Published in final edited form as:
IEEE Symp Comput Intell Bioinforma Comput Biol Proc. 2012 May ; 2012: 387–396. doi:10.1109/CIBCB.2012.6217256.

Matrix Factorization for Transcriptional Regulatory Network Inference

Michael F. Ochs and
School of Medicine, Johns Hopkins University, Baltimore, MD 21205

Elana J. Fertig
School of Medicine, Johns Hopkins University, Baltimore, MD 21205

Michael F. Ochs: mfo@jhu.edu; Elana J. Fertig: ejfertig@jhmi.edu

Abstract

Inference of Transcriptional Regulatory Networks (TRNs) provides insight into the mechanisms driving biological systems, especially mammalian development and disease. Many techniques have been developed for TRN estimation from indirect biochemical measurements. Although successful when initially tested in model organisms, these regulatory models often fail when applied to data from multicellular organisms where multiple regulation and gene reuse increase dramatically. Non-negative matrix factorization techniques were initially introduced to find non-orthogonal patterns in data, making them ideal techniques for inference in cases of multiple regulation. We review these techniques and their application to TRN analysis.

Index Terms

Matrix factorization; NMF; Transcriptional Regulatory Network; Bayesian statistics

I. Introduction

Transcriptional Regulatory Networks (TRNs) provide a method for cells to reprogram their functions as needed for survival or multicellular interactions. These networks comprise transcriptional regulators (or transcription factors, TFs) and their target genes. As one of these genes may encode an additional TF, a propagation of primary, secondary, tertiary, etc. transcription can occur, with a branching out from the activation of an initial TF to a growing set of TFs and targets. Overall, the TFs so induced form the TRN.

Following the advent of microarrays for measurement of global mRNA levels in the mid-1990's, it was realized that time series data gathered during the course of an experimental protocol could be used to look at TRNs globally within cellular systems. Early success was shown by Hartemink *et al* and Shmulevich *et al* in recovering TRNs in yeast [1]–[3] and Sabatti *et al* in bacteria [4], however translation to higher eukaryotes and mammalian systems did not prove successful. Unfortunately, it has become routine for methods that succeed dramatically within yeast to prove problematic when applied to higher organisms in other computational areas as well [5]. For example, prediction of TF targets from transcription factor binding sites (TFBS) identified through sequence similarity does not predict expression in higher organisms. Similarly, while Chromatin

ImmunoPrecipitation on microarray (ChIP-chip) technology can identify all candidate TF targets, it has not provided good predictors of gene expression in higher organisms. The reasons for this failure likely lie in the differences in gene regulation and genomic complexity.

There are three primary factors that complicate TRN prediction in multicellular organisms. First, gene reuse in multiple biological processes is dramatically increased in higher organisms. This leads to multiple regulation of genes by multiple TFs, which introduces mathematical complexity to the determination of the TF responsible for a change in expression of a target. Second, many genes are regulated post-transcriptionally, either through translational regulation or post-translational modification. For instance, many TFs require post-translational modification or cofactor binding to initiate transcription. Third, epigenetics, such as silencing by chromatin formation or DNA methylation, play a far larger role in multicellular systems than in prokaryotes and yeast. This substantially complicates the relationship between TF activity and target expression.

These three complications require new approaches and sometimes new data sources when building TRNs. The multiple regulation issue is being addressed through matrix factorization methods, which we will focus on in this review. The post-transcriptional regulation of genes leads to several issues. Perhaps most critically, it leads to many genes not being under transcriptional control, leading to substantial variance in transcript levels for these genes independent of protein level changes and functional consequences. This suggests a need to incorporate estimates of random variability in expression, which can be incorporated into individual matrix factorization techniques. The epigenetic factors influencing TFBS site access and transcriptional availability of genes requires techniques that limit the strength of priors from TFBS data to insure accurate inference in multicellular systems. In addition, integration of data measurements, such as methylation status of TFBS elements, can provide additional information to guide TRN estimation from expression data.

In this review, we focus on the development of matrix factorization in the analysis of microarray data. We highlight particularly the value of these methods to TRN prediction and address the value of including error modeling within the analyses.

II. Matrix Factorization for Expression Data

In order to address problems similar to those arising in multicellular gene expression data, new matrix factorization techniques coupled to dimensionality reduction were introduced simultaneously by ourselves in Bayesian Decomposition (BD) for spectral imaging [6] and by Lee and Seung in Nonnegative Matrix Factorization (NMF) for image processing [7]. Both techniques aimed to address the limitations of analytical methods in handling inherently positive data where the natural basis vectors to describe the data were non-orthogonal. The techniques developed to deduce the non-orthogonal basis vectors showed particular potential in inferring multiple regulation for TRN inference.

A. The Universe of Matrix Factorization

The fundamental problem of factoring a matrix to find structure to explain the physical world recurs in numerous fields, which has led to the development of similar methods under many names. Following the broader history in the development of matrix factorization techniques, the first methods that were widely used in microarray studies included the standard statistical techniques of Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) [8]. The realization of the limits of orthogonality led us to apply BD to microarray data in 2002, showing that this significantly improved inference on the yeast cell cycle [9]. Later studies demonstrated the value of BD when applied to human patient data [10], and we developed an open-source algorithm, CoGAPS, linked to R to simplify applications [11]. While BD can be considered a form of Independent Component Analysis (ICA), it is driven to inherently sparse solutions, which appears important for inference on expression data.

NMF methods, which are again similar to ICA, were applied to microarray data by Kim and Tidor in 2003 [12], and the term metagene in the NMF context was coined by Brunet *et al* in 2004 [13]. As with ICA, initial NMF variants tended to smooth solutions that appeared to limit the inference of biological processes, leading Gao and Church to introduce sparse-NMF in 2005 [14]. Additional NMF methods continue to be introduced, with constant improvements in speed.

A second Bayesian approach to matrix factorization, Bayesian Factor Regression Modeling (BFRM) [15], was applied to microarray data by Carvalho et al in 2008 [16], although it had been applied to microarray data in clinical studies prior to this publication. BFRM induces sparse solutions by removing the mean behaviors in the data, which tends to highlight differences between samples.

B. Related Techniques

There are a number of techniques closely related to matrix factorization that have been applied to expression data. Blind Source Separation (BSS) was developed to isolate the signal coming from a single source within a background of multiple sources, such as the speech of a single individual in a crowded room. BSS was applied to microarray data initially in 2004 [17]. Biclustering is a technique focused on identifying subsets of similar behavior within a matrix, where these subsets may overlap. Unlike matrix factorization, it does not lead to a decomposition into two matrices that can reconstruct the original matrix, but instead provides insight into two-dimensional correlations within the data [18].

C. Basic Nomenclature

In this section we introduce the nomenclature we will use for this paper. Different techniques have defined and normalized the computed matrices in different ways, and we will attempt to follow the statistical approach here, with **Y** representing a matrix of observations and **X** the true expression levels in the samples. In NMF, it is typical to decompose **X** into matrices **H** and **W** (**X** = **HW**), with normalization being performed on the columns of **H**. However, in gene expression, it is somewhat easier to interpret a

decomposition into **A** and **P**, with normalization on the rows of **P**. This allows interpretation of the rows of **P** as basis vectors defining the presence or absence of *Patterns* of coexpression across samples and **A** as the assignment of genes to the patterns. Figure 1 provides the relationship of the matrices together with an indication of the dimensionality reduction.

The fundamental equation therefore is

$$\mathbf{Y}=\mathbf{X}+\sigma=\mathbf{AP}+\sigma=\sum_{p}A_{np}P_{pm}+\sigma_{nm} \quad (1)$$

where the terms are described in the previous paragraph except for σ, which provides an element-level noise or error value on **Y**. The indices in the summation form are *n* indexing the genes, *m* indexing the samples (typically time series measurements for TRN estimation), and *p* indexing the patterns (i.e., rows of **P**). Because σ arises from both the biological system and technical artifacts, it is sometimes treated as two terms. However, with modern technology it is now clear that technical noise, excluding batch effects, is trivial compared to the variation seen on transcript levels in biological systems, so we use a single term and treat only “biological noise”.

Estimating the biological random noise is always part of finding **A** and **P**, even if it is implicit in the methodology. Typically, if the noise is not explicitly addressed within the mathematical model, fitting will treat the noise as Gaussian with each individual matrix element of **Y** including i.i.d. (independent, identically distributed) samples from the implied common error distribution. However, even in yeast, this assumption is not justified by the biology. As demonstrated in an experiment including 63 flasks of yeast under identical growth conditions, roughly 10% of yeast genes varied by orders of magnitude in transcript levels without any corresponding effect on phenotype [19]. Furthermore, studies in yeast [20] and human [21] showed poor correlation between mRNA levels and protein levels. This suggests that methods should explicitly include noise terms and not treat these as equal for all genes, but instead use gene-specific estimates of variance. It is hoped as data sets grow, that cell-type gene specific variance estimates may become available. Even simple models of noise identified in early work [22] have been shown to improve estimates of biological activity from microarrays [23].

It is possible that in the future this error estimation may be further improved through increased understanding of translational and regulatory processes related to individual genes. This may allow for gene-specific models that include *a priori* modeling of the probability of the translation of the mRNA to protein and potential inclusion of models of alternative splice variants and their proteins.

D. TRN Estimation

While matrix factorization will provide insight into multiple regulation by identifying coordinated patterns that can be linked to TF activity, it does not directly solve TRN structure. This relies on more complicated modeling approaches that balance adherence to

the underlying biology against computational complexity and parameter explosion. A review of the methods utilized for TRN prediction from expression data and prior knowledge is beyond the scope of this work, however Karlebach and Shamir summarized widely-used techniques previously [24].

TRN prediction often relies on two additional pieces of information beyond gene expression data. The first piece of information is TFBS data and prediction, though as noted above this should not provide too strong a prior in multicellular systems. TFBS data can be generated through ChIP-chip measurements, and coupling of this data with expression data allowed modeling of the yeast cell cycle [25]. More recently TFBS data has been generated through ChIP-seq measurements, where the immunoprecipitation pulldown is directly sequenced. The second piece of information is the result of text mining, with a focus on identification of targets of TFs as reported in the literature. While all these data sources can provide direct evidence of regulation, context often plays a role through epigenetics, so care must be taken in the use of this information as well.

Matrix factorization can provide an important insight for TRN estimation. The patterns will indicate when corresponding genes are active, and as shown in *Sections IV.B and V.A*, the genes associated with patterns can provide inference on TF activity. Since the factorization methods can be designed to appropriately utilize variance estimates and epigenetic measurements, the TF activity estimates should be more robust.

E. Issues for Matrix Factorization

The mathematics of the decomposition (or factorization) of Eqn. 1 introduces a number of issues. First, since the rows of **P** can be considered basis vectors, there are an infinite number of sets of these bases that fit the data equally well. Second, it is often the case that the actual number of dimensions that are required to fit the data within the noise is less than the smaller dimensionality of the **Y** matrix, but the minimum is often unknown. Third, the matrix **X** is unchanged by exchanges of magnitude between columns of **A** and rows of **P**.

The first issue of infinitely many solutions can be resolved by considering Fig. 2, in which we imagine a distribution in three dimensions that can be described by the standard Cartesian system (labelled *x*, *y*, *z*). Naturally, any rotation of the standard system will also describe the data, or alternatively a non-orthogonal system (labelled *a*, *b*, *c*) can be used provided $a,\vec{b},\vec{c}$ are non-colinear. Ideally these non-orthogonal bases will be related to the biological processes active within the samples. As PCA and SVD define dominant directions in the data and force orthogonality in the basis vectors along directions of maximum residual variance, each vector would then necessarily combine signals from many processes active in the biological system, confounding inference of multiple regulation in analyses relying on these methods.

One method to limit potential solutions is essentially to limit the matrix **X**. For microarray data, since gene expression is an inherently positive value as it represents an estimation of the concentration of mRNA, it is natural to treat $\mathbf{X} > 0$. This leads to the basis *abc* being natural for the factorization, as it fully captures the distribution of the positive data with

coordinates that naturally match the underlying structure in the data. It still remains to choose the best non-orthogonal bases, and this is done in different ways, as discussed below.

The second issue of dimensionality estimation remains problematic. Numerous methods that have been developed in other fields have had only limited success in applications to gene expression data. Both ad hoc methods [26] and formal methods [27] have been proposed and applied, however a definitive method remains elusive. Many of these dimensionality estimates rely on strong assumptions about the error model in **Y**, which are inapplicable to the strongly correlated and deeply structured microarray data. The best approach may be to try many dimensionalities and look for robust patterns in the results, similar to robust clustering methods.

The third issue of flexibility of solutions to exchange of flux between the **A** and **P** matrices is generally solved by normalization of either a column of **A** or a row of **P**. Either is a valid mathematical solution, and we have argued above for normalization of rows of **P** to increase biological interpretability. However, the best choice will depend on the particular application.

III. Review of Matrix Factorization Methods

Here we review the basics of matrix factorization methods applied to gene expression data, focusing on those methods that have been applied widely to gene expression data.

A. Independent Component Analysis

One method closely related to matrix factorization is independent component analysis (ICA). Like typical applications of PCA to microarray data, ICA performs matrix decomposition by projecting the data onto a lower dimensional space, using statistical independence between components rather than orthogonality. Since the observed microarray signals are a result of a mixture of underlying biological processes, the factorization of the data matrix, **Y**, can be expressed as

$$\boldsymbol{Y}=\mathbf{X}+\sigma=f(\mathbf{AP})+\sigma, \quad (2)$$

where **A** represents assignment of genes to patterns and **P** the patterns as in Eqn. 1.

For the case of linear ICA, the estimation of **P** can be formulated as

$$\mathbf{P}\approx\hat{\mathbf{P}}=\mathbf{WY}, \quad (3)$$

so that we need to find a matrix **W** (the unmixing matrix), such that the rows of matrix $\hat{\mathbf{P}}$ are statistically independent though not orthogonal. The process of finding the unmixing matrix can be performed by different algorithms, based on different metrics of statistical independence. Pournara and Wernisch provided a thorough review of ICA and other factor analysis approaches in TRN estimation [28].

B. Bayesian Decomposition and CoGAPS

Although the statistical independence requirements of ICA are not as strict as the orthogonality requirements of SVD and PCA, the assumption of independence between the underlying processes may not be fully justified in most microarray data due to multiple regulation and coordinated activation of biological processes. In order to allow bases that were not statistically independent, BD was introduced in 1999 for spectral imaging [6]. It was first applied to gene expression data in 2002 [9], and an open-source approach linked to R/Bioconductor, called CoGAPS, was created in 2010 [11].

BD applies several approaches to identify those **A** and **P** matrices that best explain the data **Y** for gene expression data. First, it applies a sparseness criterion through use of an atomic prior that is penalized for the addition of structure [29]. Second, for expression data, it applies a positive mapping from the inherently positive atomic domain to **A** and **P** limiting the solutions to positive matrices. Third, dimensionality reduction is typically used to limit the number of parameter estimates needed, so that $p \ll \min(N, M)$.

In order to determine how to place and size atoms within the atomic domain, a Markov chain Monte Carlo (MCMC) procedure is used. Atoms are created *ex vacuo* according to the prior, and atoms can be resampled, destroyed, moved, or have flux moved to neighboring atoms (see [30] for details). Using

$$\mathbf{A}=\int K_a\varphi_a \quad \mathbf{P}=\int K_p\varphi_p \quad (4)$$

convolution functions map atoms to matrix elements allowing preferred correlations between matrix elements to increase in probability. Through the convolutions, a set of values (e.g., **A**) can be constructed from a family of measures, ϕ (the atoms), using kernels, *K*. In the simplest case, an atom simply maps to a single matrix element.

The probability for each combination of **A** and **P** is determined from Bayes rule,

$$p(\mathbf{A},\mathbf{P}|\boldsymbol{Y}) \approx p(\boldsymbol{Y}|\mathbf{A},\mathbf{P})p(\mathbf{A},\mathbf{P}), \quad (5)$$

where $p(\mathbf{A}, \mathbf{P}|\mathbf{Y})$ provides the conditional probability of the model given the data (the *posterior*), $p(\mathbf{Y}|\mathbf{A}, \mathbf{P})$ the conditional probability of the data given the model (the *likelihood*), and $p(\mathbf{A}, \mathbf{P})$ the probability of the model (the *prior*).

A key feature of BD and CoGAPS, as well as most other Bayesian methods, is explicit modeling of the error distribution. In BD and CoGAPS, a least squares likelihood model is used, effectively treating the errors as Gaussian with zero mean. As the data is more closely log-normal than normal in distribution and the error appears to be multiplicative with expression level, it is usually best to work on log-transformed data. For BD and CoGAPS, the individual error elements, σ_{ij}, can be estimated from the data, which addresses the wide, phenotype-independent variance seen in expression levels in eukaryotic systems.

Furthermore, the MCMC framework can be easily extended for inference using RNA-seq measurements by including error models for count data.

CoGAPS utilizes an MCMC structure like BD, however CoGAPS incorporates more control over hyperparameters that determine sparseness. Both methods handle missing data in a parsimonious way, since the fitting of the model utilizes the likelihood, allowing missing data to be incorporated through use of high uncertainties ($\sigma_{ij} \gg 1$).

C. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) was adapted for analysis of gene expression data by Kim and Tidor [12] and Brunet *et al* [13], with the columns of **A** being referred to as metagenes. The goal of the NMF analysis is to find a small number of metagenes, effectively performing a dimensionality reduction. The expression estimates **X** are then approximated as a positive linear combination of the metagenes.

As with BD and CoGAPS, NMF provides an inherent reduction in dimensionality. In an NMF simulation, random matrices **A** and **P** are initialized according to some scheme, such as from a uniform distribution $U[0, 1]$. The matrices are updated iteratively using

$$\begin{aligned} P_{\alpha\mu} &\leftarrow P_{\alpha\mu} \frac{\sum_i A_{i\alpha} Y_{i\mu}}{\sum_i A_{i\alpha} X_{i\mu}} \\ A_{\delta\alpha} &\leftarrow A_{\delta\alpha} \frac{\sum_j Y_{\delta j} P_{\alpha j}}{\sum_j X_{\delta j} P_{\alpha j}} \end{aligned}, \quad (6)$$

which guarantees reaching a local maximum in the likelihood and minimizes

$$||Y-X||^2 = \sum_{ij} (Y_{ij} - X_{ij})^2. \quad (7)$$

Since **P** is obtained without any requirements on relationships between the rows, there are no independence criteria, as exist in SVD or ICA. The key assumptions allowing identification of a unique solution are non-negativity of the **A** and **P** matrices and the reduction of dimensionality. The absence of additional constraints does lead to a tendency for the recovery of signal-invariant metagenes that carry little or no information. Gao and Church introduced sparse NMF (sNMF), which penalized solutions based on the number of non-zero components in **A** and **P**, to address this issue [14]. Carmona-Saez *et al.* applied a similar approach in non-smooth NMF (nsNMF) through introduction of a smoothness matrix into the factorization [31].

As traditional NMF techniques do not account for uncertainty information, overfitting of the data can be an issue. In addition, the treatment of all variances as equal raises a potential problem for eukaryotic data. Least-squares NMF (lsNMF) introduced new updating rules, effectively replacing the criterion for distance minimization with a minimization of the χ^2 error [23], given by

$$\chi^2=\sum_{ij}\left(\frac{(Y_{ij}-X_{ij})}{\sigma_{ij}}\right)^2. \quad (8)$$

Using this measure, the update rules become

$$\begin{aligned} P_{\alpha\mu} &\leftarrow P_{\alpha\mu}\frac{\sum_i A_{i\alpha}\frac{Y_{i\mu}}{\sigma_{i\mu}}}{\sum_i A_{i\alpha}\frac{X_{i\mu}}{\sigma_{i\mu}}} \\ A_{\delta\alpha} &\leftarrow A_{\delta\alpha}\frac{\sum_j \frac{Y_{\delta j}}{\sigma_{\delta j}}P_{\alpha j}}{\sum_j \frac{X_{\delta j}}{\sigma_{\delta j}}P_{\alpha j}} \end{aligned}, \quad (9)$$

and the algorithm proceeds as with NMF to find a local maximum in probability according to the minimization of Eqn. 8.

All NMF methods fail to explore the posterior probability distribution, due to the inability to escape local maxima. The approach to this problem has been to routinely begin with 50 to 200 different initial random **A** and **P** matrices, then to look for the solution which provides the best fit to the data, as measured by Eqns. 7 or 8 [32]. Alternatively, robust solutions can be looked for within the factorizations, providing an indication of reliable patterns (i.e., metagenes) within the data. However, in our experience, the metagenes obtained from reasonably complex data sets can vary in terms of their χ^2 fit to the data by two orders of magnitude. As such, care must be taken to make sure that an adequate number of simulations using an NMF method have been attempted before interpreting the results.

Witten, Tibshirani, and Hastie provided a generalized penalized matrix decomposition framework [33]. This framework allows specification to a number of specific forms, including sparse SVD and non-negative sparse coding, similar to sNMF. An R package is now available for implementing many NMF approaches [34].

D. Bayesian Factor Regression Modeling

BFRM is a Bayesian MCMC technique, like BD and CoGAPS. However, it includes simultaneous solution of a linear model of the covariates, by solving

$$Y_{nm}=\mu_n+\sum_k \beta_{nk}h_{km}+\sum_p A_{np}P_{pm}+\sigma_{nm}, \quad (10)$$

where **A** can be viewed as factor loadings for latent factors **P** [16]. The **h** matrix provides known covariates in the data, and the mean vector, μ, provides a gene specific term that adjusts all genes to the same level. The patterns then are those needed after accounting for mean behavior and covariates. Like BD and CoGAPS, BFRM seeks to minimize structure in **A** and **P**.

BFRM also addresses the issue of the number of patterns through an evolutionary stochastic search. The algorithm attempts to increase the number of patterns by thresholding the

probability of inclusion of a new factor. The model is refit with the additional pattern with retention of the additional dimension if there is an improvement according to the criterion chosen. Evolution ceases when no additional factors are accepted. The BFRM software also allows running without this evolution.

E. Network Component Analysis

Network Component Analysis (NCA) uses information on the binding of TFs to DNA to reduce the possible **A** and **P** matrices [35]. Essentially, a two layer network is created, with one layer populated by transcriptional regulators and the other by their gene targets. Edges then connect each TF to the genes they regulate.

NCA handles the problem of multiple potential solutions in Eqn. 1 by including all potential solutions through

$$\boldsymbol{Y}=\mathbf{X}+\sigma=\mathbf{ATT}^{-1}\mathbf{P}+\sigma, \quad (11)$$

where **AT** includes all potential **A** matrices. By requiring **T** to be diagonal, **A** and **P** are unique up to a scaling factor. Diagonality of **T** can be guaranteed if 1) **A** is full column rank, 2) removal of a TF yields a network where **A** is still full column rank, and 3) **P** is full row rank. Criterion 3 demands that a set of TFs be considered linearly independent, which is reasonable biologically.

The solutions, **A** and **P**, are determined by minimizing Eqn. 7, just as for NMF. As with lsNMF, this could be easily extended to gene and array specific errors by inclusion of specific error terms. As with BD and CoGAPS, the rows of **P** are normalized so that each row provides the average effect of a regulator.

When applying NCA to microarray data, the relative strength of the TF in regulating its target must be determined. For each gene and each TF, the gene regulation is assumed to be proportional to the binding affinity of the TF to the promoter for a gene. Since each gene can be regulated by multiple regulators, the expression of a gene at a time point must be estimated as a combination of the regulation from different factors. For each time point, this is estimated as

$$\frac{Y_i^j}{Y_i^0}=\prod_{k=1}^{R}\left(\frac{\mathrm{TF}_k^j}{\mathrm{TF}_k^0}\right)^{\mathrm{Aff}_{ik}} \quad (12)$$

where Y_i is the expression for gene i, with the superscript indicating the time point, TF_j is the activity of the j^{th} transcriptional regulator with the superscript indicating the time point, R is the total number of regulators, and Aff_{ik} is the binding affinity for the k^{th} transcriptional regulator on the i^{th} gene. This is effectively a log-linear model where the transcriptional binding affinity is taken as a measure of the strength of gene activation, and each regulator effectively leads to a multiplicative increase in gene expression.

IV. Comparisons of Matrix Factorization Methods

The matrix factorization methods presented here have a number of tunable parameters making complete comparison difficult. Comparisons have been made through simulations, where the ground truth is well-defined but errors do not reflect likely strong correlations between genes and processes, and through analysis of data, where ground truth is generally not known but data is realistic. We present a brief summary of previous comparisons and then provide an example from our work in cancer, where several TFs were validated by Western blots using phosphoantibodies.

A. Previous Reviews

A number of reviews have focused on comparison of different NMF methods [32], the ability of NMF to recover biologically meaningful patterns [36], different methods for identifying coregulation [37], and methods for gene set analysis [38].

Devarajan notes that sparseness is a critical aspect for “parts-based” decomposition in NMF, as it provides localized patterns, so that the whole can be reconstructed from these localized features [32]. This sparseness may need to be enforced by penalizing non-sparse solutions. In addition, the use of cophenetic correlation is discussed as a method to estimate dimensionality, similar to scree plots in SVD or PCA.

Kossenkov and Ochs generated a simulated yeast cell cycle data set with errors based on real data and compared a wide variety of methods in terms of recovery of known coregulation in the background of multiple regulation [37]. BD and sNMF performed best, with AUCs of 0.98 and 0.94 respectively. NMF, lsNMF, and ICA also performed well, while PCA and clustering techniques performed relatively poorly.

In a second study, they used the Rosetta Compendium of yeast deletion mutants [19] and functional annotation to study which methods recovered biological behaviors from the data when coupled to gene set analysis of 15 well-studied processes [38]. After factorization, signatures of enrichment were identified for each method. In this case BD performed best, identifying 7 terms, with BFRM identifying 5. Clustering methods identified 4 terms, while NCA, PCA, and ICA all did more poorly. This may reflect the lack of sparseness as sNMF was not used in this study.

Overall this suggests that the Bayesian factorization methods, which naturally include sparseness, or sparse NMF methods should be the first choices for matrix factorization when the goal is TRN estimation.

B. A Mammalian Example

For this work, we reanalyzed gastrointestinal stromal tumor (GIST) cell line data using SVD, ICA, two NMF methods, and BD/coGAPS, since this data had a number of TF activities validated by direct phosphoprotein measurements [10]. In brief, the data comprised triplicate growth of GIST cells in the presence of imatinib, the targeted tyrosine kinase inhibitor of the KIT receptor tyrosine kinase. Imatinib, also known as Gleevec, is used

therapeutically in GIST patients, and the inhibition of KIT phosphorylation was validated experimentally in the cell line. Cells were harvested at nine time points and Agilent Whole Human Genome Microarrays were run on each of the three samples at each time point. Data was processed to provide mean and standard deviations for each gene at each time point. Only targets of known TFs recorded in TRANSFAC Professional Release 2008.4 were retained, providing a **Y** matrix of 1363 genes (rows) by 9 samples (columns) together with standard deviations on each element.

Analysis was then run on this data using R [39]. The implementations used included the stats package (for SVD), the fastICA package based on the fastICA algorithm [40], the NMF package [34], and the coGAPS package [11]. The algorithms applied were SVD, ICA, the Brunet *et al* NMF algorithm [13], nsNMF [31], and coGAPS. The coGAPS analysis was equivalent to the original analysis performed with BD [10], as coGAPS was created using this as a test data set. For SVD and ICA, a single run was made. For the NMF methods, 500 runs with different random initial matrices were made, and the 50 runs with smallest residuals were retained. Mean and standard deviations were generated for **A** and **P** from the 50 samples. For CoGAPS, the analysis included sampling of the posterior distribution, which was used to generate mean and standard deviation estimates.

In Fig. 3, the identified patterns are shown. In all cases, the dimensionality was set to 5, however to reduce clutter in the figures only the three patterns that were analyzed in detail are shown. It is clear that the matrix factorization methods and CoGAPS identify the same three patterns in the data. The two patterns not shown do differ between the methods, with CoGAPS producing two relatively flat patterns that capture biosynthesis and metabolic patterns (based on gene membership), and the NMF methods both producing a pattern spiking at 9 and 24 hours and a pattern with a broad peak at 12–18 hours. We refer to the three patterns shown in Fig. 3 as “Falling” (black line), “Transient” (red line), and “Rising” (blue line) respectively.

We then looked at the specific signaling process readouts used in the original publication. First, MAPK and PI3K signaling were downregulated due to imatinib suppression of KIT signaling, which appears as upregulation of the TFs downstream of MAPK and PI3K in the Falling pattern. Second, the Transient pattern showed strong upregulation of p53. Upregulation of p53 was validated through Western blots of the DNA damage response proteins and p53 across the time series. Third, the Rising pattern showed upregulation of ELK1 and STAT3, which was validated by phosphoprotein antibodies as well. To visualize the results, we converted a permutation-based *Z*-score statistic (see *Section V.A*) for each TF based on its known TRANSFAC targets to a strength of activity, with high activity reflected as bright yellow and low activity as dark blue. The results for coGAPS and the NMF methods are presented in Fig. 4. It is clear that in all cases the NMF methods do not provide either a consistent view of downstream activity (i.e., for the Falling pattern) nor accurate prediction of individual TF activity. On the other hand, CoGAPS provides consistent estimation of the full network readout downstream of KIT (i.e,, SP1, ELK1, MYC, E2F1, AP1, CREB high and FOXO low), as well as correct identification of TF activity in the other two patterns.

To verify that the results were not a driven by a larger number of significant TF activities being predicted by Co-GAPS, we generated a list of all TFs out of the 230 tested that showed high activity at a permutation $p < 0.05$ for each pattern. In all cases, the numbers of TFs determined to be significant were roughly the same between CoGAPS and the NMF methods. For example, in the Transient pattern, CoGAPS showed p53 activity at the limit of $p < 0.002$ based on the 500 permutations with 26 additional active TFs. Brunet-NMF and nsNMF detected 17 and 22 significant active TFs respectively, but p53 was not significant ($p = 0.958$ and $p = 0.788$ respectively). In order to check whether p53 might instead be associated with the additional pattern peaking at 12–18 hours, we identified the 12 and 9 TFs statistically significant in these patterns in Brunet-NMF and nsNMF respectively, but p53 was not significant ($p \sim 1$ for both algorithms).

Overall, the results of this mammalian data analysis suggest that CoGAPS is more effective at identifying correct TF activity, although both NMF methods appear equally good at determining the patterns. The reason for the failure in detection of TF activity is unclear, but it is consistent with the previous poor results in detection of biological processes in the yeast compendium data [38]. In both cases, the **A** matrix is far larger than the **P** matrix, and it is possible that solutions remain too smooth in this domain. Careful tuning of the NMF methods may create appropriate constraints to guide the methods to useful **A** matrices as well as **P** matrices, however such tuning often is counterproductive in biological studies as the methods tend to get fit to the peculiarities of individual data sets. Interestingly, BFRM, which was not used here, was also good at finding biological process signatures in the yeast data, so that it is possible that the ability of MCMC methods to more fully explore the posterior probability distribution is important.

However, while MCMC methods are designed to escape local maxima in the probability distribution and sample the distribution more fully, they are computationally expensive. For instance, CoGAPS analysis of the GIST data took several hours, while the 500 runs of the NMF methods took only a few minutes. The apparent superiority of MCMC methods in TF activity estimation suggests that recent methods developed in machine learning to replace the inherently slow MCMC approach in other disciplines may be worthy of study in expression analysis and TRN estimation (e.g., [41], [42]).

V. Incorporating Matrix Factorization into TRN Estimation

There are two approaches to incorporate improved inference of regulation from matrix factorization into TRN estimation. We refer to these as *Reverse Inference* and *Forward Inference*. The former term refers to estimation of TF activity from the behavior of known TF targets, while the latter refers to inference of TF targets given TF activity. In either case, methods that extend the known targets (see below) can be used to extend the inferences.

A. Reverse Inference

Matrix factorization will provide an indication of the activity of a TF through interpretation of the **P** and **A** matrices. Using the association of genes (columns of **A**) with patterns (rows of **P**) permits estimation of when a TF is active. This can be done through a gene set analysis

of the scores associated with a gene in a column of **A** for a set of genes that are known to be regulated by a TF. Specifically, for methods that generate error estimates for the matrix elements, a *Z*-score test can be used, where the *Z*-score for a TF within a pattern *p* is defined from the measured *Z*-scores of its targets,

$$Z_{T,p}=\frac{1}{R}\sum_{R}\frac{A_{rp}}{\sigma_{rp}}, \quad (13)$$

where *r* indexes the genes that are targets of TF *T*. This can then be compared to a random sample of genes from pattern *p* to determine the significance of the *Z*-score and the probability of activity of the TF. The amplitude of the corresponding row of **P** then gives an indication of the level of activity of the TF (see [10] for an example in cancer). This statistic was used to generate estimates in *Section IV.B*.

Alternative statistics can be generated from the **A** and **P** matrices for cases where only mean estimates of the element values are obtained. At the simplest level, these can include threshold values based on targets, although analyses based on sparse components or binary models are likely to be more robust.

B. Forward Inference

The knowledge that a TF is active, either through the inference noted above or by other predictive methods, permits one to estimate the probability that a novel target, potentially a TF itself, is activated by the TF. The most straightforward way of estimating this is to use the strength of the scores for the known targets, either the distribution of *Z*-scores for the targets of a TF or the distribution of the mean values in a pattern *p*. This then allows the other genes in pattern *p* to be compared.

One complexity is the likely activity of multiple TFs within any pattern. In the case where the activity of the TFs differ in different patterns, the behavior of the targets for a TF across all patterns can be used to isolate the genes that are tied to the specific TF. If the activity of a subset of TFs agree across all patterns, then it is necessary to have additional information to infer the probability that one of the active TFs is responsible. This would be an excellent use of comparative TFBS information, as in this focused case it is likely that the additional presence of a TFBS could resolve the potential multiple driving TFs. However, in this case the TFBS data used for forward inference must be sufficiently independent to ensure the identifiability of the inference algorithm.

VI. Extensions and Alternatives to Matrix Factorization Methods

Over the last few years the interest in NMF and other factorization methods for high-throughput biological data has increased. A number of extensions have been made to the basic techniques in order to address issues in biological data that can impact use in TRN estimation.

One recurring issue with different methods of NMF is that the basis vectors (i.e., patterns) identified may not provide a minimal representation of the expression data, in the sense of Fig. 2. While this is generally addressed through sparse matrix methods, Zhang *et al* addressed this issue instead by penalizing the angle between the basis vectors, effectively encouraging solutions fitting bases $a,\vec{b},\vec{c}$ [43].

The issue of NMF being prone to trapping in local maxima has also led to development of methods that look to improve NMF robustness. Fogel *et al* used trimmed least squares fitting and dimensionality estimation by mixture modeling to create inferential robust NMF (irNMF) [44]. The method also iteratively removes discordant observations, limiting the impact of outliers on the model.

Like NMF methods, ICA methods have suffered from the lack of sparseness and the associated lack of locality of patterns. Han and Li introduced multi-resolution ICA, which utilizes a wavelet transform prior to applying ICA, to address this issue [45].

The initial NCA implementation was computationally expensive and limited in the number of transcription factors that could be addressed in an analysis, which led to a number of modifications. Galbraith *et al* modified the original NCA to allow for more transcriptional regulators to be modeled. This was accomplished by realizing that the criteria for the number of samples, N_s, could be relaxed, so that $N_{TF} < N_s$ could be changed to $N_{TFg} < N_s$, where N_{TFg} is the number of TFs that regulate any given gene [46]. Chang *et al* introduced FastNCA that included an SVD step prior to NCA to reduce the data size [47]. In order to escape the need for full network connectivity in NCA, Chen *et al* utilized FastNCA coupled to particle swarm optimization to allow the refinement of gene targets during analysis [48].

A number of methods that serve as alternatives to NMF for TRN estimation in the case of overlapping regulation have also been introduced. Asif and Sanguinetti utilized a Hidden Markov Model (HMM) with a nonlinear likelihood to capture TRNs in yeast and bacterial systems [49]. The TF activities are treated as binary parameters with gene specific expression levels linked to the TF activity. Chuang *et al* introduced a method termed AdaFuzzy to link sequence, chromatin IP, and expression data within a constrained probabilistic sparse matrix factorization to estimate TRNs for yeast [50].

A number of extensions have focused on guiding an analysis with prior knowledge, either through inclusion of an additional matrix in the factorization or through introduction of penalties linked to known gene profiles. Yang *et al* introduced an additional matrix linked to gene ontology categories to drive the factorization toward bases that preferentially linked genes within the same ontological category [51]. Gong *et al* used DNA sequence motifs to guide clustering and determine TF activity followed by sparse component analysis to estimate the strength of regulation for the individual targets [52]. Gaujoux and Seoighe introduced marker genes to guide NMF to find patterns that linked samples that were of the same cell-type [53].

It still remains to integrate the results of matrix factorization techniques into TRN estimation formally. This will require linking two inference mechanisms in a unified framework

capable of linking inference from the data decomposition and inference on the graphical model expressing the TRN. In addition, data from TFBS studies, ChIP methods, methylation measurements, and literature mining should be integrated to improve inference.

VII. Conclusion

Determination of transcriptional regulatory networks (TRNs) provides important insight into numerous biological processes driven by cellular reprogramming. In mammals, evolution has driven high levels of gene reuse, leading to multiple regulation of genes, which greatly complicates TRN estimation. One potential path forward is to integrate matrix factorization methods, which have been developed specifically to address multiple regulation, into TRN estimation.

We have reviewed the principal methods and their extensions. Sparseness appears to be a recurring theme for obtaining good results, whether in identifying coregulation of genes as desired for TRN estimation, or for gene set analysis to identify biological processes active in the system under study. Bayesian methods implement sparseness through a prior on the potential **A** and **P** matrices, while NMF methods tend to introduce an additional matrix in the decomposition or to penalize non-sparse solutions. Either approach appears to improve recovery of coregulation, although the Bayesian methods appear somewhat superior over all, perhaps because they explore the parameter space more fully.

In the future, it will be possible to provide additional data to guide matrix factorization and TRN inference. Methylation data can provide prior distributions on the probability that a given TF target can avoid upregulation by an active TF, so that epigenetically silenced genes do not adversely affect TRN estimation. Identification of alternative splice variants, as from RNA-seq, can aid in identifying alternative TF isoforms that may have different gene targets, deconvolving potentially confusing cases where the TF targets change unexpectedly due to presently unmeasured regulatory shifts.

Recently a new set of methods have been introduced that address multiple regulation directly during TRN estimation. These techniques should be considered along side factorization methods when extending TRN models. An issue yet to be addressed is the problem of non-time series data, which will dominate the medically-relevant available data. The extension of TRN results to humans will require integration of model organism time series data with extremely sparsely sampled human clinical data.

Acknowledgments

This work was partly funded by NCI (P30CA006973).

References

1. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac Symp Biocomput. 2001:422–33. [PubMed: 11262961]

2. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics. 2002; 18(2):261–74. [PubMed: 11847074]
3. Shmulevich I, Dougherty E, Zhang W. From boolean to probabilistic boolean networks as models of genetic regulatory networks. Proceedings of the IEEE. 2002; 90(11):1778–1792.
4. Sabatti C, Rohlin L, Oh MK, Liao JC. Co-expression pattern from dna microarray experiments as a tool for operon prediction. Nucleic Acids Res. Jul; 2002 30(13):2886–93. [PubMed: 12087173]
5. Hannum G, Srivas R, Guénolé A, van Attikum H, Krogan NJ, Karp RM, Ideker T. Genome-wide association data reveal a global map of genetic interactions among protein complexes. PLoS Genet. Dec.2009 5(12):e1000782. [PubMed: 20041197]
6. Ochs MF, Stoyanova RS, Arias-Mendoza F, Brown TR. A new method for spectral decomposition using a bilinear bayesian approach. J Magn Reson. 1999; 137(1):161–76. [PubMed: 10053145]
7. Lee DD, Seung HS. Learning the parts of objects by nonnegative matrix factorization. Nature. 1999; 401(6755):788–91. [PubMed: 10548103]
8. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A. 2000; 97(18):10, 101–6.
9. Moloshok TD, Klevecz RR, Grant JD, Manion FJ, Speier WFt, Ochs MF. Application of bayesian decomposition for analysing microarray data. Bioinformatics. 2002; 18(4):566–75. [PubMed: 12016054]
10. Ochs MF, Rink L, Tarn C, Mburu S, Taguchi T, Eisenberg B, Godwin AK. Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Res. 2009; 69(23):9125–32. [PubMed: 19903850]
11. Fertig EJ, Ding J, Favorov AV, Parmigiani G, Ochs MF. Cogaps: an r/c++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics. 2010; 26(21):2792–3. [PubMed: 20810601]
12. Kim PM, Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. Jul; 2003 13(7):1706–18. [PubMed: 12840046]
13. Brunet JP, Tamayo P, Golub T, Mesirov J. Metagenes and molecular pattern discovery using matrix factorization. PNAS. 2004; 101:4164–4169. [PubMed: 15016911]
14. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics. Nov; 2005 21(21):3970–5. [PubMed: 16244221]
15. West, M. Bayesian factor regression models in the ”large p, small n” paradigm. In: Bernardo, JM.; Bayarri, MJ.; Berger, JO.; Dawid, AP., editors. Bayesian Statistics 7. Oxford: Oxford University Press; 2003.
16. Carvalho C, Chang J, Lucas J, Nevins J, Wang Q, West M. High-dimensional sparse factor modelling: Applications in gene expression genomics. J Am Stat Assoc. 2008; 103:1438–1456. [PubMed: 21218139]
17. Chiappetta P, Roubaud M, Torrésani B. Blind source separation and the analysis of microarray data. Journal of Computational Biology. 2004; 11(6):1090–1109. [PubMed: 15662200]
18. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics. 2002; 18(Suppl 1):S136–44. 1367–4803. [PubMed: 12169541]
19. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. Cell. 2000; 102(1):109–26. [PubMed: 10929718]
20. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mrna abundance in yeast. Mol Cell Biol. 1999; 19(3):1720–30. [PubMed: 10022859]
21. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ, Wei JT, Pienta KJ, Ghosh D, Rubin MA, Chinnaiyan AM. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell. 2005; 8(5):393–406. [PubMed: 16286247]
22. Rocke DM, Durbin B. A model for measurement error for gene expression arrays. J Comput Biol. 2001; 8:557–569. [PubMed: 11747612]

23. Wang G, Kossenkov AV, Ochs MF. Ls-nmf: A modified nonnegative matrix factorization algorithm utilizing uncertainty estimates. BMC Bioinformatics. 2006; 7(1):175. [PubMed: 16569230]
24. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. Oct; 2008 9(10):770–80. [PubMed: 18797474]
25. Lee SI, Batzoglou S. Application of independent component analysis to microarrays. Genome Biol. 2003; 4(11):R76. [PubMed: 14611662]
26. Bidaut G, Ochs MF. Clutrfree: cluster tree visualization and interpretation. Bioinformatics. 2004; 20(16):2869–71. [PubMed: 15145813]
27. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3(9):1724–35. [PubMed: 17907809]
28. Pournara I, Wernisch L. Factor analysis for gene regulatory networks and transcription factor activity profiles. BMC Bioinformatics. 2007; 8:61. [PubMed: 17319944]
29. Sibisi S, Skilling J. Prior distributions on measure space. Journal of the Royal Statistical Society, B. 1997; 59(1):217–235.
30. Ochs, MF. Bayesian decomposition. In: Parmigiani, G.; Garrett, E.; Irizarry, R.; Zeger, S., editors. The Analysis of Gene Expression Data: Methods and Software. New York: Springer Verlag; 2003.
31. Carmona-Saez P, Pascual-Marqui RD, Tirado F, Carazo JM, Pascual-Montano A. Biclustering of gene expression data by non-smooth non-negative matrix factorization. BMC Bioinformatics. 2006; 7:78. [PubMed: 16503973]
32. Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol. 2008; 4(7):e1000029. [PubMed: 18654623]
33. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. Jul; 2009 10(3):515–34. [PubMed: 19377034]
34. Gaujoux R, Seoighe C. A flexibile r package for nonnegative matrix factorization. BMC Bioinformatics. 2010; 11:367. [PubMed: 20598126]
35. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci U S A. 2003; 100(26):15, 522–7. [PubMed: 12509513]
36. Frigyesi A, Hoglund M. Non-negative matrix factorization for the analysis of complex gene expression data: Identifcationof clinically relevant tumor subtypes. Cancer Informatics. 2008; 6:275–292. [PubMed: 19259414]
37. Kossenkov AV, Ochs MF. Matrix factorisation methods applied in microarray data analysis. Int J Data Min Bioinform. 2010; 4(1):72–90. [PubMed: 20376923]
38. Kossenkov AV, Ochs MF. Matrix factorization for recovery of biological processes from microarray data. Methods Enzymol. 2009; 467:59–77. [PubMed: 19897089]
39. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5(10):R80. 1465–6914. (Electronic) Journal Article. [PubMed: 15461798]
40. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE transactions on neural networks/a publication of the IEEE Neural Networks Council. 1999; 10(3): 626–34. [PubMed: 18252563] hyvarinen. A IEEE Trans Neural Netw. 1999; 10(3):626–34.
41. Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of the em algorithm. Ann Statist. 1999; 27(1):94–128.
42. Kuhn E, Lavielle M. Coupling a stochastic approximation version of em with an mcmc procedure. ESAIM Probab Stat. 2004; 8:115–131.
43. Zhang J, Wei L, Feng X, Ma Z, Wang Y. Pattern expression nonnegative matrix factorization: algorithm and applications to blind source separation. Comput Intell Neurosci. 2008:168769. [PubMed: 18566689]

44. Fogel P, Young SS, Hawkins DM, Ledirac N. Inferential, robust non-negative matrix factorization analysis of microarray data. Bioinformatics. Jan; 2007 23(1):44–9. [PubMed: 17092989]
45. Han H, Li XL. Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery. BMC Bioinformatics. 2011; 12(Suppl 1):S7.
46. Galbraith SJ, Tran LM, Liao JC. Transcriptome network component analysis with limited microarray data. Bioinformatics. Aug; 2006 22(15):1886–94. [PubMed: 16766556]
47. Chang C, Ding Z, Hung YS, Fung PCW. Fast network component analysis (fastnca) for gene regulatory network reconstruction from microarray data. Bioinformatics. Jun; 2008 24(11):1349–58. [PubMed: 18400771]
48. Chen, W.; Chang, C.; Hung, Y. Transcription factor activity estimation based on particle swarm optimization and fast network component analysis. 32nd Annual International Conference of the IEEE EMBS; Buenos Aires, Argentina. September 4 2010.;
49. Asif HMS, Sanguinetti G. Large-scale learning of combinatorial transcriptional dynamics from gene expression. Bioinformatics. May; 2011 27(9):1277–83. [PubMed: 21367870]
50. Chuang CL, Hung K, Chen CM, Shieh GS. Uncovering transcriptional interactions via an adaptive fuzzy logic approach. BMC Bioinformatics. 2009; 10:400. [PubMed: 19961622]
51. Yang X, Zhou Y, Jin R, Chan C. Reconstruct modular phenotype-specific gene networks by knowledge-driven matrix factorization. Bioinformatics. Sep; 2009 25(17):2236–43. [PubMed: 19542155]
52. Gong T, Xuan J, Chen L, Riggins R, Li H, Hoffman E, Clarke R, Wang Y. Motif-guided sparse decomposition of gene expression data for regulatory module identification. BMC Bioinformatics. 2011; 12:82. [PubMed: 21426557]
53. Gaujoux R, Seoighe C. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: A case study. Infect Genet Evol. Sep.2011

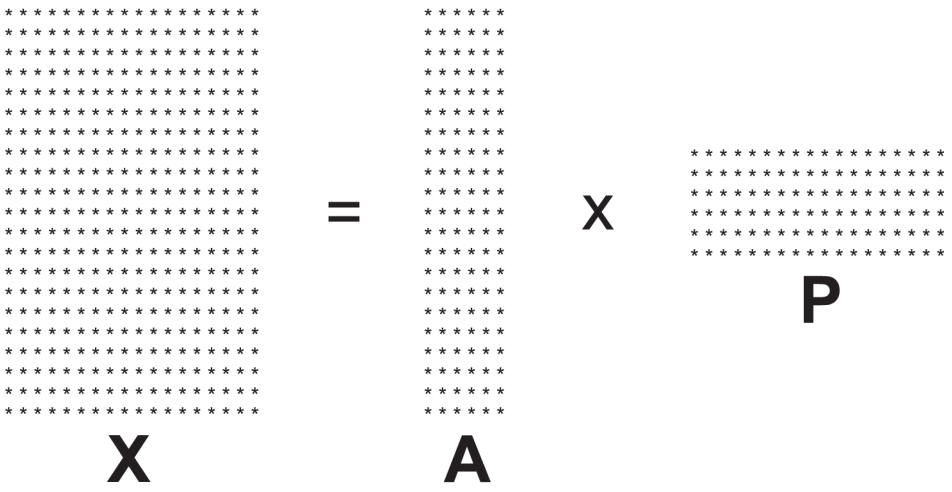

Fig. 1.

Matrix Factorization: The fundamental fitting that occurs in the techniques focused on here are the derivation of **A** and **P** matrices which combine to form a denoised model, **X**, of the measured data. **X** comprises *N* rows of genes and *M* columns of samples.

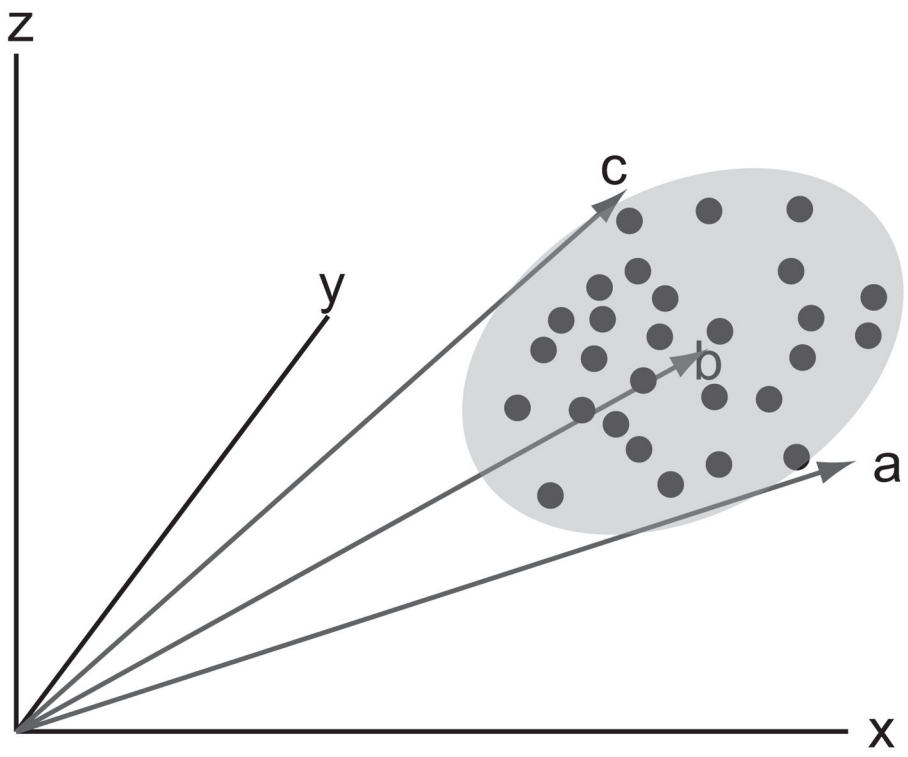

Fig. 2.

Data Distribution: The data utilized in many gene expression studies is inherently positive. PCA and SVD can be considered as methods that rotate the standard Cartesian coordinates to align with this data. BD, BFRM, and NMF search for non-orthogonal vectors to capture the distribution of the data.

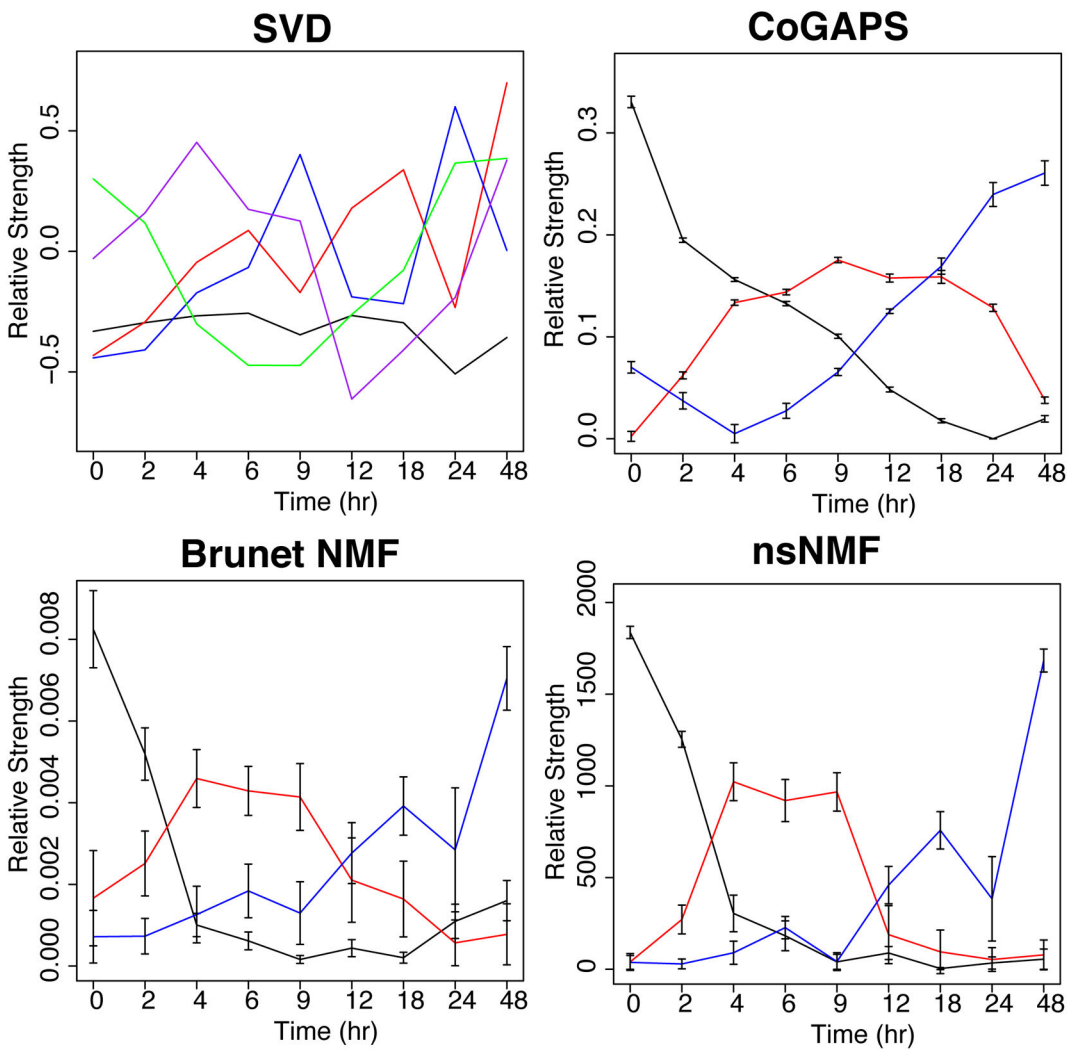

Fig. 3.

Patterns: The patterns identified in the GIST data by different algorithms are presented. For SVD, all five patterns are shown, while for other methods only the three patterns analyzed in detail previously are shown. ICA is not shown but was very similar to SVD. Error bars of one standard deviation are provided for those cases where they were measured.

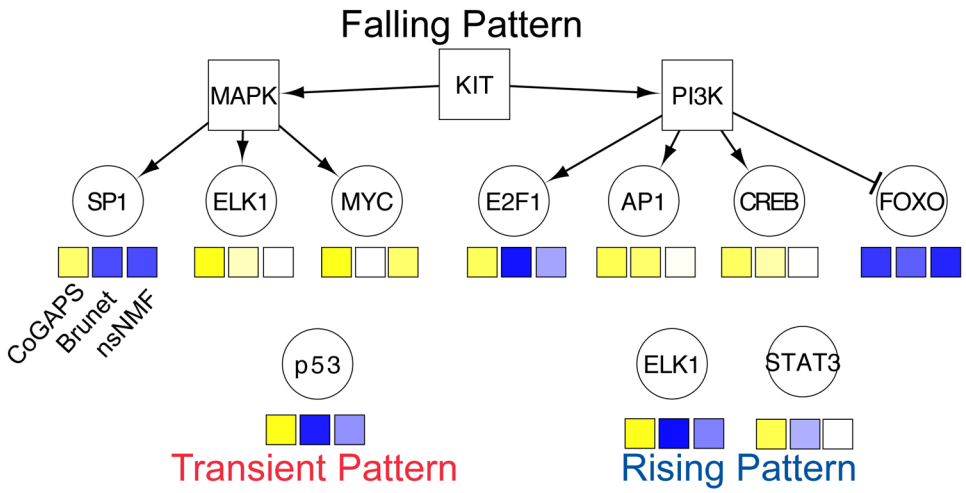

Fig. 4.

TF Activity: The estimates of TF activity based on known gene targets is presented with yellow indicating high activity and blue low activity. Under each TF, the leftmost box is for CoGAPS estimation, the middle box for the Brunet *et al* NMF estimation, and the rightmost box for nsNMF estimation.