Genomic analysis of the hierarchical structure of regulatory networks

Haiyuan Yu* and Mark Gerstein*

Departments of Molecular Biophysics and Biochemistry and Computer Science and Program in Computational Biology and Bioinformatics, Yale University, 266 Whitney Avenue, P.O. Box 208114, New Haven, CT 06520

Edited by Samuel Karlin, Stanford University, Stanford, CA, and approved July 10, 2006 (received for review October 4, 2005)

A fundamental question in biology is how the cell uses transcription factors (TFs) to coordinate the expression of thousands of genes in response to various stimuli. The relationships between TFs and their target genes can be modeled in terms of directed regulatory networks. These relationships, in turn, can be readily compared with commonplace ''chain-of-command'' structures in social networks, which have characteristic hierarchical layouts. Here, we develop algorithms for identifying generalized hierarchies (allowing for various loop structures) and use these approaches to illuminate extensive pyramid-shaped hierarchical structures existing in the regulatory networks of representative prokaryotes (*Escherichia coli*) and eukaryotes (*Saccharomyces cerevisiae*), with most TFs at the bottom levels and only a few master TFs on top. These masters are situated near the center of the protein–protein interaction network, a different type of network from the regulatory one, and they receive most of the input for the whole regulatory hierarchy through protein interactions. Moreover, they have maximal influence over other genes, in terms of affecting expression-level changes. Surprisingly, however, TFs at the bottom of the regulatory hierarchy are more essential to the viability of the cell. Finally, one might think master TFs achieve their wide influence through directly regulating many targets, but TFs with most direct targets are in the middle of the hierarchy. We find, in fact, that these midlevel TFs are "control bottlenecks" in the hierarchy, and this great degree of control for "middle managers" has parallels in efficient social structures in various corporate and governmental settings.

organization | topology | transcriptional regulation | yeast

Many biological processes can be modeled as networks, such as protein interaction, gene expression, and transcriptional regulatory networks (1–4). Networks have been used as a universal framework to model many complex systems, such as social interactions, the Internet, and ecological food webs (5–7). Individual networks have been globally characterized by a variety of graph-theoretic statistics, such as degree distribution, clustering coefficient (C), characteristic path length (L), and diameter (D) (3, 5–12). Recently, Barabasi and colleagues (7, 8) proposed a "scale-free" model in which most of the nodes have very few links, with only a few of them (hubs) being highly connected. Concurrently, Watts and Strogatz (12) found that many networks can also be described as having a "small-world" property, i.e., they are highly clustered and have small characteristic path lengths. Complex networks can be further divided into two broad categories: directed and undirected. The edges of the directed networks have a defined direction.

Previously, researchers have compared protein–protein interaction networks with social communication networks and found that protein networks share some common characteristics with them, such as scale-free and small-world properties (3, 9). However, researchers have yet to do this comparison with regulatory networks. Of all biological networks, regulatory networks are of particular interest, because to some degree they act as the master control system for the cell, tightly coordinating the expression of all genes (13–15). From a graph-theoretical point of view, regulatory networks are different from interaction networks in that they are directed. Both of these facts suggest that regulatory networks should be compared with a different type of social network, such as governmental and corporate organizations that are more oriented toward control than communication. These organizations are known to have hierarchical layouts with different levels: The stereotypical example would be a corporation with managers who supervise workers (16) (see Fig. 1).

Social hierarchical networks are often very complicated, containing many network motifs. Motifs are defined as overrepresented local network patterns (1). Four common ones in social hierarchies are shown in Fig. 1 and described below.

1. Single-input motifs (SIM), where a group of nodes (i.e., workers) are only regulated by a single node (i.e., manager).
2. Multi-input motifs (MIM), where a group of nodes together regulate another group of nodes.
3. Feed-forward loop (FFL), where a node regulates another; then, the two together regulate a third one.
4. Feed-back loop (also known as multicomponent loop; MCL), where an upstream node is regulated by a downstream one.

What makes a hierarchical structure special is that there are central control points at the top. Whether such a hierarchical structure exists in biological regulatory networks is not currently obvious. Here, we examine regulatory networks in both eukaryotes (*Saccharomyces cerevisiae*) and prokaryotes (*Escherichia coli*). We show that regulatory networks do indeed have a pyramid-shaped hierarchical structure that relates to their social counterparts. By doing so, we also identify central transcription factors (TFs) in both organisms that are on the top of the hierarchies.

Results

Building Generalized Hierarchies by Using Breadth-First Search (BFS). A simple hierarchy in a strict mathematical sense requires that the network contain no loops (i.e., it is "tree-like") (17). However, even though the concept of a simple hierarchy originally came from social studies, it is rather difficult to apply this notion to real social and biological networks, because both types of networks do indeed have prominent loops (Fig. 1*A*). In a more general sense, a hierarchy just refers to a pyramidal layered or ranked structure organized as those in social networks with few people at the top (managers) and most people at the bottom (workers). Consequently, for this study we want to create a precise construction of

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, ''Frontiers in Bioinformatics: Unsolved Problems and Challenges,'' held October 15–17, 2004, at the Arnold and Mabel Beckman Center of the National Academies of Science and Engineering in Irvine, CA. Papers from this Colloquium will be available as a collection on the PNAS web site. See the introduction to this Colloquium on page 13355 in issue 38 of volume 102. The complete program is available on the NAS web site at www.nasonline.org/bioinformatics.

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: TF, transcription factor; BFS, breadth-first search; BFS-level, BFS to define level.

*To whom correspondence may be addressed. E-mail: mark.gerstein@yale.edu or haiyuan.yu@yale.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0508637103

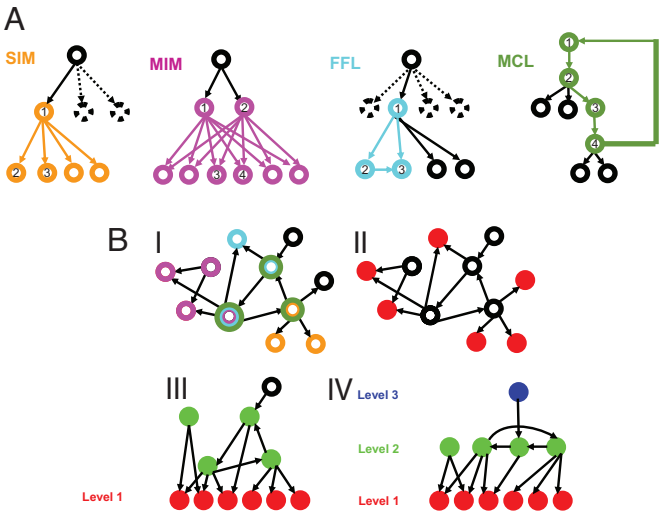

Fig. 1. Illustration of network motifs and the BFS-level method. (*A*) Four common network motifs in social networks. Different colors represent different motifs. All four schematics came from real social networks shown in Fig. 17, which is published as supporting information on the PNAS web site. (*I*) Single-input motifs (SIM). For example, node 1 is a professor or a director, and nodes 2 and 3 are his/her students or assistants, respectively. In the yeast regulatory network, node 1 is *NDD1*, and nodes 2 and 3 are *STB5* and *MCM21*, whose only regulator is *NDD1*. (*II*) Multi-input motifs (MIM). Nodes 1 and 2 can be professors, and nodes 3 and 4 can be two students that they coadvise. In Fig. 17*B*, nodes 1 and 2 are Senior Director and Executive Director, and nodes 3 and 4 are different departments that they cosupervise. In the yeast regulatory network, nodes 1 and 2 are *FKH1* and *FKH2*. Together, they regulate node 3 (*DBF2*) and node 4 (*HDR1*). (*III*) Feed-forward loop (FFL). For example, node 1 is the chairman of a department, node 2 is a professor in the department, and node 3 is a shared secretary. In yeast regulatory network, node 1 (*MBP1*) regulates node 2 (*SWI4*). Then, they collectively regulate node 3 (*SPT21*). (*IV*) Multicomponent loops (MCL). In Fig. 17*D*, node 1 is a chairman, node 2 is a director, node 3 is a coordinator, and node 4 is a scientist. Then some of the scientists form an advisory committee that oversees the chairman. In yeast regulatory network, node 1 is *REB1*, node 2 is *SIN3*, node 3 is *UME6*, and node 4 is *HSF1*. (*B*) Illustration on how to determine a generalized hierarchy using our BFS-level method. (*I*) A toy example with all four motifs mentioned in *A*. Each color represents a motif (color coding is the same as in *A*). (*II*) Finding all of the bottom (terminal) nodes in the network. A TF is a bottom node if and only if it does not regulate other TFs. TFs that only regulate themselves (i.e., autoregulation) are also considered as bottom nodes. All bottom nodes in the network are colored red. (*III*) Finding midlevel nodes. One performs a one-level deep BFS search starting at each of the bottom nodes to find what regulates them. Direct regulators of all bottom nodes are considered as level-2 nodes, which are in green. (*IV*) Finding topmost nodes. The procedure in the previous step (*III*) is repeated until all levels are determined. We call this overall process BFS-level. In this toy example, there are only three levels, and the node at the top level is in blue. However, in the yeast regulatory network, there are four levels.

"generalized hierarchies" that matches our social intuition and allows for loops. In essence, we assign a level number to each TF in the regulatory network to determine which TFs are at the top and which are at the bottom.

We call this construction method BFS to define level ("BFS-level"). As shown in Fig. 1*B*, it is based on a straightforward application of BFS: We first identified all TFs at the bottom level (i.e., level 1). A TF is at the bottom level if and only if it does not regulate other TFs. TFs that only regulate themselves (i.e., autoregulation) are also placed at the bottom. Starting from each bottom TF, we then performed a BFS to convert the whole network into a "breadth-first tree" (18) (see Fig. 2*A* and Table 1). In other words, we define the level of a nonbottom TF in the hierarchy as its shortest distance from a bottom one. Here, the construction procedure is only focused on interregulation between TFs (or officials in social networks). A top TF could directly regulate non-TF target genes (or a higher-ranked official could have an assistant with no managerial responsibility), but this structure will not affect the constructed hierarchy. If the resulted layered structure has a pyramidal shape (i.e., few nodes at the top and most nodes at the bottom), we then considered it as a generalized hierarchy.

Note a few features about this construction.

1. It is mathematically precise. There is only one unique solution for a given network, and a node is unambiguously placed at a single level.
2. It subsumes simple hierarchies. If a network does not contain loops, the BFS-level method would assign levels to nodes according to the perfect simple hierarchy of the network.
3. It does not change the network topology or connections (i.e., it does not "amputate" the network). In particular, it preserves all loops and takes into account all connections in assigning level.
4. It makes biological and social sense in that it builds from the ground up. One could imagine doing a similar BFS from the top down (see *Supporting Text* and Figs. 7–14, which are published as supporting information on the PNAS web site). However, we believe that this approach does not match our social intuition (e.g., putting the owner of a small business at the same level of hierarchy as the president of a country).
5. It is not trivial to construct a hierarchy for any given directed network. There are a number of possible variations as discussed below and in *Supporting Text*.

Pyramidal Regulatory Hierarchies and Their Nonmonotonic Out-Degree Distributions. Fig. 2*A* and Table 1 clearly show that the yeast regulatory network has a four-layer pyramid-shaped hierarchical structure; i.e., the number of TFs on each level is smaller than that of the previous level. A similar pyramidal hierarchy was also observed in *E. coli* (see Fig. 2*C* and Table 2, which is published as supporting information on the PNAS web site).

This hierarchical structure is actually very similar to that in social networks. Fig. 2*B* shows a representative social hierarchy: the Macao government. (This example was chosen because, although it is realistic, it is sufficiently simple to represent on a single page.) In Fig. 2*B*, there is only one chief executive (i.e., the president). Five secretaries are at the level immediately below the chief executive. There is a clear inverse relationship between the level in the hierarchy and the number of people at each level.

Intuitively, one might expect that the out-degree distribution at each level should parallel the pyramidal structure of hierarchy. For instance, it could increase uniformly as one goes from the bottom to the top, because, as one goes up, there is more to regulate. However, this possibility is not the case for social hierarchies. It has been shown that a typical organization scheme for companies is that middle managers supervise the most people, not those at the bottom or top of the hierarchies (16), as illustrated by Fig. 2*B*.

We then examined the average number of targets for TFs at different levels of the regulatory hierarchies for both *S. cerevisiae* and *E. coli*. We found the same relationship, i.e., TFs at the second level have the most targets, whereas those at the bottom and higher levels all have fewer targets by and large (see Fig. 2 *A* and *C*).

We also tested the robustness of our results by adding, deleting, or rearranging 20% of the regulatory interactions at random. All results remained the same, suggesting that the global conclusions from our calculations would be largely unaffected by noise in the data sets (see *Supporting Text*). It is also noteworthy that there might be hidden organizational structures because there are many within-level regulations, which is a possible direction for future analysis.

Bottlenecks of the Hierarchies Lie in the Middle. Fig. 2*A* and *C* clearly shows that the regulatory information in the hierarchies is passed

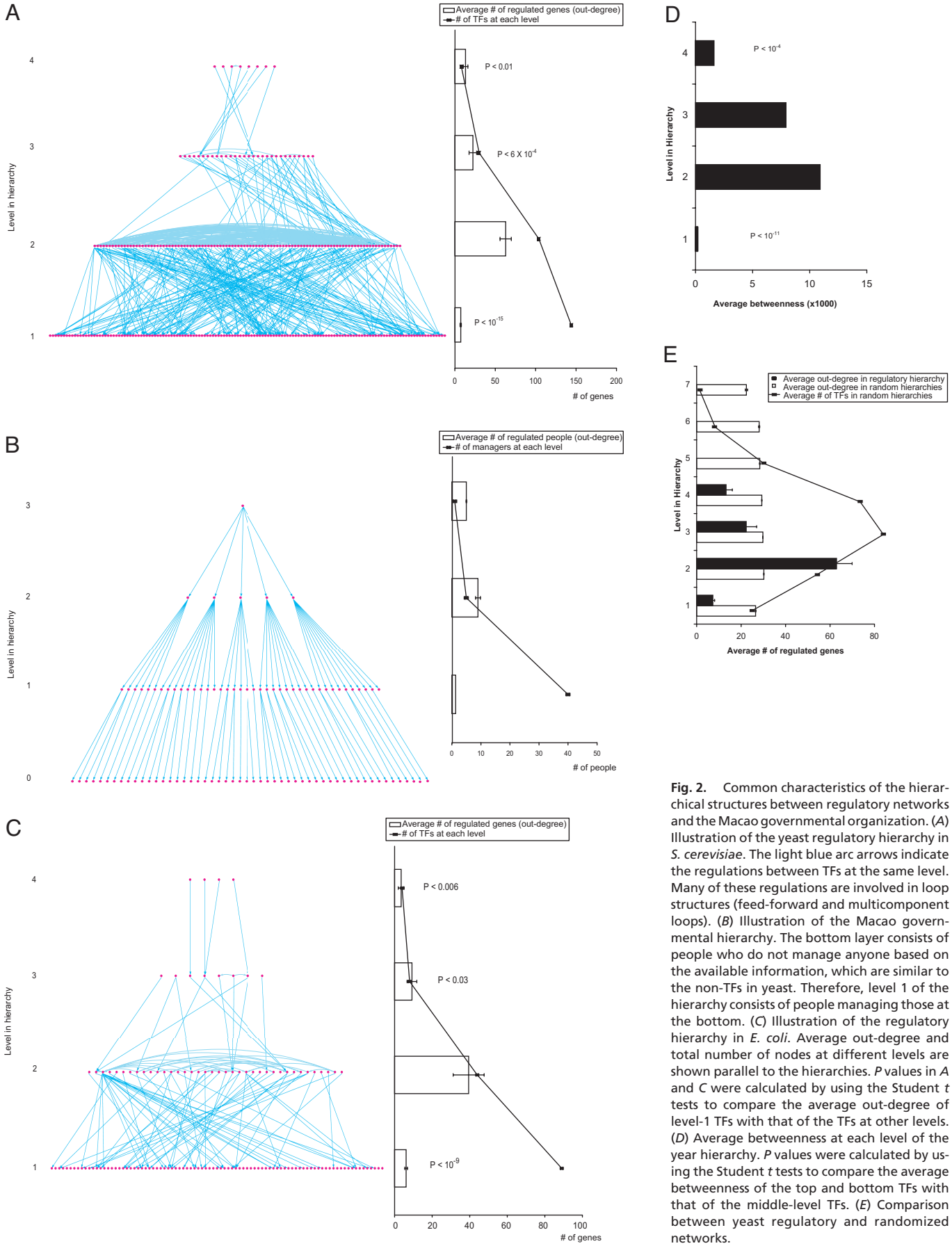

Fig. 2. Common characteristics of the hierarchical structures between regulatory networks and the Macao governmental organization. (*A*) Illustration of the yeast regulatory hierarchy in *S. cerevisiae*. The light blue arc arrows indicate the regulations between TFs at the same level. Many of these regulations are involved in loop structures (feed-forward and multicomponent loops). (*B*) Illustration of the Macao governmental hierarchy. The bottom layer consists of people who do not manage anyone based on the available information, which are similar to the non-TFs in yeast. Therefore, level 1 of the hierarchy consists of people managing those at the bottom. (*C*) Illustration of the regulatory hierarchy in *E. coli*. Average out-degree and total number of nodes at different levels are shown parallel to the hierarchies. *P* values in *A* and *C* were calculated by using the Student *t* tests to compare the average out-degree of level-1 TFs with that of the TFs at other levels. (*D*) Average betweenness at each level of the year hierarchy. *P* values were calculated by using the Student *t* tests to compare the average betweenness of the top and bottom TFs with that of the middle-level TFs. (*E*) Comparison between yeast regulatory and randomized networks.

Table 1. Hierarchy of *S. cerevisiae* regulatory network

Level	Genes									
4	*SPT23*	*HIR3*	*ADA2*	*GAT1*	*NGG1*	*DAT1*	*MOT3*	*GZF3*		
3	*MIG2*	*ZMS1*	*SWI3*	*SET2*	*IMP2′*	*MIG1*	*HFI1*	*XBP1*	*RTG3*	*ZAP1*
	SIR2	*SIR4*	*HAP1*	*DAL80*	*CYC8*	*ARO80*	*PHO80*	*SUI2*	*PHO2*	*SPT20*
	GAT3	*BDF1*	*NOT5*	*RIM101*	*SIN3*	*OPI1*	*CDC47*	*MSN4*	*HPR1*	*HMRA2*
2	*SMP1*	*INO2*	*CLN3*	*SIR3*	*SUT1*	*HAC1*	*SNF5*	*IME1*	*SKN7*	*RGT1*
	CUP9	*RFX1*	*YOX1*	*TUP1*	*YAP6*	*CIN5*	*HIR2*	*YFL044C*	*YML081W*	
	HSF1	*HAP3*	*HCM1*	*PHO4*	*NDD1*	*FKH1*	*CLN1*	*UME6*	*CAD1*	*REB1*
	MET4	*ASK10*	*FAR1*	*TOS4*	*CRZ1*	*SPT16*	*STP2*	*SUM1*	*DOT6*	*LEU3*
	GAL4	*MATA1*	*HAP4*	*GCN4*	*RAP1*	*RLM1*	*KTI11*	*FKH2*	*IXR1*	*YHP1*
	YAP1	*MBP1*	*TYE7*	*FZF1*	*POG1*	*NRG1*	*MET32*	*HMLALPHA1*		*STE12*
	ASH1	*HMLALPHA2*		*SPT5*	*NHP6A*	*GAL11*	*OAF1*	*HAP5*	*SWI5*	*DIG1*
	HMS2	*SET1*	*SOK2*	*BCK2*	*SNT2*	*PDR3*	*PDR1*	*PHD1*	*ACE2*	*ADR1*
	CBF1	*RTG1*	*CAT8*	*CSE2*	*MCM1*	*ROX1*	*SWI6*	*PAF1*	*KSS1*	*SWI1*
	RME1	*ABF1*	*ATS1*	*TEC1*	*SFP1*	*MAC1*	*ALPHA1*	*GLN3*	*AZF1*	*FHL1*
	SW14	*MET31*	*HAL9*	*STB1*	*TOS8*	*NAB3*	*YAP5*			
1	*HAA1*	*ARG81*	*RSC3*	*UPC2*	*THI3*	*SSN2*	*RDR1*	*DST1*	*MED8*	*PDC2*
	DAL82	*CHA4*	*EAF3*	*RGA1*	*CDC36*	*SNF1*	*YAP3*	*PPR1*	*ARG80*	*NOT3*
	MAF1	*ARR1*	*YJL206C*	*IWS1*	*YDR520C*	*GCR2*	*RCO1*	*FLO8*	*TOA1*	*NDT80*
	AFT2	*SDS3*	*SNF6*	*CT16*	*CDC73*	*GIS1*	*PGD1*	*SRB7*	*MED2*	*MGA2*
	CAF4	*SPT3*	*THI2*	*SPT4*	*SKO1*	*SSU72*	*SPT7*	*RSF1*	*LYS14*	*YPL230W*
		CAF16	*HAP2*	*TPO1*	*WAR1*	*SSN8*	*STB4*	*ITC1*	*ROX3*	*NUT2*
	MBF1	*MSS11*	*NUT1*	*RAD9*	*STE5*	*MIG3*	*RFA1*	*ACA1*	*RSC2*	*RDS3*
	MET28	*MAL13*	*STB5*	*SMK1*	*CDC39*	*CAF130*	*YRR1*	*TFA2*	*MSN1*	*PIP2*
	HST1	*BAS1*	*CAF40*	*PUT3*	*YKU70*	*NRD1*	*RDS1*	*CDC50*	*MGA1*	*CST6*
	KAR4	*RFA2*	*RAD50*	*MF(ALPHA)2*		*GTS1*	*RPH1*	*GCR1*	*CLN2*	*RAD18*
	STP1	*NRG2*	*MSN2*	*RCS1*	*YDR026C*	*SFL1*	*HIR1*	*RPI1*	*TOA2*	*RLR1*
	NHP6B	*RIM4*	*WHI2*	*HMS1*	*PHO23*	*MF(ALPHA)1*		*IME4*	*PLM2*	*SIP4*
	MAL33	*RPN4*	*WTM1*	*RDS2*	*STP4*	*STO1*	*MET18*	*RSC1*	*TFA1*	*TIS11*
	CUP2	*ECM22*	*STB2*	*UME1*	*RGM1*	*MOT2*	*SPT8*	*SRB4*	*SRD1*	*SPT21*
	CUP2	*ECM22*	*STB2*	*UME1*	*RGM1*	*MOT2*	*SPT8*	*SRB4*	*SRD1*	*SPT21*
	HOG1	*SPT2*	*UGA3*	*DAL81*	*SET3*	*HTZ1*	*STD1*			

from the top to the bottom. A path in the regulatory network represents a specific regulation (activation or inhibition) of a downstream TF by an upstream one. If any intermediate TF along this path is disabled, the regulation is broken. If we consider each path as a unique flow of regulatory information, the number of paths through each node is thus how much flow it controls. In graph theory, "betweenness" is an important topological parameter that describes precisely this concept. The betweenness of a node is defined as the number of shortest paths going through this node. If there is more than one shortest path between a pair of nodes, each path is given equal weight so that the overall weight of all paths is unity (10, 19). We call nodes with the highest betweenness "bottlenecks," in analogy to heavily used intersections leading to major highways or bridges in social transportation systems. Because TFs in the middle of the hierarchy not only pass information directly to their targets but also carry the information flows from the top TFs to the bottom ones, it is quite intuitive to see that these TFs should be the bottlenecks that control the most information flows.

We calculated the average betweenness of all TFs at each level in the hierarchy. Our results agree well with our expectation (see Fig. 2*D*): The TFs in levels 2 and 3 have significantly higher betweenness than those at the top or bottom of the hierarchy. Similar results were also observed in the *E. coli* hierarchy (see Figs. 15 and 16, which are published as supporting information on the PNAS web site). To some degree, these results also validate the way we constructed hierarchies by using our BFS-level method. Because the calculation of betweenness is only based on the connectivity of the network, completely independent of how we placed the nodes into layers within the hierarchy, the fact that the calculated results agree with our expectation confirms the plausibility of our method. Please note that one should not take the betweenness calculation as a definitive measure of the information flow, because it does not take into account some other possible contributing factors (e.g., gene expression and protein abundance).

Regulatory Hierarchies Are Well Organized. Next, we investigated random networks to see whether a similar hierarchical organization could be achieved by chance. We randomly rewired the edges between TFs and their targets within the whole yeast regulatory network (see *Materials and Methods*). Fig. 2*E* clearly shows that the pyramid-shaped hierarchical structure does not exist in random networks, whose layered structures consist of many more levels (on average 7.2 levels) than real hierarchies ($P < 0.001$). Furthermore, the average out-degree is almost constant between different levels of random networks. Similar results were also found for randomly rewiring the *E. coli* hierarchy (see Fig. 15).

In a social context, it has been shown that flatter hierarchies give managers at each level more freedom (20). Moreover, the number of levels in a hierarchy is determined by the degree of standardization of the work processes. In a corporation where workers perform similar tasks (e.g., in an auto-assembly plant), hierarchies tend to be flatter (21). In a similar fashion, different types of genes are known to cooperate to carry out a certain function. Therefore, it is quite reasonable for the regulatory hierarchies to be flatter than random expectation.

It has also been found that the number of people supervised by each manager is determined by the nature of the job (21). In a situation where workers under the same manager perform different tasks and need more mutual accommodation (e.g., in a law firm), the average number of people supervised by a single manager is very small (22–24). A similar situation exists in the cell. At the top of the regulatory hierarchies, interplay between top-level and

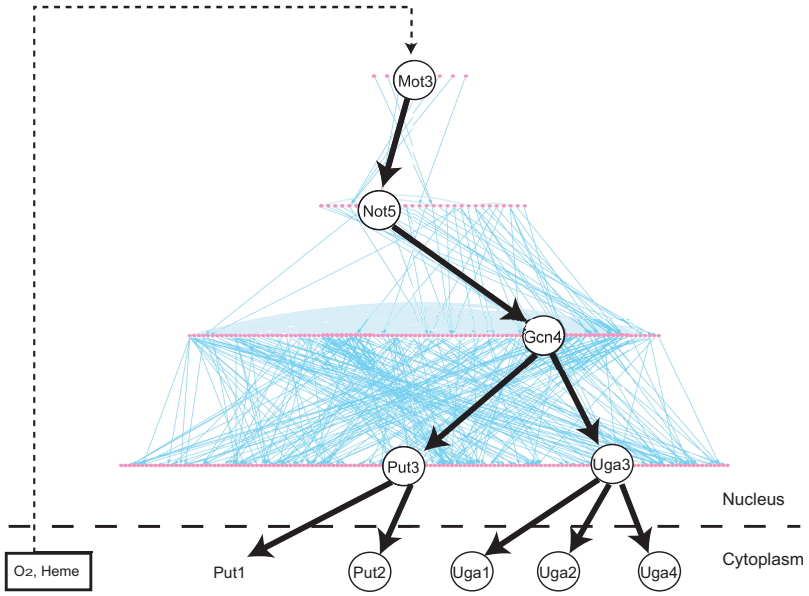

Fig. 3. A biological example to illustrate the multistep cogitation processes in the regulatory hierarchy, showing aerobic growth mediated by Mot3. We divided the image into two parts, nucleus and cytoplasm, because TFs only function in the nucleus, whereas other proteins (such as the enzymes Put1, Put2, Uga1, Uga2, and Uga3) normally function in the cytoplasm.

downstream TFs is needed to initiate a process. Furthermore, top-level TFs tend to regulate TFs associated with many different pathways and functions (see below). Therefore, it is quite reasonable that the average out-degree at the topmost level is small. After commitment, however, middle-level TFs can turn on massive expression of many genes in response to stimuli, reflecting their larger average out-degree. At the bottom of the hierarchy, TFs regulate only few specific target genes.

Decision-Making Schemes in Regulatory Hierarchies. We further analyzed the regulatory hierarchies in *S. cerevisiae* and *E. coli*. We observed two distinct types of regulatory processes in them. These processes are readily understandable as different decision-making schemes, given that we know gene expression is regulated in response to various internal and external stimuli.

"Reflex" processes. A nonnegligible number of TFs (52 in *S. cerevisiae*; 30 in *E. coli*) neither regulate other TFs nor are regulated by other TFs. They respond to specific stimuli turning on (or shutting down) the expression of their targets. We call this type of decision-making a reflex process. Regulation of the *trp* operon in *E. coli* is a perfect example. The *trp* operon encodes genes for the synthesis of tryptophan. TrpR is a repressor that, when activated by tryptophan binding, represses expression of the *trp* operon. TrpR is not regulated by any other TFs (25). In *S. cerevisiae*, a similar example is Arg-81, a TF involved in arginine metabolism (26). Upon the presence of arginine, Arg-81 shuts down the expression of many enzymes involved in arginine biosynthesis, such as *ARG*1, *ARG*3, and *ARG*8 (27). (Note that some of the reflex assignments to TFs may result from incompleteness of the known regulatory data sets.)

"Cogitation" processes. The majority of TFs in both regulatory networks are regulated by other TFs. Most of these regulate other TFs as well. Thus, TFs at the top become the global modulators for all downstream ones. The decision is amplified and executed while being passed down. We call this a cogitation process.

Cogitation processes have some nice parallels to the overall description of decision making in apoptosis. Apoptosis consists of three phases: decision, commitment, and execution (28–30). In the first phase, the cell senses proapoptotic signals and determines whether it should die. This step is reversible. In the commitment phase, however, the cell makes an unstoppable decision to die, which leads to the execution phase, where the actual destructive process is carried out (28–30). Such a multistep decision-making scheme has two advantages: (*i*) it can work as a signal amplifier to rapidly increase the magnitude of the response just like the cAMP-cascade in glycogen metabolism (31); and (*ii*) it can act like a noise filter to convert continuous inputs into all-or-none switch-like outputs (32).

We can see clear examples of cogitation processes in the yeast regulatory hierarchy. In particular, the expression of *MOT*3, a top-level TF involved in aerobic growth, is activated by heme and oxygen (33, 34) (Fig. 3), representing the decision phase. Mot3 in turn activates the expression of *NOT*5 and *GCN*4 (1, 35), which are both midlevel TF hubs with a large number of targets. Once their expression is turned on, the cell is committed. Finally, in execution, Gcn4 activates two specific bottom-level TFs, Put3 and Uga3, which trigger the expression of enzymes in proline and nitrogen utilization, respectively (36, 37).

Note that the distinction between the cogitation and reflex processes is purely based on the topology of the regulatory network. It is of course the case that even a reflex process could be very slow if the nontranscriptional events that underlie it are exceptionally time-consuming. However, transcriptional processes are normally much slower than nontranscriptional ones (e.g., phosphorylation). So it is quite reasonable to believe that our conclusions based on network topology reflect the actual timing of the processes.

Top-Level TFs Receive Signals Through Protein–Protein Interactions. In our analyses, we have shown that regulation of gene expression in the cell normally happens in a multistep fashion starting from the top TFs. Because of the following, we hypothesized that the TFs at the top of the hierarchy would receive most of the stimulating signals and thus should have more interaction partners: (*i*) the cell regulates expression of its thousands of genes in response to internal and external stimuli, and (*ii*) TFs receive these signals through interactions with other molecules, mainly other proteins, because they usually function within the nucleus.

Fig. 4*A* clearly shows that top-level TFs on average interact with more proteins than the others, confirming our hypothesis. Furthermore, we examined another important topological quantity, closeness, defined as the inverse of the sum of the distances

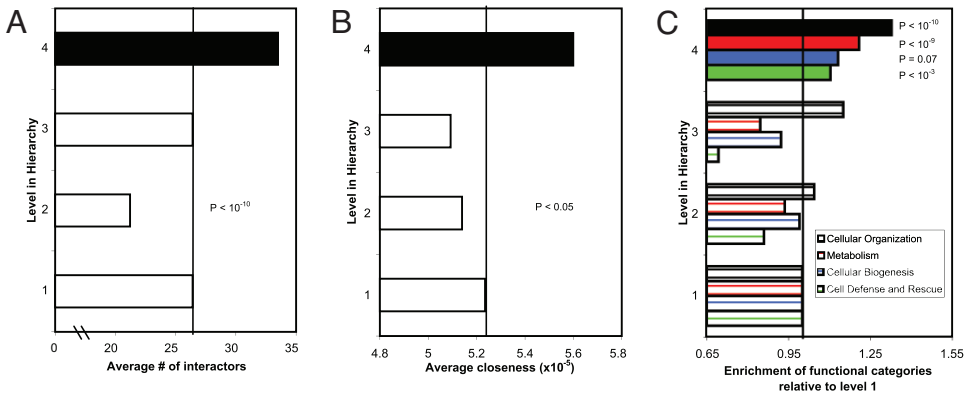

Fig. 4. Correlations between levels in the hierarchy and other topological and functional properties. (*A* and *B*) Average number of interaction partners (*A*) and average closeness (*B*) for TFs at each level. *P* values were calculated with Student's *t* tests to compare the top bar with the sum of the test bars. (*C*) Enrichment of functional categories relative to level 1. For each functional category in the Munich Information Center for Protein Sequences (MIPS) functional classification schemes, we calculated the percentage of interaction partners of TFs that have this function. The percentage of a certain category was then normalized against the corresponding one at level 1. Thus, all bars at level 1 have a value of 1. Because we were analyzing the transcriptional regulatory networks, we ignored the functional category ''transcription.'' *P* values were calculated with cumulative binomial distributions to compare the statistical significance of enrichment at level 4 to that of the sum of the other levels (see *Supporting Text*).

from a certain node to all other nodes (19). Fig. 4*B* shows that the top TFs by and large have significantly higher closeness in the interaction network than all other TFs, indicating that these TFs are at the center of the interaction network (i.e., close to all proteins) (19). This result further confirms our hypothesis that these TFs receive signals through protein–protein interactions. The signals are then processed and passed onto lower-level TFs along the hierarchy. Finally, we analyzed the functional composition of the interaction partners of the TFs at each level of the hierarchy by using the Munich Information Center for Protein Sequences (MIPS) functional classification schemes (38). As shown in Fig. 4*C*, we found that three functional categories are significantly enriched within the interaction partners of the top TFs compared with those of the bottom ones ($P < 0.05$). They are as follows.

1. Cellular organization: Most of the proteins in this category are localized to different organelles within the cell to keep their integrity.
2. Metabolism: The cell utilizes these proteins to respond to the nutrition changes in the environment, such as during the diauxic shift when the yeast cell switches from using glucose to ethanol as a carbon source (39).
3. Cell defense and rescue: Obviously, most proteins in this category carry out defenses against various types of stress that the cell may sustain.

A good example is the protein Ire1 (see Fig. 5), which belongs to all three categories. It is a transmembrane protein on the endoplasmic reticulum (ER) membrane, with serine–threonine kinase and endoribonuclease activities (40, 41). It is one of the main factors involved in the unfolded protein response and myo-inositol metabolism (40, 41). Upon the presence of unfolded proteins, Ire1 activates the SAGA complex (comprising Ada2, Gcn5, Hfi1, Ngg1, Spt20, Spt3, and Spt7) through directly interacting with Ada2 to enhance transcriptional induction of ER stress-responsive genes (42). In the available regulatory network, one possible path is that Ada2 successively turns on the expression of three TFs: Rtg3, Hmra1, and Ime4. Ime4 then induces the expression of 18 other genes. For example, Egd2 is a subunit of the heteromeric nascent polypeptide-associated complex that binds unfolded proteins in the ER to help them form secondary structures (43); Vik1 is involved

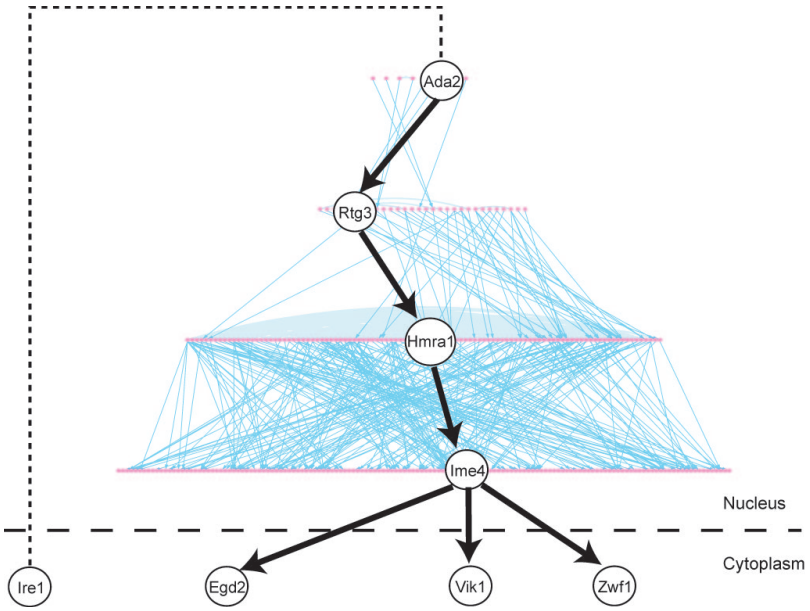

Fig. 5. A biological example to illustrate that the top-level TFs receive internal and external signals through protein–protein interaction, showing unfolded protein response mediated by Ire1.

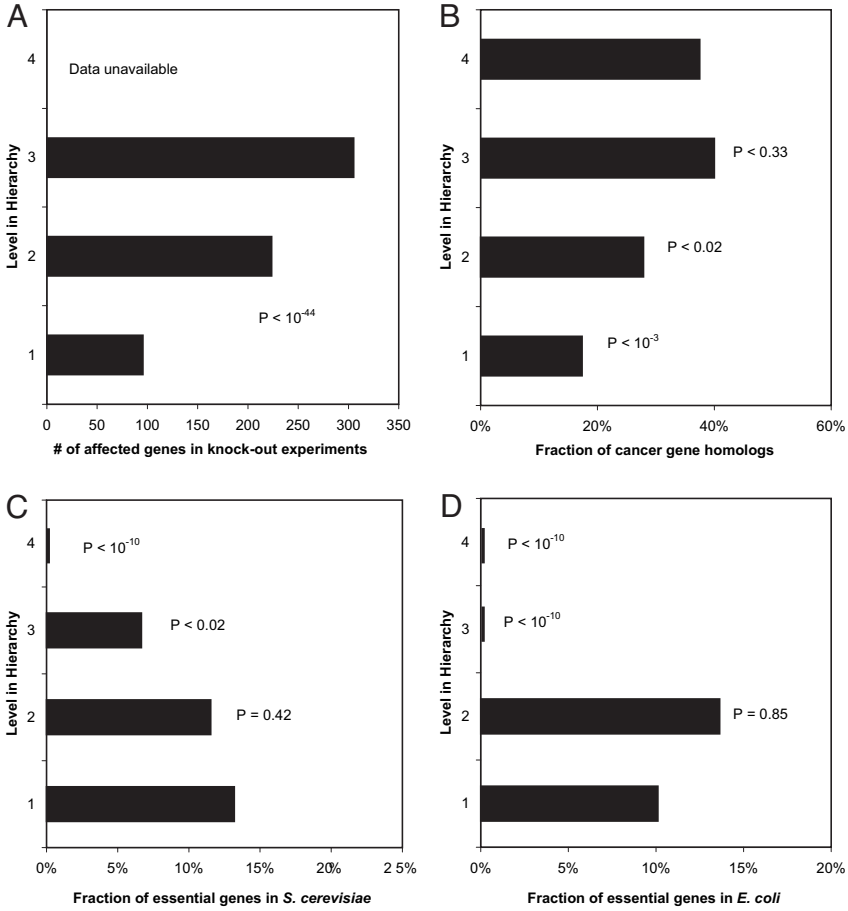

Fig. 6. Correlations between levels in the hierarchy and other biological properties. (*A*) Deletion of TFs at higher levels disrupts the expression of more genes. A gene is defined as disrupted if *P* is <0.05 determined by Rosetta knockout experiments (47). Because the knockout experiments were only performed on 41 TFs, *t* tests cannot be performed to examine the statistical significance of the differences between the average numbers of affected genes across different levels. Therefore, we performed a χ^2 test and found that deletion of TFs at higher levels disrupts the expression of more genes, which is statistically significant when compared with random expectation ($P < 10^{-45}$; see *Supporting Text*). (*B*) TFs at higher levels in the hierarchy have a strong tendency to have human homologs associated with cancer. *P* values measure the statistical significance between the fractions of human cancer gene homologs among TFs at a certain level with that at level 4. (*C*) TFs at the bottom of the yeast hierarchy have a strong tendency to be essential genes. *P* values measure the statistical significance between the fractions of essential genes among TFs at a certain level with that at level 2 and were calculated by using cumulative binomial distributions (see *Supporting Text*). (*D*) TFs at the bottom of the *E. coli* hierarchy have a strong tendency to be essential genes. All calculations are similar to those in *C*.

in ER organization and biogenesis (44); and Zwf1 is required for oxidative stress response and fatty acid metabolism (45, 46).

One might think that most top-level TFs are involved in chromatin-remodeling complexes, because these complexes affect a large number of transcriptional events and their components have high degrees in the interaction network. However, this assumption is, in fact, not the case (for detailed descriptions of functions of top-level TFs, see Table 3, which is published as supporting information on the PNAS web site). Even though there is no strong functional pattern for the top-level TFs, most of them seem to be global modulators that respond to various cellular stresses (e.g., anomalous levels of nitrogen or glucose).

Paradox of Influence and Essentiality. ***Higher-level TFs are more influential.*** We next examined the influence of each TF by using the Rosetta knockout experiments (47). Fig. 6*A* shows that deletions of genes at higher levels of the hierarchy affect more genes than deletions of those at the bottom; i.e., higher-level TFs are more influential. (Note that because the Rosetta knockout experiments were only performed on 276 genes, no genes at level 3 were tested in the experiments.)

Furthermore, we investigated the influence of TFs in terms of the ability of their human homologs to initiate disease, especially cancer. We calculated the fraction of TFs at different layers that have cancer-related homologs in humans. Our calculations show that human homologs of TFs at higher levels have a higher tendency to be cancer related (see Fig. 6*B*), further confirming the influence of high-level TFs in the hierarchy.

Lower-level TFs are more essential. Because we have shown that TFs at higher levels are more influential, it is reasonable to assume that these TFs should also be more essential (i.e., lethal) (48). However, based on our calculations in yeast, we found that TFs at the lower levels of the network have a much higher tendency to be essential (Fig. 6*C*). A similar result was also obtained in *E. coli* (Fig. 6*D*). One possible explanation for the separation of the influence from essentiality may be that TFs at the top of the hierarchy act more like modulators coordinating gene expression across different pathways (e.g., Mot3); therefore, all pathways remain functional upon deletion of these TFs, even though the precise expression between most pathways will not be well organized. On the other hand, TFs at the bottom are in charge of specific pathways (e.g., Put3 and Uga3). Upon their deletion, certain pathways will cease operating, causing the cell to die.

Discussion

In general, our results show that there is a pyramid-shaped hierarchical structure in regulatory networks, which is well organized in a clearly nonrandom manner. The major decision-making scheme in this hierarchy is a cogitation-like multistep process, where the TFs at the top receive signals from internal and external stimuli through protein–protein interactions. These TFs strongly influence those below (in terms of the overall fraction of cellular genes affected). However, surprisingly, the TFs at the bottom are more essential to the viability of the cell.

Because bottom TFs are relatively easy to define in regulatory networks, our BFS-level method is a reasonable way to turn the network into a tree in graph theory (18). However, as mentioned above, it is not trivial to construct a hierarchy for any given directed network; an assortment of possible variations readily comes to mind. In particular, our method essentially assigns the lowest possible level of each TF as its level in the hierarchy because it is shortest-path based. Alternatively, one could calculate the longest

path from a TF to a bottom node and assign this number as its level. For simple hierarchies, both methods will produce exactly the same results. For networks containing loops, the constructed hierarchies will be slightly different. Our BFS-level method has problems solving feed-forward type of situations, whereas the longest-path method has problems solving feed-back type of situations. It is difficult to argue which method is better. In *Supporting Text*, we describe implementing this variant and other related ones. Our results show that, in fact, most variations have similar global trends, confirming the validity of our conclusions.

Furthermore, as shown in Fig. 2*E*, our BFS-level method could assign a level number to every node in any directed network, even one that is randomly generated. However, the key aspect of a generalized hierarchy is its pyramidal shape. As we showed in Fig. 2, regulatory hierarchies have a similar pyramidal shape to social ones. We are also able to show that the topological features of the regulatory hierarchy correspond well to aspects associated with efficiency in its social counterparts. As discussed in detail above, these features are completely different from those in random networks, suggesting their functional implications.

Moreover, previous studies have examined the relationships between the essentiality of a TF and its number of descendants (i.e., out-degree). It has been shown that TFs regulating more targets tend to be more essential (49).

Materials and Methods

Regulatory Networks. We constructed the *S. cerevisiae* regulatory network by combining the results of various genetic, biochemical, and ChIP-chip experiments in yeast (1, 2, 50–54). To ensure the quality of the network, we manually examined the network and removed all questionable ORFs and DNA-binding enzymes (e.g., PolIII). The final network contained 8,371 regulatory interactions involving 286 TFs and 3,369 targets. The *E. coli* regulatory network was constructed in a similar manner, which consisted of 2,370 regulatory interactions between 145 TFs and 1,063 genes (55, 56).

Yeast Interaction Network. The interaction network was created by combining various databases and large-scale experiments (38, 49, 57–63). Because large-scale experiments are known to be error-prone (64, 65), we only considered protein pairs with multiple sources of support [using the likelihood ratio of ≥300 criteria from Jansen *et al.* (66)]. The final network contained 23,294 interactions involving 4,743 proteins.

Generation of Random Networks. We first generated random networks by randomly connecting TFs with target genes, while keeping the total numbers of TFs (286), target genes (3,369), and edges (8,371) constant. Then, we ran the BFS-level method to build the layered structure from the randomized network and repeated all calculations. This procedure was repeated 1,000 times. The results were averaged and are shown in Fig. 2*E*. We also performed similar calculations for the *E. coli* regulatory network and found similar results (see Fig. 15).

H. Y. thanks Dr. Philip Kim for many stimulating discussions. M. G. was supported by the National Institutes of Health.

1. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al.* (2002) *Science* 298:799–804.
2. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M (2002) *Genes Dev* 16:3017–3033.
3. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) *Nature* 411:41–42.
4. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M (2001) *J Mol Biol* 314:1053–1066.
5. Albert R, Barabasi AL (2002) *Rev Mod Phys* 74:47–97.
6. Albert R, Jeong H, Barabasi AL (1999) *Nature* 401:130–131.
7. Albert R, Jeong H, Barabasi AL (2000) *Nature* 406:378–382.
8. Barabasi AL, Albert R (1999) *Science* 286:509–512.
9. Amaral LA, Scala A, Barthelemy M, Stanley HE (2000) *Proc Natl Acad Sci USA* 97:11149–11152.
10. Girvan M, Newman ME (2002) *Proc Natl Acad Sci USA* 99:7821–7826.
11. Huberman BA, Adamic LA (1999) *Nature* 401:131.
12. Watts DJ, Strogatz SH (1998) *Nature* 393:440–442.
13. Yu H, Luscombe NM, Qian J, Gerstein M (2003) *Trends Genet* 19:422–427.
14. Ihmels J, Levy R, Barkai N (2004) *Nat Biotechnol* 22:86–92.
15. Jansen R, Greenbaum D, Gerstein M (2002) *Genome Res* 12:37–46.
16. Woodward J (1980) *Industrial Organization: Theory and Practice* (Oxford Univ Press, Oxford).
17. Whyte LL, Wilson AG, Wilson D (1969) *Hierarchical Structures* (Elsevier, New York).
18. Cormen HT, Leiserson EC, Rivest LR (1993) *Introduction to Algorithms* (MIT Press, Boston).
19. Freeman LC (1977) *Sociometry* 40:35–41.
20. Ivancevich JM, Donnelly JH, Jr (1975) *Administrative Sci Q*, 272–280.
21. Mintzberg H (1979) *The Structuring of Organizations* (Prentice–Hall, Englewood Cliffs, NJ).
22. Wilensky HL (1967) *Organizational Intelligence* (Basic Books, New York).
23. Urwick LF (1956) *Harvard Business Rev*, 39–47.
24. Filley AC, House RJ (1969) *Managerial Process and Organizational Behaviour* (Scott Foresman, Glenview, IL).
25. Zhang RG, Joachimiak A, Lawson CL, Schevitz RW, Otwinowski Z, Sigler PB (1987) *Nature* 327:591–597.
26. De Rijcke M, Seneca S, Punyammalee B, Glansdorff N, Crabeel M (1992) *Mol Cell Biol* 12:68–81.
27. Svetlov VV, Cooper TG (1995) *Yeast* 11:1439–1484.
28. Thompson CB (1995) *Science* 267:1456–1462.
29. Kroemer G, Petit P, Zamzami N, Vayssiere JL, Mignotte B (1995) *FASEB J* 9:1277–1287.
30. Jacob M, McCarthy N (2002) *Apoptosis: The Molecular Biology of Programmed Cell Death* (Oxford Univ Press, Oxford).
31. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson J (1994) *Molecular Biology of the Cell* (Garland, New York).
32. Ferrell JE, Jr, Machleder EM (1998) *Science* 280:895–898.
33. Csank C, Costanzo MC, Hirschman J, Hodges P, Kranz JE, Mangan M, O'Neill K, Robertson LS, Skrzypek MS, Brooks J, *et al.* (2002) *Methods Enzymol* 350:347–373.
34. Abramova N, Sertil O, Mehta S, Lowry CV (2001) *J Bacteriol* 183:2881–2887.
35. Hodges PE, McKee AH, Davis BP, Payne WE, Garrels JI (1999) *Nucleic Acids Res* 27:69–73.
36. Talibi D, Grenson M, Andre B (1995) *Nucleic Acids Res* 23:550–557.
37. Siddiqui AH, Brandriss MC (1989) *Mol Cell Biol* 9:4706–4712.
38. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B (2002) *Nucleic Acids Res* 30:31–34.
39. DeRisi J, Iyer V, Brown P (1997) *Science* 278:680–686.
40. Welihinda AA, Tirasophon W, Kaufman RJ (1999) *Gene Express* 7:293–300.
41. Welihinda AA, Tirasophon W, Green SR, Kaufman RJ (1998) *Mol Cell Biol* 18:1967–1977.
42. Welihinda AA, Tirasophon W, Kaufman RJ (2000) *J Biol Chem* 275:3377–3381.
43. George R, Beddoe T, Landl K, Lithgow T (1998) *Proc Natl Acad Sci USA* 95:2296–2301.
44. Wright R, Parrish ML, Cadera E, Larson L, Matson CK, Garrett-Engele P, Armour C, Lum PY, Shoemaker DD (2003) *Yeast* 20:881–892.
45. Minard KI, Jennings GT, Loftus TM, Xuan D, McAlister-Henn L (1998) *J Biol Chem* 273:31486–31493.
46. Juhnke H, Krems B, Kotter P, Entian KD (1996) *Mol Gen Genet* 252:456–464.
47. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai HY, He YDD, *et al.* (2000) *Cell* 102:109–126.
48. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, *et al.* (1999) *Science* 285:901–906.
49. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) *Trends Genet* 20:227–231.
50. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO (2001) *Nature* 409:533–538.
51. Lieb JD, Liu X, Botstein D, Brown PO (2001) *Nat Genet* 28:327–334.
52. Hahn JS, Hu Z, Thiele DJ, Iyer VR (2004) *Mol Cell Biol* 24:5249–5256.
53. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, *et al.* (2001) *Nucleic Acids Res* 29:281–283.
54. Guelzim N, Bottani S, Bourgine P, Kepes F (2002) *Nat Genet* 31:60–63.
55. Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, *et al.* (2004) *Nucleic Acids Res* 32:D303–D306.
56. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) *Nat Genet* 31:64–68.
57. Yu H, Zhu X, Greenbaum D, Karro J, Gerstein M (2004) *Nucleic Acids Res* 32:328–337.
58. Bader GD, Betel D, Hogue CW (2003) *Nucleic Acids Res* 31:248–250.
59. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) *Nucleic Acids Res* 30:303–305.
60. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y (2000) *Proc Natl Acad Sci USA* 97:1143–1147.
61. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.* (2000) *Nature* 403:623–627.
62. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al.* (2002) *Nature* 415:180–183.
63. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.* (2002) *Nature* 415:141–147.
64. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) *Nature* 417:399–403.
65. Jansen R, Lan N, Qian J, Gerstein M (2002) *J Struct Funct Genomics* 2:71–81.
66. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) *Science* 302:449–453.