# Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns

Timothy R. Lezon*, Jayanth R. Banavar*, Marek Cieplak†, Amos Maritan‡, and Nina V. Fedoroff§¶‖

*Department of Physics, 104 Davey Laboratory, and §Department of Biology and Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802; ¶Santa Fe Institute, Santa Fe, NM 87501; †Institute of Physics, Polish Academy of Science, Aleja Lotnikow 32/48, 02-668 Warsaw, Poland; and ‡Dipartimento di Fisica, Consorzio Nazionale Interuniversitario per le Scienze Fisiche della Materia and Istituto Nazionale di Fisica Nucleare, Università di Padova, Via Marzolo 8, 35131 Padova, Italy

Contributed by Nina V. Fedoroff, October 17, 2006 (sent for review September 9, 2006)

**We describe a method based on the principle of entropy maximization to identify the gene interaction network with the highest probability of giving rise to experimentally observed transcript profiles. In its simplest form, the method yields the pairwise gene interaction network, but it can also be extended to deduce higher-order interactions. Analysis of microarray data from genes in *Saccharomyces cerevisiae* chemostat cultures exhibiting energy metabolic oscillations identifies a gene interaction network that reflects the intracellular communication pathways that adjust cellular metabolic activity and cell division to the limiting nutrient conditions that trigger metabolic oscillations. The success of the present approach in extracting meaningful genetic connections suggests that the maximum entropy principle is a useful concept for understanding living systems, as it is for other complex, nonequilibrium systems.**

gene interactions | network inference | signaling | metabolic oscillations

The application of techniques for sampling expression levels of all of an organism's genes through time has yielded large amounts of data on the activity states of cellular genomes. Microarray data have been analyzed by using a variety of statistical tools to detect significant differences in gene expression levels and identify meaningful subgroups of genes exhibiting similar expression patterns (1, 2). Correlations and other statistical measures that group genes by profile similarity identify functionally interconnected groups of genes because proteins encoded by genes involved in the same biological process are often coregulated (3, 4). However, correlation measures do not provide direct insight into the identity or nature of the gene interactions that give rise to the observed expression patterns. Much effort is being devoted to the reconstruction of gene interaction networks using a variety of modeling approaches, ranging from simple Boolean networks through dynamical models of cellular processes (5–8). Various types of Bayesian network models (9, 10), graphical Gaussian models (11, 12), and relevance networks (13) have been developed to extract information about gene interactions directly from expression profiles. However, even for a simple linear model, the system is underdetermined because the number of genes sampled in a microarray experiment is invariably much larger than the number of samples, with the consequence that myriad networks can reproduce the observed data with fidelity. Efforts to constrain the model space by incorporating additional information from interventions and perturbations, other types of molecular data, or literature mining are useful on a small scale but rapidly become unwieldy with increasing gene numbers (14–17). Alternative approaches make simplifying assumptions about network topology or postulate that the microarray data are drawn randomly from a Gaussian distribution (11, 12, 18).

To avoid such assumptions, which are often either untestable or untenable, and address the underdetermination problem, we have developed an approach to gene network inference from gene expression data that relies on Boltzmann's concept of entropy maximization to support statistical inference with minimal reliance on the form of missing information (19, 20). Entropy maximization has proved powerful in the analysis of both complex equilibrium systems and, more recently, such nonequilibrium systems as neural networks and global climate (21–25). The underlying rationale is that each macroscopically observable state of a complex system corresponds to a number of microscopic states. Because the number of ways of realizing a given macroscopic state can vary widely, the most likely state of the system as a whole is the one that corresponds to the largest number of microscopic states. Here we explore the utility of the maximum entropy principle in extracting information about gene interactions from microarray data. We formulate a procedure to identify the pairwise genetic interaction network that has the highest probability of giving rise to the macrostate captured in the observed expression data. As pointed out by Shannon (20), information and entropy are interlinked: the more information one has, the lower the entropy. The logic of our approach is to determine the probability distribution governing the microarray data subject to the entropy-reducing constraint that the available information on gene expression levels, such as their pairwise and higher-order correlations, is faithfully encoded. Because the resulting network is selected by the maximum entropy principle and assumes nothing about missing information, any system with a lower entropy requires more information than is available from the microarray data. Moreover, the network obtained is necessarily in agreement with the actual network of molecular interactions (22).

We assess the ability of the maximum entropy approach to extract relevant genetic relationships by analyzing microarray expression data from the well studied eukaryote *Saccharomyces cerevisiae* growing under conditions that the support energy metabolic oscillations (26, 27). We report that the strongest gene interactions inferred in our analysis of the genes exhibiting the largest fluctuations in transcript levels during metabolic oscillations identify a network of genes coding for key proteins known to be involved in the several interconnected signaling and regulatory processes that adjust the cellular metabolic state and the cell cycle to the nutrient supply. Inclusion of genes showing smaller fluctuations under the same experimental conditions identifies important genes involved in such fundamental cellular

Author contributions: T.R.L., J.R.B., M.C., A.M., and N.V.F. performed research.

The authors declare no conflict of interest.

Abbreviation: TOR, target of rapamycin.

‖To whom correspondence should be addressed at: Pennsylvania State University, 219 Wartik Laboratory, University Park, PA 16802. E-mail: nvf1@psu.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0609152103/DC1.

processes as mitochondrial maintenance, pH regulation and cell wall biosynthesis, DNA replication and repair, and transcription. These results demonstrate that interconnections among cellular processes are reflected in interconnections among genes and indicate that it may be possible to retrieve more relevant information about cellular signaling and regulatory pathways directly from gene expression data than previous methods have yielded.

## Results

**Network Calculation.** To infer the most likely network, we selected subsets of the genes exhibiting the highest profile variance to minimize the contribution of experimental noise. We centered each profile at a mean value of 0 and normalized the expression profiles to unit variance to focus on the influence of the shape of the gene profile rather than its amplitude. We constructed the covariance matrix $C$, with the matrix element $C_{ij}$ representing the correlation between the normalized expression profiles of gene $i$ and gene $j$. For data normalized in this way, $C$ is exactly the matrix of Pearson correlations between gene profiles. As detailed in *Methods* and in the calculations in supporting information (SI) *SI Text*, we obtained the matrix of pairwise gene interactions $M$ that maximizes the system entropy by inverting $C$ in the space spanned by its non-zero eigenvectors.

**Analyzing Microarray Data.** We applied the maximum entropy method to infer gene interactions from genome-wide gene expression data derived from a well characterized eukaryotic organism, the yeast *S. cerevisiae*, cultures of which exhibit highly coordinated metabolic fluctuations, gene expression patterns, and cell division cycles under certain conditions (26, 27). Because of its importance in a variety of both traditional and contemporary biotechnological applications, as well as its use as a model eukaryote, *S. cerevisiae* has been studied extensively under carefully controlled conditions. There is already an abundance of genetic, physiological, biochemical, and molecular information about its response to nutrient conditions which can be queried to determine whether the genes and genetic interactions identified by the present method play important roles in the physiological oscillations that occur under limiting nutrient conditions.

We first analyzed data from a recent study that monitored changes in transcript levels in yeast cultures exhibiting energy metabolic oscillations of ≈40-min duration (26). The fluctuations in raw expression levels over the course of several metabolic oscillations, as measured by the standard deviation, varied among the 4,670 genes monitored from 5.2, which can be considered as a measure of the noise in the data, to >1,800, as shown in Fig. 1. Because microarray data are often noisy, we focused our analysis on subsets of genes exhibiting high expression profile variance. The first two subsets comprise the 582 genes with raw profile standard deviations greater that 400.8, the smallest of which is ≈77 times the magnitude of the noise, and 1,008 genes, whose smallest profile standard deviation is still ≈47 times the noise level. (see *SI Text* for a full description of the criteria used in selecting these subsets.) It should be underscored that irrespective of the size of the subset considered, the deduced interactions arise from the influence of all variables. Hence even if a gene is not explicitly considered in the network calculation, its effect is nonetheless integrated into the interactions among the remaining genes. The analysis yields a measure of the magnitude and sign of the interactions most likely to give rise to the observed data.

**Relative Magnitude of Pairwise and Higher-Order Interactions.** To determine the relative contributions of pairwise and higher-order gene interactions, we used perturbation theory to compute the strengths of all possible three-gene interactions for the

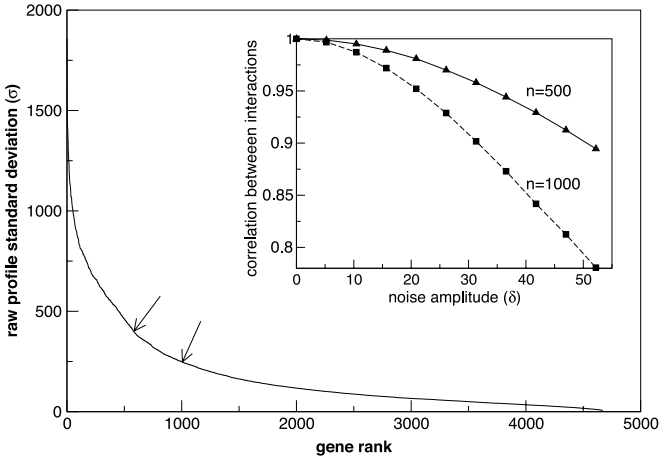


**Fig. 1.** A plot of the rank-ordered standard deviations ($\sigma$) of raw expression profiles for the 4,670 genes in the short-period data set. Arrows indicate cutoff values for 582 genes and 1,008 genes. (*Inset*) The Spearman correlation between the calculated interactions and those that result when noise of a fixed amplitude is added to the raw profiles. Random noise from a Gaussian distribution of width $\delta$ is added to each of the raw expression profiles, and the network for the noise-enhanced data was calculated and compared with the network from the raw data by using the Spearman correlation. When the 500 genes with largest profile amplitudes are retained (solid line), noise of $\delta$ = 52.2, ≈10 times the estimated background level, does not significantly change the network. The network is more sensitive to noise when 1,000 genes are retained (dashed line), but the correlation is still 0.9 between the original network and that with noise added at six times the estimated background.

582-gene subset (see *SI Text*). Fig. 2 shows the distribution of pairwise and three-gene interaction strengths. The magnitude of triplet interaction strengths is generally much smaller than that of pairwise interactions. Indeed, only 151 of the more than 32 million possible three-gene interactions among the 582 highest-variance genes have a magnitude >0.03. These rare higher-order interactions may prove important and are discussed below. However, most triplet interactions are very small, indicating that pairwise interaction network captures a majority of the important genetic couplings in the sampled yeast cells.

**The Pairwise Interaction Network.** To assess the structure of the pairwise interaction network inferred using the present method, we

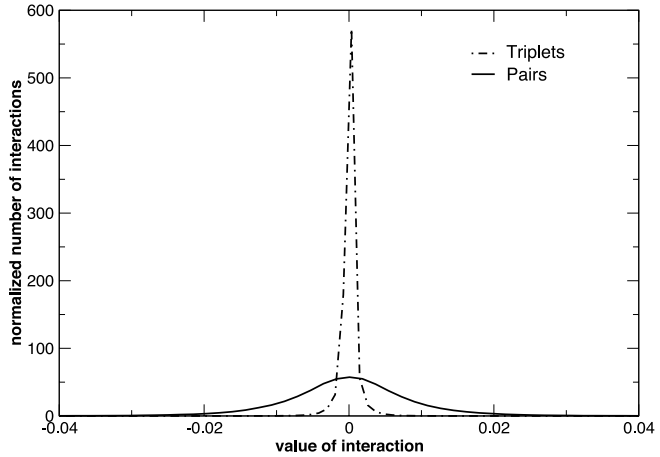


**Fig. 2.** The distribution of the relative strengths of pairwise and three-gene interactions among the 582 genes in the short-period data set showing the largest fluctuations during metabolic oscillations. Each curve is normalized to unit area.

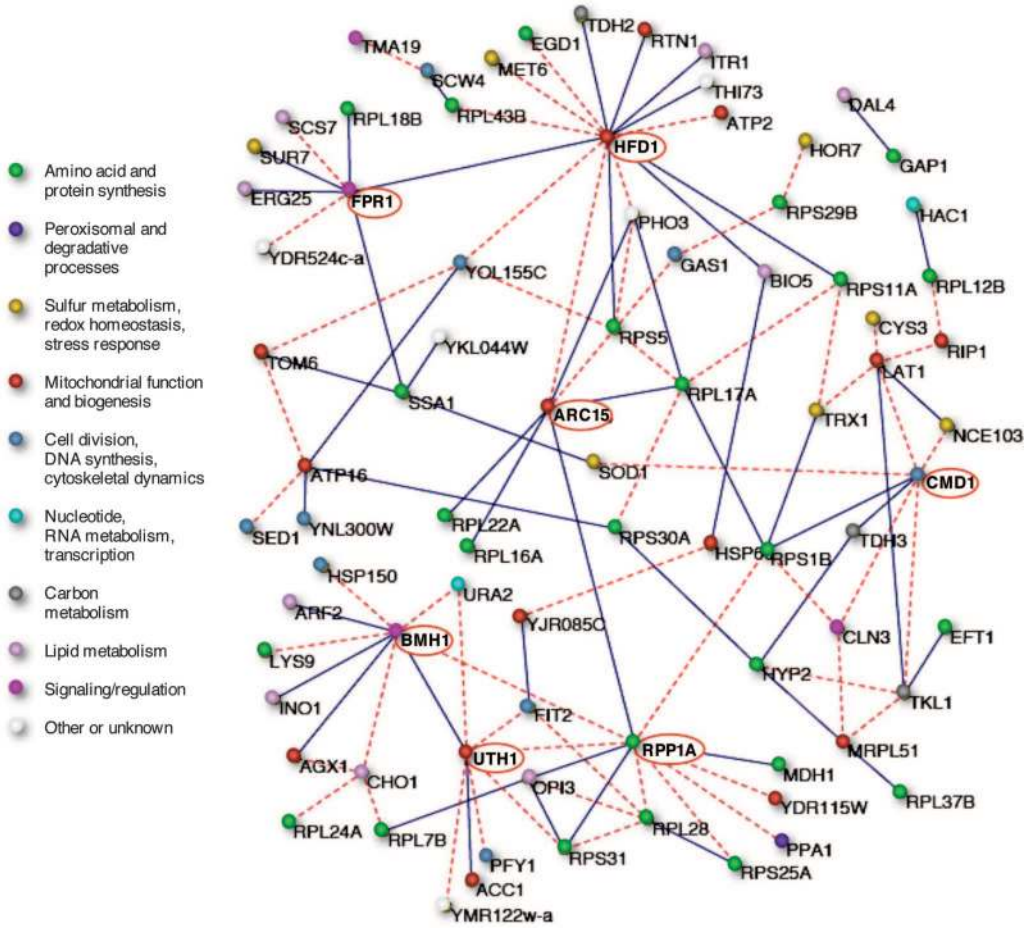


**Fig. 3.** The network of the strongest 110 pairwise interactions inferred by entropy maximization using the 582 genes showing the most marked fluctuations in transcript levels in the data set from yeast chemostat cultures showing 40-min metabolic oscillations (26). Nodes are identified by gene names and color-coded to indicate the cell process in which they participate (there is some ambiguity in assigning genes to categories). The solid blue lines denote positive couplings, and the dashed red lines denote negative couplings. The identity of the hubs circled in red is discussed in the text.

visualized the subnetwork exhibiting the strongest 110 interactions (Fig. 3 and SI Table 1) from the full network comprising 169,071 pairwise interactions among the 582 genes exhibiting the largest profile fluctuations. The number of interactions selected for this analysis is somewhat arbitrary and the general features of the strongly interacting part of the graph do not change significantly when this number is modestly altered. The gene interaction network comprising the genes showing the strongest couplings is highly interconnected. The single pair of genes (*Dal4* and *Gap1*) not connected to the rest of the network in Fig. 3 becomes connected if a slightly larger subset of genes is included in the graph. Moreover, the network nodes vary substantially in their connectivity, with some genes, designated hubs, exhibiting strong pairwise interactions with many genes. The highly interconnected network structure is observed for the genes exhibiting the strongest interactions, while a comparable graph of the weakest 110 pairwise interactions among the 582 genes is largely disconnected (SI Fig. 5), as are graphs both from random networks and from networks deduced from randomized data using the maximum entropy method, as illustrated in SI Fig. 6 (28).

The maximum entropy network identifies connections between genes involved in diverse cellular processes. To emphasize this diversity, the genes participating in the strongest pairwise interactions have been color-coded by metabolic function in Fig. 3. This diversity of interconnected functions stands in marked contrast to the results obtained with widely used clustering approaches based on profile similarity (29). Correlation clustering identifies genes involved in common functions: the expression levels of genes involved in mitochondrial functions and protein synthesis, for example, exhibit well correlated peaks of expression at different points in the yeast metabolic oscillations (26, 27).

Yeast strongly prefers glucose or fructose over other carbon sources, rapidly fermenting either sugar to ethanol even under aerobic conditions, while also storing energy in the form of glycogen and trehelose (30). When sugar is abundant, genes encoding enzymes required for utilization of other carbon sources are repressed, as are genes encoding proteins of the mitochondrial tricarboxylic acid cycle, and gluconeogenesis, while genes encoding glycolytic enzymes, hexose transporters and ribosomal protein genes are activated (31). Conversely, when a yeast culture growing on a glucose-containing medium depletes it of glucose, it up-regulates genes encoding enzymes involved in respiration and other mitochondrial functions and down-regulates genes involved in other cellular functions, such as protein synthesis (32). At low rates of nutrient supply, yeast growing in chemostat cultures become synchronized and oscillate between primarily fermentative and oxidative metabolic states with a regular period (33). These alternations entail profound changes in the machinery for making proteins, the activity of mitochondria, transcription, translation and DNA replication (34). As illustrated in Fig. 4, the partially overlapping target of rapamycin (TOR) and protein kinase A (PKA) path-

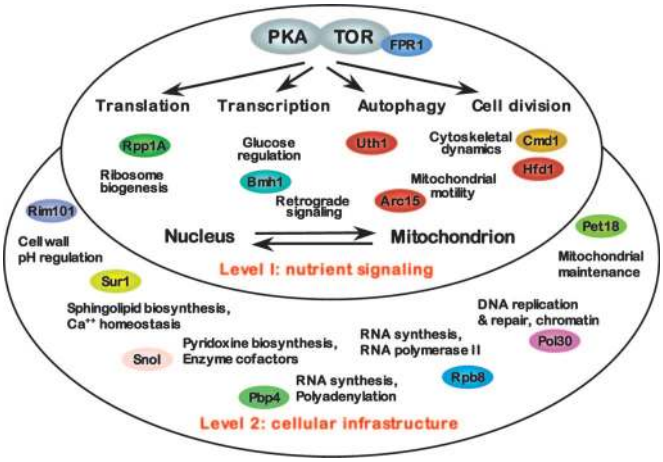


**Fig. 4.** A diagrammatic representation of the cellular processes identified by the network hubs among the 582 (level 1) and the 1,008–2,000 (level 2) genes exhibiting the most marked fluctuations in transcript levels during 40-min metabolic oscillations. PKA and TOR represent the PKA and TOR nutrient signaling pathways; other ovals contain the designations of hub genes identified as described in the text and color-coded by cell process as in Fig. 3 (see *SI Text* for details and references on the hub genes).

ways are primary mediators of nutrient signaling in yeast (35, 36). They can be regarded as "master" regulators, controlling transcription, translation, mRNA stability, nutrient uptake, communication between the mitochondrion and the nucleus and cell division in response to changes in carbon and nitrogen nutrient supplies (35, 36). TOR and PKA signaling are, in turn, mediated by a variety of proteins specific to each cellular process.

**Network Hubs Encode Key Cellular Proteins in Nutrient Signaling.** Strikingly, the hubs in the pairwise gene interaction network shown in Fig. 3 encode proteins involved in the critical processes that tune cell growth and division to the nutrient supply (Fig. 4). Among the seven genes with more than six edges subjected to detailed analysis, three encode proteins involved in TOR signaling (Fpr1, Bmh1, and Uth1), two are outer mitochondrial membrane proteins (Hfd1 and Arc15), one is a ribosomal protein (Rpp1A), and one encodes calmodulin (Cmd1) (see *SI Text* for additional details and references for the hub genes). Briefly, Bmh1, Fpr1, and the mitochondrial protein Uth1 interconnect the TOR pathway with the metabolic and physical state of the mitochondrion, as well as with the retrograde signaling system that adjusts expression of nuclear genes encoding mitochondrial proteins in response to changes in nutrient supply (37–39). Rpp1A is a component of the ribosomal stalk and may be a translational regulatory protein; transcription and stability of both its mRNA and those of other ribosomal proteins is regulated through the TOR signaling pathway (36, 40). Cmd1 and the mitochondrial Arc15 and Hfd1 proteins are involved in the actin cytoskeletal dynamics that are essential for endocytosis, cell division and mitochondrial motility; these interconnect with the TOR signaling pathway through the Fpr1 protein (41, 42). The strongest pairwise gene interaction detected in the subset of 582 genes (SI Table 1) genes is between *Fpr1* and *Ssa1*, a gene that encodes a key regulator of the overlapping PKA nutrient signaling pathway (43).

**Including More Genes.** We asked how the network structure changes when more genes are included in the analysis. When the number of genes is expanded to 1,008, still well above the noise level, hubs representing such fundamental cellular processes as pH regulation and cell wall biosynthesis (*Rim101*), DNA replication (*Pol30*), pyridoxine biosynthesis (*Sno1*), mitochondrial organization, and biogenesis (*Pet18*) are added to the network, although all of the original seven hubs are still represented among the genes showing the strongest interactions (SI Fig. 7; see *SI Text* for additional details and references for hub genes). Further expansion of the gene set to 1,500 and 2,000 adds genes involved in mRNA biogenesis (*Pbp4* and *Rpb8*) and sphingolipid biosynthesis (*Sur1*). Because of the initial ranking of genes by the magnitude of the transcript fluctuations during metabolic oscillations, expansion of the analyzed subset incorporates progressively more genes that show less marked variation in transcript abundance. These progressive expansions add genes whose genetic and physiological analysis shows them to be important in the more basic cellular processes of DNA replication, transcription and metabolism (Fig. 4). Not surprisingly, the genes encoding proteins involved in adjusting the cells' immediate metabolic state to the nutrient supply show the greatest variation in transcript abundance, while genes encoding proteins involved in cellular infrastructure show less marked fluctuations in the course of the metabolic adjustments.

**Three-Gene Interactions.** A study of the strongest three-gene interactions also identified genes that encode proteins likely to be important in regulating metabolic activity. The *Pnc1* gene, which is the most highly interconnected hub in the triplet network and is involved in 74 of the top 100 three-gene interactions, encodes a nicotinamide deaminase that plays a major role in yeast lifespan extension in response to caloric restriction, precisely the conditions of the experiments from which the data set was derived (44). The second most highly interconnected gene, participating in 66 of the strongest 100 three-gene interactions, is the *Tma19* gene, the yeast homolog of the well studied mammalian translationally controlled tumor protein (TCTP) gene, a calcium-binding protein that interacts with microtubules, regulates translation and exerts an apoptotic effect. The yeast Tma19 protein, which interacts with microtubules, exhibits redox-dependent translocation to mitochondria under stress conditions, and influences lifespan, may be a similar multifunctional protein (45).

**Gene Networks at Different Oscillatory Frequencies.** Metabolic oscillations of markedly different periodicities have been reported under different regimes of nutrient dilution and oxygen supply (33, 46). To determine whether the different periods are associated with similar or different states of the genetic and cellular network, we compared the data set obtained from cultures exhibiting a 40-min period of oscillation (26) with transcript data obtained from cultures exhibiting a 5-h oscillatory period (27). Correlation clustering yields superficially similar results, identifying groups of coexpressed genes that encode proteins involved in amino acid and protein synthesis, RNA metabolism, sulfur metabolism, DNA replication and mitosis, as well as in mitochondrial structure and function (26, 27). However, although the categories are the same and roughly equally represented in both data sets, there is little overlap in the genes represented in each category (SI Fig. 8*a*). Moreover, pairs of genes whose expression patterns are highly correlated in one data set are not necessarily correlated in the other (SI Fig. 8*b*). The genetic network inferred from the long-period data set using the entropy maximization method described here also differs from that extracted from short-period data set (SI Fig. 9 and SI Table 2). However, the *Rpp1A* hub is common to both networks; this and several additional ribosomal protein gene hubs in the long-period network are all regulated through the TOR signaling pathway (47). Moreover, mitochondrial protein genes, albeit different ones, constitute hubs in both short- and long-period networks (see *SI Text* for additional details and references for hub genes). We conclude that although some of the same signaling pathways are in-

volved, rather different states of the gene network support the observed short- and long-period metabolic oscillations.

## Discussion

The novelty of the present work lies in the ability of our method to identify genes that code for important cellular signaling and regulatory proteins controlling yeast nutrient responses from gene expression data alone. That is, the most strongly interacting and highly interconnected genes of the inferred pairwise gene interaction network for the short-period data set encode key control proteins. This contrasts markedly with the results of the "clustering" methods widely used today to analyze microarray data. Such correlation-based methods identify genes whose expression profiles are similar; these can be thought of as "members of the same choir," under the direction of common regulator or "conductor." The present network inference method identifies the conductors. Correlation-based analytical methods were used to identify coordinately regulated groups of ribosomal protein and mitochondrial genes in the data derived from yeast cultures exhibiting short-period metabolic oscillations (26). By contrast, the *Fpr1* and *Bmh1* hub genes of the network derived here from the same data set encode key components of the molecular machinery that regulates expression of all ribosomal protein genes and multiple mitochondrial genes, respectively (37, 38). For example, the rapamycin-binding *Fpr1*-encoded FK506-binding protein 12 (FKBP12) mediates the direct interaction of Tor1 kinase with chromatin to regulate transcription of both ribosomal protein and RNA genes (48, 49). Evidence is accumulating that the Tor kinase and prolyl isomerases, such as FKBP12, associate with and directly modulate histone acetylases and deacetylases at Tor target genes (48–50) (also see *SI Text*).

Perhaps the most striking result of the present analysis is that interconnections among the several cellular processes that mediate the concerted periodic genetic and metabolic shifts observed in nutrient-limited yeast chemostat cultures are reflected in gene interactions. That is, the present method can detect couplings between genes coding for proteins involved in different cellular processes, such as protein synthesis, cell division, and mitochondrial motility, which must be coordinated in response to nutrient availability. These observations reveal that there is more information about system dynamics in gene expression profiles than had been extracted previously, underscoring the integration of the cellular and genetic aspects of cell function. Our methodology is therefore likely to be useful in identifying key players in cellular networks of systems that are less well characterized than yeast. By facilitating analysis of the intact networks, the methodology we have developed should also make it possible to monitor the impact of subtle modifications of, for example, key signaling components on network function. Finally, the success of the present approach in extracting meaningful genetic connections indicates that the entropy maximization concept will be useful in understanding living systems, as it has been for other complex, nonequilibrium systems.

## Methods

Let the state vector $\mathbf{x} = (x_1, \ldots, x_N)$ denote the expression levels of the $N$ genes that are probed in a microarray experiment, and let $\rho(\mathbf{x})$ denote the probability that the genome is in the arbitrary state $\mathbf{x}$. We determine $\rho(\mathbf{x})$ by maximizing the Shannon entropy, $S = -\sum_{\mathbf{x}} \rho(\mathbf{x}) \ln\rho(\mathbf{x})$, subject to the constraint that $\rho(\mathbf{x})$ is normalized and that its first moment, $\langle x_i \rangle$, and second moment, $\langle x_i\, x_j \rangle$, coincide with those derived from the expression data. This procedure leads to a Boltzmann-like distribution $\rho(\mathbf{x}) \sim e^{-H}$, where $H = \frac{1}{2} \sum_{ij} x_i\, M_{ij}\, x_j$ plays the role of the energy function in conventional statistical mechanics. Thus, the matrix element $M_{ij}$ has the natural interpretation of the interaction between genes $i$ and $j$. The general result for linear systems, the derivation of which is given in *SI Text*, is that the matrix of interactions between genes can be obtained by inverting the matrix of their covariances, $M_{ij}^{-1} = C_{ij} = \langle x_i\, x_j \rangle - \langle x_i \rangle \langle x_j \rangle$, where the average of any generic quantity $z$ is defined as $\langle z \rangle = \int d^N x\, \rho(\mathbf{x})\, z$ and the integral is over the space spanned by the expression levels of N genes.**

The covariance matrix ($C_{ij}$) can readily be obtained from the gene expression data. However, the number of microarray samples in a typical microarray data set is much smaller than the number of genes, and therefore the covariance matrix is noninvertible. We use spectral decomposition to get around this difficulty, taking $M$ to be the inverse of $C$ in the non-zero eigenspace corresponding to the subspace spanned by the gene expression data, yielding $M_{ij} = \sum_k \omega_k^{-1}\, v_i^k\, v_j^k$, where $\omega_k$ is the $k$th eigenvalue of $C$, $v^k$ is its corresponding eigenvector, and the sum is over all of the non-zero eigenvalues. The matrix $C$ can be expressed as $C_{ij} = \sum_k \omega_k\, v_i^k\, v_j^k$. It should be noted that the eigenvectors with large eigenvalues contribute the most to $C$ but have little effect on $M$. The gross features of the data are captured in these eigenvectors, and therefore such general features indicate little about the nature of the couplings between genes. On the other hand, the eigenvectors with small eigenvalues dominate the calculation of $M$. These eigenvectors correspond to the residual fluctuations in expression levels that remain when the common, large-scale fluctuations are removed.

The elements of the matrix $M$ are, by definition, the effective pairwise gene interactions that reproduce the gene profile covariances exactly while maximizing the entropy of the system. The method is readily generalizable to higher-order interactions in perturbation theory (see *SI Text*). The strength and the sign of the interaction represent the mutual influence on each other of the expression levels of a pair of genes. This is necessarily indirect, because gene interactions are mediated by proteins. The magnitude of the element $M_{ij}$ is a measure of the strength of the net interaction between genes $i$ and $j$. The sign of the interaction indicates the nature of the coupling: a negative coupling between genes indicates that a change in expression level of either gene is accompanied by a similar change in the expression level of the other gene. Conversely, a positive coupling indicates that a change in one is accompanied by an opposite change in the other. The diagonal element $M_{ii}$ provides a measure of the influence that gene $i$ has on the whole network. Nodes with large diagonal values have strong couplings with several other nodes, whereas nodes with smaller diagonal elements generally have couplings of lesser magnitude. The gene couplings integrate all of the influences not considered as part of the network (see SI Fig. 10). It should be noted, however, that the nature of the correlation between the expression profiles of two genes cannot be deduced directly from their coupling.

**This is a robust result for linear systems and can be derived in several ways. An alternative way of arriving at this result without invoking the maximization of entropy follows from the assumptions that $\ln \rho(\mathbf{x})$ peaks at $\mathbf{x}^{(0)}$, is normalizable, and is a smooth function that can be expressed in a Taylor expansion up to quadratic order: $\ln \rho(\mathbf{x}) = \ln \rho(\mathbf{x}^{(0)}) - (1/2)\Sigma_{ij} (x_i - x_i^{(0)})\, M_{ij}\, (x_j - x_j^{(0)}) + \ldots$, where the neglected terms are of cubic order in $(x_i - x_i^{(0)})$ and $-M$, the matrix of the second derivative of $\ln\rho(\mathbf{x})$ with respect to $\mathbf{x}$, is negative definite. Note that $\mathbf{x}^{(0)} = \langle \mathbf{x} \rangle$. Within this Gaussian approximation, one again obtains the result that $M$ is the inverse of $C$. Not surprisingly, this same result is found in the graphical Gaussian model, in which expression level data are assumed to be drawn from a Gaussian distribution (12).

This work was supported in part by Ministero per l'Università e per la Ricerca Scientifica e Tecnologica Programma di Ricerca Cofinanziato 2005, Istituto Nazionale di Fisica Nucleare, National Aeronautics and Space Administration Exploration Systems Mission Directorate, National Science Foundation Integrative Graduate Education and Research Traineeship DGE-9987589, Ministry of Science in Poland Grant 2P03B-03225, and the Willaman Professorship endowment.

1. Pan W (2002) *Bioinformatics* 18:546–554.
2. D'Haeseleer P, Liang S, Somogyi R (2000) *Bioinformatics* 16:707–726.
3. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) *Proc Natl Acad Sci USA* 95:14863–14868.
4. Saldanha AJ, Brauer MJ, Botstein D (2004) *Mol Biol Cell* 15:4089–4104.
5. Liang S, Fuhrman S, Somogyi R (1998) *Pac Symp Biocomput*, 18–29.
6. Akutsu T, Miyano S, Kuhara S (2000) *J Comput Biol* 7:331–343.
7. Chen KC, Csikasz-Nagy A, Gyorffy B, Val J, Novak B, Tyson JJ (2000) *Mol Biol Cell* 11:369–391.
8. Shmulevich I, Dougherty ER, Kim S, Zhang W (2002) *Bioinformatics* 18: 261–274.
9. Friedman N (2004) *Science* 303:799–805.
10. Friedman N, Linial M, Nachman I, Pe'er D (2000) *J Comput Biol* 7:601–620.
11. Toh H, Horimoto K (2002) *Bioinformatics* 18:287–297.
12. Schafer J, Strimmer K (2005) *Bioinformatics* 21:754–764.
13. Butte AJ, Kohane IS (2000) *Pac Symp Biocomput*, 418–429.
14. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) *Science* 292:929–934.
15. Tegner J, Yeung MK, Hasty J, Collins JJ (2003) *Proc Natl Acad Sci USA* 100:5944–5949.
16. Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) *Science* 301:102–105.
17. Li S, Wu L, Zhang Z (2006) *Bioinformatics* 22:2143–2150.
18. Yeung MK, Tegner J, Collins JJ (2002) *Proc Natl Acad Sci USA* 99:6163–6168.
19. Boltzmann L (1964) *Lectures on Gas Theory* (Cambridge Univ Press, London).
20. Shannon CE (1948) *Bell Syst Tech J* 27:379–423.
21. Jaynes ET (1957) *Phys Rev* 106:620–630.
22. Jaynes ET (1957) *Phys Rev* 108:171–190.
23. Dewar R (2003) *J Phys A* 36:631–641.
24. Dewar RC (2005) *J Phys A* 38:L371–L381.
25. Schneidman E, Berry MJ, II, Segev R, Bialek W (2006) *Nature* 440:1007–1012.
26. Klevecz RR, Bolen J, Forrest G, Murray DB (2004) *Proc Natl Acad Sci USA* 101:1200–1205.
27. Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) *Science* 310:1152–1158.
28. Albert R, Barabasi AL (2002) *Rev Mod Phys* 74:47–95.
29. Bono H, Okazaki Y (2002) *Curr Opin Struct Biol* 12:355–361.
30. Futcher B (2006) *Genome Biol* 7:107.
31. Gelade R, Van de Velde S, Van Dijck P, Thevelein JM (2003) *Genome Biol* 4:233.
32. DeRisi JL, Iyer VR, Brown PO (1997) *Science* 278:680–686.
33. Richard P (2003) *FEMS Microbiol Rev* 27:547–557.
34. Xu Z, Tsurugi K (2006) *FEBS J* 273:1696–1709.
35. Rohde JR, Cardenas ME (2004) *Curr Top Microbiol Immunol* 279:53–72.
36. Chen JC, Powers T (2006) *Curr Genet* 49:281–293.
37. Lorenz MC, Heitman J (1995) *J Biol Chem* 270:27531–27537.
38. Bertram PG, Zeng C, Thorson J, Shaw AS, Zheng XF (1998) *Curr Biol* 8:1259–1267.
39. Camougrand N, Kissova I, Velours G, Manon S (2004) *FEMS Yeast Res* 5:133–140.
40. Santos C, Ballesta JP (2005) *Mol Microbiol* 58:217–226.
41. Boldogh IR, Yang HC, Nowakowski WD, Karmon SL, Hays LG, Yates JR, III, Pon LA (2001) *Proc Natl Acad Sci USA* 98:3162–3167.
42. Schaerer-Brodbeck C, Riezman H (2003) *FEMS Yeast Res* 4:37–49.
43. Geymonat M, Wang L, Garreau H, Jacquet M (1998) *Mol Microbiol* 30:855–864.
44. Anderson RM, Bitterman KJ, Wood JG, Medvedik O, Sinclair DA (2003) *Nature* 423:181–185.
45. Rinnerthaler M, Jarolim S, Heeren G, Palle E, Perju S, Klinger H, Bogengruber E, Madeo F, Braun RJ, Breitenbach-Koller L, *et al.* (2006) *Biochim Biophys Acta* 1757:631–638.
46. Parulekar SJ, Seamons GB, Rolf MJ, Lim HC (1986) *Biotechnol Bioeng* 28:700–710.
47. Rohde J, Heitman J, Cardenas ME (2001) *J Biol Chem* 276:9583–9586.
48. Rohde JR, Cardenas ME (2003) *Mol Cell Biol* 23:629–635.
49. Tsang CK, Bertram PG, Ai W, Drenan R, Zheng XF (2003) *EMBO J* 22:6045–6056.
50. Li H, Tsang CK, Watkins M, Bertram PG, Zheng XF (2006) *Nature* 442:1058–1061.