Versatility and Connectivity Efficiency of Bipartite Transcription Networks

Mark P. Brynildsen, Linh M. Tran, and James C. Liao
Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, California

ABSTRACT The modulation of promoter activity by DNA-binding transcription regulators forms a bipartite network between the regulators and genes, in which a smaller number of regulators control a much lager number of genes. To facilitate representation of gene expression data with the simplest possible network structure, we have characterized the ability of bipartite networks to describe data. This has led to the classification of two types of bipartite networks, versatile and nonversatile. Versatile networks can describe any data of the same rank, and are indistinguishable from one another. Nonversatile networks require constraints to be present in data they describe, which may be used to distinguish between different network topologies. By quantifying the ability of bipartite networks to represent data we were able to define connectivity efficiency, which is a measure of how economic the use of connections is within a network with respect to data representation and generation. We postulated that it may be desirable for an organism to maximize its gene expression range per network edge, since development of a regulatory connection may have some evolutionary cost. We found that the transcriptional regulatory networks of both *Saccharomyces cerevisiae* and *Escherichia coli* lie close to their respective connectivity efficiency maxima, suggesting that connectivity efficiency may have some evolutionary influence.

INTRODUCTION

Bipartite networks have been used to represent many biological systems and engineering tasks, including gene expression regulation (1–6), signal processing (7,8), image processing (9–11), and spectrum analysis (12,13). These networks consist of a layer of sources connected to a layer of outputs, where every connection (edge) represents the influence of a source on an output (Fig. 1 *A*). In some cases, the output nodes are fully connected to the sources, for example, microphones recording simultaneous speeches in the same location. In others, the outputs are sparsely connected to the source signals, such as in transcriptional regulatory networks.

In general, it is advantageous to describe data with the simplest structure possible, both for interpretation and mechanistic reasons (14,15). However, conventional bipartite network analyses such as principal component analysis (PCA) and independent component analysis, assume that networks are fully connected. For systems governed by sparsely connected networks, this assumption could lead to the deduction of unrealistic source signals (4,14,16,17). A variation of PCA, called Sparse PCA, has been developed that acknowledges this issue and attempts to alleviate it (16,17). However, Sparse PCA like its precursor, PCA, requires deduced source signals to be mutually orthogonal. Such a mathematical constraint without any phenomenological justification may hinder the ability to provide simple representation, especially if the simplest structure may require oblique source signals (14,15). A complementary approach, network component analysis (NCA), takes into account known network connectivity in deducing source signals and allows for orthogonal and oblique source signals (4). However, if the a priori network connectivity has some degree of uncertainty, as in the case of ChIP-chip data being used to analyze DNA-microarray data, there may be simpler connectivities capable of describing the same data. Alternatively, exploratory factor analysis attempts to simplify structure by performing orthogonal or oblique rotations on a factorization. While the goal of this technique is to achieve simplicity of structure, the implementation has had difficulty with situations where the complexity of the simplest network exceeds that of maximal sparsity (one connection per output to the source layer) (15). To facilitate data representation with the simplest structure possible, we have characterized the ability of bipartite networks to describe data.

The ability of bipartite networks to describe data may be limited by network connectivity. In some cases, such as fully connected networks, any data within the span of the network can be described, while in other cases, such as sparsely connected networks, certain elements of the data may be required to lie on a single line or hyperplane. This leads to the classification of two types of bipartite networks, those networks whose output range is not limited by their connectivity, which we will term ''versatile'', and those networks whose output range is hindered by their connectivity, which we will term ''nonversatile''. Intuitively, one might think that any missing edge from a network might compromise its ability to describe data, and therefore any network besides a fully connected network will be nonversatile. However, this is not true, and there are networks that are not fully connected that can represent data equally as well as fully connected networks. These networks are also versatile and are not limited by their connectivity. The very existence of these networks demonstrates that there is no justification from data alone to conclude more than minimal versatile connectivity.

Submitted February 2, 2006, and accepted for publication June 12, 2006.

Address reprint requests to J. C. Liao, Tel.: 310-825-1656; E-mail: liaoj@ucla.edu.

© 2006 by the Biophysical Society

0006-3495/06/10/2749/11 $2.00

doi: 10.1529/biophysj.106.082560

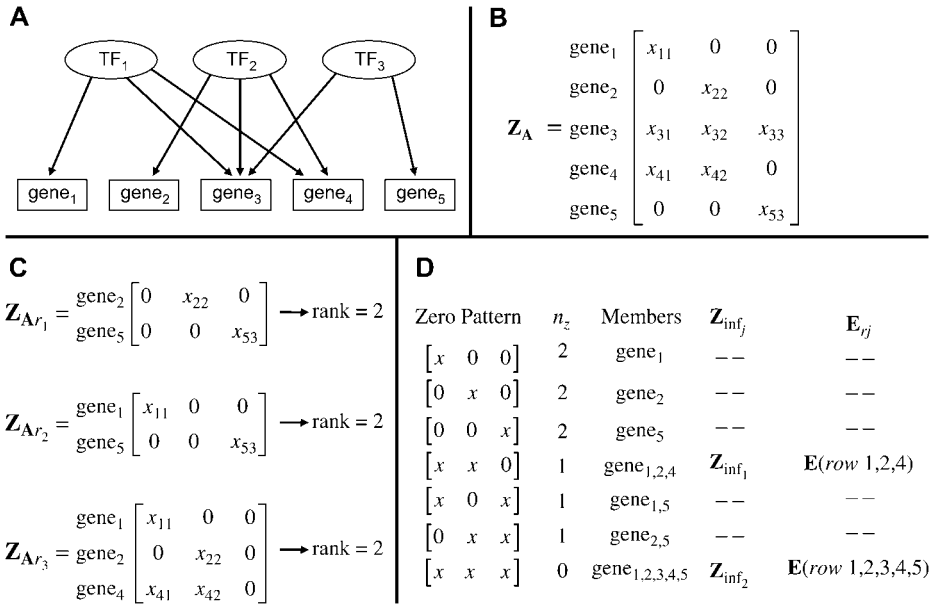

FIGURE 1 (*A*) Bipartite network depicting a hypothetical transcriptional regulatory network. (*B*) $\mathbf{Z}_A$ corresponding to network in panel *A*. (*C*) $\mathbf{Z}_{\mathbf{A}r_i}$ created from $\mathbf{Z}_A$ in panel *B*. (*D*) Table of $\mathbf{Z}_{v_j}$, $\mathbf{Z}_{inf_j}$, n_z, and $\mathbf{E}_{rj}$ from $\mathbf{Z}_A$ in panel *B*.

Thereby, the most complex structure ever needed to describe data is the minimal versatile connectivity. Nonversatile networks, on the other hand, have their own utility, since their constraints are often present in datasets. Since nonversatile networks are often sparser than versatile networks they would provide the simplest representation under many circumstances.

In this article we define the minimal connectivity to achieve versatility, define the constraints present in nonversatile networks, discuss the implications of versatile and nonversatile networks, and suggest possible applications for their use. To demonstrate the utility of these concepts we examined the transcriptional regulatory networks of *Saccharomyces cerevisiae* and *Escherichia coli*. We recognized that for bipartite networks the ability to represent data is equivalent to the ability to generate data. With this in mind, we defined connectivity efficiency, which is a measure of how economic the use of connections is within a network with respect to data representation/generation ability. We then analyzed the connectivity efficiencies of the transcriptional regulatory networks of *S. cerevisiae* and *E. coli*. We postulated that it may be biologically desirable for organisms to maximize their gene expression range (breadth of possible gene expression profiles) per network edge, since development of a regulatory connection may have some evolutionary cost. Subsequently, we found that both networks lay close to their respective connectivity efficiency maxima, suggesting that connectivity efficiency may have some evolutionary influence.

BACKGROUND

We are interested in the ability of bipartite networks to represent data. A bipartite network represents an output $e_i(t)$ by the linear mixing of sources, $p_j(t)$, through a mixing rule described by

$$e_i(t) = \sum_{j=1}^{L} a_{ij} p_j(t), \tag{1}$$

where a_{ij} values are the connectivity strengths. The mixing rule can be written in a matrix form,

$$\mathbf{E} = \mathbf{AP}, \tag{2}$$

where $\mathbf{E}$ is the output data ($N \times M$), $\mathbf{A}$ is the matrix of network connectivity strengths ($N \times L$), and $\mathbf{P}$ is the collection of source signals ($L \times M$). Bipartite network representation can further be generalized by considering only the connectivity pattern of matrix $\mathbf{A}$,

$$\mathbf{Z}_A = \{\mathbf{A} \in \mathbb{R}^{NxL} | a_{ij} = 0, \text{ for a given set of } (i,j)\}, \tag{3}$$

where the values of the nonzero a_{ij} are left unconstrained and can take on any value—positive, negative, or zero. For the purpose of this article, networks with varying connectivity strengths but the same connectivity pattern, $\mathbf{Z}_A$, will be discussed identically.

Versatile networks

Ideally, we would prefer to represent data with the simplest network connectivity possible. In the context of this work the simplicity and sparsity of networks will be synonymous. Thus, we seek to find the sparsest network connectivity that can reliably represent data. Naturally, we begin by considering networks that can represent any data. These networks are termed versatile, and are characterized by the following theorem.

Theorem 1

A linear bipartite network with connectivity pattern $\mathbf{Z}_A$ ($N \times L$) can describe any data within $\mathbb{R}^{\mathbf{L}}$, if all reduced forms of $\mathbf{Z}_A$, $\mathbf{Z}_{\mathbf{Ar}_i}(z_i \times L)$, are full row rank.

Here, $\mathbf{Z}_{\mathbf{Ar}_i}$ is defined as the rows of $\mathbf{Z}_A$ which contain zeros in the i^{th} column of $\mathbf{Z}_A$, where z_i is the number of zeros in the i^{th} column of $\mathbf{Z}_A$. To test this, consider the nonzero entries of $\mathbf{Z}_{\mathbf{Ar}_i}$ as nonzero random values that cannot combine on their own to produce a rank deficiency.

To demonstrate use of Theorem 1 we have provided a hypothetical transcriptional regulatory network in Fig. 1 *A*, transformed the network into $\mathbf{Z}_A$ form (Fig. 1 *B*), and determined all $\mathbf{Z}_{\mathbf{Ar}_i}$ (Fig. 1 *C*). Both $\mathbf{Z}_{\mathbf{Ar}_1}$ and $\mathbf{Z}_{\mathbf{Ar}_2}$ are full row rank, but $\mathbf{Z}_{\mathbf{Ar}_3}$ is not, and therefore the network in Fig. 1 *A* does not satisfy Theorem 1. For a network that would satisfy Theorem 1, simply connect TF_3 to the first, second, or fourth gene. The proof of this theorem along with examples is presented in Appendix A.

A consequence of the versatility theorem is that all connectivity patterns that satisfy the required criterion will represent data equally. This means that there may exist a minimal connectivity that satisfies the criterion, which may be used to represent data created from denser network structures. To determine the minimal connectivity (sparsest network) to achieve versatility we must find the limit of the criterion. To do so we recognize that $\mathbf{Z}_{\mathbf{Ar}_i}$ can only be full row rank if $z_i < L$ for every column of $\mathbf{Z}_A$. Therefore, the minimal connectivity to achieve versatility contains $L(L-1)$ missing edges, specifically $(L-1)$ per column of $\mathbf{Z}_A$. However, not all network connectivities with $(L-1)$ missing edges per column are versatile. Any network must still be in compliance with the above criterion to be versatile, even if it has the same number of, or a lesser number of missing edges than the minimal connectivity to achieve versatility.

Minimal connectivity for versatility is maximal connectivity for NCA-compliance

Interestingly, there exists a relationship between the minimal connectivity to achieve versatility and NCA. To guarantee the uniqueness of NCA solutions there are three criteria that must be satisfied. The second criterion in Liao et al. (4) deals with the connectivity pattern, $\mathbf{Z}_A$, and the ranks of its reduced forms, which are essentially identical to the reduced forms described here. It states that the rank of every reduced form, $\mathbf{Z}_{\mathbf{Ar}_i}$, must be $(L-1)$. The maximum rank for any $\mathbf{Z}_{\mathbf{Ar}_i}$ is $(L-1)$, and can only be achieved if $z_i \geq (L-1)$. Therefore, a necessary condition for NCA-compliance is that a network must have a minimum of $(L-1)$ zeros per column.

Versatility requires that all $\mathbf{Z}_{\mathbf{Ar}_i}$ be full row rank. The maximum row rank of $\mathbf{Z}_{\mathbf{Ar}_i}$ is $(L-1)$, and can only be achieved if $z_i = (L-1)$. This corresponds to the minimal connectivity to achieve versatility described previously. Thus, the minimum connectivity to achieve versatility requires all $\mathbf{Z}_{\mathbf{Ar}_i}$ to have $z_i = (L-1)$ and be of rank $(L-1)$, and the maximum connectivity (largest number of nonzero connections) to be NCA-compliant requires all $\mathbf{Z}_{\mathbf{Ar}_i}$ to have $z_i = (L-1)$ and be of rank $(L-1)$. Therefore, the minimum connectivity to achieve versatility is equivalent to the maximum NCA-compliant connectivity. To illustrate, examples have been provided in Appendix A.

Nonversatile networks

Nonversatile networks are those connectivity patterns that do not satisfy the versatility criterion. These networks have a reduced ability to represent data compared to that of versatile networks. Fig. 2 illustrates this concept, where the *y* axis is a measure of data representation capability (versatility index) that will be defined in the next section, and the *x* axis is the number of edges in the network. The reduced ability of nonversatile networks to represent data is due to connectivity constraints that dictate the type of data the network is able to describe. However, these constraints are often present in datasets, leading to the possibility of data representation with simpler structures than versatile networks. Therefore, we have characterized these constraints in the following theorem, such that they may aid in simplifying network structures used to represent data.

Definitions

1. A zero pattern is a $1 \times L$ vector that indicates, by the position of zero entries, which transcription factors (TF) do not control expression of a gene. The number of zero entries in a zero pattern is designated by n_z. A system with three TFs has seven possible zero patterns, which are shown in Fig. 1 *D*. The zero pattern $[x \quad x \quad 0]$ indicates that a gene is not controlled by TF_3.

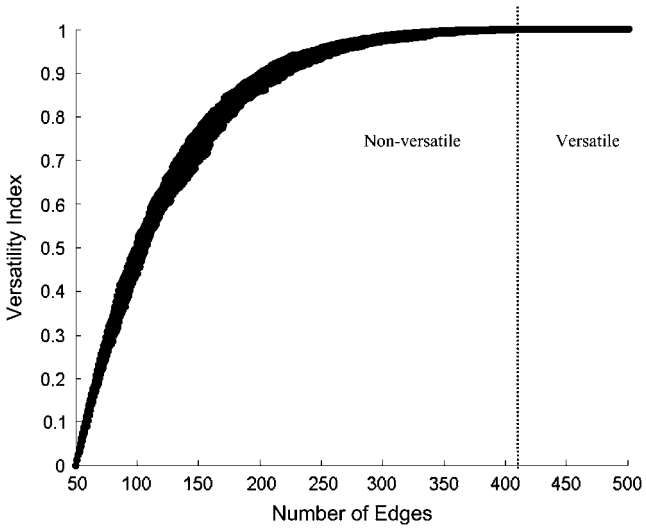

FIGURE 2 Plot of versatility index versus number of edges, for >10,000 networks with 50 outputs and 10 sources.

2. Any gene that satisfies the definition of a zero pattern is a member of that zero pattern. For instance, the zero pattern $[x \quad x \quad 0]$ requires genes to not be regulated by TF_3, therefore, $gene_{1,2,4}$ are all members.
3. An informative zero pattern, $\mathbf{Z}_{\mathrm{inf_j}}$, is any zero pattern with $> L - n_{\mathrm{zj}}$ members, where n_{zj} is equal to the number of zeros in $\mathbf{Z}_{\mathrm{inf_j}}$. Fig. 1 *A* has two $\mathbf{Z}_{\mathrm{inf_j}}$, $\mathbf{Z}_{\mathrm{inf_1}} = [x \quad x \quad 0]$ and $\mathbf{Z}_{\mathrm{inf_2}} = [x \quad x \quad x]$.
4. $\mathbf{E}_{\mathrm{rj}}$ is a matrix composed of the genes (rows of $\mathbf{E}$) that are members of $\mathbf{Z}_{\mathrm{inf_j}}$. From Fig. 1, $\mathbf{E}_{\mathrm{r1}} = \mathbf{E}$(rows 1, 2, 4).

Theorem 2

Any dataset, $\mathbf{E}$ ($N \times L$), may be represented by a linear bipartite network characterized by connectivity pattern $\mathbf{Z}_{\mathrm{A}}$ ($N \times L$) if every $\mathbf{E}_{\mathrm{rj}}$ has rank $\leq (L - n_{\mathrm{zj}})$.

Fig. 1 *D* summarizes all items needed to evaluate Theorem 2. As one can see, only two $\mathbf{Z}_{\mathrm{inf_j}}$ ($[x \quad x \quad 0]$, $[x \quad x \quad x]$) exist and need to be evaluated by $\mathbf{E}_{\mathrm{rj}}$. For a dataset to be represented by the network in Fig. 1 *A*, $\mathbf{E}_{\mathrm{r1}}$ must have a rank ≤ 2, and $\mathbf{E}_{\mathrm{r2}}$ must have rank ≤ 3. The proof of this theorem along with examples is presented in Appendix B. The theorem identifies bipartite connectivity constraints from $\mathbf{Z}_{\mathrm{A}}$ that must be present within $\mathbf{E}$ for $\mathbf{Z}_{\mathrm{A}}$ to represent it. Theorem 2 may be used to check whether a dataset can be represented by a network. The procedure to use Theorem 2 is presented in Table A1 of Appendix B along with an example to illustrate its use.

It should be noted that Theorem 2 is general and can be applied to any bipartite network. In fact, if one were to check whether a dataset could be represented by a versatile network, $\mathbf{Z}_{\mathrm{A}}$ ($N \times L$), only one $\mathbf{Z}_{\mathrm{inf_j}}$ would be found that had $>(L - n_{\mathrm{zj}})$ members. This $\mathbf{Z}_{\mathrm{inf_j}}$ would not have any zero entries and would check whether the dataset was contained within $\mathbb{R}^{\mathbf{L}}$, a condition present in Theorem 1.

Implications of nonversatile networks

Although a dataset may satisfy Theorem 2 for a particular nonversatile network, the dataset may still contain additional constraints. This is due to the fact that constraints from nonversatile networks are nonunique. In fact, any network that can be created from another network by edge deletion (which we call the offspring networks) will have the same set of constraints or a larger set that contains the previous network's constraints. This means that the nonversatility criterion does not identify the minimal nonversatile connectivity to represent data, but simply identifies whether a dataset may be represented by a particular nonversatile network. To deduce the minimal nonversatile connectivity to represent data a method must be developed that can efficiently search for constraints in data, rather than see if data fits the constraints of a nonversatile network. This leads to the question of network reconstruction from constraints embedded in the data, which we will leave for the Discussion.

Connectivity efficiency

For bipartite networks the ability to represent data is equivalent to the ability to generate data. For transcriptional regulatory networks the ability to generate data would be the ability to generate gene expression. Knowing that transcriptional regulatory networks are generally sparse and that versatile networks of the same size would be fairly dense, we knew that transcriptional regulatory networks would not be versatile, and thus not have the maximal capability. With this in mind, we postulated that it may be desirable for organisms to maximize their gene expression ability per connection of the network, since it is safe to assume that there could be an evolutionary cost associated with the development of every regulatory interaction in the network.

First, we needed to define an index which could give us an indication of how close a network is to being versatile. We wanted the index to range from 0 to 1, where any network with a value of 1 would be versatile and any network with a value of 0 would be the most nonversatile (one connection per output to the source layer). We also required that if a network failed Theorem 1 for every $\mathbf{Z}_{\mathbf{A}\mathrm{r_i}}$, any edge deletion within the network would decrease its index. We require that the nonversatile network fail every $\mathbf{Z}_{\mathbf{A}\mathrm{r_i}}$, because those $\mathbf{Z}_{\mathbf{A}\mathrm{r_i}}$ that comply with Theorem 1 correspond to columns of $\mathbf{Z}_{\mathrm{A}}$ that are versatile in nature, and thus edge deletion may not change that if they have $<(L - 1)$ zeros. With these conditions in mind we defined the versatility index,

$$VI(\mathbf{Z}_{\mathbf{A}}) = 1 - \frac{\mathbf{Z}_{\mathbf{A}}^{\mathrm{c}} - (N - L)}{\max(\mathbf{Z}^{\mathrm{c}}) - (N - L)}, \tag{4}$$

where $VI(\mathbf{Z}_{\mathrm{A}})$ is the versatility index of $\mathbf{Z}_{\mathrm{A}}$, $\mathbf{Z}_{\mathrm{A}}^{\mathrm{c}}$ are the constraints imposed by $\mathbf{Z}_{\mathrm{A}}$, $\max(\mathbf{Z}^{\mathrm{c}})$ are the constraints from the most nonversatile network the same size as $\mathbf{Z}_{\mathrm{A}}$, N is equal to the number of outputs, and L the number of regulators. The method to determine $\mathbf{Z}_{\mathrm{A}}^{\mathrm{c}}$ and $\max(\mathbf{Z}^{\mathrm{c}})$ can be found within Appendix D. Both are based off of the principles detailed in Theorem 2. Subsequently, we can define the connectivity efficiency ($CE(\mathbf{Z}_{\mathrm{A}})$), as

$$CE(\mathbf{Z}_{\mathbf{A}}) = \frac{VI(\mathbf{Z}_{\mathbf{A}})}{\text{Number of edges in } \mathbf{Z}_{\mathbf{A}}}. \tag{5}$$

Connectivity efficiency (CE) is an average measure of how much each edge in a network contributes to the ability of that network to represent/generate data. We calculated the connectivity efficiency for the transcriptional regulatory networks of *S. cerevisiae* (CE = 4.7e-5) and *E. coli* (CE = 1.9e-4). While this might not appear significant, when plots of the versatile efficiencies from networks of the same size (same number of genes and regulators) and edge distribution are created, the versatile efficiency for *S. cerevisiae* is 87% of the maximum, and that of *E. coli* is 55% of the maximum, and both lie on the same shoulder of their respective maxima as depicted in Fig. 3. That shoulder represents networks that are sparser than the maximum, and its significance will be address in detail within the Discussion.

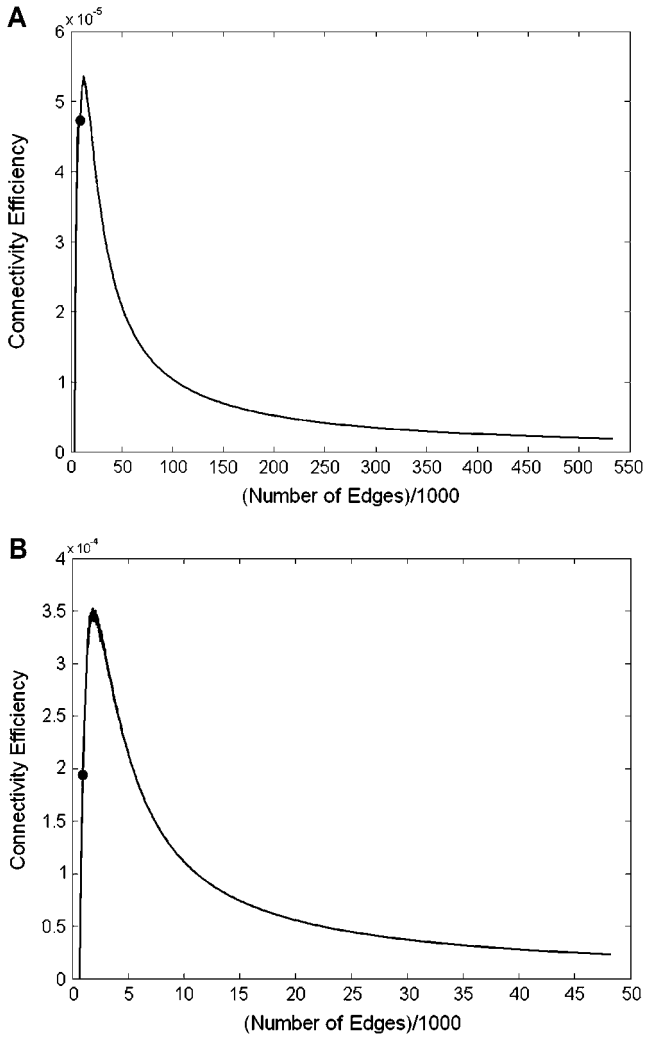

FIGURE 3 (*A*) Connectivity efficiency plot for the transcriptional regulatory network of *S. cerevisiae* (*circle*) plotted against networks of the same size (same number of regulators and genes), sampled from the same edge distribution, with a varying degree of edge density (*line*). (*B*) Connectivity efficiency plot for the transcriptional regulatory network of *E. coli* (*circle*) plotted against networks of the same size (same number of regulators and genes), sampled from the same edge distribution, with a varying degree of edge density (*line*).

DISCUSSION

Generally, it is desirable to describe data in the simplest possible manner. For systems governed by bipartite networks, this translates into describing data with the simplest possible structure. It has long been argued that simplicity of structure has more physical meaning than other considerations, such as orthogonality, during data representation (14). In fact, it has been shown that such abstract constraints yield erred results (4). In this work we have characterized the ability of bipartite networks to describe data, so as to facilitate data representation with the simplest possible structure. As we have shown, the ability of bipartite networks to describe data is dependent upon the network connectivity. Here we have classified bipartite networks into two categories based on their connectivity, versatile networks that do not have any restrictions imposed by their connectivity on the type of data they can describe, and nonversatile networks that do. This distinction gives rise to exclusive properties of each class that have implications for data representation, data compression, and network and source signal reconstruction.

Versatile networks can describe any data, and do not need to be fully connected. Therefore, the maximal connectivity necessary to describe any data would be the minimal versatile connectivity. This signifies the ability of some versatile networks to explain output generated from denser network structures. Theoretically, this would provide data compression capability superior to that of PCA. However, this capability comes at a cost. Since versatile networks are equally capable there is no way to discern the true network and source signals from data generated by versatile networks. Even if one were to assume that the true network was the minimal versatile connectivity, this would identify a whole class of networks that satisfy Theorem 1. The connections within the network would have no physical meaning since they could be rearranged in many different ways without impacting the system. This would be undesirable for situations where the actual arrangement of connections was of importance, such as in transcriptional regulatory networks. However, nonversatile networks do not share this deficiency.

Nonversatile networks are capable of describing a limited set of data. Restrictions that match those dictated by their network connectivity must be present in datasets for representation by them. This limitation, however, has its utility—since output created from nonversatile networks carry the connectivity restrictions derived from the original network. This enables network and source signal reconstruction on their outputs, and lends credibility to physical meanings attributed to their connections. Though reconstruction remains possible and seems plausible, efficient search algorithms must be designed to probe for connectivity restrictions from nonversatile networks. Whether these concepts will be incorporated into current techniques or form the basis of novel approaches, the additional complication of noise must be hurdled. While versatile networks can describe any data, including data riddled with noise, the restrictions left by nonversatile networks may be obscured by noise and more difficult to locate. This however, is an unavoidable complication when attempting to decipher underlying mechanisms, and does not change the basic principles of versatile and nonversatile network representation.

In addition, the concept of network versatility has been applied to the transcriptional regulatory network of *S. cerevisiae* and *E. coli*. Connectivity efficiency, which is an economic measure of connection usage, was calculated for the transcriptional regulatory networks of *S. cerevisiae* and *E. coli* and plotted against the connectivity efficiencies of other networks of the same size and sampled from the same distribution. It was found that the connectivity efficiencies of

S. cerevisiae and *E. coli* were 87% and 55% of the maximum of their respective plots, and that both were found on the same shoulder of their maxima. That shoulder represents networks that have fewer edges than the maximal efficient network. This is an important feature because the transcriptional networks of *S. cerevisiae* and *E. coli* are more likely to be missing connections than containing erred edges. Therefore, the true transcriptional networks of these organisms should approach the maximal versatile efficiency. In fact, Harbison et al. (18) claimed that the 203 transcription factors they performed genome-wide location analysis on is most likely to comprise all of the DNA-binding transcriptional regulators in *S. cerevisiae*, and that the false-positive rate of their analysis should be ~96% while the false-negative rate should be ~24%. Combined with the fact that the majority of open reading frames in *S. cerevisiae* have been found after its genome sequencing the size of the transcriptional network should not change much. Therefore, any addition of edges to the transcriptional network of yeast will invariably push the network toward the maximal versatile efficiency. For *E. coli*, since an analogous genome-wide location analysis has never been done, the likelihood for missing connections over erred connections seems to be even higher. These findings suggest that connectivity efficiency may be a quantity that transcriptional networks evolve to maximize.

In conclusion, we have characterized the ability of bipartite networks to represent data, which has led to the concepts of versatility and nonversatility. Both of these concepts have been derived, described, and discussed in detail. Lastly, we demonstrated the utility of these concepts by analyzing the connectivity efficiencies of *S. cerevisiae* and *E. coli*, which suggested that measures derived from these concepts, may have some biological or evolutionary importance.

METHODS

Transcriptional networks

S. cerevisiae: Using a p-value threshold of 1×10^{-3}, transcriptional regulatory networks were obtained from the ChIP-chip data of Lee et al. (19) and Harbison et al. (18) (YPD and all conditions). The networks were then merged to obtain a network comprised of all transcription factor-promoter binding relationships known through ChIP-chip experimentation.

Escherichia coli: The network was obtained by combining information from RegulonDB version 4 (20), Ver. 1.1 of Shen-Orr et al. (21), and Pernestig et al. (22). CsrA was included as a transcriptional regulator since small regulatory RNAs can be incorporated into bipartite networks without a loss of generality.

Network processing

Due to the size of the transcriptional networks of *S. cerevisiae* and *E. coli*, it was necessary to use the versatility index shortcut calculation described in Appendix D. To utilize this calculation, every regulator in the system must have a gene it solely controls. Not all regulators in the transcriptional networks of *S. cerevisiae* and *E. coli* have this attribute. Therefore, those regulators without this attribute along with all of the genes they participate in controlling were removed from the networks. The remaining networks (*S. cerevisiae*: 3630 genes, 147 regulators; *E. coli*: 680 genes, 71 regulators) were then analyzed as described in Appendix D.

Versatility index plot

Networks were created from an algorithm whose initial $N \times L$ network had one edge per output and the same number of edges per regulator. For every iteration an edge was randomly added to the network of the previous step. The algorithm concluded when the network was fully connected. A versatility index was calculated at every iteration for the network of that step. To ensure use of the versatility index shortcut calculation, an output for every regulator was required to contain a single edge, until the remaining $N - L$ outputs were fully connected. Then edges were added at random to the remaining L outputs until the network was fully connected.

APPENDIX A

Proof of Theorem 1

Definition

The connectivity pattern, $\mathbf{Z}_A$, can be defined as

$$\mathbf{Z_A} = \{\mathbf{A} \in \mathbb{R}^{N\times L} | a_{ij} = 0, \text{for a given set of } (i,j)\}, \quad \text{(A1)}$$

where the values of the nonzero a_{ij} are left unconstrained and can take on any value, positive, negative, or zero. $\mathbf{Z}_A$ characterizes a class of networks that all have the same zero pattern, but varying connectivity strengths (nonzero a_{ij}).

Theorem 1

A linear bipartite network with connectivity pattern $\mathbf{Z}_A$ ($N \times L$) can describe any data within $\mathbb{R}^{\mathbf{L}}$, if all reduced forms of $\mathbf{Z}_A$, $\mathbf{Z}_{\mathbf{A}r_i}$ ($z_i \times L$), are full row rank.

Here, $\mathbf{Z}_{\mathbf{A}r_i}$ is defined as the rows of $\mathbf{Z}_A$ which contain zeros in the i^{th} column of $\mathbf{Z}_A$, where z_i is the number of zeros in the i^{th} column of $\mathbf{Z}_A$.

Proof

If a connectivity pattern, $\mathbf{Z}_A$ ($N \times L$), can linearly describe any data, $\mathbf{E}$ ($N \times M$), within $\mathbb{R}^{\mathbf{L}}$, there exists a matrix, $\mathbf{A}$ ($N \times L$), characterized by $\mathbf{Z}_A$, that can provide an exact decomposition of the data, which is equivalent to the singular value decomposition

$$\mathbf{E} = \mathbf{AP} = \mathbf{USV^T}, \quad \text{(A2)}$$

where $\mathbf{E}$ is the output data ($N \times M$), $\mathbf{A}$ is the matrix ($N \times L$) defined by the zero pattern $\mathbf{Z}_A$, $\mathbf{P}$ is the linear system solution ($L \times M$) to $\mathbf{E}$ and $\mathbf{A}$, $\mathbf{S}$ is the diagonal matrix ($L \times L$) of the first L singular values of $\mathbf{E}$ oriented in decreasing order, and $\mathbf{U}$ ($N \times L$) and $\mathbf{V}$ ($L \times M$) are unitary matrices of the right and left singular vectors of the elements in $\mathbf{S}$. It follows that the component matrices of the decompositions will be related as follows:

$$\mathbf{A} = \mathbf{UX} \in \mathbf{Z}_\mathbf{A} \tag{A3}$$

$$\mathbf{P} = \mathbf{X}^{-1}\mathbf{S}\mathbf{V}^\mathbf{T}. \tag{A4}$$

For $\mathbf{X}$ to be invertible it must be full rank. The ranks of a matrix and matrix multiplication are governed by

$$\text{rank}(\mathbf{M}) \leq \{\min(N, L) | \mathbf{M} \in \mathbb{R}^{\text{N}\times\text{L}}\} \tag{A5}$$

$$\text{rank}(\mathbf{MB}) \leq \min\{\text{rank}(\mathbf{M}), \text{rank}(\mathbf{B})\}. \tag{A6}$$

Since $\mathbf{X}$ is full rank and

$$\text{rank}(\mathbf{U}) = L \tag{A7}$$

$$\text{rank}(\mathbf{SV}^\mathbf{T}) = L, \tag{A8}$$

the ranks of $\mathbf{A}$ and $\mathbf{P}$ must be allowed to be

$$\text{rank}(\mathbf{P}) = \text{rank}(\mathbf{SV}^\mathbf{T}) = L \tag{A9}$$

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{U}) = L. \tag{A10}$$

While $\mathbf{P}$ is unrestricted, $\mathbf{A}$ is restricted by $\mathbf{Z}_\text{A}$ and may not be allowed to satisfy Eq. A10. The positioning of zeros in $\mathbf{Z}_\text{A}$ may lead to rank deficiencies in $\mathbf{A}$. To check we can consider the nonzero entries of $\mathbf{Z}_\text{A}$ as nonzero random values that cannot combine on their own to produce a rank deficiency. We can then check the rank of $\mathbf{Z}_\text{A}$ directly. However, allowing $\mathbf{A}$ to satisfy Eq. A10 is a necessary but insufficient condition to satisfy Eq. A3. For a necessary and sufficient condition, one can break up Eq. A3 as

$$\mathbf{a} = \mathbf{uX} \in \mathbf{Z}_\mathbf{a}, \tag{A11}$$

where $\mathbf{a}$ ($j \times L$) is a collection of j rows from $\mathbf{A}$ where j can be any number of rows from 1 to N, $\mathbf{u}$ ($j \times L$) is the collection of rows from $\mathbf{U}$ corresponding to $\mathbf{a}$, and $\mathbf{Z}_\text{A}$ ($j \times L$) is the collection of rows from $\mathbf{Z}_\text{A}$ corresponding to $\mathbf{a}$. $\mathbf{X}$ must still be invertible, so to satisfy Eq. A11, $\mathbf{a}$ must be allowed to satisfy

$$\text{rank}(\mathbf{a}) = \text{rank}(\mathbf{u}). \tag{A12}$$

To satisfy Eq. A3, all possible $\mathbf{a}$ must be allowed by $\mathbf{Z}_\text{a}$ to satisfy Eq. A12. One can now see that Eq. A10 is a special case of Eq. A12, where $i = N$. Here

$$\text{rank}(\mathbf{u}) \leq \min(j, L). \tag{A13}$$

Since $\mathbf{u}$ can be full rank, min(j,L), $\mathbf{a}$ must be allowed by $\mathbf{Z}_\text{a}$ to be full rank. Analogous to $\mathbf{A}$ and $\mathbf{Z}_\text{A}$, the positioning of zeros in $\mathbf{Z}_\text{a}$ may lead to rank deficiencies in $\mathbf{a}$. However, it is unnecessary to check all possible $\mathbf{Z}_\text{a}$ for rank deficiencies.

We notice that rank deficiencies appear when rows of $\mathbf{a}$ contain zeros in the same column/columns. To capture all possible rank deficiencies, we define $\mathbf{Z}_{\mathbf{A}\text{r}_\text{i}}$ ($z_\text{i} \times L$), as the rows of $\mathbf{Z}_\text{A}$, which contain zeros in the i^{th} column of $\mathbf{Z}_\text{A}$, where z_i is the number of zeros in the i^{th} column of $\mathbf{Z}_\text{A}$. If we consider $\mathbf{Z}_{\mathbf{A}\text{r}_\text{i}}$ for every column of $\mathbf{Z}_\text{A}$, all rank deficiencies can be accounted for. If all $\mathbf{Z}_{\mathbf{A}\text{r}_\text{i}}$ are full rank (same check process as $\mathbf{Z}_\text{A}$ above), then $\mathbf{a}$ will be allowed to satisfy Eq. A12, and thus Eq. A3 will be satisfied. However, since $\mathbf{Z}_{\mathbf{A}\text{r}_\text{i}}$ will always have a zero column by definition, $\mathbf{Z}_{\mathbf{A}\text{r}_\text{i}}$ can only be full rank if it is full row rank.

Examples

To illustrate the criterion for network versatility, consider the Network A and B shown in Fig. A1, which can be represented by the connectivity pattern

$$\mathbf{Z}_\mathbf{A} = \begin{bmatrix} x & 0 & 0 \\ x & x & 0 \\ x & x & x \\ x & 0 & x \\ 0 & x & x \\ 0 & x & 0 \end{bmatrix} \quad \mathbf{Z}_\mathbf{B} = \begin{bmatrix} x & 0 & 0 \\ x & x & 0 \\ x & x & x \\ x & 0 & x \\ 0 & x & x \\ 0 & x & x \end{bmatrix},$$

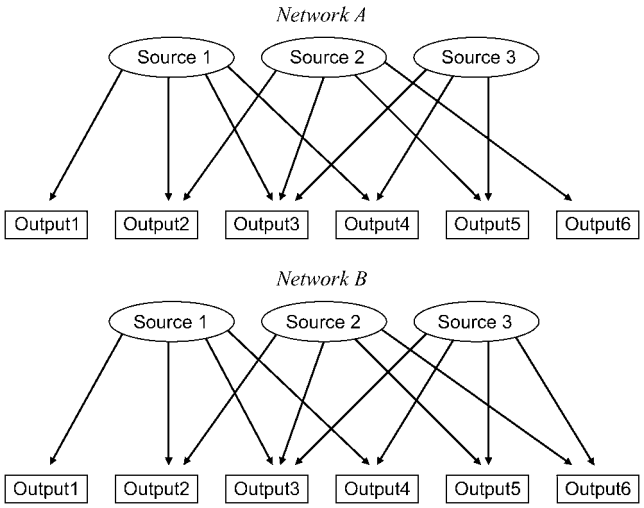

FIGURE A1 Diagram of two bipartite networks (A and B) that have (L – 1) missing edges per regulator. Network A is nonversatile, and Network B is versatile.

where the reduced matrices of $\mathbf{Z}_\mathbf{A}$ and $\mathbf{Z}_\mathbf{B}$ are

$$\mathbf{Z}_{\mathbf{A}\text{r}_1} = \begin{bmatrix} 0 & x & x \\ 0 & x & 0 \end{bmatrix}, \text{ rank of 2} \quad \mathbf{Z}_{\mathbf{B}\text{r}_1} = \begin{bmatrix} 0 & x & x \\ 0 & x & x \end{bmatrix}, \text{ rank of 2}$$

$$\mathbf{Z}_{\mathbf{A}\text{r}_2} = \begin{bmatrix} x & 0 & 0 \\ x & 0 & x \end{bmatrix}, \text{ rank of 2} \quad \mathbf{Z}_{\mathbf{B}\text{r}_2} = \begin{bmatrix} x & 0 & 0 \\ x & 0 & x \end{bmatrix}, \text{ rank of 2}$$

$$\mathbf{Z}_{\mathbf{A}\text{r}_3} = \begin{bmatrix} x & 0 & 0 \\ x & x & 0 \\ 0 & x & 0 \end{bmatrix}, \text{ rank of 2} \quad \mathbf{Z}_{\mathbf{B}\text{r}_3} = \begin{bmatrix} x & 0 & 0 \\ x & x & 0 \end{bmatrix}, \text{ rank of 2}$$

The rank of these structurally specified matrices can be determined by allowing random nonzero values to occupy the nonzero positions. Here $\mathbf{Z}_{\mathbf{A}\text{r}_1}$ and $\mathbf{Z}_{\mathbf{A}\text{r}_2}$ are full row rank, while $\mathbf{Z}_{\mathbf{A}\text{r}_3}$ is not. Therefore, Network A is not versatile. On the other hand, all $\mathbf{Z}_{\mathbf{B}\text{r}_\text{i}}$ are full row rank, and therefore Network B is versatile.

Minimal connectivity for versatility is maximal connectivity for NCA-compliance

To illustrate this boundary, consider the networks shown in Fig. A2. Network A is versatile since every $\mathbf{Z}_{\mathbf{A}\text{r}_\text{i}}$ is full row rank, but not NCA-compliant, since $\mathbf{Z}_{\mathbf{A}\text{r}_3}$ has a rank of (L–2). Network B is both versatile and NCA-compliant since every $\mathbf{Z}_{\mathbf{B}\text{r}_\text{i}}$ is full row rank and of rank (L–1). Network C is nonversatile since $\mathbf{Z}_{\mathbf{C}\text{r}_3}$ is not full row rank, but it is NCA-compliant since all $\mathbf{Z}_{\mathbf{C}\text{r}_\text{i}}$ are of rank (L–1). This example illustrates that networks that are versatile can also be NCA-compliant, but that this can only happen if the network is of the minimum connectivity to achieve versatility.

APPENDIX B

Proof of Theorem 2

Definitions

1. A zero pattern is a $1 \times L$ vector that indicates, by the position of zero entries, which transcription factors (TF) do not control expression of a gene. The number of zero entries in a zero pattern is designated by n_z. A system with three TFs has seven zero patterns, which are shown in

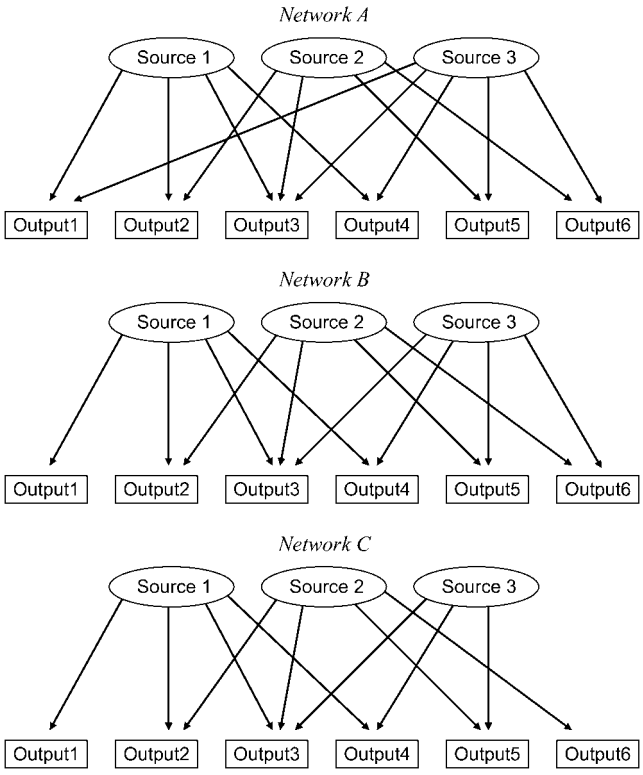

FIGURE A2 Diagram of three bipartite networks (*A–C*) that demonstrate the relationship between NCA and versatility. Network A is versatile but not NCA-compliant, Network B is both versatile and NCA-compliant, and Network C is NCA-compliant but not versatile.

Fig. 1 *D*. The zero pattern $[x \quad x \quad 0]$ indicates that a gene is not controlled by TF_3.

2. Any gene that satisfies the definition of a zero pattern is a member of that zero pattern. For instance, the zero pattern $[x \quad x \quad 0]$ requires genes to not be regulated by TF_3, therefore, $gene_{1,2,4}$ are all members
3. An informative zero pattern, $\mathbf{Z}_{\text{inf}_j}$, is any zero pattern with $>L - n_{zj}$ members, where n_{zj} is equal to the number of zeros in $\mathbf{Z}_{\text{inf}_j}$. Fig. 1 *A* has two $\mathbf{Z}_{\text{inf}_j}$, $\mathbf{Z}_{\text{inf}_1} = [x \quad x \quad 0]$ and $\mathbf{Z}_{\text{inf}_2} = [x \quad x \quad x]$.
4. $\mathbf{E}_{rj}$ is a matrix composed of the genes (rows of $\mathbf{E}$) that are members of $\mathbf{Z}_{\text{inf}_j}$. From Fig. 1, $\mathbf{E}_{r1} = \mathbf{E}(\text{rows } 1, 2, 4)$.

Theorem 2

Any dataset, $\mathbf{E}$ ($N \times L$), may be represented by a linear bipartite network characterized by connectivity pattern $\mathbf{Z}_A$ ($N \times L$) if every $\mathbf{E}_{rj}$ has rank $\leq (L - n_{zj})$.

Proof

If a connectivity pattern, $\mathbf{Z}_A$ ($N \times L$), can linearly describe a dataset, $\mathbf{E}$ ($N \times M$), within $\mathbb{R}^{\mathbf{L}}$, there exists a matrix, $\mathbf{A}$ ($N \times L$), characterized by $\mathbf{Z}_A$, such that

$$\mathbf{E} = \mathbf{AP}. \tag{B1}$$

It follows that $\mathbf{E}$ may be described implicitly as a function of $\mathbf{A}$, if $\mathbf{A}$ has full rank. If $\mathbf{A}$ has full rank, $\mathbf{A}$ can be partitioned into $\mathbf{A}_1$ ($(N - L) \times L$) and $\mathbf{A}_2$ ($L \times L$) such that $\mathbf{A}_2$ is invertible. After partitioning and substituting for $\mathbf{P}$, one obtains

$$\mathbf{E_1} = \mathbf{A_1}\mathbf{A_2^{-1}}\mathbf{E_2}. \tag{B2}$$

Note that multiple partitions of $\mathbf{A}$ exist, since the only requirement of Eq. B2 is that $\mathbf{A}_2$ be invertible. Consequences of this point will be discussed shortly.To determine whether restrictions originate from $\mathbf{Z}_A$ that are required to exist in $\mathbf{E}$ for Eq. B1 to be satisfied, we make the following transformation:

$$\mathbf{E_1} = \mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}\mathbf{E_2}. \tag{B3}$$

For generalization purposes, we substitute $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$ for $\mathbf{A_1}\mathbf{A_2^{-1}}$. This notation, which will be used for the remainder of the text, signifies that we only know that $\mathbf{A}$ ($N \times L$) is characterized by $\mathbf{Z}_A$ and that all nonzero values are considered unknown. Analogous to Eq. B2, for Eq. B3 to hold true, $\mathbf{Z}_{\mathbf{A}_2}$ has to have full rank, and thus $\mathbf{Z}_A$ has to have full rank. The rank can be calculated by considering the nonzero entries of $\mathbf{Z}_A$ as nonzero random values that cannot combine on their own to produce a rank deficiency. Eq. B3 represents a relationship solely between the data, $\mathbf{E}$, and network, $\mathbf{Z}_A.\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$ is defined analogous to $\mathbf{Z}_A$, where the positions of the zero entries are known and the nonzero entries are left unconstrained. However, unlike $\mathbf{Z}_A$, where a zero entry indicates the absence of a connection between a source and output, zeros within $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$ indicate constraints on how outputs may be related to one another. For instance, if the following $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$ were obtained,

$$\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}} = \begin{bmatrix} x & 0 & 0 \\ x & x & x \\ x & x & x \end{bmatrix} \quad \mathbf{E_1} = \begin{bmatrix} x & 0 & 0 \\ x & x & x \\ x & x & x \end{bmatrix} \mathbf{E_2},$$

then the first row of $\mathbf{E}_1$ would need to be a multiple of the first row of $\mathbf{E}_2$, while the other rows of $\mathbf{E}_1$ could be any linear combination of all of the rows of $\mathbf{E}_2$.

Zeros within $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$ dictate how the outputs of $\mathbf{E}$ may be related, and thus represent connectivity constraints from $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$. However, the partition from Eqs. B2 and B3 is inherently nonunique, since there can be multiple partitions of $\mathbf{Z}_A$ that meet the full rank requirement of $\mathbf{Z}_{\mathbf{A}_2}$. As one might expect, different selections of $\mathbf{Z}_{\mathbf{A}_2}$ generate different zero patterns in $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$. Thus, separate connectivity constraints are identified by different network partitions. To properly define the output limits of a network, all of the constraints must be considered. This requires an understanding of how zeros propagate from $\mathbf{Z}_A$ to $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$.

Since the nonzero elements in $\mathbf{Z}_A$ are left unconstrained and can take on any value, the determination of $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2^{-1}}}$ is not a case of simple linear algebra. Therefore, we have derived a set of rules that that describe how zeros propagate through structural multiplication ($\mathbf{Z_A}\mathbf{Z_B}$) and structural inverse ($\mathbf{Z}_{\mathbf{A}^{-1}}$) operations. These operations are analogous to their linear algebra counterparts, except that instead of being defined for fully specified matrices, their operations are designed for networks defined analogous to $\mathbf{Z}_A$.

Rule 1

Zeros can only be created by multiplication ($\mathbf{Z_A}\mathbf{Z_B}$), if a row of $\mathbf{Z}_A$ is structurally perpendicular to a column of $\mathbf{Z}_B$. For a row to be structurally perpendicular to a column, they must have zeros in complementary positions.

Rule 2

The number of zeros that propagate through a structural multiplication, $\mathbf{Z_A}\mathbf{Z_B}$, where $\mathbf{Z_B}$ is invertible, is limited by: $\text{Zeros}_{\mathbf{Z_A}\mathbf{Z_B}} \leq \text{Zeros}_{\mathbf{Z_A}}$, where $(\mathbf{Z_A} \in \mathbb{R}^{N \times L}, \mathbf{Z_B} \in \mathbb{R}^{L \times L})$, $\text{Zeros}_{\mathbf{Z_A}\mathbf{Z_B}}$ is the number of zeros in $\mathbf{Z_A}\mathbf{Z_B}$, and $\text{Zeros}_{\mathbf{Z_A}}$ is the number of zeros in $\mathbf{Z}_A$.

Rule 3

Zeros can only exist in $\mathbf{Z}_{\mathbf{A}^{-1}}$, if singular minors can be created from $\mathbf{Z}_A$.

Rule 4

The number of zeros that propagate through a structural inverse is limited by $\text{Zeros}_{\mathbf{Z}_{\mathbf{A}^{-1}}} \leq \text{Zeros}_{\mathbf{Z}_{\mathbf{A}}}$, where $(\mathbf{Z}_{\mathbf{A}^{-1}}, \mathbf{Z}_{\mathbf{A}} \in \mathbb{R}^{\mathrm{LxL}})$and $\text{Zeros}_{\mathbf{Z}_{\mathbf{A}^{-1}}}$ is the number of zeros in $\mathbf{Z}_{\mathbf{A}^{-1}}$.

Rule 5

Zeros in $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2}^{-1}}$ are created from members of $\mathbf{Z}_{\text{inf}_j}$. If exactly $(L - n_{zj})$ $\mathbf{Z}_{\text{inf}_j}$ members are partitioned into $\mathbf{Z}_{\mathbf{A_2}}$, the $\mathbf{Z}_{\text{inf}_j}$ members in $\mathbf{Z}_{\mathbf{A_1}}$will be structurally perpendicular to zeros in $\mathbf{Z}_{\mathbf{A}_2^{-1}}$ created from $\mathbf{Z}_{\text{inf}_j}$ members in $\mathbf{A_2}$.

Proofs for these Rules can be found in Appendix C.

According to Rule 5, zeros within $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2}^{-1}}$ occur when $\geq (L - n_{zj})$ members of $\mathbf{Z}_{\text{inf}_j}$are in $\mathbf{Z}_{\mathbf{A}}$, and exactly $(L - n_{zj})$ members are partitioned into $\mathbf{Z}_{\mathbf{A_2}}$. We recognize that it does not matter which members of $\mathbf{Z}_{\text{inf}_j}$ are partitioned into $\mathbf{Z}_{\mathbf{A_2}}$, and that the zeros in the remaining members of $\mathbf{Z}_{\text{inf}_j}$ will be conserved in $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2}^{-1}}$. Finally, by rearranging Eq. B3, we obtain

$$\mathbf{0} = \begin{bmatrix} \mathbf{I} & \mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2}^{-1}} \end{bmatrix} \begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \end{bmatrix}. \tag{B4}$$

To satisfy Eq. B4, for every row_i of $[\mathbf{I} \quad \mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2}^{-1}}]$ the matrix composed of those rows of $\begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \end{bmatrix}$ that multiply against nonzero entries in row_i of $[\mathbf{I} \quad \mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2}^{-1}}]$ should be of rank $\leq (L - n_{zj})$. Therefore, $\mathbf{E}_{rj}$ created from collecting all of the rows of $\mathbf{E}$ that correspond to members of $\mathbf{Z}_{\text{inf}_j}$ should have rank $\leq (L - n_{zj})$ if $\mathbf{Z}_{\mathrm{A}}$ can represent $\mathbf{E}$. This will be a requirement for all possible $\mathbf{Z}_{\text{inf}_j}$.

Example

To illustrate the use of Theorem 2 and Table A1, consider the network in Fig. A3 and corresponding connectivity pattern:

$$\mathbf{Z}_{\mathbf{A}} = \begin{bmatrix} x & 0 & 0 \\ x & x & 0 \\ x & x & 0 \\ x & 0 & x \\ 0 & x & x \\ 0 & x & 0 \end{bmatrix}$$

For this $\mathbf{Z}_{\mathrm{A}}$ we can construct Table A2. Only two zero patterns are informative, $\mathbf{Z}_{\text{inf}_1} = [x \quad x \quad 0]$ and $\mathbf{Z}_{\text{inf}_2} = [x \quad x \quad x]$. To demonstrate zero generation in $\mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2}^{-1}}$, we partition $(L - n_{zj})$ $\mathbf{Z}_{\text{inf}_j}$ members into $\mathbf{Z}_{\mathbf{A_2}}$,

$$\mathbf{Z}_{\mathbf{A}_1} = \begin{bmatrix} x & 0 & 0 \\ x & x & 0 \\ 0 & x & x \end{bmatrix} \begin{matrix} \text{gene 1} \\ \text{gene 2} \\ \text{gene 5} \end{matrix}$$

$$\mathbf{Z}_{\mathbf{A}_2} = \begin{bmatrix} x & x & 0 \\ x & 0 & x \\ 0 & x & 0 \end{bmatrix} \begin{matrix} \text{gene 3} \\ \text{gene 4} \\ \text{gene 6} \end{matrix}.$$

It follows that

TABLE A1 Procedure used to determine whether a dataset, E, contains the connectivity constraints dictated by $\mathbf{Z_A}$

Procedure for using Theorem 2
1. Identify all possible $1 \times L$ zero patterns.
2. Determine those zero patterns that have $>(L - n_z)$ members. This will be the list of $\mathbf{Z}_{\text{inf}_j}$.
3. Create $\mathbf{E}_{rj}$ for every $\mathbf{Z}_{\text{inf}_j}$ and check whether all $\mathbf{E}_{rj}$ have rank $\leq (L - n_z)$.

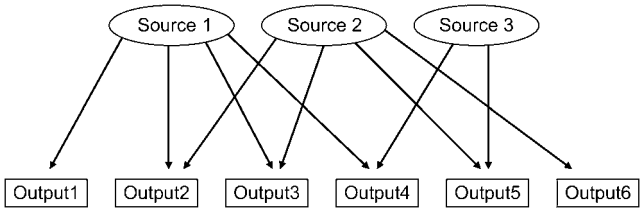

FIGURE A3 Diagram of a bipartite network used for deduction of connectivity constraints from Table A1.

$$\mathbf{Z}_{\mathbf{A}_1}\mathbf{Z}_{\mathbf{A}_2^{-1}} = \begin{bmatrix} x & 0 & x \\ x & 0 & x \\ x & x & x \end{bmatrix}.$$

After rearranging Eq. B3 and substituting for the current example, we obtain

$$\mathbf{0} = \begin{bmatrix} \mathbf{I} & \mathbf{Z}_{\mathbf{A_1}}\mathbf{Z}_{\mathbf{A_2}^{-1}} \end{bmatrix} \begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & x & 0 & x \\ 0 & 1 & 0 & x & 0 & x \\ 0 & 0 & 1 & x & x & x \end{bmatrix} \begin{bmatrix} \text{row 1 of } E \\ \text{row 2 of } E \\ \text{row 5 of } E \\ \text{row 3 of } E \\ \text{row 4 of } E \\ \text{row 6 of } E \end{bmatrix}. \tag{B5}$$

Eq. B5 states that:

1. The matrix formed by rows 1, 3, 6 of $\mathbf{E}$ must have rank ≤ 2.
2. The matrix formed by rows 2, 3, 6 of $\mathbf{E}$ must have rank ≤ 2.
3. The matrix formed by rows 5, 3, 4, 6 of $\mathbf{E}$ must have rank ≤ 3.

Note that Statement 3 simply checks if $\mathbf{E}$ is in $\mathbb{R}^{\mathbf{L}}$, and that Statements 1 and 2 require that

$$\begin{bmatrix} \text{row 1 of } E \\ \text{row 2 of } E \\ \text{row 3 of } E \\ \text{row 6 of } E \end{bmatrix} \equiv \mathbf{E}_{\mathrm{r1}}$$

has a rank $\leq (L - n_{zj}) = 2$. For a dataset to be represented by the network in Fig. A3, $\mathbf{E}_{\mathrm{r1}}$ must have a rank ≤ 2, and $\mathbf{E}_{\mathrm{r2}}$ must have a rank ≤ 3.

APPENDIX C

Structural linear algebra proofs

The first two rules deal with properties of structural multiplications. The first rule states that for a zero to be created in $\mathbf{Z}_{\mathrm{A}}\mathbf{Z}_{\mathrm{B}}$ a row of $\mathbf{Z}_{\mathrm{A}}$ must be

TABLE A2 Example of using Theorem 2 and the procedure from Table A1 for bipartite network in Fig. A3

Zero pattern	n_z	Members	$\mathbf{Z}_{\text{inf}_j}$ Pattern
$[x \quad 0 \quad 0]$	2	gene_1	
$[0 \quad x \quad 0]$	2	gene_6	
$[0 \quad 0 \quad x]$	2	—	
$[x \quad x \quad 0]$	1	$\text{gene}_{1,2,3,6}$	$\mathbf{Z}_{\text{inf}_1} = [x \quad x \quad 0]$
$[x \quad 0 \quad x]$	1	$\text{gene}_{1,4}$	
$[0 \quad x \quad x]$	1	$\text{gene}_{5,6}$	
$[x \quad x \quad x]$	0	$\text{gene}_{1,2,3,4,5,6}$	$\mathbf{Z}_{\text{inf}_2} = [x \quad x \quad x]$

structurally perpendicular to a column of $\mathbf{Z}_B$. Be reminded that the nonzero entries of both $\mathbf{Z}_A$ and $\mathbf{Z}_B$ are left unconstrained, and thus may take on any value. Therefore, we must allow the product of any two nonzero entries to also be left unconstrained. So for any two vectors to be perpendicular, when both vectors are structurally defined, every nonzero entry of one vector must multiply against a zero entry of the other vector. This is the definition of structurally perpendicular.

As a consequence of Rule 1, a vector $\vec{y}$ ($1xL$) can be structurally perpendicular to only as many vectors of an L basis as $\vec{y}$ has zero entries. To illustrate, consider a vector $\vec{y}$ ($1 \times L$) and an invertible matrix $\mathbf{B}$ ($L \times L$), where both are structurally defined and $\mathbf{B}$ is a basis of L space:

$$\vec{u} = \vec{y}\mathbf{B}. \tag{C1}$$

To produce a zero within $\vec{u}$ ($1xL$), $\vec{y}$ must be structurally perpendicular to a column vector of $\mathbf{B}$, a vector of an L basis. The sparsest possible structurally defined basis, $\mathbf{B}$, is diagonal or a permutation thereof. In that case, $\vec{y}$ would be structurally perpendicular to as many vectors of $\mathbf{B}$ as $\vec{y}$ has zero entries. However, if $\mathbf{B}$ is not diagonal or a permutation thereof, $\vec{y}$ can be structurally perpendicular to only as many vectors of $\mathbf{B}$ as $\vec{y}$ has zero entries, but may be less, depending on the structure of $\vec{y}$ and $\mathbf{B}$. To demonstrate, consider the following $\vec{u}$, $\vec{y}$ and basis, $\mathbf{B}$:

$$\vec{u} = \vec{y}\mathbf{B}\begin{bmatrix} x & x & 0 \end{bmatrix} = \begin{bmatrix} x & x & 0 \end{bmatrix}\begin{bmatrix} x & 0 & 0 \\ 0 & x & 0 \\ 0 & x & x \end{bmatrix},$$

$$\vec{u} = \vec{y}\mathbf{B}\begin{bmatrix} x & x & x \end{bmatrix} = \begin{bmatrix} x & x & 0 \end{bmatrix}\begin{bmatrix} x & 0 & 0 \\ 0 & x & x \\ 0 & 0 & x \end{bmatrix}.$$

Both bases have the same number of nonzero entries; however, the structure of the first basis allows the number of zeros in $\vec{y}$ to propagate to $\vec{u}$, while the second basis does not. Therefore, the number of zeros in $\vec{u}$ will always be less than or equal to the number of zeros in $\vec{y}$. Since $\mathbf{Z}_A$ is a collection of stacked $\vec{y}$ vectors the same holds true for all the rows of $\mathbf{Z}_A$.

Rules 3 and 4 deal with the properties of structural inverses. Inverses are defined as

$$a_{ij}^{-1} = \frac{1}{|\mathbf{A}|}C_{ij}, \tag{C2}$$

$$C_{ij} = -1^{(i+j)}M_{ij}, \tag{C3}$$

for i,j = 2,3

$$M_{23} = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}. \tag{C4}$$

When M_{ij} is singular you will see a zero at position (j,i) of $\mathbf{A}^{-1}$. However, to reiterate, the nonzero entries of $\mathbf{Z}_A$ are left unconstrained. This means that assumptions cannot be made about the values of the nonzero entries, and thus zeros within $\mathbf{Z}_{\mathbf{A}^{-1}}$ must come from minors that are singular irrespective of the nonzero entries. As it turns out, any possible row zero pattern that may be found in $\mathbf{Z}_A$, can create singular minors in $\mathbf{Z}_{\mathbf{A}^{-1}}$ if there are exactly ($L - z_j$) members in $\mathbf{Z}_A$. Rule 4 states that the number of zeros in $\mathbf{Z}_{\mathbf{A}^{-1}}$ will always be less than or equal to the number of zeros in $\mathbf{Z}_A$. To explain, consider the following linear algebra operation:

$$\mathbf{A} = (\mathbf{A}^{-1})^{-1}. \tag{C5}$$

An analogous operation can be defined for structural linear algebra operations,

$$\mathbf{Z}_\mathbf{A} = (\mathbf{Z}_{\mathbf{A}^{-1}})^{-1}, \tag{C6}$$

only for those invertible $\mathbf{Z}_A$ that have zero entries that all contribute to singular minors for $\mathbf{Z}_{\mathbf{A}^{-1}}$. Otherwise,

$$\mathbf{Z}_\mathbf{A} \neq (\mathbf{Z}_{\mathbf{A}^{-1}})^{-1}, \tag{C7}$$

and the number of zeros in $\mathbf{Z}_{\mathbf{A}^{-1}}$ is less than that in $\mathbf{Z}_A$. To illustrate, consider the following:

$$\mathbf{Z}_\mathbf{A} = \begin{bmatrix} x & x & x \\ x & 0 & 0 \\ 0 & x & x \end{bmatrix} \qquad \mathbf{Z}_\mathbf{B} = \begin{bmatrix} x & x & x \\ x & 0 & 0 \\ x & 0 & x \end{bmatrix}$$

$$\mathbf{Z}_{\mathbf{A}^{-1}} = \begin{bmatrix} 0 & x & 0 \\ x & x & x \\ x & x & x \end{bmatrix} \qquad \mathbf{Z}_{\mathbf{B}^{-1}} = \begin{bmatrix} 0 & x & 0 \\ x & x & x \\ 0 & x & x \end{bmatrix}$$

$$(\mathbf{Z}_{\mathbf{A}^{-1}})^{-1} = \begin{bmatrix} x & x & x \\ x & 0 & 0 \\ x & x & x \end{bmatrix} \quad (\mathbf{Z}_{\mathbf{B}^{-1}})^{-1} = \begin{bmatrix} x & x & x \\ x & 0 & 0 \\ x & 0 & x \end{bmatrix}$$

$$\mathbf{Z}_\mathbf{A} \neq (\mathbf{Z}_{\mathbf{A}^{-1}})^{-1}, \qquad \mathbf{Z}_\mathbf{B} = (\mathbf{Z}_{\mathbf{B}^{-1}})^{-1}.$$

Both $\mathbf{Z}_A$ and $\mathbf{Z}_B$ have the same number of zero entries, except those in $\mathbf{Z}_B$ all contribute to singular minors whereas those in $\mathbf{Z}_A$ do not. Therefore, the number of zeros in $\mathbf{Z}_\mathbf{A}$ equals the number in $\mathbf{Z}_{\mathbf{B}^{-1}}$, and the number in $\mathbf{Z}_A$ is less than the number in $\mathbf{Z}_{\mathbf{A}^{-1}}$.

The final rule is a combination of the knowledge from the first four rules. By following structural linear algebra Rules 1–4 we realize that zeros within $\mathbf{Z}_{\mathbf{A}_1}\mathbf{Z}_{\mathbf{A}_2^{-1}}$ are generated when there are $>(L–n_{zj})$ members of $\mathbf{Z}_{\text{inf}_j}$ in $\mathbf{Z}_A$, and ($L - n_{zj}$) are contained within $\mathbf{Z}_{\mathbf{A}_2}$. This is because any member of $\mathbf{Z}_{\text{inf}_j}$ will be structurally perpendicular to zeros in $\mathbf{Z}_{\mathbf{A}_2^{-1}}$created from its fellow members.

APPENDIX D

Determining Z_A^c and max(Z^c)

Both $\mathbf{Z}_A^c$ and max($\mathbf{Z}^c$) can be determined from Theorem 2. The number of constraints ($\mathbf{Z}_A^c$) imposed by $\mathbf{Z}_A$ on a dataset $\mathbf{E}$, is

$$\mathbf{Z}_\mathbf{A}^c = \sum_{j=1}^{n}(\text{members}(\mathbf{Z}_{\text{inf}\,j}) - (L - n_{zj})), \tag{D1}$$

where $\mathbf{Z}_{\text{inf}_j}$ and n_{zj} are defined from Theorem 2, and n is equal to the total number of $\mathbf{Z}_{\text{inf}_j}$ that have $>(L - n_{zj})$ members.The most nonversatile network that is the same size of $\mathbf{Z}_A$ will be the sparsest network, and thus have the largest number of missing edges. The network that has the largest number of missing edges and is the same size as $\mathbf{Z}_A$ will be a network that has one edge per row. However, the one edge per row criteria classifies a large number of networks that all have the same sparsity. So the question arises, which one contains max($\mathbf{Z}^c$)? The answer is that all $\mathbf{Z}_A$ ($N \times L$) that have N edges, have one edge per row, and are of the same size have the same number of constraints, and thus may be used to calculate max($\mathbf{Z}^c$). A shortcut calculation can be derived from the above equation by realizing that there cannot be any zero columns of $\mathbf{Z}_A$. Therefore, one can calculate max($\mathbf{Z}^c$), from the following formula with only knowledge of the network size, $N \times L$, and not the structure:

$$\max(\mathbf{Z}^c) = N \times 2^{(L-1)} - \sum_{i=0}^{L-1}\left(\frac{L!}{(L-i)!i!} \times (L-i)\right). \tag{D2}$$

The first term of Eq. D2 is the equivalent to the first term in the summation of Eq. D1 when all rows only have one edge, and analogously the second term of Eq. D2 is equivalent to the second term in the summation of Eq. D1 when all rows only have one edge.
Since

$$L \times 2^{(L-1)} = \sum_{i=0}^{L-1} \left(\frac{L!}{(L-i)!i!} \times (L-i) \right),$$

we can make the following substitution:

$$\max(\mathbf{Z}^c) = (N-L) \times 2^{(L-1)}. \quad \text{(D3)}$$

A similar formula to calculate $\mathbf{Z}_A^c$ can be obtained under situations when there is at least one row per column that is controlled by only that column,

$$\begin{aligned} \mathbf{Z}_A^c &= \sum_{i=1}^{L} (n_i \times 2^{(L-i)}) - \sum_{i=0}^{L-1} \left(\frac{L!}{(L-i)!i!} \times (L-i) \right) \\ &= \sum_{i=1}^{L} (n_i \times 2^{(L-i)}) - L \times 2^{(L-1)}, \end{aligned} \quad \text{(D4)}$$

where n_i is equal to the number of rows with i nonzero entries.It should be noted that $\mathbf{Z}^c$ may be different for different $\mathbf{Z}_A$ even though they may have the same number of edges. This is because even though two rows with three edges each, have the same number of edges as three rows with two edges each, there is a difference between $2 \times 2^3 = 16$ and $3 \times 2^2 = 12$.

This work has been supported by the Center for Cell Mimetic Space Exploration and NASA University Research, Engineering and Technology Institute under award No. NCC 2-1364, National Science Foundation No. ITR CCF-0326605, and the University of California at Los Angeles-Department of Energy Institute for Genomics and Proteomics.

REFERENCES

1. Alter, O., P. O. Brown, and D. Botstein. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*. 97:10101–10106.
2. Alter, O., P. O. Brown, and D. Botstein. 2003. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl. Acad. Sci. USA*. 100:3351–3356.
3. Holter, N. S., M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA*. 97:8409–8414.
4. Liao, J. C., R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. 2003. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA*. 100:15522–15527.
5. Liebermeister, W. 2002. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*. 18:51–60.
6. Yeung, M. K., J. Tegner, and J. J. Collins. 2002. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA*. 99:6163–6168.
7. Ham, F., N. Faour, and J. Wheeler. 1999. 21st Seismic Research Symposium. Las Vegas, NV. 133–140.
8. Vigario, R., J. Sarela, V. Jousmaki, M. Hamalainen, and E. Oja. 2000. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Trans. Biomed. Eng*. 47:589–593.
9. Kasprzak, W., and A. Cichocki. 1996. Proceedings of ICPR '96. Vienna, Austria.
10. Lin, Q., F. Yin, and H. Liang. 2005. International Symposium of Neural Networks 2005. Chongqing, China.
11. Park, S., and F. Ham. 2003. Proceedings of the 25th Annual international Conference of the IEEE EMBS. Cancun, Mexico.
12. Steinbock, O., B. Neumann, B. Cage, J. Saltiel, S. Muller, and N. Dalal. 1997. A demonstration of principal component analysis for EPR spectroscopy: identifying pure component spectra from complex spectra. *Anal. Chem*. 69:3708–3713.
13. Uy, D., and A. O'Neill. 2005. Principal component analysis of Raman spectra from phosphorus-poisoned automotive exhaust-gas catalysts. *J. Raman Spectrosc*. 36:988–995.
14. Thurstone, L. 1947. The Simple Structure Concept in Multiple Factor Analysis: A Development and Expansion of The Vectors of Mind. The University of Chicago Press, Chicago, IL.
15. Browne, M. 2001. An overview of analytic rotation in exploratory factor analysis. *Multivariate Behav. Res*. 36:111–150.
16. Chennubhotla, C., and A. Jepson. 2001. Eighth International Conference on Computer Vision. Vancouver, Canada.
17. Zou, H., T. Hastie, and R. Tibshirani. 2004. Sparse Principal Component Analysis. Technical report, Department of Statistics, Stanford University. Http://www-stat.stanford.edu/~hastie/papers/sparsepc.pdf
18. Harbison, C. T., D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 431:99–104.
19. Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*. 298:799–804.
20. Salgado, H., S. Gama-Castro, A. Martinez-Antonio, E. Diaz-Peredo, F. Sanchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jimenez-Jacinto, A. Santos-Zavaleta, C. Bonavides-Martinez, and J. Collado-Vides. 2004. RegulonDB (Ver. 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*. 32:D303–D306.
21. Shen-Orr, S. S., R. Milo, S. Mangan, and U. Alon. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet*. 31:64–68.
22. Pernestig, A. K., D. Georgellis, T. Romeo, K. Suzuki, H. Tomenius, S. Normark, and O. Melefors. 2003. The *Escherichia coli* BarA-UvrY two-component system is needed for efficient switching between glycolytic and glucogenic carbon sources. *J. Bacteriol*. 185: 843–853.