Results: We have improved NCA for transcription factor activity (TFA) estimation, based on the observation that most genes are regulated by only a few TFS. This observation leads to the derivation of a new identifiability criterion which is tested during numerical iteration that allows us to decompose data when the number of TFS is greater than the number of experiments. To show that our method works relatively few data points. On the other hand, such methods would inherently increase numerical challenges leading to multiple solutions. Therefore, solutions to both the problems are needed.

1 INTRODUCTION

1.1 Background

Transcription factors (TFS) act alone or in combination on target promoters to control gene expression. TF regulatory activity is controlled by higher-level cellular functions, such as signaling pathways to reflect cellular physiology and environment, typically through post-transcriptional regulation, such as phosphorylation, oligomerization, ligand binding or subcellular translocation. Activated TFS control transcriptional initiation through interaction with DNA or RNA polymerase. Thus TFAs in principle can be inferred from the transcript levels of the genes they control. In the simplest case, where one TF controls one gene, the activity of this TF is directly proportional to the gene it controls. In reality, however, TF-promoter connectivity is more complicated requiring non-trivial mathematics for deconvolution.

Network component analysis (NCA) (Liao et al., 2003; Tran et al., 2005) is a model-based decomposition method to deduce such information, namely transcription factor activity (TFA) and regulation control strength (CS), from transcriptome data and transcription factor (TF)-promoter connectivity network. Connectivity networks are constructed by preprocessing ChIP-chip data, DamID methylation profiling or extensive literature search (Lee et al., 2002; van Steensel and Henikoff, 2000). To allow unique mathematical decomposition, three identifiability criteria need to be satisfied (Liao et al., 2003; Tran et al., 2005), which govern the applicability of this method. In particular, one of the criteria requires that the number of data points be greater than the number of TFS. In this work, we relax this theoretical limitation to allow analysis based on limited data points.

Other approaches to deduce TFA include the REDUCE (and equivalently MATRIX REDUCE) methods that fit a linear model of gene regulation to mRNA expression data by regression (Bussemaker et al., 2001; Gao et al., 2004). These methods are ‘cluster free’ and only require a input promoter sequence (or ChIP-chip binding data) and mRNA expression data. From this set REDUCE evaluates all possible candidate motifs, where each motif corresponds to a known or unknown (but shared sequence) regulatory factor. The motifs are fit individually across genes and the authors select a final set of regulatory motifs as those that have the largest reduction of variance of the corresponding gene expressions. Here normalized motif binding copy number is taken as the regulation strength of the TF for that gene, and TFA profiles are obtained from gene expression data by linear regression.
The advantage of REDUCE is that it is not limited by the identifiability criteria governing NCA, since its CS is fixed. However, because the method uses the copy number of binding sites as a surrogate for regulatory strength, it does not distinguish between positive and negative regulation of the same TF on different genes and may compromise the accuracy of TFA estimates.

Other methods to deduce underlying regulatory signals (not TFA per se) include principal components analysis (PCA) and independent components analysis (ICA). These methods explain the collection of gene expression profiles by a set of underlying basis profiles (vectors) termed eigengenes or expression modes (Alter et al., 2000). In PCA the eigengenes are orthonormal. In ICA the expression modes are statistically independent and non-Gaussian (Lee and Batzoglou, 2003; Liebermeister, 2002). In this case independence is a stronger assumption than uncorrelatedness. The top k significant basis vectors describe the majority of the observed gene expression variance, and they have been shown to correspond to major biological phenomena in the cell cycle (Alter et al., 2000).

However, both ICA and PCA approaches fail to incorporate the biological structure of the network, and accordingly the description is based on statistical properties. To provide biological insight, one may integrate these basis vectors with expression profiles by robust regression (Yeung et al., 2002) or with binding site data (ChIP-chip) by pseudo-inverse projection (Alter and Golub, 2004). The latter provides a dynamic temporal view of TF-gene binding activity in select cell-cycle and replication initiation factors.

That PCA and ICA methods prove useful for mapping highly coordinated transcriptional responses, such as the cell cycle, is encouraging because it suggests that data decompositions can find independent features of related biological processes. However, orthogonal (or independent) basis vectors do not necessarily have to correspond to underlying biological features and in fact may depend on rotation. We argue that it is the topology (or structure) of the network that is important for determining TFA because TFs may work together to regulate gene expression. This insight is the motivation for NCA, a method that takes network topology into account in order to compute the component TFAs and control strengths from a dynamic non-linear model (Kao et al., 2005; Kao et al., 2004; Liao et al., 2003; Tran et al., 2005).

1.2 Network component analysis

NCA formulates gene expression as the product of the contribution of each regulating TFA using a combinatorial power-law model, which can be viewed as a log-linear approximation of any non-linear kinetic system in multiple dimensions (Almeida and Voit, 2003; Savageau, 1976; Torres and Voit, 2002; Voit and Almeida, 2004). It captures some non-linear synergistic effects and yet still remains mathematically tractable and generally applicable to most genes.

The model used in NCA includes the synthesis and degradation of mRNA, which are assumed to be in a quasi-steady state:

\[
\frac{d[mRNA_i]}{dt} = V_{\text{synthesis}} - V_{\text{degradation}} = k_p \prod_{j=1}^{k} TFA_j^{a_j} - k_d [mRNA_i].
\]  

Note that the quasi-steady state of mRNA does not imply that mRNA levels are constant. It only means that the rate of mRNA change is much smaller than its synthesis rate or degradation rate. At the quasi-steady state, the synthesis rate is about the same as the degradation rate.

TFA itself is driven by interactions between proteins (P), metabolites (M), and other internal and external parameters (Θ). The change in activity per time can be represented mathematically as a non-linear function:

\[
\frac{dTFA_j}{dt} = F_j(P(t), M(t), Θ).
\]

These functions are unknown and are not explicitly considered in the model. However, we assume that the time scale of TFA change is longer than that of mRNA. Therefore, mRNAs can reach a quasi-steady state (within 10 min) while TFAs are ‘drifting’ in a time scale of hours (cell division time).

Since DNA microarray is commonly expressed in log-ratio with respect to a reference state, we derive the following log-linearized form of Equation (1) at steady-state:

\[
\log [mRNA_i(t)] = \sum_{j=1}^{k} a_j \log TFA_j.
\]  

This system can be expressed in canonical matrix form as:

\[
E = AP + Γ
\]  

The log ratios of gene expression are expressed as a linear combination of the log ratios of TFAs weighted by their control strengths [Equation (4)]. NCA is a decomposition of the data matrix E (containing the log expression ratios) into two matrices A (containing the control strengths) and P (containing the log TFA ratios) via Equation (3), where Γ is the residual to be minimized.

NCA differs from PCA and ICA in that it constrains solutions by the known connectivity between the TF and promoters (i.e. the A matrix) and the known zero TFAs (i.e. the P matrix) from mixed wild-type and knockout datasets. The connectivity constraints are derived from genome-wide DNA binding assays or extensive literature search. The presence of potential TF-promoter interaction indicates non-zero control strength to be estimated as an element in A. Otherwise, the corresponding elements in A are constrained to zero.

Note that A and P are both unknown which makes the objective of NCA decomposition similar to PCA and ICA. However, the results of PCA and ICA yield full connectivity between TF and genes. The orthogonal or independent assumption and the resulting full connectivity are not always satisfied by biological networks. Yet without additional constraints (orthogonal or independence), the solution to the matrix decomposition problem is not unique.

The data matrix E can be a composite of parallel experiments from different strains, different conditions, in addition to time series, the corresponding columns of P represents the TFA ratio under that condition or in that strain. Thus, if the TFA ratio is known to be unchanged between the experimental condition and the reference (e.g. TF knockouts), the log TFA ratio in the corresponding P elements can also be constrained to zero. If the zero patterns of A and P satisfy a set of mathematical criteria, then the decomposition of E to A and P, is unique (Liao et al., 2003; Tran et al., 2005).
2 RESULTS

2.1 The revised third criterion for NCA

To ensure unique solutions to the matrix decomposition problem, NCA requires that three criteria be satisfied (Liao et al., 2003, Tran et al., 2005). In particular, one criterion requires that the P matrix must have full row rank. This implies that the number of TFA analyzed must be smaller or equal to the number of data points. This criterion significantly limits the number of TFAs that can be derived from microarray data. To alleviate this problem, we noted that most genes are only regulated by a much smaller number of TFs than the total TFs of interest. Based on this observation, the third criterion is revised as follows (see proof in Section 5).

Given a data matrix E, the decomposition $E = A P + \Gamma$, s.t. $A \in Z_A$, $P \in Z_P$ is unique to scaling factors (or essentially unique) for a given residual $\Gamma$ if each reduced matrix $P_i (i = 1, \ldots, N)$ has full row rank, and $P \in Z_P$. The other two criteria remain the same (see Section 5).

Here $E$ is an $N \times M$ matrix of gene expression data with $N$ genes and $M$ conditions, $A$ is defined by a suitable zero pattern $Z_A$ and is $N \times L$, where $L$ is the number of TFAs ($N < L$) and $P$ is $L \times M$, also defined by a zero pattern $Z_P$ (Tran et al., 2005). $P_i$ is defined as a matrix composed of rows of $P$ corresponding to the non-zero elements in row $i$ of $A$. This new criterion is checked during iteration gene by gene. The revised criterion implies that the number of data points for each gene must be greater than or equal to the number of TFs regulating that gene. This improves the applicability of NCA, because it is likely that for most organisms, the maximal number of TFs regulating a gene is <5 or 6. For example, at least in Saccharomyces cerevisiae genes regulated by a high number of TFs probably occupy a small portion of the whole genome (Lee et al., 2002).

According to the revised third criterion, the TFs regulating each gene must be linearly independent in the available experiments. This mathematical condition cannot be satisfied a priori, but can be tested during the numerical procedure. In general, analyzing multiple dissimilar experimental conditions is likely to generate linearly independent TFAs. In practice, the condition number is checked instead of the rank of the matrix, as noisy data may have full row rank but a condition number that is high. Therefore, the condition number of each $P_i$ must be lower than a set threshold. This condition number must be checked during numerical iteration, which minimizes the Frobenius norm of $\Gamma$, $\|\Gamma\|_F$, subject to the specific zero constraints in $A$ and $P$:

$$ (A, P) = \text{arg min} (\|E - A \cdot P\|_F) s.t. A \in Z_A, P \in Z_P. \quad (5) $$

Many minimization techniques can be used, but we used the bi-linear QR factorization (Tran et al., 2005) to successively iterate $A$ and $P$ from a set of initial guesses until convergence.

2.2 Probabilistic aspects of identification

Since $A$ and $P$ are unknown we seek essentially unique decompositions of $E$ which are related by an invertible scaling matrix $X$ that preserves the zero pattern of $A$ and $P$. Thus, $A = X^{-1} \Gamma B$ and $P = X^{-1} \Gamma P$ represent all equivalent solutions. If $X$ is diagonal, it represents legitimate scaling factors. If $X$ is not diagonal then any non-diagonal elements in $X$ will recombine contributions between different TFs and degenerate the results.

If NCA criteria are violated, then it is possible to have a non-diagonal $X$ such that $A$ is still in $Z_A$, but the $X$ values are degenerate. The result is an infinite number of decompositions of $E$ into $A$ and $P$ beyond simple scaling. This is problematic for data decomposition because when $X$ is not diagonal, the solutions may be linear combinations of the columns of $A$ and rows of $P$. This problem is common to any matrix decomposition to two unknown matrices and also manifests itself in multivariate regression and factor analysis methods as colinearity, and is generally addressed by ridge regression or by the requirement of orthogonality (Anderson, 1984). Note that the degeneracy of solutions does not represent stochastic behavior of the system. It is a result of the ill-posed model.

In a Bayesian framework, the identifiability problem can be treated using a probabilistic approach. Degenerate solutions owing to violation of the identifiability criteria may in principle be distinguished using informative priors (Sabatti and James, 2005). However, uninformative priors, as in most common cases, result in ‘variance inflation’ in the posterior distribution of parameters that violates the identifiability criteria (Gelman, 2004). This is readily apparent from a simple example [Equation (6)].

Consider the following structured $A$ matrix (genes $\times$ TFs) which describes the regulation of five genes by five TFs.

$$ A = \begin{pmatrix} * & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ 0 & * & * & 0 & 0 \\ 0 & 0 & 0 & + & 0 \\ 0 & 0 & 0 & 0 & + \end{pmatrix}. \quad (6) $$

Figure 1a details a compatible $X$ that preserves the zero pattern ($Z_A$) of Equation (6). Since non-diagonal $X$ exist, some column vectors of $A$ are now linear combinations of $A$, and we cannot determine the true values of $A$ or $P$. This leads to multiple solutions not of the same family (Fig. 1b).

To investigate the significance of this problem, Figure 1c details what happens when we increment a non-diagonal term of $X$ in simulated data. The result is a uniform spread of the values of each point estimate of TFA. Thus, in a probabilistic method such as Bayesian NCA (Sabatti and James, 2005), variance inflation will be seen in posteriors of $A$ and $P$ when the NCA identifiability criteria are not satisfied.

2.3 Enforcing the third criterion

It is possible that the revised third criterion is violated during iteration. In this case, we propose to orthogonalize each reduced matrix (subnetwork) whenever the regulating TFs of each gene violate criterion 3. This renders each co-regulating TFA independent, reducing the condition number and satisfying the third criterion. The new TFAs are then used to continue the NCA iteration.

Let $P_{i+}$ represent the reduced TFA matrix for a gene $i$. The symmetric orthogonalization of $P_{i+}$ is expressed by the following equation:

$$ \tilde{P}_{i+} = (P_{i+} P_{i+})^{-1/2} P_{i+}. \quad (7) $$

Here the inverse square root of $(P_{i+} P_{i+})^{-1/2}$ is found by eigenvalue decomposition of $P_{i+} P_{i+}$ by taking the inverse square root of $P_{i+} P_{i+}$.
Each component eigenvalue. In this form no vector in $P_{ri}$ is favored over another.

For speed and numerical stability, the algorithm can be expressed without matrix inversion as follows:

$$P_{ri} = \frac{1}{2} P_{ri}^T - \frac{1}{2} P_{ri}^T P_{ri}$$

The norm in Equation (2) can be interpreted as the Frobenius norm, and $[.]$ indicates assignment. The above equation is iterated until convergence. The convergence property and the mathematical proof are detailed in Hyvarinen (1999) and Hyvarinen and Oja (2000). For clarity, what is proved is that after a suitable number of iterations, the matrix $P_{ri} P_{ri}^T$ converges to the identify matrix. We then return to recalculate matrix $A$ until the bilinear optimization converges. Note that the TFAs not included in $P_{ri}$ are not affected here. Symmetric orthogonalization (SO) methods are fundamental mathematical developments in matrix theory that have been extensively used in ICA decompositions (Hyvarinen and Oja, 2000).

### 2.4 Whitening transformation

Computation time is a significant factor for algorithms that analyze large regulatory networks. For NCA faster convergence may be obtained by preprocessing the microarray data by a whitening procedure. This method, common in neural network and data preprocessing applications, has the principal feature of flattening noise and amplifying the signal. Optimization can be performed in this transformed feature space, and the solution in the original space can be obtained by the inverse procedure as with ICA methods (Hyvarinen and Oja, 2000).

A zero-mean random vector $x$ is white if its elements are uncorrelated and unit variance. Any input data matrix can be made white by a linear transformation that decorrelates the experimental input vectors and scales them to unit variance (Hyvarinen and Oja, 2000). For microarray data, the whitening transformation can be computed from eigenvalue decomposition of the variance–covariance matrix of $E$. Let $\Sigma_E (M \times M) \equiv E^T E$ represent the variance–covariance matrix of the microarray data $E$. Since $\Sigma_E$ is positive definite and symmetric, it can always be represented as $\Sigma_E = U D U^T$ where $U$ is an orthogonal matrix having $e_1, \ldots, e_n$ eigenvectors of $\Sigma_E$ and the diagonal eigenvalues of $\Sigma_E$. Then one possible whitening transformation for $E$ is:

$$W = D^{-1/2} \cdot U^T$$

$$E = E \cdot W^T.$$  

The whitening matrix $W(M \times M)$ is positive symmetric and non-singular, and $E$ is the whitened microarray data. NCA decomposition is then performed on the new whitened matrix $E = A \cdot P$, with $A$ constrained by the same zero pattern as $A$. De-whitening of matrix $P$ (or transformation back to the original feature space) is easily accomplished by $P = PW^{-1}$. After de-whitening $P$ we then calculate the matrix $A$ from the original microarray data matrix $E$ by NCA decomposition. For this reason, the NCA identifiability criteria are preserved when $A$ is calculated.

### 3 APPLICATION AND RESULTS

#### 3.1 Network component analysis of the S.cerevisiae cell cycle

To test the biological applicability of our new algorithm, we analyzed $S.cerevisiae$ cell-cycle data (Spellman et al., 1998). We restricted the analysis to $\alpha$-factor, cdc15, elutriation and cdc28 experiments (73 data points). We generate the connectivity network from ChIP-chip binding data by selecting TF–gene interactions with those genes whose ChIP-chip binding $p$-value $< 0.001$ (Lee et al., 2002). A TF-gene pair is connected if the binding $p$-value $< 0.001$ and disconnected (set to zero in the $A$ matrix otherwise). Note that the presence of a connection does not indicate its final value. An edge may go to zero after decomposition, or to some regulation strength value. The initial values of the non-zero connections in $A$ may be set randomly. This procedure results in a connectivity network ($N_i$) that covers 96 TFs over $\sim 2200$ genes. The remaining 17 TFs were excluded owing to NCA condition 2 violations. Note that this network is data limited and not identifiable (more TFs $> $ experiments) if we were to use the previous NCA criteria (Liao et al., 2003; Tran et al., 2005). However, most genes are regulated by $< 7$ TFs so we can apply our new criterion and numerical procedure. We ran NCA using whitening and the SO
procedure (as described above) starting with a random initial guess for the non-zero CS values in the connectivity network.

3.2 Comparison of Methods that Compute TFA: NCA and REDUCE
To compare these results with REDUCE (e.g. MATRIX REDUCE) we obtained the TFAs reported by Bussemaker et al. (2001) and Gao et al. (2004). These data consist of 37 TFAs reported from the analysis of 113 TFs covering ~6000 genes over ~750 experimental data points. The authors determined that only 37 transcription factor ChIP binding patterns out of 113 were significant predictors of gene expression in one or more experiments. As REDUCE analyzes each microarray experiment individually, we can extract cell cycle TFA data points without effecting the results. The common experiments include alpha factor arrest, cdc15, elutriation experiments (Spellman et al., 1998). In addition, Gao et al. (2004) analyzed additional cell cycle data from Zhu et al. (2000) that includes fkh1/fkh2 double knockout in a-factor treatment, as well as fkh1/fkh2 double knockout cell cycle experiments. The cdc15 data include a replicate at time t = 160, so we averaged these two data points for further analysis.

The S.cerevisiae cell cycle is periodic with cell divisions approximately every 60 mins. An interesting question to ask is which TFs have periodic TFAs as these may indicate regulated/coordinate transcriptional responses and possibly provide additional evidence that a TF is in fact a cell cycle regulator. We tested the TFAs for periodicity in each individual cell cycle experiment using a robust estimator (Ahdesmaki et al., 2005) with the standard false discovery rate (FDR) adjusted p-value < 0.05. We did not detect periodicity in any of the exclusive experiments (cdc28, fkh1/fkh2 knockouts), however, in the common experiments we did detect periodic TFAs for both REDUCE and NCA (Table 1).

In the case of NCA some TFAs were periodic in two or more experiments. In REDUCE most TFAs were statistically periodic in one cell cycle experiment only. For the cdc15 data we adopted an additional Bonferroni p-value correction (0.05/23) because the data might be noisy owing to periodic stress responses induced by heat-shock arrest method (Zhang, 1999).

Figure 2 shows plots of NCA and REDUCE computed TFAs for a few common cell cycle TFs. Overall the results suggest that the REDUCE TFAs (Fig. 2a-f, dotted lines) are noisier than NCA TFAs (Fig. 2a-f solid lines), but qualitatively are in agreement with each other. For example, Figure 2a shows that the result of NCA and REDUCE are similar for the well-known cell cycle regulator ACE2. Both methods predict periodic phases that peak near cell division and this peak is characteristic of ACE2 regulation of early G1, specific gene transcription (Spellman et al., 1998). Figure 2b shows that the TFAs of SWI4 exhibits two major peaks within a cell cycle that both correspond to late G1 transcriptional activity (Spellman et al. 1998). Thus unlike the original periodicity measure by (Spellman et al., 1998) we are able to extract and detect multiple periodic peaks that correlate with the cell cycle.

The remaining REDUCE TFAs (Fig. 1d–f) are not statistically periodic when the corresponding NCA TFAs are. The most probable reason for this is that the REDUCE introduces numerical error due to fitting a fixed CS value. Of interest is the TFA of DIG1. Figure 1c shows the NCA TFA which peaks in approximately one cell cycle. This peak correlates with the current known function of DIG1; a TF that mediates cell cycle arrest in response to mating factor (α-factor) at Start (Mendenhall and Hodge, 1998).

3.3 What is the effect of false connectivity?
False connectivity includes incorrect edges (false positives) and the non-detection of true edges (false negatives) by binding site analysis or experimental methods. The sensitivity of TFAs to false connectivity in the S.cerevisiae cell cycle was investigated previously (Yang et al., 2005). The authors altered up to 10% of the connectivity for each network by randomly deleting and inserting connections between TFs and genes. They found that most TFAs, especially the cell cycle regulators, are robust and not affected by perturbation.

In this work, we assessed TFA variability by using two connectivity networks constructed from ChIP-chip data using different p-value cut-offs. Here network N1 is selected from ChIP-chip data with a binding p-value<0.001 (as described previously), and network N2 is selected analogously with ChIP-chip binding p-values<0.1. We eliminate any genes and TFs that do not satisfy NCA condition 2. In N1 we found 31 periodic TFAs out of the total 96 (FDR p-value<0.05), 23 are known cell-cycle regulators (Table 1). In N2 we found 29 periodic TFAs out of 112 (FDR p-value<0.05), 14 are known cell-cycle regulators. Between these two networks, we identified nine well-known cell cycle regulators and computed their correlation coefficients between the two networks N1 and N2 (Table 2). These results are consistent with previous results obtained with NCA by subnetwork decomposition on similar cell cycle data (Yang et al., 2005), and with the results of other statistical methods to identify cell cycle regulated TFs computed by a non-parametric test of differential expression between sets of genes bound or not bound by a ChIP-chip (Tsai et al., 2005). Thus TFA computations are robust to connectivity network selection provide the NCA criteria are satisfied.

3.4 Statistical significance of TFA
The researcher may be interested in identifying experiments where one or more TFAs are significantly perturbed with respect to the reference. To determine the statistical significance we construct a null hypothesis distribution for each experimental data point using the network connectivity pattern and shuffling (scrambling) the genes. The distribution is built by multiple iterations of permutation testing (~100). From this distribution we compute the Z-score of the original TFA for each experimental data point. The result is a p-value at each data point that determines if the computed TFA is different then the null hypothesis.

4 DISCUSSION
NCA is a method to deduce TFAs and CS from gene expression data and a connectivity network. However, previous mathematical requirements limit the total number of TFAs inferred in any single analysis to the experiment sample size. In this work, we developed a new identifiability criterion that relaxes this requirement such that the number of data points only has to exceed the maximal number of TFs regulating any single gene. Since most genes are regulated by less than five TFs, one needs only five transcriptome data points for NCA, but can analyze far more TFs than the number of data points.
4.1 NCA compared with linear methods for computing TFAs

Another method similar to NCA is REDUCE which computes TFAs from gene expression data and a connectivity network. Similar to REDUCE, NCA uses the connectivity network as a model of the global regulatory network. The significant difference between these methods is that NCA models CS as a function of gene expression data whereas REDUCE fixes CS to binding affinity (as determined by ChIP-chip assay or from promoter binding site analysis). Thus unlike NCA, REDUCE does not distinguish between positive and negative regulation of the same TF on different genes and this may affect the accuracy of TFA estimates. It is well established that binding strength does not correlate with regulation activity. Thus, using binding data as the fixed CS may compromise TFA estimates.

On the other hand, NCA is also similar to PCA or ICA methods in that it specifies constraints that guarantee a unique bi-linear data decomposition. These constraints are derived from the connectivity network. From a bipartite network point of view, PCA and ICA, assume full connectivity of TFs and genes. In addition, PCA and ICA require that each basis vector is decorrelated or independent from the others. Thus these decompositions represent a statistical interpretation of gene regulation. We argue that integrating the connectivity network directly rather than by projection methods or regression methods provides a more detailed description of biological phenomena at the TF level and thus more accurate quantification of TFAs and CS.

4.2 The importance of identifiability

NCA identifiability violations result in infinite solutions that represent linear combinations of TFAs. Therefore, these degenerative results are not interpretable from a biological standpoint. One possible way to circumvent this problem is to use a Bayesian approach.

Table 1. Periodic TFAs deduced in the cell cycle

<table>
<thead>
<tr>
<th>NCA</th>
<th>REDUCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE2</td>
<td>CIN5</td>
</tr>
<tr>
<td>ADR1**</td>
<td>CRZ1</td>
</tr>
<tr>
<td>DIG1**</td>
<td>FZF1</td>
</tr>
<tr>
<td>FKH1**</td>
<td>GCN4</td>
</tr>
<tr>
<td>FKH2**</td>
<td>HS1</td>
</tr>
<tr>
<td>MBP1**</td>
<td>NDD1</td>
</tr>
<tr>
<td>MCM1**</td>
<td>PHO4</td>
</tr>
<tr>
<td>MSA1**</td>
<td>REB1</td>
</tr>
<tr>
<td>RME1**</td>
<td>RFX1</td>
</tr>
<tr>
<td>STB1**</td>
<td>RME1</td>
</tr>
<tr>
<td>SUM1**</td>
<td>SMP1</td>
</tr>
<tr>
<td>SWI4**</td>
<td>STB1</td>
</tr>
<tr>
<td>YAP1**</td>
<td>STE12</td>
</tr>
<tr>
<td>BAS1</td>
<td>SW15</td>
</tr>
<tr>
<td>CAD1</td>
<td>SW16</td>
</tr>
</tbody>
</table>

Double asterisks (**) indicate periodicity detected in two or more experiments. Boldface highlights known cell-cycle regulators in the literature. REDUCE data is reported from Gao et al. (2004).
approach that uses different prior probabilities for different solutions (Sabatti and James, 2005). In this framework, the equivalent solutions due to non-identifiability are distinguished based on given prior probability distributions. This approach may work well if informative prior distributions are available. However, without such information, violation of the identifiability criteria leads to variance inflation, and the ability to distinguish among equivalent solutions is degenerated. This situation is particularly severe if a large fraction of the network violates the identifiability criteria. Therefore, it is important to detect whether the problem is well-posed before analysis.

In conclusion, we note that it is important to determine the identifiability of bi-linear data decomposition such as NCA. In this regard, this article provides a new criterion that allows realistic TFAs to be determined from less data in both deterministic and probabilistic approaches.

5 PROOF OF CRITERION 3 FOR NETWORK COMPONENT ANALYSIS

**DEFINITION 1.** All \( N \times L \) matrices (defined as \( A \)) in which a given set of positions are zero are said to have the same zero pattern. Such matrices belong to the set \( Z_A \), which is defined as

\[
Z_A \equiv \{ A \in \mathbb{R}^{N \times L} | A_{ij} = 0 \text{ for a given set of } (i,j) \}. \tag{10}
\]

The elements not specified by the zero pattern can take positive, negative or zero values. Similarly, the zero pattern for \( P \ (L \times M) \) is defined as

\[
Z_P \equiv \{ P \in \mathbb{R}^{L \times M} | P_{kl} = 0 \text{ for a given set of } (k,l) \} \tag{11}
\]

**DEFINITION 2.** To define the reduced matrices \( G_{ij} \) \( (j = 1, \ldots, L) \), which is used in criterion 2 below, we will first build the combination matrices \( G_j((N \times MN) \times (L + 1)) \) from \( A \ (N \times L) \) and \( P(L \times M) \). \( G_j \) is mathematically described as

\[
G_j = \begin{bmatrix}
A_{ij} & A \\
\vdots & \vdots \\
P_{ij} & Q(P_{ij}^T)
\end{bmatrix}, \tag{12}
\]

where \( M_{ij} \) stands for the column vector \( j \) of matrix any \( M \), (which could be either \( A \) or \( P \)) and \( Q(v) \) stands for the block diagonal matrix with each block being vector \( v \). For example, consider a matrix \( P \),

\[
P = \begin{pmatrix}
2 & 0 & 1 \\
0 & 5 & 4
\end{pmatrix}
\tag{13}
\]

the matrix \( Q(P_{ij}^T) \) is

\[
Q(P_{ij}^T) = \begin{pmatrix}
2 & 0 \\
0 & 0 \\
1 & 0 \\
0 & 2 \\
0 & 0 \\
0 & 1
\end{pmatrix}
\tag{14}
\]

The reduced matrix \( G_{ij} \) is now defined as matrix \( G_j \) without rows corresponding to non-zero elements in the left most column.

**THEOREM.** Given a matrix \( E \ (N \times M) \), if the following conditions are satisfied, then the decomposition of \( E = AP + \Gamma \), where \( A(N \times L) \in Z_A \), \( P(L \times M) \in Z_P \) is unique up to a scaling non-singular diagonal matrix \( X(L \times L) \).

In other words, for the same allowable \( \Gamma \), any alternative decompositions of \( E = \tilde{A}P + \tilde{\Gamma} \) where \( \tilde{A}(N \times L) \in Z_A \), and \( \tilde{P}(L \times M) \in Z_P \) are different only by a diagonal matrix \( X(L \times L) \) i.e.

\[
\tilde{A} = AX^{-1} \tag{15}
\]

\[
P = XP \tag{16}
\]

Criterion 1: \( A \) has full column rank.

Criterion 2: Each reduced matrix \( G_{ij} \) \( (j = 1, \ldots, L) \), defined in Tran et al. (2005) and in Definition 2, has a rank of \( L - 1 \).

Criterion 3: Each row of \( E \) is regulated by rows of \( P \) that are linearly independent.

Conditions 1 and 2 are the same as those proven previously in Tran et al. (2005), and the Condition 3 will be proved in the following section.

5.1 Proof of the third criterion:

Suppose we find \( A, P, \tilde{A} \) and \( P \) such that

\[
E - \Gamma = AP = \tilde{AP}. \tag{17}
\]

We want to show that there exists an invertible matrix \( X(L \times L) \), such that

\[
\tilde{A} = AX^{-1}, \tag{18}
\]

\[
P = XP, \tag{19}
\]

and the invertible matrix must be a diagonal matrix under certain conditions. Since \( \tilde{A} \) has full column rank (Criterion 1) we can write:

\[
\tilde{A}^TAP = \tilde{A}^T \tilde{AP}, \tag{20}
\]

or

\[
P^T = P^T X^T. \tag{22}
\]
From above equations we obtain the following:

\[ AP = AXP \]  \hspace{1cm} (23) 

which can be written as

\[ (A - AX)\mathbf{P} = 0. \]  \hspace{1cm} (24) 

There are two cases that satisfy the above equation.

i. \( \mathbf{P} \) has full row rank, which is condition 3 previously used in Liao et al. (2003); Tran et al. (2005), and Equation (24) implies that \( A - AX = 0 \), which is equivalent to

\[ \mathbf{A} = \mathbf{AX}. \]  \hspace{1cm} (25) 

In Tran et al. (2005), it was proved that if the all \( G_{ij} (j = 1, \ldots, L) \) derived from \( \mathbf{A} \) and \( \mathbf{P} \) satisfy Condition 2, \( \mathbf{X} \) must be diagonal so that Eq. A.20 can be satisfied with both \( \mathbf{A} \) and \( \mathbf{X} \) belonging to \( \mathbf{Z}_\mathbf{A} \). Otherwise, \( \mathbf{A} \in \mathbf{Z}_\mathbf{A} \) and \( \mathbf{A} \in \mathbf{Z}_\mathbf{A} \) that can relate to each other by any non-singular matrix \( \mathbf{X} \).

ii. \( \mathbf{P} \) has rank of \( r \leq \min (L,M) \), and thus \( A - AX \in N(P^j) \), where \( N(P^j) \) is the left null space of \( \mathbf{P} \) and be described by \( \mathbf{N} \in \mathbb{R}^{q \times L} \) with \( q = L - r \):

\[ \mathbf{NP} = 0 \]  \hspace{1cm} (26) 

Since \( A - AX \in N(P^j) \), we can write:

\[ A - AX = \mathbf{WN}, \]  \hspace{1cm} (27) 

where \( \mathbf{W} \in \mathbb{R}^{q \times L} \) is a linear coefficient matrix. If \( \mathbf{X} \) is a non-singular diagonal matrix, then

\[ \mathbf{A} = (\mathbf{A} - \mathbf{WN})\mathbf{X}^{-1}. \]  \hspace{1cm} (28) 

If there exists a non-zero coefficient matrix \( \mathbf{W} \) such that \( \mathbf{WN} \in \mathbf{Z}_\mathbf{A} \), another solution family \( \mathbf{A} \) cannot be scaled directly to the first solution family \( \mathbf{A} \) by Equation (28). Therefore, we need to find the condition of \( \mathbf{N} \), so that nonzero matrix \( \mathbf{W} \) does not exist.

If there exist a nontrivial \( \mathbf{W} \), such that \( \mathbf{WN} \in \mathbf{Z}_\mathbf{A} \), then for each row (gene) vector \( i \) of \( \mathbf{Z}_\mathbf{A} \), \( \mathbf{A}_i (1 \times L) \), can be written as the product of row vector \( i \) of \( \mathbf{W} \), \( \mathbf{W}_i (1 \times q) \), and matrix \( \mathbf{N}(q \times L) \):

\[ \mathbf{A}_i = \mathbf{W}_i \mathbf{N} \]  \hspace{1cm} (29) 

Since we only know the zero elements \( j \) in row vector \( \mathbf{A}_i \), Eq. A.24 is reduced to

\[ 0 = \mathbf{WN}_j, \]  \hspace{1cm} (30) 

where reduced matrix \( \mathbf{N}_j(q \times z) \) consists only columns \( j \) that correspond to \( z \) zero elements in \( \mathbf{A}_i \). However, if \( \mathbf{W}_i \) has rank \( q \), the row vector \( \mathbf{W}_i \) must be zero.

Therefore, the necessary condition is \( z \geq q \). If \( r = M = \min (L,M) \), then \( z \geq L - M \), and hence the number of non-zero element in \( \mathbf{A}_i \) (i.e. the number of transcription factors regulating gene \( i \), \( l = L - z \), must be less than \( M \) (i.e. the number of experiments).

Note that the left null matrix \( \mathbf{N} \in \mathbb{R}^{q \times L} \) having rank \( q \) represents \( q \) sets of linear constraints on the dependent rows of \( \mathbf{P} \), and each column \( j \) of \( \mathbf{N} \) defines linear coefficient of row \( j \) with other rows in \( \mathbf{P} \). That means if row \( j \) of \( \mathbf{P} \) is linearly independent to all other rows, the column \( j \) of \( \mathbf{N} \) will be zero. For example, if row 1 and 2 of \( \mathbf{P} (5 \times 4) \) are linearly dependent, the left null matrix \( \mathbf{N} \) is

\[ \mathbf{N} = \begin{bmatrix} n_1 & n_2 & 0 & 0 & 0 \end{bmatrix}, \]  \hspace{1cm} (31) 

where \( n_i \) represent the linear coefficient of row 1 and 2 of \( \mathbf{P} \) such that

\[ n_1P_{1k} + n_2P_{2k} = 0 \] \hspace{1cm} for \( k = 1, \ldots, M \) \hspace{1cm} (32) 

Because the reduced matrix \( \mathbf{N}_{ij} \) is full row rank (\( = q \)), then the non-zero members in each rows of \( \mathbf{N} \) (corresponding to dependent rows of \( \mathbf{P} \)) cannot be eliminated all together, which means the dependent rows of \( \mathbf{P} \) will not combine together in regulating row \( i \) of \( \mathbf{E} \) (Criterion 3). For the above example, \( \mathbf{N}_{ij} \) is rank deficiency when both the first and second columns are eliminated. That means the corresponding row \( \mathbf{A}_i \) has non-zero elements in both the first and second columns. Therefore, each row \( i \) of \( \mathbf{E} \) must be regulated by rows of \( \mathbf{P} \) that are linearly independent.

**ACKNOWLEDGEMENTS**

This work was supported by NSF grant CCF-0326605 and the UCLA-DOE Genomics and Proteomics Institute. S.I.G. was partially supported by a UCLA-IGERT bioinformatics traineeship (NSF-IGERT 9987641). L.M.T. was supported by a UCLA-IGERT bioinformatics traineeship (NSF-IGERT 9987641).

**Conflict of Interest:** none declared.

**REFERENCES**


Yang, Y.L. et al. (2005) Inferring yeast cell cycle regulators and interactions using transcription factor activities. BMC Genomics, 6, 90.