Published in final edited form as:
Methods Enzymol. 2009 ; 467: 59–77. doi:10.1016/S0076-6879(09)67003-8.

Matrix Factorization for Recovery of Biological Processes from Microarray Data

Andrew V. Kossenkov and
The Wistar Institute, 3601 Spruce Street, R214, Philadelphia, PA, 19104, Phone: 215-495-6898, Fax: 215-898-4521, akossenkov@wistar.org

Michael F. Ochs*
The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, 550 North Broadway, Suite 1103, Baltimore, MD 21205

Abstract

We explore a number of matrix factorization methods in terms of their ability to identify signatures of biological processes in a large gene expression study. We focus on the ability of these methods to find signatures in terms of gene ontology enhancement and on the interpretation of these signatures in the samples. Two Bayesian approaches, Bayesian Decomposition (BD) and Bayesian Factor Regression Modeling (BFRM), perform best. Differences in the strength of the signatures between the samples suggest that BD will be most useful for systems modeling and BFRM for biomarker discovery.

INTRODUCTION

Microarray technology introduced a new complexity into biological studies through the simultaneous measurement of thousands of variables, replacing a technique (the Northern blot) that typical measured at most tens of variables. Traditional analysis focused on measurements with minimal statistical complexity, but direct application of such tests (e.g., the t-test) to microarrays resulted in massive numbers of 'significant' differentially regulated genes, when reality suggested far fewer. There were a number of reasons for the failure of these tests, including the small number of replicates leading to chance detection when tens of thousands of variables were measured (1), the unmodeled covariance arising from coordinated expression (2), and non-gene specific error models (3). While a number of statistical issues have now been successfully addressed (4), two aspects of the biology of gene expression raise difficulties for many analyses.

The issues can be noted in a simple model of signaling in the yeast *S. cerevisiae*. In Figure 1, the three overlapping MAPK pathways are shown. The pathways share a number of upstream regulatory components (e.g., Ste11), and regulate sets of genes divided here into five groups (A – E), with a few of the many known targets shown. The Fus3 mating response MAPK protein activates the Ste12 transcription factor, leading to expression of groups A and B. The Kss1 filamentation response MAPK protein activates the Ste12-Tec1 regulatory complex, leading to expression of groups B, C, and D. The Hog1 high-osmolarity response MAPK protein activates the Sko1 transcription factor, leading to expression of groups D and E. The standard methods used in microarray analysis will look for genes that are differentially expressed between two states. If we imagine those two states as mating

*Phone: 410-955-8830, FAX: 410-955-0859, mfo@jhu.edu.

activation and filamentation activation, we identify genes associated with each process, but we do not identify all genes associated with either process. Alternatively, clustering in an experiment where each process is independently active will lead to identification of five clusters (one for each group A – E) even though only three processes are active. Naturally, the complexity is substantially greater as there is no true isolation of a single biological process, as any system with only a single process active would be dead, and any measurement is convolved with measurements of ongoing biological behavior required for survival, homeostasis, or growth. These processes use many of the same genes, due to borrowing of gene function that has occurred throughout evolution. [Note: for S. cerevisiae, plain text Ste12 indicates the protein, while italic text *ste12* indicates the gene.]

Essentially, this example shows the two underlying biological principles that need to be addressed in many analyses of high-throughput data – multiple regulation of genes due to gene reuse in different biological processes and non-orthogonality of biological process activity arising from the natural simultaneity of biological behaviors. Mathematically we can state the problem as a matrix factorization problem,

$$D_{ij} = \sum_{k=1}^{P} A_{ik} P_{kj} + \varepsilon_{ij} \quad (1)$$

where **D** is the data matrix comprising measurements on *N* genes (or other entities) indexed by *i* across *M* conditions indexed by *j*, **P** is the pattern matrix for *P* patterns indexed by *k*, **A** is the amplitude or weighting matrix that determines how much of each gene's behavior can be attributed to each pattern, and **ε** is the error matrix. **P** is essentially a collection of basis vectors for the factorization into *P* dimensions, and as such it is often useful to normalize the rows of **P** to sum to 1. This makes the **A** matrix similar to loading or score matrices, such as in principal component analysis. It is useful to note there that the nonindependence of biological processes is equivalent to non-orthogonality of the rows of **P**, indicating the factorization is ideally into a basis space that reflects underlying biological behaviors but is not orthonormal.

We introduced Bayesian Decomposition, a Markov chain Monte Carlo algorithm, to address these fundamental biological issues in microarray studies (5), extending our original work in spectroscopy (6). Kim and Tidor introduced non-negative matrix factorization (NMF), created by Lee and Seung (7), into microarray analysis (8,9), for the same reason. Subsequently it was realized that sparseness aids in identifying biologically meaningful processes, and sparse NMF was introduced (10). Fortuitously, due to its original use in spectroscopy, sparseness was already a feature of Bayesian Decomposition through its atomic prior (11). More recently, Carvalho and colleagues introduced Bayesian Factor Regression Modeling (BFRM), an additional Markov chain Monte Carlo method, for microarray data analysis (12).

Targeted methods that directly model multiple sources of biological information have been introduced as well. Liao and Roychowdhury introduce network component analysis (NCA), which relied on information about the binding of transcriptional regulators to help isolate the signatures of biological processes (13). The use of information on transcriptional regulation can also aid in sparseness, as shown by its inclusion in BD as prior information (14).

These methods have been developed and applied primarily to microarray data, as it was the first high-throughput biological data that included dynamic behavior, in contrast to sequence data. Microarrays were developed independently by a number of groups in the 1990s (15,16), and their use is now widespread. A number of technical issues plagued early arrays, and error rates were high. The development of normalization and other preprocessing

procedures improved data reproducibility and robustness (17-19), leading to studies that demonstrated the ability to produce meaningful data sets from arrays run in different laboratories at different times (20). Data can be accessed, though not always with useful metadata, in the GEO and ArrayExpress repositories (21,22).

However, the methods discussed here are also suitable for other high-throughput data where the fundamental assumptions of multiple overlapping sets within the data and non-orthogonality of these sets across the samples holds. In the near future, these data are likely to include large scale proteomics measurements and metabolite measurements.

We have previously undertaken a study of some of these methods to determine their ability to solve equation 1 using simulations of the cell cycle (23). This study did not address the recovery of biologically meaningful patterns from real data, where numerous unknowns exist. Most of these relate to the fundamental issue that separates biological studies from those in physics and chemistry – in biology we are unable to isolate variables of interest away from other unknowns, as to do so is to kill the organism under study. Instead, we must perform studies in a background of incomplete knowledge of the activities a cell is undertaking and incomplete knowledge of the entities (e.g., genes, proteins) associated with these processes. In addition, sampling is difficult and therefore tends to be limited (i.e., large N, small P), and the data remain prone to substantial variance, perhaps due to true biological variation instead of technical issues.

We have undertaken a new analysis of the Rosetta compendium, a data set of quadruplicate measurements of 300 yeast gene knock-outs and chemical treatments (3), to determine how well various matrix factorization methods recover signatures of biological processes. The Rosetta study included 63 control replicates of wild-type yeast grown in rich media, allowing a gene-specific error model. One interesting result to emerge from this work is that roughly 10% of yeast genes appear to be under limited transcriptional regulation, so that their transcript levels vary by orders of magnitude without a corresponding variation in protein levels or phenotype. This has obvious implications for studies where whole genome transcript levels are measured on limited numbers of replicates.

Using known biological behaviors that are affected by specific gene knock-outs, we compared a number of methods from clustering through the matrix factorization methods discussed above to determine how well such methods recover biological information from microarray measurements. We first give a brief description of each method, then we present the data set and results of our analyses.

OVERVIEW OF METHODS

Clustering Techniques

To provide a baseline for comparison, we applied two widely used clustering techniques to the data set, as well as an approach where genes were assigned to groups at random. Hierarchical clustering was introduced for microarray work by Eisen and colleagues (24), and because of easy-to-use software and its lead as the first technique, it has seen significant use and is available in desktop tools (25). Hierarchical clustering, as performed by most users, is done in an agglomerative fashion, using a metric to determine inter-gene and inter-cluster distances. Metrics used in microarray studies include Pearson correlation, which captures the shape of changes across the samples, and Euclidean distance, which captures the magnitude of changes. Hierarchical clustering creates a tree of distances (a dendrogram) and groups the genes based on the nodes of this tree. As such, different numbers of clusters can be created by cutting at different levels on the tree, however each specific set of clusters is the most parsimonious for that level and that metric.

K-means (or K-medians) clustering has also been widely used in microarray studies, and it relies on an initial random assignment of genes to *P* clusters. Genes are then moved between clusters based on gene-cluster distances in an iterative fashion. The same metrics are typically used as in hierarchical clustering, and since the number of clusters is defined *a priori*, there is no necessity of choosing a tree level as in hierarchical clustering. However, a tree can be created after clustering is complete if desired.

Traditional Statistical Approaches

The factorization implied by equation 1 can be accomplished in a number of ways. One of the most widely used is singular value decomposition (SVD) or its relative, principal component analysis (PCA). These methods create *M* new basis vectors from the data in **D**, and these new basis vectors are orthonormal. SVD is an analytic procedure that decomposes **D** into the product of a left singular matrix **U**, a diagonal matrix of ordered values *S* referred to as the singular values, and a right singular matrix $\mathbf{V}^T$, i.e.,

$$\mathbf{D}=\mathbf{USV}^T. \quad (2)$$

Alter and colleagues introduced SVD to microarray studies, and defined the rows of $\mathbf{V}^T$ as eigengenes, and the columns of **U** as eigenarrays (26). The eigengenes are similar to the concept of patterns for equation 1. PCA performs a similar decomposition, however the analysis proceeds from the covariance matrix, so that the principal components (PCs) follow the variance in the data. The first PC is aligned with the axis of maximum variance in the *M*-dimensional space of the data matrix, with each additional PC chosen to be orthogonal to the previous PCs and in the direction that maximizes variance among all orthogonal directions. This creates a new orthonormal basis space in which the PCs represent directions of maximum variance. The singular values are now referred to as scores, and the value of the score provides the amount of variance explained by the corresponding PC. In most applications of PCA and SVD to microarray data, the matrices are truncated so that only the strongest eigengenes or PCs are retained. This is a form of dimensionality reduction, which in the case of PCA, retains the maximum amount of variance across the data at each possible dimension.

The orthogonality conditions of SVD and PCA were realized to be overly constraining for microarray data. Lin and colleagues and Liebermeister independently introduced independent component analysis (ICA) to microarray analysis to address this issue (27,28). As with typical applications of PCA, ICA projects the data onto a lower dimensional space. In linear ICA, the goal is to solve equation 1 by finding **P**, such that

$$\mathbf{P}:\boldsymbol{Y}=\mathbf{WD}, \quad (3)$$

through the identification of the unmixing matrix, **W**. The unmixing matrix is designed to make the rows of **Y**, and therefore **P**, as statistically independent as possible. A number of measures of independence can be used, such as maximizing negentropy or non-gaussinity (29). Because ICA is not strictly constrained like PCA or SVD, it is possible to obtain multiple solutions for **Y** from the same data. As such, sometimes multiple applications must be performed and a rule applied to pick the best **Y** (30).

Matrix Factorization Techniques

The desire to escape both the exclusivity of gene assignment to a single cluster occurring in clustering and the independence criteria of statistical methods such as PCA led to the introduction of two techniques from other fields that addressed these issues. Naturally, these

methods require constraints, as equation 1 is degenerate, allowing an infinite number of equally good solutions in the absence of a constraint, such as the one provided by an orthonormal basis in PCA. The methods are distinguished by the methods of constraint and the search algorithm for finding an optimal solution to equation 1 within these constraints. All these methods also rely on dimensionality reduction, so that the number of elements in the matrices **A** and **P** are less than those in **D**.

Bayesian Decomposition (BD) applies a positivity constraint within an atomic prior to limit the possible **A** and **P** matrices. The atomic prior relies on implementation of an additional domain, an atomic domain modeling an infinite one dimensional space upon which atoms are placed, and mappings between it and the **A** and **P** matrices. This provides great flexibility, as the mappings, in the form of convolution functions, can distribute an atom to a complex distribution encoding additional prior knowledge (e.g., the form of a response curve, a coordinated change in multiple genes). The atomic domain comprises a positive additive distribution (11), and an Occam's Razor argument (i.e., parsimony) penalizes excessive structure through the prior distribution on the atoms. The resulting posterior distribution that combines this prior with the likelihood determined from the fit to the data is sampled by a Markov chain Monte Carlo Gibbs sampler (31). This approach allows patterns to be constrained in multiple ways, permitting the rows of **P** to be nonorthogonal, while still identifying unique solutions. Even with unique directions defined by the rows of **P**, there is still flexibility in the equation that allows amplitude in rows of **P** to be transferred to columns in **A** without changing **D**. As such, the rows of **P** are normalized to sum to 1. For the work presented here, a simple convolution function that maps each atom to a single matrix element is used, as this only enforces positivity on **A** and **P**, similar to NMF.

The posterior distribution sampled by BD is generated from the prior and the likelihood through Bayes' equation,

$$p(A,P|\mathbf{D}) = \frac{p(\mathbf{D}|A,P)\,p(A,P)}{p(\mathbf{D})} \tag{3}$$

where $p(\mathbf{A,P} \mid \mathbf{D})$is the posterior distribution, $p(\mathbf{D} \mid \mathbf{A,P})$ is the likelihood, $p(\mathbf{A,P})$ is the prior, and $p(\mathbf{D})$ is the marginal likelihood of the data, which is also known as the evidence. The likelihood is the probability distribution associated with a χ^2 distribution, and BD therefore uses the estimates of error during modeling, which can be very powerful given the large variation in uncertainty across different genes in a microarray experiment. This also permits seamless treatment of missing values, as they can be estimated at a typical value (background level) with a large uncertainty, thus not affecting the likelihood. The evidence is not used by BD, as Gibbs sampling requires only relative estimates of the posterior distribution, however it has been proposed that it can be used for model selection, which in this case would be determining the correct number of dimensions, *P*, in equation 1 (32). Presently, BD requires a choice of *P*.

Non-negative matrix factorization (NMF) applies positivity and dimensionality reduction to find the patterns of **P**, each of which is defined as a positive linear combination of rows of **D**. Each row of **D** is therefore a linear combination of patterns, with the weight given by the corresponding element in **A**. As with BD, the choice of *P* must be made before applying the algorithm. In an NMF simulation, random matrices **A** and **P** are initialized according to some scheme, such as from a uniform distribution. The two matrices are then iteratively updated with

$$P_{\alpha\mu} = P_{\alpha\mu}\frac{\sum_i A_{i\alpha} D_{i\mu}}{\sum_i A_{i\alpha} M_{i\mu}}$$
$$A_{\delta\alpha} = P_{\delta\alpha}\frac{\sum_j D_{\delta j} P_{\alpha j}}{\sum_i M_{\delta j} P_{\alpha j}} \quad (4)$$

which guarantees reaching a local maximum in the likelihood. The updating rules climb a gradient in likelihood, which does lead to the problem of becoming trapped in a local maximum in the probability space. In general, application of NMF therefore is done multiple times from different initial random points, and the best fit to the data is used. The fits obtained from repeated runs on complex microarray data can vary significantly in some cases, due to the complex probability structure that appears typical for biological data. MCMC techniques tend to be more resistant to this problem, as they are designed specifically to escape local maxima, although they are prone to miss sharp local maxima in relatively flat spaces, however this has not yet appeared to be a problem in biological data. The absence of constraints beyond positivity in NMF does lead to a tendency for the recovery of signal-invariant metagenes that carry little or no information, and the failure to include error estimates can lead to genes with large variance being overweighted during fitting. These issues have been addressed in the extensions to NMF discussed below.

Network component analysis (NCA) uses information on the binding of transcriptional regulators to DNA and dimensionality reduction to reduce the possible **A** and **P** matrices. The concept is to create a two layer network with one layer populated by transcriptional regulators and the other by the genes they regulate, with edges connecting regulators to target genes. NCA addresses the degeneracy of equation 1 through

$$\mathbf{D} = \mathbf{AXX}^{-1}\mathbf{P} + \varepsilon \quad (5)$$

where **AX** includes all possible **A** matrices and $\mathbf{X}^{-1}\mathbf{P}$ all possible **P** matrices. By demanding that **X** be diagonal, **A** and **P** are uniquely determined to a scaling factor (i.e., the rows of **P** require normalization just as in BD). The diagonality of **X** requires that the transcriptional regulators be independent. The solution of equation 5 is found by minimizing

$$|| \mathbf{D} - \mathbf{AP}||^2, \quad (6)$$

which is equivalent to maximizing the likelihood with an assumption of uniform Gaussian errors.

For the application of NCA, the relative strength of the transcription of a gene by a regulator must be determined. This is done by measuring the binding affinity of a transcription factor to the promoter for a gene. Since each gene can be regulated by multiple regulators, the expression of a gene in a given condition must be estimated as a combination of the regulation from different factors. A log-linear model is used, so each additional binding of a regulator leads to a multiplicative increase in expression. However, it is not clear that the affinity of binding of a transcription factor is the dominant issue in determining transcript abundance, especially in eukaryotes.

Bayesian factor regression modeling (BFRM) is a Markov chain Monte Carlo technique that solves

$$D_{ij}=\mu_i+\sum_{k=1}^{r}\beta_{ik}h_{kj}+\sum_{p=1}^{P}A_{ip}P_{pj}+\varepsilon_{ij} \quad (7)$$

where **A** can be viewed as factor loadings for latent factors **P** (12). The **h** matrix provides a series of known covariates in the data, which are then treated using linear regression with coefficients **β**. The mean vector, **μ**, provides a gene specific term that adjusts all genes to the same level, while the **ε** matrix provides for noise, treated as normally distributed with a pattern specific variance. The latent factors here are then those that remain after accounting for covariates. This model has also been extended by inclusion in **D** of response variables as additional columns. This extends the second summation in equation 7 to *P*+*Q*, where *Q* is the number of latent factors tied to response variables. In both cases, the model also aims for sparse solutions, equivalent to the Occam's razor approach of BD.

BFRM also attempts to address the issue of the number of patterns or latent factors. This is done through an evolutionary stochastic search. Essentially, the algorithm attempts to change *P* to *P+1* by thresholding the probability of inclusion of a new factor. The model is refit with the additional factor, and the factor is retained if it improves the model by some criterion. In actuality, the algorithm can suggest multiple additional latent factors at each step and choose to keep multiple factors. Evolution ceases when no additional factors are accepted. The BFRM software allows turning off of the evolution, which we have done here to allow direct comparison with other methods at the same *P*.

Extensions to Non-negative Matrix Factorization

NMF has become widely used in a number of fields, including analysis of high-throughput biological data. Unlike BD and BFRM, there is no inherent sparseness criterion applied in NMF. This is not surprising, as the original application to imaging argues against sparseness (7), since images tend to have continuous elements. Sparseness is added to NMF in sparse NMF (sNMF), which penalizes solutions based on the number of non-zero components in **A** and **P** (10). A similar approach is presented in non-smooth NMF (nsNMF), which created a sparse representation of the patterns by introducing a smoothness matrix into the factorization (33). Addressing the lack of error modeling, least-squares NMF (lsNMF) converts equation 4 to a normalized form, adjusting the D_{ij} and M_{ij} terms by the specific uncertainty estimates at each matrix element (34). It also introduces stochastic updating the matrix elements, in an attempt to limit the problem of trapping in local maxima.

APPLICATION TO THE ROSETTA COMPENDIUM

The sample data set for this study is generated from experiments on the yeast, *S. cerevisiae*, which has been studied in depth for a number of biological processes, including the eukaryotic cell cycle, transcriptional and translational control, cell wall construction, mating, filamentous growth, and response to high osmolarity. There is substantial existing biological data on gene function, providing a large set of annotations for analysis (35,36). In addition, there is a rich resource, the Saccharomyces Genome Database, maintained by the community that includes sequence and expression data, protein structure, pathway information, and functional annotations (37).

The Rosetta compendium provides a large set of measurements of expression in *S. cerevisiae* including 300 deletion mutants or chemical treatments targeted at disrupting specific biological functions (3). The 300 experimental conditions were all probed by microarray four times, with dye flips (technical replicates) of two biological replicates. Control experiments involved 63 wild-type yeast grown in rich medium and then analyzed

by microarrays. The gene-specific variation seen in these “identical” cultures were combined with variance measured from quadruplicate measurements of each mutant or chemical treatment to produce a gene-specific error model. This error model provided the estimate of the uncertainty for those algorithms utilizing such an estimate.

The data were downloaded from Rosetta Inpharmatics and filtered to remove experiments where less than two genes underwent three-fold changes and to remove genes that did not change by three-fold across the remaining experiments. The resulting data set comprised 764 genes and 228 experiments with associated error estimates.

All algorithms were applied to the data using default settings, with a Pearson correlation metric and average linkage used for clustering procedures and maximum iterations parameter of 1000 for NMF, sNMF, lsNMF, ICA and NCA. Patterns for clusters in clustering methods were calculated as average of sum-1 normalized expression profiles for each gene from a cluster. BD and lsNMF were run using the PattTools Java interface (available from the authors). NMF and sNMF were run using the same code base, with sNMF sparseness set to 0.8 (38). BFRM was run using version 2 of the BFRM software (12), and BD and BFRM both sampled 5000 points from the posterior distribution using default settings on hyperparameters.

Clustering methods (HC, KMC) naturally assigned a gene to a single cluster. For methods that provided uncertainty estimates for values in the **A** matrix (BD, lsNMF), we used threshold of 3σ to decide if a gene belonged to a pattern. Note that this permitted a gene to be assigned to multiple patterns, each of which explained part of the overall expression at a significant level. An additional conversion step was done for methods that provide continuous-values for elements in the **A** matrix without uncertainty measurements (NMF, sNMF, ICA, NCA, BFRM). For these methods we assigned a gene to a group based on the absolute value of the corresponding element in matrix **A** being above average of the absolute values for the gene, as in (14).

The original publication applied biclustering to the data and reported on a number of clusters tied to specific biological processes at varying levels of significance (3). Clusters were found for mitochondrial function, cell wall construction, protein synthesis, ergosterol biosynthesis, mating, MAPK signaling, rnr1/HU genes, histone deacetylase, isw genes, vacuolar APase/iron regulation, sir genes, and the tup1/ssn6 global repressor. We converted these strong signatures to Munich Information Center for Protein Sequences (MIPS) categories from the Comprehensive Yeast Genome Database (35,36). These categories are detailed in Table 1. We added MIPS class 38, transposable elements, to the list to look for methods that could distinguish the mating response from the filamentation response (39).

We searched for signatures of these processes in the results of the analyses. In order to keep the analysis simple and less biased, we looked only for these specific processes. However, it is important to remember that real biological interpretation often relies on identification of coordinated changes in sets of related biological processes (e.g., mating, meiosis, cell fate, etc.). For all techniques, we focused on 15 patterns or clusters, as we have previously identified this as providing the best dimensional estimation (39). Analysis was performed using ClutrFree, which calculates enrichment and hypergeometric test values for all patterns for each MIPS term (40).

RESULTS OF ANALYSES

Although the fundamental goal of the non-clustering methods is the optimal solution to equation 1, albeit potentially with covariates as in equation 7, the methods differ substantially in their treatment of the data. BD, as applied here, and the NMF methods

require positivity in **A** and **P**, while NCA, ICA, PCA, and BFRM allow negative values. The **A** matrix is still easily interpreted in terms of enrichment of the gene ontology terms from table 1, however the **P** matrix can vary greatly in its information content. Therefore we focused on recovery of the signatures identified in the original study in terms of a hypergeometric *p*-value when the genes were assigned to patterns as described above. In addition, we focused on the **P** matrix for the strong mating pattern recovered with good *p*-values by all methods to determine what type of information can be recovered.

Table 2 provides *p*-values determined using the gene ontology categories in table 1. The table does not include any NMF methods, as these produced no *p*-values under 0.50. This may reflect the known problem that NMF tends to spread signal over many elements, and in this case even the sparse method failed to isolate a signature, though this may reflect a conservative sparseness parameter. We identified all patterns with an uncorrected *p*-value under 0.05 or the strongest *p*-value present under 0.5, if no *p*-values reached the 0.05 threshold. In addition we show how many of the eight terms in which each method found at least one significant pattern. The Bayesian Markov chain Monte Carlo methods performed best in this regard, and it appears that the other matrix factorization methods captured less of these signatures than the clustering methods, although it is the case that these specific signatures were chosen based on their inclusion in the original paper that relied on clustering.

The original paper reported on deletion mutants that were associated with these patterns, however it is difficult to use this information with matrix factorization methods. For instance, in Bayesian Decomposition, while the pattern associated with protein synthesis does include all mutants mentioned in the paper, it also includes all other deletion mutants. This was taken to indicate that protein synthesis is vital to all growing yeast (39), a true though not overly useful insight. On the other hand, BFRM shows two patterns associated with this term, one has only the *bub2* deletion mutant (indicated by *bub2Δ*) and the other only *ste4Δ* with any strength in the pattern matrix. This may reflect the strong sparseness that BFRM enforces on the data, thus indicating that in terms of differences between deletion mutants, these two are the most significant in terms of protein synthesis. No other matrix factorization methods had a significant *p*-value for this term.

The mating term was deemed significant by all methods. Mating and filamentation are strongly coupled in yeast, with the main difference in transcriptional response to pathway activation being the use of the Tec1 cofactor. Tec1 is the driver of transposon activity, so we expect the filamentation signature to include the Transposable Elements category, even though it may include the Mating category due to sharing of genes between these two processes as indicated in figure 1. We use the Transposable Elements term to choose a mating pattern and a filamentation pattern for BD and NCA, where two patterns appear associated with mating.

For BD, we assign pattern E to mating and pattern D to filamentation. Looking at the associated rows of the **P** matrix for deletion mutants associated with this pattern in the original paper, we find that most deletion mutants show signal for pattern E, however *ste11Δ*, *ste7Δ*, *ste18Δ*, *ste12Δ*, *fus3Δ*, *kss1Δ* /*fus3Δ* show no signal, *ste4Δ* and *ste5Δ* show low signal. In addition, *tec1Δ* shows high signal, as expected for mating. We see almost identical signals in pattern D, however here there is low signal for *tec1Δ*, which is expected as knocking out the key regulator of filamentation should eliminate the transcriptional pattern from the data.

For BFRM, we assign pattern E to mating and pattern G to filamentation. Pattern E shows no signal except in the *dig1Δ* /*dig2Δ* double deletion mutants and *sst2Δ*. This makes sense

again in terms of strong sparseness, as the Dig1 and Dig2 repressors suppress the mating response. With their absence, there should be a strong mating response signal, although it is also expected that there will also be a strong filamentation response signal. Pattern G is more complicated, including a number of deletion mutants, such as *anp1Δ*, *gas1Δ*, and *swi4Δ*. As it does not include the *dig1Δ* /*dig2Δ* double mutants, it is likely that the filamentation and mating signature are combined in pattern E, which is not unusual in the analyses of this data set.

For NCA, we assign pattern E to mating and D to filamentation, although both are enhanced in each and show very similar patterns in **P**. For pattern E, *ste4Δ* shows low signal and *kss1Δ* /*fus3Δ* gives a negative signal, while all other *ste* deletion mutants appear similar to other mutants. In pattern D, there is a sharp negative peak for the *ssn6Δ*, and not any strong variation for the *ste* mutants. This is somewhat difficult to interpret in terms of the mating pathway.

For ICA, three patterns are associated with the Mating term and none with Transposable Elements term. In all patterns the *ste* deletion mutants and other mutants on key pathway members are low, however this is true in almost all patterns. ICA recovers patterns that are very sparse generally, so there are few mutants strongly associated with these or other patterns. In all three patterns, the *tec1Δ* mutant also has no signal. PCA is equally difficult to interpret for the one pattern, D, associated with the Mating and Transposable Elements terms. This arises from the orthogonality required in PCA. However, for pattern D the *dig1Δ* /*dig2Δ* mutants show a small positive signal, and the *ste* deletion mutants show very low signal, as does the *tec1Δ* mutant. However, many mutants show significant signal, and many are zero.

The clustering methods essentially report a few genes associated with the clusters found for mating and filamentation. These include the *dig1Δ* /*dig2Δ* mutants and *sst2Δ*, also detected by BFRM. For the filamentation patterns, hierarchical clustering produces a fairly even pattern across all deletion mutants, while K-means clustering shows a great deal of variation. There is no clear signature related to filamentation in either case.

DISCUSSION

The matrix factorization methods discussed in this work have significantly different designs, and this affects their value for different types of analysis. PCA is very fast and decomposes the data into a series of PCs that capture maximum variance at each potential dimensionality. This can be very powerful for denoising data if only the strongest PCs are retained, although this has not been successful for high-throughput biological data. In addition, PCA can provide insight to the strongest patterns in the data. The orthogonality demanded of additional patterns can mix biological behaviors, however, so it is not capable of isolating strongly overlapping signatures in data.

ICA provides a higher order statistical measure for independence, when compared to PCA, however it performs very poorly in terms of capturing signatures in the data. This may reflect the tight coupling of biological processes in living systems, making it difficult to identify true statistical independence in the data. In contrast, NMF methods appear to have difficulty isolating signatures due to the strong overlap, and the bias toward smooth solutions results in a spreading of signal across the patterns. While this is a strength in many applications, it can limit their value for high-throughput biological data. However, sparse methods are likely to overcome some of this difficulty, but tuning is required and may lead to overfitting.

Network component analysis relies heavily on affinity data for the binding of transcriptional regulators to their targets. Unfortunately, such data assumes that the targets of regulators are known, as well as that affinities have been measured or can be reliably computed. This remains rare, especially in mammalian systems that are the focus of most studies. As such, NCA is at a disadvantage in comparisons such at these, so its failure to find many patterns is not surprising. In cases where such data is available, it can be a very useful technique to apply.

The Bayesian methods, BD and BFRM, performed best among the matrix factorization methods. In one sense, this is not surprising as the methods were created to address biological data and include sparseness in their design. BD appears to require less sparseness, leading to its ability to identify continuous distributions in the patterns (e.g., all mutants except those in the mating pathway). However, it does not therefore identify samples that strongly distinguish the gene sets, as BFRM does. In contrast, BFRM is unable to recover cases where the continuous distribution is desirable. This suggests that BD is most useful when trying to find solutions to equation 1 that reproduce the full system without an emphasis on strong differences, while BFRM is most useful when the goal is to find those conditions that are the most dissimilar in terms of signatures. This suggests the use of BFRM for biomarker discovery, including the ability to handled covariates as in equation 7, and the use of BD for systems modeling.

REFERENCES

1. Tusher VG, Tibshirani R, Chu G. Proc Natl Acad Sci U S A 2001;98:5116–21. [PubMed: 11309499]
2. Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ, Churchill GA. Statistica Sinica 2002;12:203–18.
3. 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. Cell 2000;102:109–26. [PubMed: 10929718]
4. Allison DB, Cui X, Page GP, Sabripour M. Nat Rev Genet 2006;7:55–65. [PubMed: 16369572]
5. Moloshok TD, Klevecz RR, Grant JD, Manion FJ, Speier W. F. t. Ochs MF. Bioinformatics 2002;18:566–75. [PubMed: 12016054]
6. Ochs MF, Stoyanova RS, Arias-Mendoza F, Brown TR. J Magn Reson 1999;137:161–76. [PubMed: 10053145]
7. Lee DD, Seung HS. Nature 1999;401:788–91. [PubMed: 10548103]
8. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Proc Natl Acad Sci U S A 2004;101:4164–9. [PubMed: 15016911]
9. Kim PM, Tidor B. Genome Res 2003;13:1706–18. [PubMed: 12840046]
10. Gao Y, Church G. Bioinformatics 2005;21:3970–5. [PubMed: 16244221]
11. Sibisi S, Skilling J. Journal of the Royal Statistical Society, B 1997;59:217–35.
12. Carvalho CM, Chang J, Lucas J, Nevins JR, Wang Q, West M. J. Am. Stat. Assoc 2008;103:1438–56.
13. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Proc Natl Acad Sci U S A 2003;100:15522–7. [PubMed: 14673099]
14. Kossenkov AV, Peterson AJ, Ochs MF. Stud Health Technol Inform 2007;12
15. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Nat Biotechnol 1996;14:1675–80. [PubMed: 9634850]
16. Schena M, Shalon D, Davis RW, Brown PO. Science 1995;270:467–70. [PubMed: 7569999]
17. Cheng L, Wong WH. Proc Natl Acad Sci U S A 2001;98:31–36. [PubMed: 11134512]
18. Bolstad BM, Irizarry RA, Astrand M, Speed TP. Bioinformatics 2003;19:185–93. [PubMed: 12538238]

19. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Nucleic Acids Res 2003;31:e15. [PubMed: 12582260]
20. English SB, Butte AJ. Bioinformatics 2007;23:2910–7. [PubMed: 17921495]
21. Edgar R, Domrachev M, Lash AE. Nucleic Acids Res 2002;30:207–10. [PubMed: 11752295]
22. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S, Brazma A. Nucleic Acids Res 2005;33:D553–5. [PubMed: 15608260]
23. Kossenkov AV, Ochs MF. Int J Data Mining Bioinfo. In Press.
24. Eisen MB, Spellman PT, Brown PO, Botstein D. Proc Natl Acad Sci U S A 1998;95:14863–8. [PubMed: 9843981]
25. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. Methods Enzymol 2006;411:134–93. [PubMed: 16939790]
26. Alter O, Brown PO, Botstein D. Proc Natl Acad Sci U S A 2000;97:10101–6. [PubMed: 10963673]
27. Liebermeister W. Bioinformatics 2002;18:51–60. [PubMed: 11836211]
28. Lin, SM.; Liao, X.; McConnell, P.; Vata, K.; Carin, L.; Goldschmidt, P. Methods of Microarray Data Analysis II. Lin, SM.; Johnson, KE., editors. Kluwer Academic Publishers; Boston: 2002.
29. Hyvrinen, A.; Karhunen, J.; Oja, E. Independent Component Analysis. John Wiley & Sons; New York: 2001.
30. Frigyesi A, Veerla S, Lindgren D, Hoglund M. BMC Bioinformatics 2006;7:290. [PubMed: 16762055]
31. Geman S, Geman D. IEEE Transactions on Pattern Analysis and Machine Intelligence 1984;PAMI-6:721–41.
32. Skilling, J. Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics; 2006;
33. Carmona-Saez P, Pascual-Marqui RD, Tirado F, Carazo JM, Pascual-Montano A. BMC Bioinformatics 2006;7:78. [PubMed: 16503973]
34. Wang G, Kossenkov AV, Ochs MF. BMC Bioinformatics 2006;7:175. [PubMed: 16569230]
35. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A. Nucleic Acids Res 2004;32:D41–4. [PubMed: 14681354]
36. Guldener U, Munsterkotter M, Kastenmuller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE, Michael H, Kaps A, Talla E, Dujon B, Andre B, Souciet JL, De Montigny J, Bon E, Gaillardin C, Mewes HW. Nucleic Acids Res 2005;33:D364–8. [PubMed: 15608217]
37. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM. Nucleic Acids Res 2004;32:D311–4. [PubMed: 14681421]
38. Hoyer P. J Mach Learning Res 2004;5:1457.
39. Bidaut G, Suhre K, Claverie JM, Ochs MF. BMC Bioinformatics 2006;7:99. [PubMed: 16507110]
40. Bidaut G, Ochs MF. Bioinformatics 2004;20:2869–71. [PubMed: 15145813]

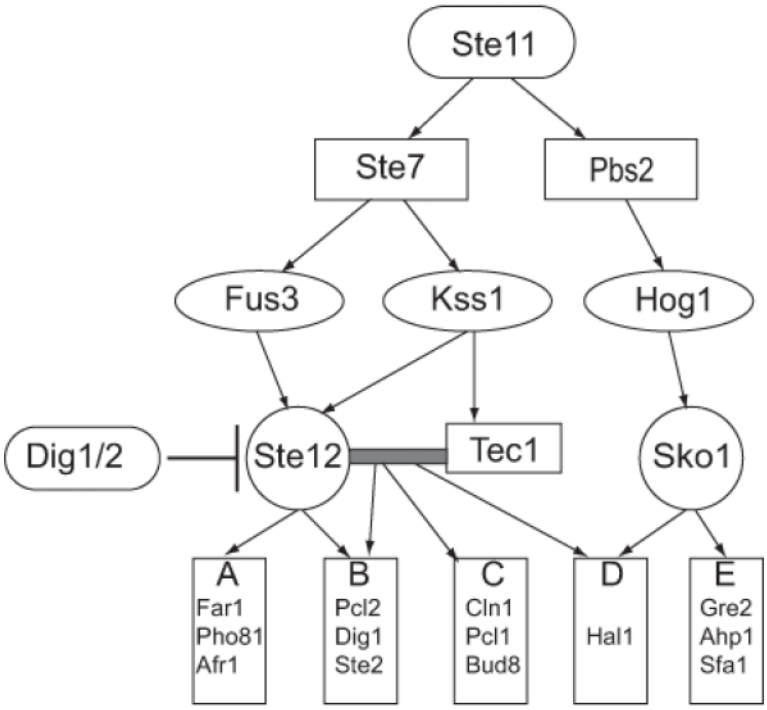

Figure 1.
The tightly coupled MAPK pathways in *S. cerevisiae*. Activation of the pathways lead to transcriptional responses, which produce overlapping sets of transcripts that would be measured in a gene expression experiment. This multiple regulation, which is ubiquitous in eukaryotic biology, motivates the use of matrix factorization methods in high-throughput biological data analysis.

Table 1

The mapping of originally reported processes identified by two dimensional clustering to MIPS categories together with the number of proteins in each category in MIPS and in the analyzed data set. Transposable elements have been added to track the difference between mating and filamentation, as filamentation requires transposable element activation

Original Report	MIPS Number	MIPS Name	Proteins	In Data
Mitochondrial function	02.45	Energy conversion and regeneration	44	4
Cell wall	42.01	Biogenesis of cell wall	214	33
Protein synthesis	12	Protein synthesis	480	16
Protein synthesis	12.01.01	Ribosomal proteins	246	9
Mating	41.01.01	Mating	69	15
MAPK activation	30.01.05.01.03	MAPKKK cascade	27	5
Histone deacetylase	10.01.09.05	DNA conformation modification	187	5
--	38	Transposable elements	120	14

Table 2

Hypergeometric p-values for enrichment in gene ontology terms for different methods: BD – Bayesian Decomposition, BFRM – Bayesian Factor Regression Modeling, NCA – Network Component Analysis, ICA – Independent Component Analysis, PCA – Principal Component Analysis, HC – Hierarchical Clustering, KMC – Kmeans Clustering. The random clustering and pattern recognition methods, as well as all Non-negative Matrix Factorization methods, are not included as they generated no *p*-values under 0.50. Letters are assigned to patterns or clusters as they appear within a column in order to allow the reader to identify repeated uses of the same pattern

MIPS Name	BD	BFRM	NCA	ICA	PCA	HC	KMC
Energy generation (ATP synthase)	A 0.029	A 0.17	A 0.39	A 0.19	A 0.47	A 0.28	A 0.16
Biogenesis of cell wall	B 0.015	B 0.050	B 0.14	A 0.083	B 0.18	B 0.069	A 0.15
Protein synthesis	A 0.0076	C 0.016 D 0.021	B 0.37	B 0.12	C 0.084	C 0.0028 D 0.045 E 0.0016	B 0.009 C 0.04
Ribosomal proteins	C 0.017 A 0.016	C 0.038	C 0.33	A 0.074	C 0.020	D 0.015 E 0.038	B 0.0008
Mating	D 0.0001 E $<10^{-5}$	E $<10^{-5}$	D 0.0018 E 0.015	C 0.0041 D 0.14 E 0.0026	D 0.0018	F $<10^{-5}$	D $<10^{-5}$
MAPKKK cascade	F 0.04	D 0.19	E 0.43	B 0.18	E 0.28	E 0.17	D 0.17
DNA conformation modification	-	F 0.069	F 0.092	F 0.27	F 0.073	F 0.051	E 0.19
Transposable elements	D $<10^{-5}$ G $<10^{-5}$	G $<10^{-5}$ E 0.0004	D $<10^{-5}$ E 0.0002	-	D 0.025	G $<10^{-5}$	F $<10^{-5}$
Terms with Significant Enrichment	7	5	2	1	3	4	4