Published in final edited form as:
Int J Data Min Bioinform. 2010 ; 4(1): 72–90.

Matrix Factorization Methods Applied in Microarray Data Analysis

Andrew V. Kossenkov and
The Wistar Institute, 3601 Spruce Street, Philadelphia, PA 19104 USA

Michael F. Ochs
Department of Oncology, Johns Hopkins University, 550 North Broadway, Suite 1103, Baltimore, MD 21205 USA

Andrew V. Kossenkov: akossenkov@wistar.org; Michael F. Ochs: mfo@jhu.edu

Abstract

Numerous methods have been applied to microarray data to group genes into clusters that show similar expression patterns. These methods assign each gene to a single group, which does not reflect the widely held view among biologists that most, if not all, genes in eukaryotes are involved in multiple biological processes and therefore will be multiply regulated. Here, we review several methods that have been developed that are capable of identifying patterns of behavior in transcriptional response and assigning genes to multiple patterns. Broadly speaking, these methods define a series of mathematical approaches to matrix factorization with different approaches to the fitting of the model to the data. We focus on these methods in contrast to traditional clustering methods applied to microarray data, which assign one gene to one cluster.

Keywords

microarray; matrix factorization; statistics; gene expression; mRNA

1 Introduction

The development of microarray technology [54,38] in the mid-1990s ushered in an era of massively parallel exploration of the transcriptional response in many biological systems (e.g., [1,10]). Initial work on interpreting the transcriptional data focused on the application of statistical techniques for identifying genes linked to conditions [26,59] and clustering techniques linking genes by overall transcriptional profiles [11]. These techniques have been reviewed recently by Quackenbush [50]. However, it was understood quite early that such approaches overlooked the expected multiple regulation of genes [2,44].

In this review, we look at five methods that have been developed to identify patterns of behavior across conditions and to link genes to these patterns. Mathematically, the methods reduce to matrix factorization as in

$$\mathbf{D}=\mathbf{M}+\varepsilon=\mathbf{AP}+\varepsilon, \quad (1)$$

Correspondence to: Michael F. Ochs, mfo@jhu.edu.

where the matrix **D** is the measured data, **M** is the reconstruction of the data matrix from the factorization, **A** provides the distribution of the genes into the patterns, **P** provides a measure of behaviors across the conditions (i.e., the patterns), and ε is the error. The measured data, **D**, are the estimates of the expression levels for each gene in each condition (see Figure 1), and these would typically be deduced from replicated microarray measurements using a statistical technique. For some of the methods discussed here, the additional information available in the form of uncertainties on these measurements is used to improve the estimation of **A** and **P**. This approach to estimate **A** and **P** is similar to Factor Analysis [32] and Blind Source Separation [9]. There are also methods that utilize fuzzy clustering techniques [16,14] and biclustering [58] to assign genes to multiple groups or to find patterns in subsets of the data, however we focus only on the matrix factorization methods here.

We discuss five algorithms that have been applied with varying success to microarray data. Principal Component Analysis (PCA) and its sibling, Singular Value Decomposition (SVD), are statistical techniques that use orthogonality conditions to define a new set of basis vectors for multidimensional data. Independent Component Analysis (ICA) can be considered a variation of PCA, where the orthogonality condition is converted into a more flexible independence criterion. Network Component Analysis (NCA) utilizes prior information in the form of knowledge of the targets of transcriptional regulators to restrict the possible solutions to equation 1. Nonnegative Matrix Factorization (NMF) and extensions, such as sparse NMF (sNMF) and least squares NMF (lsNMF), enforce positivity and other constraints on the **A** and **P** matrices in Equation 1. Bayesian Decomposition (BD) uses Markov chain Monte Carlo sampling and integrates prior distributions that can introduce constraints to limit the solution space.

2 Algorithms for Matrix Factorization

2.1 Singular Value Decomposition and Principal Component Analysis

Singular Value Decomposition (SVD) and the closely related Principal Component Analysis (PCA) methods were first introduced to microarray analysis by Alter *et al.* in their analysis [2] of the Stanford yeast cell cycle data [57]. Later, these methods were applied to many other data, for example genetic profiling in leprosy [5], the analysis of gene expression in Down's syndrome [39], the analysis of human fibroblast data [21], breast tumor classification [19], and the identification of tissue specific gene expression patterns [42].

SVD operates on the data matrix, **D**, of N genes over M conditions with rank r as shown in Figure 2. In this case D_{ij} is the expression level of the i^{th} gene in the j^{th} condition. The elements of the i^{th} row of **D** form an M-dimensional vector, $g_i = D_{i\bullet}$, which is referred to as the transcriptional response or expression profile of the i^{th} gene. Alternatively, the elements of the j^{th} column of **D** form an N-dimensional vector, $a_j = D_{\bullet j}$, which is referred to as the expression profile of the j^{th} condition. SVD of the matrix **D** produces two orthonormal bases, one defined by right singular vectors and the other by left singular vectors, such that

$$\mathbf{D}=\mathbf{U}\mathbf{S}\mathbf{V}^T, \tag{2}$$

where columns of **U** are the left singular vectors, which form an orthonormal basis for the condition expression profiles, **S** is a diagonal matrix of ordered singular values, and the rows of $\mathbf{V}^T$ are the right singular vectors corresponding to ordered singular values, which form an orthonormal basis for the gene transcriptional responses (see Figure 2). Therefore, gene transcriptional response $D_{i\bullet}$ can be described as a linear combination of the right singular

vectors, also called eigengenes. Alternatively, sample expression profile $D_{\bullet j}$ can be presented as a linear combination of the left singular vectors called eigenassays. Eigengenes represent expression patterns found in the data, and eigenassays define what genes are associated with a condition.

In some cases it may be practical to reduce the dimensionality of the matrices to $p < M$, then only the p largest singular values are calculated, while the rest of the matrix is discarded, as in Figure 3. This way, only p expression patterns will be found in the process. The truncated SVD is not an exact decomposition of the original data matrix. Nevertheless, the approximation may be sufficient for practical applications. This approach is used in many fields to filter out noise, but it has not yet proven useful for removing noise from microarray data.

PCA is sometimes incorrectly treated as a synonym for SVD, but it is a special case that operates on the covariance matrix instead of the original data matrix. PCA projects data into directions with the most data variance via a linear transformation. A new coordinate system is selected in such a way that the greatest variance of the data is located on the first coordinate (called the first principal component, or first PC), the second greatest variance on the second coordinate, etc. All PCs are required to be orthogonal to each other. Retaining only the top p PCs as expression patterns allows matrix dimensionality reduction, while retaining the maximum amount of variance in the data at any dimension.

One of the most valuable features of SVD and PCA is that it is possible to determine a number of expression patterns that explain the data by filtering out the eigengenes that represent noise or experimental artifacts [2]. Using this property, PCA can be used to obtain the true dimensionality of the data for methods that require a defined number of patterns prior to analysis [49]. Also, SVD is capable of detecting biologically meaningful patterns of expression, even when clustering methods fail due to weak signals in the data [60].

The major issue with PCA and SVD is the restriction of orthogonality that they impose on underlying expression patterns. If the observed signals (e.g., PCs) are not orthogonal by nature, which is generally true for biological data, PCA can fail to produce biologically meaningful expression patterns. In addition, these methods do not take into account any error measures associated with the data points. This can lead to the propagation of noise across PCs due to the orthogonality constraint, as once noise is modeled in a PC, additional PCs must have structure orthogonal to this noise. A modification that accounts for noise in PCA has been introduced by Sanguinetti *et al.* [53].

2.2 Independent Component Analysis

The application of Independent Component Analysis (ICA) to gene expression data was introduced independently by Lin *et al.* [37] and Liebermeister [36]. Liebermeister compared ICA with PCA, demonstrating that introduction of a nonorthogonal basis for dimensionality reduction is both more meaningful biologically and more capable of accounting for high-order dependencies in the data. The work by Lin *et al.* analyzed the Rosetta yeast deletion mutant data [24], while the work by Liebermeister [36] focused on yeast cell cycle data [57] and B-cell lymphoma data [1]. In more recent studies, ICA has also been applied to the classification of ovarian cancer [40], the study of endometrial cancer [52], and the diagnosis of human cancer types [63].

Like typical applications of PCA to microarray data, ICA performs matrix decomposition by projecting the data onto a lower dimensional space. However, by removing all linear correlations, ICA allows a nonorthogonal basis for such a decomposition, although it still requires statistical independence between components. Since the observed microarray

signals are a result of a mixture of underlying biological processes in the cell, the factorization of the data matrix, **D**, can be expressed as

$$\mathbf{D}=f(\mathbf{AP}) \tag{3}$$

where matrix **P** includes statistically independent biological processes, and matrix **A** is a mixture matrix showing for each gene what biological processes contribute to the expression profile of the gene. For the case of linear ICA,

$$\mathbf{D}=f(\mathbf{AP})=\mathbf{AP}. \tag{4}$$

In order to estimate matrix **P**, the problem for linear ICA may be formulated as

$$\mathbf{P} \approx \boldsymbol{Y}=\mathbf{WD}, \tag{5}$$

so that we need to find a matrix **W** (the unmixing matrix), such that the rows of matrix **Y** are as statistically independent as possible. In this case, **Y** will be an approximation for **P**.

The process of finding the unmixing matrix can be performed by different algorithms, based on different metrics of statistical independence. For example, maximum likelihood estimation, a statistical approach for finding estimates of unknown parameters that result in the highest probability for the observations [25], can be applied. Another approach is to maximize negentropy (or, equivalently, minimize mutual information), by maximizing

$$J(\boldsymbol{Y})=H(\boldsymbol{Y}_{gauss}) - H(\boldsymbol{Y}), \tag{6}$$

where

$$H(x)= - \int f(x)\log f(x)\, dx, \tag{7}$$

and $\mathbf{Y}_{gauss}$ is a random Gaussian matrix with the same covariance as **Y**.

Maximum non-gaussianity can also be used as a measure of independence by using the Kurtosis metric,

$$\text{kurt}\,(\boldsymbol{Y})=E\{\boldsymbol{Y}^4\} - 3(E\{\boldsymbol{Y}^2\})^2 \tag{8}$$

where $E(\mathbf{Y}^4)$ and $E(\mathbf{Y}^2)$ are the 4th and 2nd moments of **Y** respectively.

As has been mentioned earlier, ICA has an advantage over PCA, since it does not impose an orthogonality requirement. It is able, therefore, to recover mixed signals. ICA also has been shown to outperform PCA, K-means clustering and the Plaid model [34] on combined yeast cell-cycle [57], yeast stress [17], *C. elegans* [29], and human normal tissue [23] data.

Although the statistical independence requirements of ICA are not as strict as the orthogonality requirements of SVD and PCA, the assumption of the independence between the underlying processes may not be fully justified in most microarray data. The method also does not take into account any error measures associated with microarray measurements. In addition, relaxing the strict criteria used in SVD and PCA allows some ICA methods to identify multiple sets of components, by becoming trapped in local maxima in the probability space defining the relative probability of different potential **A** and **P** matrices. As such, multiple applications to the data may be necessary, and a decision must then be made concerning which components to retain [13], similar to the choice of factors in factor analysis.

2.3 Network Component Analysis

Introduced by Liao and Roychowdhury, Network Component Analysis (NCA) uses information on the binding of transcriptional regulators to DNA to reduce the possible **A** and **P** matrices [35]. The concept is to create a two layer network with one layer populated by transcriptional regulators and the other by genes they regulate. Edges then connect each regulator to the genes they regulate.

NCA handles the degeneracy problem in Equation 1 by including all potential solutions through

$$\mathbf{D}=\mathbf{AXX}^{-1}\mathbf{P}+\varepsilon, \tag{9}$$

where **AX** includes all potential **A** matrices. The solutions are limited by requiring **X** to be diagonal, which determines **A** and **P** uniquely up to a scaling factor. The diagonality of **X** can be guaranteed if 1) **A** is full column rank, 2) removal of a regulatory node (transcriptional regulator) yields a network where **A** is still full column rank, and 3) **P** is full row rank. Effectively this last criterion demands that the transcriptional regulators be independent, such that no regulator can be described as a linear combination of other regulators within the data set.

The solutions, **A** and **P**, are determined by minimizing

$$\left\|\mathbf{D}-\mathbf{AP}\right\|^2 \tag{10}$$

which is equivalent to maximizing the likelihood with an assumption of uniform Gaussian errors. As seen below (Equation 15), this could be easily extended to gene and array specific errors by inclusion of specific error terms. The rows of **P** are normalized so that each row provides the average effect of a regulator.

For the application of NCA to gene expression data, the relative strengths of the regulation of each gene must be determined. For each gene and each regulator, the gene regulation is assumed to be proportional to 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 (e.g., time point) must be estimated as a combination of the regulation from different factors. For each time point, this is estimated as

$$\frac{E_i^t}{E_i^0} = \prod_{j=1}^{R} \left(\frac{\mathrm{TF}_j^t}{\mathrm{TF}_j^0} \right)^{\mathrm{Aff}_{ij}} \tag{11}$$

where E_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 or regulators, and Aff_{ij} is the binding affinity for the j^{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.

A similar approach can be used without the constraint of Equation 11, with a more flexible limitation being applied on the configuration of the network. Yu and Li extended factor analysis (FA) to include a sparse matrix encoding knowledge of links between transcriptional regulators and the genes they regulate in a connectivity matrix [62]. FA has been applied in many fields, and it is widely used in analysis of social and medical data, where the goal is to identify independent factors that are related to phenotype or response [32]. FA often relies on techniques such as PCA to isolate patterns that can explain the variance in the data, but then applies a multidimensional rotation to create mixtures of the principal components that relate to interpretable attributes in the system. However, other initial analyses have been used, such as hierarchical clustering [48]. Yu and Li initiated the connectivity matrix with high probability relationships, but allowed new connections to be formed between regulators and genes, predicting novel regulation. This is particularly useful as our knowledge of links between regulators and genes, and how these links vary in context-specific ways (e.g., in different cell types, in different environments), is quite limited.

2.4 Nonnegative Matrix Factorization

First introduced by Lee and Seung for feature recognition in images [33], non-negative matrix factorization (NMF) was adopted for analysis of gene expression data by Kim and Tidor [28] and Mesirov and colleagues [6]. Studies using NMF analysis have included yeast mutants [28], leukemia [6], toxicology [12], and lung squamous cell carcinoma [27].

The goal of the NMF analysis is to find a small number of metagenes, each defined as a positive linear combination of n genes. The mRNA level estimates across conditions for each gene can be approximated then as a positive linear combination of these metagenes, which appear in the analysis as rows in matrix **P**.

Unlike PCA and SVD, NMF provides an inherent reduction in dimensionality, with the dimensionality being a required input parameter, K. Then, the RHS of Equation 1 becomes a sum over K metagenes, such that

$$M_{ij} = \sum_{K} A_{ik} P_{kj}. \tag{12}$$

The value of element P_{kj} indicates the strength of metagene k in condition j, while the value of element A_{ik} provides the strength of the assignment of gene i to metagene k.

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 two matrices are then iteratively updated with

$$P_{\alpha\mu} \leftarrow P_{\alpha\mu} \frac{\sum_i A_{i\alpha} D_{i\mu}}{\sum_i A_{i\alpha} M_{i\mu}}$$
$$A_{\delta\alpha} \leftarrow A_{\delta\alpha} \frac{\sum_j D_{\delta j} P_{\alpha j}}{\sum_j M_{\delta j} P_{\alpha j}}, \tag{13}$$

which guarantees reaching a local maximum in the Likelihood and minimizes

$$\left\| D - M \right\|^2 = \sum_{ij} (D_{ij} - M_{ij})^2. \tag{14}$$

In comparison to PCA and ICA, NMF is capable of finding smaller, more localized patterns, as well as global patterns [28], since it is does not restrict the recovered metagenes by an independence criterion. The key assumption is non-negativity of the underlying signals, which is reasonable for single color expression data and non-log transformed ratio expression data, since there are no negative copies of mRNA and no negative transcription. The absence of constraints does lead to a tendency for the recovery of signal-invariant metagenes that carry little or no information. This problem was addressed by Gao and Church using sparse NMF (sNMF), which penalized solutions based on the number of non-zero components in **A** and **P** [15]. Carmona-Saez *et al.* used a similar approach in non-smooth NMF (nsNMF), which created a sparse representation of the metagenes by introducing a smoothness matrix into the factorization [8].

NMF techniques do not account for uncertainty information, leading to potential overfitting of the data, just as with PCA and ICA. We recently created a modified NMF, least-squares NMF (lsNMF), by introducing new updating rules and replacing the criterion for distance minimization with a minimization of the χ^2 error [61], given by

$$\chi^2 = \sum_{ij} \left(\frac{(D_{ij} - M_{ij})}{\sigma_{ij}} \right)^2, \tag{15}$$

where σ_{ij} is the error estimate for data element D_{ij}. With this measure, the update rules become

$$P_{\alpha\mu} \leftarrow P_{\alpha\mu} \frac{\sum_i A_{i\alpha} \frac{D_{i\mu}}{\sigma_{i\mu}}}{\sum_i A_{i\alpha} \frac{M_{i\mu}}{\sigma_{i\mu}}}$$
$$A_{\delta\alpha} \leftarrow A_{\delta\alpha} \frac{\sum_j \frac{D_{\delta j}}{\sigma_{\delta j}} P_{\alpha j}}{\sum_j \frac{M_{\delta j}}{\sigma_{\delta j}} P_{\alpha j}}, \tag{16}$$

and the algorithm proceeds as with NMF to find a local maximum in probability according to the minimization of Equation 15.

All NMF methods suffer from an inability to explore the probability distribution (i.e., the solution space) adequately, due to the inability to escape local maxima in the probability distributions. The solution to this issue has been to routinely begin with up to 100 different initial random **A** and **P** matrices, then to look for the solution which provides the best fit to the data, as measured by Equation 14 or 15. In our experience, the metagenes obtained 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.

2.5 Bayesian Decomposition

Introduced just before NMF, Bayesian Decomposition (BD) was developed by us for applications in spectral imaging [46], and we later adapted it for gene expression data [44]. It has been used as a supervised method to isolate gene expression related to tissue type [43], as a discovery method to isolate biomarkers related to lung cancer subtype [30], and as a method to isolate coregulated genes for promoter analysis in Plasmodium falciparum [47]. It was successfully applied recently for identifying signaling activity in yeast deletion mutants from microarray data [4].

Like lsNMF, BD uses constraints and dimensionality reduction to limit the possible **A** and **P** matrices in the solution space. In addition, it applies an Occam's Razor argument, similar to sNMF, to penalize excessive structure in the estimates of **A** and **P**. In BD, these constraints are encoded through a prior distribution in a Bayesian statistical framework [56]. This approach allows patterns to be constrained in multiple ways, permitting the rows of **P** the freedom to be nonorthogonal and nonindependent.

The incorporation of prior knowledge into the analysis is implemented through an atomic domain and mappings between it and the model domain (i.e., the **A** and **P** matrices), as shown in Figure 4. The atomic domain forms a positive additive distribution, and mappings to the model domain provide constraints [55]. For instance, in Figure 4, the mapping shown on the left enforces positivity on the matrix **A**, as with NMF, by simply placing the amplitude of the atom into a single matrix element. In contrast, the mapping on the right enforces correlated transcriptional levels between genes through information on shared transcriptional regulators. The atomic domain consists of two infinitely divisible one-dimensional spaces (actually 2^{32} points), one corresponding to the **A** matrix and one to the **P** matrix. Atoms are created along these two lines according to a prior distribution that is uniform in position and exponential in amplitude, which encourages a minimization of structure. Using

$$\mathbf{A}=\int K_w \varphi_w, \tag{17}$$

convolution functions map atoms to matrix elements allowing preferred correlations between matrix elements to increase in probability. Through the convolutions, a set of values, here the matrix **A**, can be constructed from a family of measures, φ, (the atoms) using kernels, K. In the simplest case (left in Figure 4), an atom simply maps to a single matrix element.

In order to determine how to place and size atoms, a Markov chain Monte Carlo (MCMC) procedure is used. MCMC techniques allow exploration of probability distributions [20], here the solution space of possible **A** and **P** matrices. Atoms are created *ex vacuo* according to the prior. A unit flux is mapped to the model domain using Equation 17, and the affect on the likelihood (i.e., the fit to the data) is computed. The amplitude of the atom is then

adjusted based on the parameters of the change of the likelihood, to maximize the exploration of the probability distribution. The probability for each point in the solution space can be determined from Bayes rule,

$$p(\mathbf{A}, \mathbf{P}|\mathbf{D}) \approx p(\mathbf{D}|\mathbf{A}, \mathbf{P})p(\mathbf{A}, \mathbf{P}), \tag{18}$$

where $p(\mathbf{A}, \mathbf{P}|\mathbf{D})$ is the conditional probability of the model given the data (the *posterior*), $p(\mathbf{D}|\mathbf{A}, \mathbf{P})$ is the conditional probability of the data given the model (the *likelihood*), and $p(\mathbf{A}, \mathbf{P})$ is the probability of the model (the *prior*). The posterior distribution is the solution space for our problem, since it measures the probability of the model (**A** and **P**) in light of the data. Here we are ignoring a normalizing parameter, $p(\mathbf{D})$, the *evidence* or *marginal distribution*, since it does not affect the Markov chain sampling. The relative probability at a point in the solution space is determined completely by the likelihood, which is easily determined by comparing the model to the data (i.e., **M** and **D**), and the prior, which is the probability of the model independent of the data. In effect, the sampler is using relative probability measurements, provided through a Bayesian approach, for determining the acceptability of a change in the model.

Bayesian methods have been applied to gene expression analysis as well in the form of mixture models for validating clusters from hierarchical methods [45,41], though these retain the problem of the assignment of each gene to a single cluster.

The sampling of the probability distributions of the elements of the **A** and **P** matrices provide both a mean solution and uncertainty estimates, unlike NMF, which provides a single local best solution. In addition, solutions differing by more than the uncertainties (e.g., two distinct sets of expression signatures that both reconstruct the data) can be identified. MCMC techniques also are less prone to becoming trapped in local maxima, since they have an ability to escape these regions. In addition, BD uses simulated annealing [18] during equilibration to decrease the probability of becoming trapped in a broad local maxima.

In BD, the sampling from the posterior distribution and the encoding of the prior are done using a customized bilinear form of the Massive Inference™ Gibbs sampler from Maximum Entropy Data Consultants, Cambridge, England [55]. Nevertheless, other approaches to sampling that rely on Equations 1, 17, and 18 should provide similar results.

BD has a significant advantage over many methods applied to microarray data, in that it handles missing data in a parsimonious way. Because of the atomic domain and the prior upon it, all proposed changes to the model are made *ex vacuo*, so that there is no feedback from the data. Therefore, missing values can be ignored by setting them to some minimal value (1 for ratio measurements, 0 for absolute measurements), and providing a large uncertainty estimate. In this way, missing values have no impact upon the model and do not need to be estimated, which, except for some time series data, can be extremely problematic.

3 Application to a Sample Data Set

In order to compare the performance of different methods, we created an artificial dataset that simulates the yeast cell cycle. The dataset contains expression levels for 288 genes across 48 time points, which includes two full passes through the four phases of the cell cycle. There are five patterns in the data: four modeled on cell cycle phases and one modeled on a metabolic oscillator, encoded in the **P** matrix of Equation 1. The expression profile of each gene is a mixture of 1 or more expression patterns linked to a phase or the

oscillator. While it would be better to use a real data set, comparison of the methods requires a gold standard, and these do not exist except in very limited cases and with minimal information.

The data was generated using

$$D_{ij}=\frac{1}{4}\sum\left\{\mathbf{N}(0,\sigma_a)+x_{ij}e^{\mathbf{N}(0,\sigma_b)}\right\} \tag{19}$$

where $x_{ij} = \Sigma A_{ip}P_{pj}$ is noiseless simulated data arising from known **A** and **P** matrices. This simulates a microarray experiment with four-fold replication and additive and multiplicative noise according to the widely accepted model [51]. Multiplicative noise was 29% of signal, while additive noise was 16% of peak signal.

We then applied a number of methods to see how well the original **A** and **P** matrices were recovered. For comparison purposes, we first applied clustering techniques, hierarchical clustering (HC), K-means clustering (KMC), and random clustering (by randomly assigning genes to 5 groups, RND). For these methods, the assignment of genes to patterns is automatic. We applied nonnegative matrix factorization (NMF), its sparse form (sNMF), its least squares form (lsNMF), and Bayesian Decomposition. To assign genes to groups, we applied a threshold such that the strength of the assignment of a gene to a pattern (i.e., value in the appropriate **A** element) had to be above the threshold times the average value of assignment of genes to that pattern (i.e., column average). We applied principal component analysis (PCA), independent component analysis (ICA), and network component analysis (NCA). To assign genes to groups, we used the same threshold but took the average of the absolute values in the columns, as these methods can return negative values.

The code used for calculations was obtained from multiple sources. NCA was run in Matlab (Mathworks, Inc.) using code from J. Liao [35] with random initial states. Inclusion of partial prior information on the network appeared to not improve the results. NMF and sNMF calculations relied on code from P. Hoyer [22], with a sparse matrix setting of 0.8. Calculations for lsNMF were done using code from G. Wang [61], and Bayesian Decomposition utilized code from our group [44]. Standard Matlab routines were used for all other methods. When applicable, standard deviations of the mean (i.e., standard errors) were provided to the algorithms from the four simulated replicated arrays.

The accuracy was calculated for each run of each method, based on the number of true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) assignments. True positives are counted as the number of genes in a pattern in the simulated **A** matrix recovered in the pattern in the reconstructed **A** matrix. False positives are genes associated with the pattern in the recovered **A** matrix that are not in the simulated **A** matrix. The true and false negatives are similarly calculated. The accuracy represents the fraction of correct assignments, thus

$$\mathrm{Acc}=\frac{TP+TN}{TP+TN+FP+FN}. \tag{20}$$

Since there are 5! = 120 possible combination of assignment of recovered groups to real groups, we chose the best assignment in terms of accuracy values for each result of each method. Ten runs for methods that are based on random initial step or random walks were

performed, and average and maximum accuracy is reported in Table 1. An ROC analysis, which looks at the capabilities of a classification method across all possible thresholds was done, and the results are presented as area under the curve (AUC) values in Table 1. An AUC of 1 indicates a perfect test.

4 Conclusion

The general goal of microarray data analysis is to identify signatures of biological significance. In the simplest form, these may be biomarkers that allow prediction of a phenotypic response, such as inhibition of tumor growth by a therapeutic compound. Such biomarkers allow the refinement of treatment of individuals based on their specific molecular profile. For biomarkers, the overall goal of analysis is a strong signal that is related directly to class or condition. Another goal of microarray data analysis may be the identification of changes in a biological process under study. For instance, we may wish to know how the activity in signaling networks within a cell changes during a perturbation. In these cases, we need to disentangle the many variations across conditions in the transcriptional response and relate them to multiple biological processes. The techniques discussed here are often more useful for this second type of analysis, which reduces mathematically to the problem of matrix factorization. The factorization produces patterns that provide insight into how conditions are linked, together with an assignment of genes to these patterns.

Unlike simple clustering, matrix factorization allows the assignment of each gene to multiple coexpression groups, reflecting the biological reality of multiple regulation. These coexpression groups form patterns, referred to in various ways by the authors of the different methods (e.g., principal components, metagenes). All of the methods presented here, SVD, PCA, ICA, NCA, NMF, and BD, use different contraints to limit the possible patterns, since mathematically there are an infinite number of solutions to the matrix decomposition. The methods fall roughly into two groups. SVD and PCA are analytic approaches that constrain the solutions by requiring orthogonality between the patterns. ICA, NCA, NMF, and BD rely on dimensionality reduction and iterative fitting of the data, albeit by different methods. These latter techniques must deal with the problem of multiple possible solutions. ICA and NMF address this by requiring multiple applications of the algorithms to the data with different initial conditions, looking for the best local maximum in probability, which ideally mirrors the true global maximum. NCA uses a maximum likelihood approach with a constraint on the connection between regulators and the genes they regulate. BD uses a Markov chain Monte Carlo approach with simulated annealing to minimize the probability of becoming trapped in a local maximum. BD has the added advantage that as an MCMC technique, it samples the probability distribution instead of making a point estimate. However, all three techniques require care in application to avoid selection of matrices that do not reflect the probability distribution. As can also be seen from Table 1, BD more reliably finds the correct matrices on repeated runs (lower standard deviation of the accuracy) due to the MCMC exploration. A summary of some strengths and weaknesses for the techniques in the analysis of microarray data is provided in Table 2.

The determination of the correct dimensionality for interpretation in SVD and PCA or for fitting for the other techniques remains an open problem. We have used a heuristic approach based on the consistency of the assignment of genes in **A** and of conditions in **P** [3,4], however estimates can still vary for a complex data set. For PCA, numerous methods have been proposed, but little progress has been made for microarray data. Cangelosi and Goriely recently proposed a heuristic information measure that appears to set an upper limit on dimensionality for microarray data [7]. In combination with other approaches, this may provide at least a range of dimensions suitable for analysis. Since PCA is a computationally

inexpensive, dimensionality estimates from a PCA analysis could guide the choice of dimensionality for the other approaches discussed here. Nevertheless, even a minor misestimation of the dimensionality of the data may lead to loss of an important signal in microarray analysis.

Data overfitting is a significant danger with these techniques, and it can be exacerbated when no error model is included in the analysis. The original PCA, SVD, ICA, NCA, and NMF methods do not have tools to handle variance measurements on the microarray data, however modifications of PCA and NMF that target this issue were shown to improve performance of these methods. Nevertheless, error models for microarray data remain poorly defined, and the standard measures of the fit to the data, which rely on these models (e.g., χ^2) may not be reliable in a quantitative sense.

Because of the large volume of detailed information developed in molecular biology, the ability to guide analysis with prior knowledge can be valuable to pattern recognition methods for microarray analysis. Linking of data points (such as by known coregulation) can also help to address the issue of the high noise level in the data, by the “borrowing of statistical power” across data points. Prior knowledge can take a simple form, such as minimization of structure, which is used in modified versions of NMF. It can also be more complex, such as the linking of conditions or genes based on known classification or coregulation, as used by NCA and BD [31].

The methods reviewed here provide different approaches to the problem of matrix factorization of microarray data. Such factorization is a useful compliment to statistical tests and clustering, especially when the goal of analysis is the dissection of the complex interactions occuring between biological processes.

Acknowledgments

The authors wish to acknowledge funding from the National Library of Medicine (LM009382, LM008932) and the Maryland Tobacco Restitution Fund. They also wish to acknowledge helpful comments from the reviewers, especially concerning NCA.

References

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu JL, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 2000;403(6769):503–11. [PubMed: 10676951]
2. 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):10101–6. [PubMed: 10963673]
3. Bidaut G, Ochs MF. Clutrfree: cluster tree visualization and interpretation. Bioinformatics 2004;20(16):2869–71. [PubMed: 15145813]
4. Bidaut G, Suhre K, Claverie JM, Ochs MF. Determination of strongly overlapping signaling activity from microarray data. BMC Bioinformatics 2006;7(1):99. [PubMed: 16507110]
5. Bleharski JR, Li H, Meinken C, Graeber TG, Ochoa MT, Yamamura M, Burdick A, Sarno EN, Wagner M, Rollinghoff M, Rea TH, Colonna M, Stenger S, Bloom BR, Eisenberg D, Modlin RL. Use of genetic profiling in leprosy to discriminate clinical forms of the disease. Science 2003;301(5639):1527–30. [PubMed: 12970564]
6. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 2004;101(12):4164–9. [PubMed: 15016911]
7. Cangelosi R, Goriely A. Component retention in principal component analysis with application to cdna microarray data. Biol Direct 2007;2(1):2. [PubMed: 17229320]

8. 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]

9. Chiappetta P, Roubaud MC, Torresani B. Blind source separation and the analysis of microarray data. J Comput Biol 2004;11(6):1090–109. [PubMed: 15662200]

10. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998;2(1):65–73. [PubMed: 9702192]

11. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95(25):14863–8. [PubMed: 9843981]

12. Fogel P, Young SS, Hawkins DM, Ledirac N. Inferential, robust non-negative matrix factorization analysis of microarray data. Bioinformatics 2007;23(1):44–9. [PubMed: 17092989]

13. Frigyesi A, Veerla S, Lindgren D, Hoglund M. Independent component analysis reveals new and biologically significant structures in micro array data. BMC Bioinformatics 2006;7:290. [PubMed: 16762055]

14. Futschik, ME.; Kasabov, NK. Fuzzy clustering of gene expression data. 2002 IEEE International Conference on Fuzzy Systems; Honolulu, HI. IEEE; 2002. p. 414-419.

15. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics 2005;21(21):3970–5. [PubMed: 16244221]

16. Gasch AP, Eisen MB. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biol 2002;3(11):RESEARCH0059. [PubMed: 12429058]

17. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 2000;11(12):4241–57. [PubMed: 11102521]

18. Geman S, Geman D. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 1984;6:721–741.

19. Ghosh D. Singular value decomposition regression models for classification of tumors from microarray experiments. Pac Symp Biocomput 2002;7:18–29. [PubMed: 11928474]

20. Gilks, WR.; Richardson, S.; Spiegelhalter, DJ. Interdisciplinary statistics. Chapman Hall; London: 1996. Markov chain Monte Carlo in practice.

21. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci U S A 2000;97(15):8409–14. [PubMed: 10890920]

22. Hoyer P. Non-negative matrix factorization with sparseness constraints. J Mach Learning Res 2004;5:1457–1469.

23. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR. A compendium of gene expression in normal human tissues. Physiol Genomics 2001;7(2):97–104. [PubMed: 11773596]

24. 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]

25. Hyvrinen, A.; Karhunen, J.; Oja, E. Independent Component Analysis. John Wiley Sons; New York: 2001.

26. Ideker T, Thorsson V, Siegel AF, Hood LE. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J Comput Biol 2000;7(6):805–17. [PubMed: 11382363]

27. Inamura K, Fujiwara T, Hoshida Y, Isagawa T, Jones MH, Virtanen C, Shimane M, Satoh Y, Okumura S, Nakagawa K, Tsuchiya E, Ishikawa S, Aburatani H, Nomura H, Ishikawa Y. Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. Oncogene 2005;24(47): 7105–13. [PubMed: 16007138]

28. Kim PM, Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 2003;13(7):1706–18. [PubMed: 12840046]
29. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. A gene expression map for caenorhabditis elegans. Science 2001;293(5537):2087–92. [PubMed: 11557892]
30. Kossenkov AV, Bidaut G, Ochs MF. Genes associated with prognosis in adenocarcinomas across studies at multiple institutions. Methods of Microarray Data Analysis IV 2005:239–253.
31. Kossenkov AV, Peterson AJ, Ochs MF. Determining transcription factor activity from microarray data using bayesian markov chain monte carlo sampling. Medinfo 2007;12
32. Lawley, DN.; Maxwell, AE. Factor analysis as a statistical method. 2. American Elsevier; New York: p. 1971
33. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401(6755):788–91. [PubMed: 10548103]
34. Lee SI, Batzoglou S. Application of independent component analysis to microarrays. Genome Biol 2003;4(11):R76. [PubMed: 14611662]
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):15522–7. [PubMed: 14673099]
36. Liebermeister W. Linear modes of gene expression determined by independent component analysis. Bioinformatics 2002;18(1):51–60. [PubMed: 11836211]
37. Lin, SM.; Liao, X.; McConnell, P.; Vata, K.; Carin, L.; Goldschmidt, P. Using functional genomic units to corroborate user experiments with the rosetta compendium. In: Lin, SM.; Johnson, KE., editors. Methods of Microarray Data Analysis II. Kluwer Academic Publishers; Boston: 2002.
38. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14(13):1675–80. [PubMed: 9634850]
39. Mao R, Zielke CL, Zielke HR, Pevsner J. Global up-regulation of chromosome 21 gene expression in the developing down syndrome brain. Genomics 2003;81(5):457–67. [PubMed: 12706104]
40. Martoglio AM, Miskin JW, Smith SK, MacKay DJ. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. Bioinformatics 2002;18(12):1617–24. [PubMed: 12490446]
41. Medvedovic M, Yeung KY, Bumgarner RE. Bayesian mixture model based clustering of replicated microarray data. Bioinformatics 2004;20(8):1222–32. [PubMed: 14871871]
42. Misra J, Schmitt W, Hwang D, Hsiao LL, Gullans S, Stephanopoulos G, Stephanopoulos G. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. Genome Res 2002;12(7):1112–20. [PubMed: 12097349]
43. Moloshok TD, Datta DJ, Kossenkov AV, Ochs MF. Bayesian decomposition classification of the project normal data set. Methods of Microarray Data Analysis III 2003:211–232.
44. 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]
45. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 2001;8(1):37–52. [PubMed: 11339905]
46. Ochs MF, Stoyanova RS, Arias-Mendoza F, Brown T. A new method for spectral decomposition using a bilinear bayesian approach. J Magn Reson 1999;137:161–176. [PubMed: 10053145]
47. Peterson AJ, Kossenkov AV, Ochs MF. Linking gene expression patterns and transcriptional regulation in plasmodium falciparum. Methods of Microarray Data Analysis V 2007:137–56.
48. Peterson LE. Factor analysis of cluster-specific gene expression levels from cdna microarrays. Comput Methods Programs Biomed 2002;69(3):179–88. [PubMed: 12204446]
49. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001;2(6):418–27. [PubMed: 11389458]

50. Quackenbush J. Computational approaches to analysis of dna microarray data. Methods Inf Med 2006;45(Suppl 1):91–103.
51. Rocke DM, Durbin B. A model for measurement error for gene expression arrays. J Comput Biol 2001;8:557–569. [PubMed: 11747612]
52. Saidi SA, Holland CM, Kreil DP, MacKay DJ, Charnock-Jones DS, Print CG, Smith SK. Independent component analysis of microarray data in the study of endometrial cancer. Oncogene 2004;23(39):6677–83. [PubMed: 15247901]
53. Sanguinetti G, Milo M, Rattray M, Lawrence ND. Accounting for probe-level noise in principal component analysis of microarray data. Bioinformatics 2005;21(19):3748–54. [PubMed: 16091409]
54. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science 1995;270(5235):467–70. [PubMed: 7569999]
55. Sibisi S, Skilling J. Prior distributions on measure space. Journal of the Royal Statistical Society, B 1997;59(1):217–235.
56. Sivia, DS. Data analysis: a Bayesian tutorial. Oxford University Press; Oxford: 1996.
57. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998;9(12):3273–97. [PubMed: 9843569]
58. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics 2002;18(Suppl 1):S136–44. [PubMed: 12169541]
59. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98(9):5116–21. [PubMed: 11309499]
60. Wall, ME.; Rechtsteiner, A.; Rocha, LM. Singular value decomposition and principal component analysis. In: Berrar, D.; Dubitzky, W.; Granzow, M., editors. A Practical Approach to Microarray Data Analysis. Kluwer; Norwell, MA: 2003. p. 91-109.
61. Wang G, Kossenkov AV, Ochs MF. Ls-nmf: a modified nonnegative matrix factorization algorithm utilizing uncertainty estimates. BMC Bioinformatics 2006;7:175. [PubMed: 16569230]
62. Yu T, Li KC. Inference of transcriptional regulatory network by two-stage constrained space factor analysis. Bioinformatics 2005;21(21):4033–8. [PubMed: 16144806]
63. Zhang XW, Yap YL, Wei D, Chen F, Danchin A. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. Eur J Hum Genet 2005;13(12): 1303–11. [PubMed: 16205741]

Biographies

Andrew Kossenkov holds a masters degree in Applied Mathematics from Moscow Engineering Physics Institute (Moscow, Russia) and Ph.D in Biomedical Science from Drexel University (Philadelphia, PA). He currently is a postdoctoral fellow in laboratory of Dr. Louise Showe at The Wistar Institute.

Michael Ochs is an Associate Professor of Oncology at the Sidney Kimmel Cancer Center at Johns Hopkins (Baltimore, MD). His main interests are in Bayesian methods for analysis of biological data and computational modeling of cell signaling.

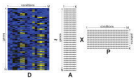

Figure 1.
Matrix Factorization. The data matrix, **D**, is modeled as arising from the multiplication of a set of patterns, the rows of **P**, and the assignment of genes to those patterns with varying strengths, the columns of **A**.

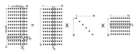

Figure 2.
Singular Value Decomposition. The initial matrix **D** is decomposed into the product of the left singular matrix **U**, the diagonal matrix of ordered singular values **S**, and the right singular matrix $\mathbf{V}^T$. The vectors v_k are the eigenassays, while vectors u_I are the eigengenes.

Figure 3.
Truncated Singular Value Decomposition. It is possible to discard smaller singular values, keeping only the first *p* singular values that keep most of the expression information.

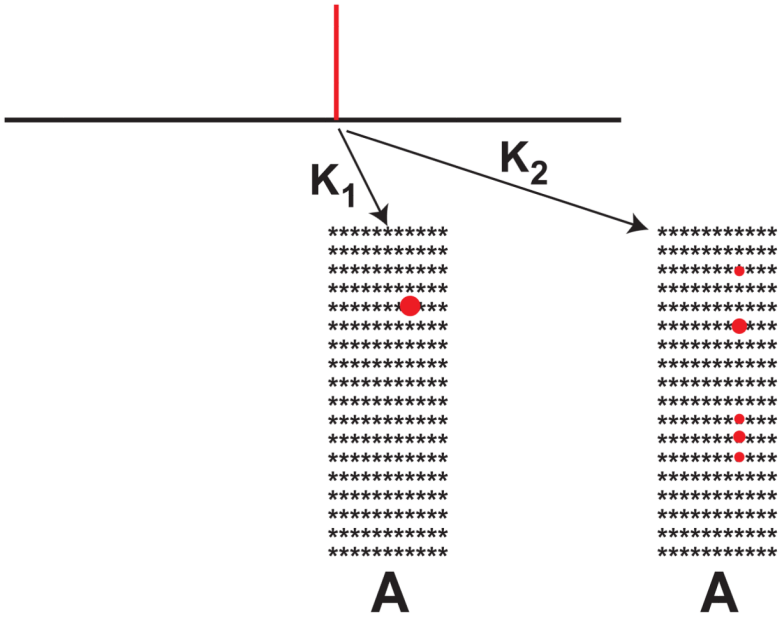

Figure 4.

Atomic Domain. Bayesian Decomposition utilizes two atomic domains, one that maps to the **A** matrix and one to the **P** matrix. The kernel functions, *K*, allow an atom to map to multiple matrix elements, introducing correlations in the model. Here, K_1 is a simple mapping, with all the amplitude of an atom going to a single matrix element, while K_2 represents a kernel for a transcriptional regulator, with five genes responding in a correlated manner.

Table 1

Results of Application of Matrix Factorization Methods on Simulated Data

Method	#TP	#TN	#FP	#FN	Mean Acc	StdDev	Max	Min	AUC
RND	295	471	427	248	**0.532**	0.007	0.538	0.516	0.55
HC	275	885	12	267	**0.806**	0	–	–	–
KMC	244	853	44.5	300	**0.761**	0.004	0.770	0.758	–
PCA	256	829	68	287	**0.753**	0	–	–	0.80
ICA	476	823	74	67	**0.902**	0.007	0.912	0.892	0.93
NCA	334	813	83	209	**0.797**	0.03	0.833	0.736	0.84
NMF	504	824	76.3	38.7	**0.920**	0.03	0.962	0.895	0.92
sNMF	514	824	73.1	29.4	**0.929**	0.03	0.959	0.895	0.94
lsNMF	489	811	86.5	54.4	**0.902**	0.003	0.908	0.897	0.92
BD	524	847	50.1	19.2	**0.952**	0.0007	0.953	0.951	0.98

Table 2

Methods: Advantages and Issues for Microarray Data Analysis

Method	Advantages	Issues
PCA	Strong statistical basis, analytic method, fast, potentially able to determine dimensionality of data	Orthogonality requirement tends to yield patterns that do not reflect biology, error can propagate in lower order patterns
ICA	Strong statistical basis, flexible independence criteria	Independence requirement still limiting for biological data, potential trapping in local maximum in probability space
NCA	Incorporates prior knowledge of transcriptional regulation, handles degeneracy parsimoniously	Does not yet include error modeling, context-specific regulation may limit usefulness of prior knowledge
NMF	Directly factors matrices into biological meaningful patterns, extended to error modeling and sparse matrix methods	No escape from local maxima in potentially complex probability space, sparseness may not yet be adequate in many analyses
BD	Markov chain Monte Carlo explores probability distribution best among methods, flexibly incorporates prior knowledge	Computationally expensive, data size limited to few thousand genes, error models still limited