)	145 (19.0%)	141 (17.0%)	80 (20.7%)
	Total	4,654	406	301	763	829	386

Figure 1 Dynamic representation of the transcriptional regulatory network and standard statistics. **a**, Schematics and summary of properties for the endogenous and exogenous sub-networks. **b**, Graphs of the static and condition-specific networks. Transcription factors and target genes are shown as nodes in the upper and lower sections of each graph respectively, and regulatory interactions are drawn as edges; they are coloured by the number of conditions in which they are active. Different conditions use distinct sections of the network. **c**, Standard statistics (global topological measures and local network motifs) describing network structures. These vary between endogenous and exogenous conditions; those that are high compared with other conditions are shaded. (Note, the graph for the static state displays only sections that are active in at least one condition, but the table provides statistics for the entire network including inactive regions.)

associate with transcription factors replacing a larger fraction of their interactions. Its histogram reveals a uni-modal central distribution with two groups of extreme outliers (Fig. 2b). At one extreme ($I \leq 10\%$), 12 transcription factors retain all interactions across multiple states. At the other end ($I \geq 90\%$), 27 transcription factors replace all interactions in switching conditions. Many of these are so extreme that they only regulate genes in a single condition and are inactive otherwise. These include six transient hubs of known importance for the cell cycle and stress response. Most transcription factors interchange only part of their interactions ($10\% < I < 90\%$). This group comprises most of the hubs; surprisingly, permanent hubs interchange interactions as often as transient ones, but over a larger number of conditions. Furthermore, transcription factors in this group often regulate genes of distinct functions in different conditions, thus shifting regulatory roles. For example, the permanent hub Abf1 regulates cell growth during endogenous conditions, but refocuses to intracellular transport during stress response (in addition to its maintained core functions).

The rewiring highlighted by the interchange index allows transcription factors to be active in many conditions. Indeed, 95 of the 142 transcription factors are used in more than one process (Fig. 1b). Specifically, within endogenous conditions 53 of 92 transcription factors overlap between cell cycle and sporulation (Fig. 2c), and there is a similar overlap for exogenous conditions (Supplementary Information). With so much intersection in the repertoire of active transcription factors, the precise regulation of a condition cannot arise from the specificity of individual transcription factors. As others have observed[24], combinatorial transcription factor usage

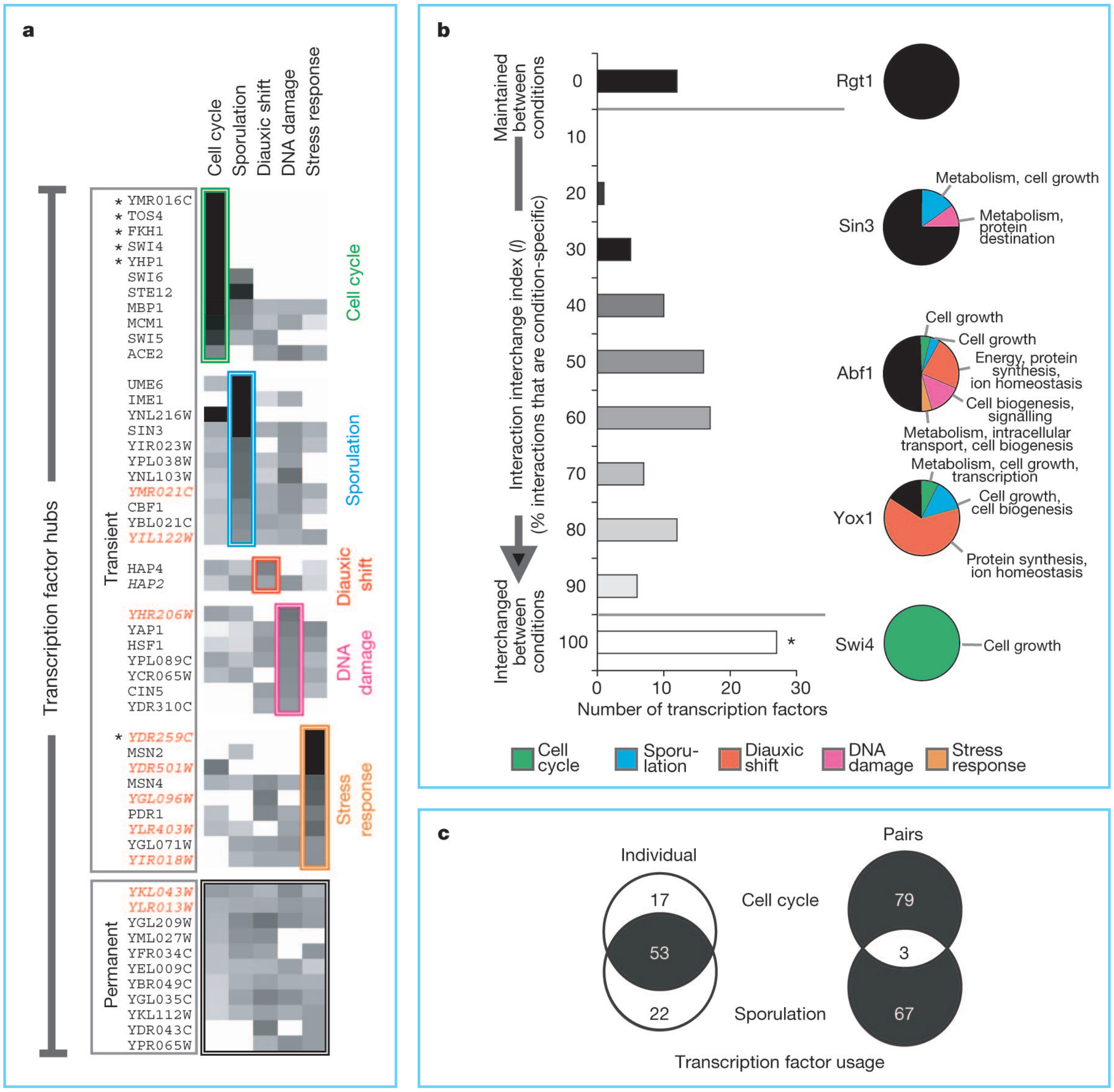


Figure 2 Derived 'follow-on' statistics for network structures. **a**, Transcription factor hub usage in different cellular conditions. The cluster diagram shades cells by the normalized number of genes targeted by transcription factor hubs in each condition. One cluster represents permanent hubs and the others condition-specific transient hubs. Genes are labelled with four-letter names when they have an obvious functional role in the condition, and seven-letter open reading frame names when there is no obvious role. Of the latter, gene names are red and italicised when functions are poorly characterized. Starred hubs show extreme interchange index values, $I = 1$. **b**, Interaction interchange (I) of transcription factors between conditions. A histogram of I for all active transcription factors shows a uni-modal distribution with two extremes. Pie charts show five example transcription factors with different proportions of interchanged interactions. We list the main functions of the distinct target genes regulated by each example transcription factor. Note how the transcription factors' regulatory functions change between conditions. **c**, Overlap in transcription factor usage between conditions. Venn diagrams show the numbers of individual transcription factors (large intersection) and pair-wise transcription factor combinations (small intersection) that overlap between the two endogenous conditions.

seems to be the key. We calculated that there are 360 unique pairwise transcription factor combinations (that is, two transcription factors regulating the same target) used in at least one condition. In contrast to individual transcription factors, only a minor proportion of pairs (51 of 360) participate in multiple processes and just 3 of 149 pairs overlap between endogenous conditions (Fig. 2c).

Thus far we have focused on the large dynamic changes occurring between different cellular conditions. However, dynamic transitions also take place within individual processes. Earlier, SANDY defined endogenous sub-networks by their long paths and high clustering. We can study the source of these observations by looking at the full scope of inter-regulation between transcription factors during the cell cycle (Fig. 3). This is possible because in a previous study[13] expression-level measurements were made throughout the cell cycle and differentially expressed genes were assigned to one of five phases (early G1, late G1, S, G2 and M). We then back-tracked from the classified genes to identify active sub-networks during each phase (Methods).

A cluster diagram (Fig. 3a) shows that most transcription factors that are active in the cell cycle operate only in a particular phase (for example, Swi4 in late G1). Additionally, a sizeable minority of transcription factors is ubiquitously active throughout the whole cycle. We uncover two major forms of transcription factor inter-regulation. In serial inter-regulation[25] (Fig. 3b), the phase-specific transcription factors regulate each other in a sequential manner to drive the cell cycle forward. In fact, we detect complete loops of interactions within the complex circuitry, and the resulting regulatory cascades undoubtedly create the long paths. We also introduce the concept of parallel inter-regulation (Fig. 3c), in which the ubiquitous transcription factors control the phase-specific ones in a two-tiered system. This effectively provides a stable signal to aid the transition between phases. Furthermore, because about a third of ubiquitous transcription factors comprise permanent hubs, they may provide a channel of communication to relate the cell-cycle progression with house-keeping functions. Similar observations apply to sporulation (Supplementary Information).

SANDY presents an approach to examining biological network dynamics. In applying it to the yeast regulatory system, it becomes apparent that many observations made in the static state are not applicable to the condition-specific sub-networks. However, in refocusing to a dynamic perspective, we uncover substantial topological changes in network structure, and we capture the essence of the transcriptional regulatory data in a new way. Because of limitations in current data sets, we can examine this only through integrating gene-expression information. However, we anticipate future experiments to determine condition specific interactions directly. Given the robustness of the observations to large perturbations (Methods), we expect our approach and findings to remain valid for these new data sets. Furthermore, we anticipate that many of the concepts we have introduced could be readily transferred to other types of biological networks and complex sub-systems in multicellular organisms, such as those directing the circadian cycle[26] and cellular development. □

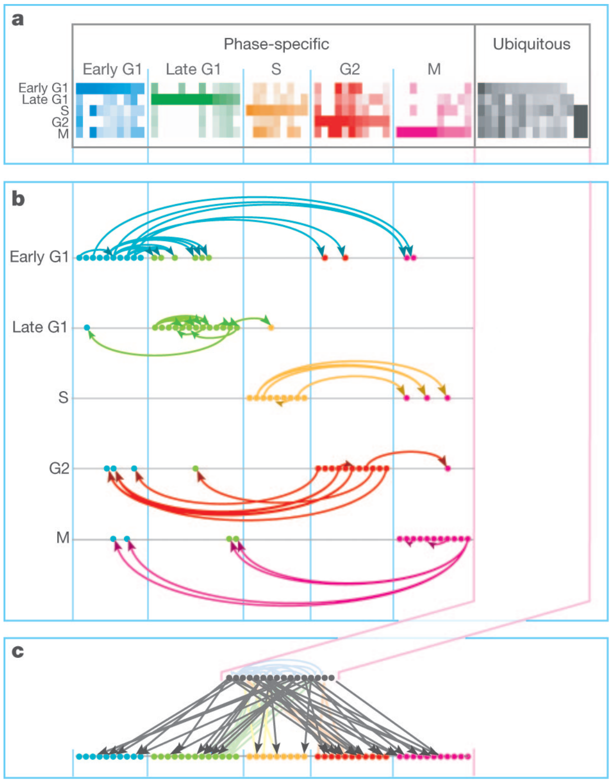


Figure 3 Transcription factor inter-regulation during the cell cycle. **a**, The 70 transcription factors active in the cell cycle. The diagram shades each cell by the normalized number of genes targeted by each transcription factor in a phase. Five clusters represent phase-specific transcription factors and one cluster is for ubiquitously active transcription factors. Transcription factor names are given in the Supplementary Information. **b**, Serial inter-regulation between phase-specific transcription factors. Network diagrams show transcription factors that are active in one phase regulate transcription factors in subsequent phases. In the late phases, transcription factors apparently regulate those in the next cycle. **c**, Parallel inter-regulation between phase-specific and ubiquitous transcription factors in a two-tiered hierarchy. Serial and parallel inter-regulation operate in tandem to drive the cell cycle while balancing it with basic house-keeping processes.

Methods

Detailed descriptions of the methods are in the Supplementary Information and at http://sandy.topnet.gersteinlab.org.

Data sets

The transcriptional regulatory network was assembled from the results of genetic, biochemical and ChIP-chip experiments[9–11]. The gene-expression data were compiled from 240 microarray experiments for five conditions[12–16]. We identified the following numbers of genes with differential expression: cell cycle, 455; sporulation, 477; diauxic shift, 1,823; DNA damage, 1,718; and stress response, 866.

Back-tracking algorithm

We used the following algorithm to define sections of the regulatory network used in each condition: (1) identify transcription factors as being ‘present’ in a condition if they have sufficiently high expression levels; (2) flag differentially expressed genes that appear in the regulatory network; (3) mark as ‘active’ the regulatory links between present transcription factors and differentially expressed genes; and (4) search for any other present transcription factors that are linked to a transcription factor with an already active link and make this connection active. The last step is repeated until no more links are made active. The same procedure identifies sub-networks that are active in particular phases of cell cycle and sporulation.

SANDY

This extends the methodology used by the TopNet software tool[27] and it evaluates each sub-network with the following: (1) standard statistics, including global measures of topology (k_{in}, k_{out}, l, and c)[6] and local motif occurrence (SIM, MIM and FFL)[4]; (2) follow-on statistics, including permanent and transient hub identification, interchange index (I) and counting the overlap in transcription factor usage (individual and pairs) across multiple conditions. (Hubs are transcription factors in the top 30%, by number of target genes, in at least one condition. The number of target genes is normalized to measure the relative influence of a transcription factor hub in a particular process.) In all cases, regulatory functions are obtained from the Saccharomyces Genome Database[28] and are current as of June 2004; and (3) a comparison of observations and random expectation by simulating sub-graphs that are similar in size to each sub-network, and calculating standard and follow-on statistics for them. Simulated sub-graphs sample the same number of differentially expressed genes and back-track through the static network. We also tested the sensitivity of our observations to noise by randomly perturbing the static networks by 30% (random addition, deletion and replacement of interactions), back-tracking from the original differentially expressed genes and then recalculating the statistics.

Received 15 January; accepted 24 June 2004; doi:10.1038/nature02782.

1. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **407,** 651–654 (2000).
2. Guelzim, N., Bottani, S., Bourgine, P. & Kepes, F. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genet.* **31,** 60–63 (2002).

3. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298,** 824–827 (2002).
4. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.* **31,** 64–68 (2002).
5. Oltvai, Z. N. & Barabasi, A. L. Systems biology. Life's complexity pyramid. *Science* **298,** 763–764 (2002).
6. Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5,** 101–113 (2004).
7. Milo, R. *et al.* Superfamilies of evolved and designed networks. *Science* **303,** 1538–1542 (2004).
8. Teichmann, S. A. & Babu, M. M. Gene regulatory network growth by duplication. *Nature Genet.* **36,** 492–496 (2004).
9. Svetlov, V. V. & Cooper, T. G. Review: compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast* **11,** 1439–1484 (1995).
10. Horak, C. E. *et al.* Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.* **16,** 3017–3033 (2002).
11. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298,** 799–804 (2002).
12. 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).
13. Cho, R. J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2,** 65–73 (1998).
14. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705 (1998).
15. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257 (2000).
16. Gasch, A. P. *et al.* Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell* **12,** 2987–3003 (2001).
17. Odom, D. T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303,** 1378–1381 (2004).
18. Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113,** 395–404 (2003).
19. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393,** 440–442 (1998).
20. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* **268,** 1803–1810 (2001).
21. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. & Gerstein, M. Genomic analysis of essentiality within protein networks. *Trends Genet.* **20,** 227–231 (2004).
22. Martinez-Antonio, A. & Collado-Vides, J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.* **6,** 82–489 (2003).
23. Madan Babu, M. & Teichmann, S. A. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* **31,** 1234–1244 (2003).
24. Pilpel, Y., Sudarsanam, P. & Church, G. M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.* **29,** 153–159 (2001).
25. Simon, I. *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106,** 697–708 (2001).
26. Ueda, H. R. *et al.* A transcription factor response element for gene expression during circadian night. *Nature* **418,** 534–539 (2002).
27. Yu, H., Zhu, X., Greenbaum, D., Karro, J. & Gerstein, M. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res.* **32,** 328–337 (2004).
28. Christie, K. R. *et al.* Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.* **32,** D311–D314 (2004).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank P. Bertone, N. Domedel-Puig, E. Hovig, R. Jansen, K. Kleivi, G. Koentges, E. Koonin, B. Lenhard, A. Paccanaro, J. Rozowsky, J. Tegner and A. Todd for comments on the paper. N.M.L. thanks the Anna Fuller Fund and the MRC LMB Visitor's Program. M.M.B. acknowledges financial support from the Cambridge Commonwealth Trust, Trinity College, Cambridge and the MRC LMB. M.G. is supported by the NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.M.L., M.M.B., S.A.T. & M.G. (sandy@bioinfo.mbb.yale.edu).

Author Queries

JOB NUMBER: 2782

JOURNAL: Nature

Table count = 0 *Figure count* = 3

Q1 AUTHOR: Please check that the display items are as follows (doi:10.1038/nature02782): Figures 1, 2, 3 (colour); Tables: None; Boxes: None. Please check all figures (and tables if any) very carefully as they have been relabelled, re-sized and adjusted to Nature's style. Please also make sure that any error bars are defined (as s.e.m. or s.d. etc).Article classification: broad area 1, 15.

Author's corrections – Page 1	*Author's corrections – Page 2*
Author's corrections – Page 3	*Author's corrections – Page 4*

Author's corrections – Page 5	*Author's corrections – Page 6*
Author's corrections – Page 7	*Author's corrections – Page 8*