BMC Bioinformatics

BioMed Central

Methodology article

LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates

Guoli Wang, Andrew V Kossenkov and Michael F Ochs*

Address: Division of Population Science, Fox Chase Cancer Center, Philadelphia, PA, USA

Email: Guoli Wang - guoli.wang@fccc.edu; Andrew V Kossenkov - andrew.kossenkov@fccc.edu; Michael F Ochs* - m_ochs@fccc.edu
* Corresponding author

Published: 28 March 2006

*BMC Bioinformatics*2006, **7**:175 doi:10.1186/1471-2105-7-175

Received: 06 October 2005
Accepted: 28 March 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/175

Abstract

Background: Non-negative matrix factorisation (NMF), a machine learning algorithm, has been applied to the analysis of microarray data. A key feature of NMF is the ability to identify patterns that together explain the data as a linear combination of expression signatures. Microarray data generally includes individual estimates of uncertainty for each gene in each condition, however NMF does not exploit this information. Previous work has shown that such uncertainties can be extremely valuable for pattern recognition.

Results: We have created a new algorithm, least squares non-negative matrix factorization, LS-NMF, which integrates uncertainty measurements of gene expression data into NMF updating rules. While the LS-NMF algorithm maintains the advantages of original NMF algorithm, such as easy implementation and a guaranteed locally optimal solution, the performance in terms of linking functionally related genes has been improved. LS-NMF exceeds NMF significantly in terms of identifying functionally related genes as determined from annotations in the MIPS database.

Conclusion: Uncertainty measurements on gene expression data provide valuable information for data analysis, and use of this information in the LS-NMF algorithm significantly improves the power of the NMF technique.

Background

Because of their ability to link genes that behave similarly across conditions in a gene expression study, pattern recognition and clustering are widely used for data analysis of microarray data (see [1] for a review). Numerous methods have been adopted to cluster genes or samples including hierarchical clustering [2], maximum likelihood clustering [3], the cluster affinity search technique [4], quality threshold clustering [5], and fuzzy K-means clustering [6], among others. Some methods specifically aim to identify subsets of behaviors within the data, where genes behave similarly only over a subset of samples. Such methods include two-way clustering [7] and biclustering [8]. Other methods allow genes to belong to multiple patterns, reflecting biological roles where genes function in multiple cellular processes. Such methods include Bayesian Decomposition [9,10], principal component analysis [11], independent component analysis [12], and nonnegative matrix factorization [13-15].

By creation of a constrained model (e.g., a model with only positive points), non-negative matrix factorization (NMF) shares with Bayesian Decomposition (BD) the potential to more accurately identify sets of genes that together provide function. Both aim to recover two matri-

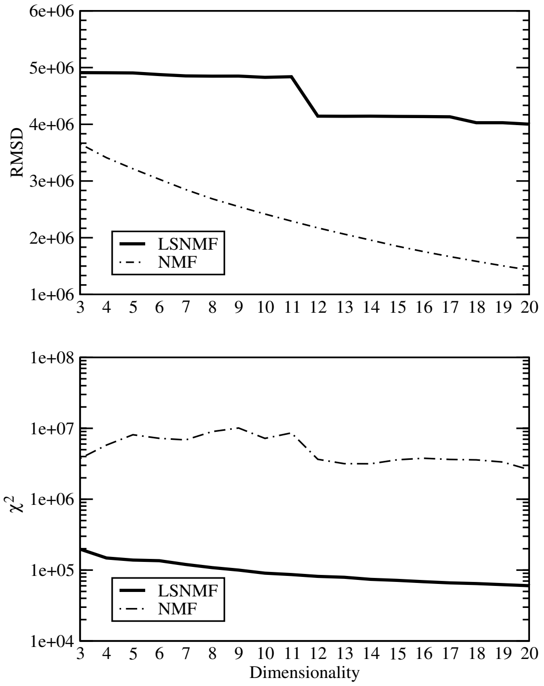

Figure 1

Fits to the Data. The comparison of LS-NMF and NMF for fitting data is shown for the Rosetta Compendium. NMF is based on root mean squared deviation (RMSD) fitting, and as the number of dimensions increases, the fit improves with potential overfitting possible. The χ^2 measurement provides the standard measure of fit to the data and shows how the LS-NMF method fits data much better once individual uncertainty measurements are considered.

ces, **A** and **P**, that reproduce the data within the uncertainty as in

$$\mathbf{D} = \mathbf{M} + \varepsilon = \mathbf{A} \cdot \mathbf{P} + \varepsilon. \quad (1)$$

In the case of NMF, this constraint is positivity in the **A** and **P** matrices, while in BD the constraint is provided through relationships between model points in the form of convolution functions during matrix reconstruction [16]. For microarray data, the matrix **D** provides the estimates of transcriptional levels, such that each column corresponds to the estimate for a single condition, with each matrix element in a column corresponding to the estimate for a single gene (or probe set) in that condition. A row of **D** corresponds to the processed intensity for a single gene across all conditions. If **D** has dimension of $I \times J$, then **A** has dimensions $I \times K$, and **P** has dimensions $K \times J$, where K is the dimensionality (i.e., rank). There is no accepted method yet to choose *K a priori*, however many possible *K* can be tested in order to find an optimal value [17].

The overall simulation in the original NMF algorithm aims to minimize the difference between **M** and **D** in Equation 1, with every element in **D** given the same weight in evaluating the difference between the two matrices. However, most microarray measurements are replicated and, in addition, statistical techniques have been developed to estimate uncertainty measurements for all data points in **D** [18-21]. This uncertainty information is extremely valuable for identifying the best model [10], and it has been used in microarray analysis in other contexts as well [22]. Inclusion of this information in fitting the **A** and **P** matrices should improve the NMF algorithm by allowing it to more precisely fit the reconstructed **M** to the data, **D**. Such approaches have also been used successfully within supervised methods, such as LS-SVM [23].

Results

Data sets

We applied the least squares nonnegative matrix factorization (LS-NMF) algorithm to two publicly available microarray datasets, one with no individual estimates of uncertainties for each data point and one with such information. The first set is yeast cell cycle data from cultures synchronized with a temperature sensitive cdc28-mutant [24], which has a single Affymetrix GeneChip measurement at each time point. The second set is the yeast deletion mutant compendium from Rosetta Inpharmatics [25], which comprises microarray measurements of mRNA levels from yeast cultures containing either clones of *S. cerevisiae* with gene deletions or chemical treatments. Both data sets were preprocessed as described in Methods to create the estimates of mRNA levels for the data matrix, **D**.

Algorithm performance

Data fitting

To measure the ability of LS-NMF to fit the data, we measured the χ^2 fit (as in Equation 6) between the model of the data provided by **M** and the data in **D**, as in Equation 1. The original NMF algorithm used the root mean square distance (RMSD) between these two matrices as a measure of fit, however this is incapable of taking into account the uncertainty information used by LS-NMF. Essentially, NMF ignores variations in the precision of the measurements. The χ^2 measure provides a standard approach to determining the fit between a model of a data set and the data set itself. Figure 1 shows the difference between the **D** and **M** matrices using LS-NMF and NMF for dimensionalities of K = 3 - 20 on the Rosetta data set. Figure 1 demonstrates that the χ^2 error value decreased with increasing dimensionality for LS-NMF, but not for the NMF simulation. At all dimensions, LS-NMF consistently fits the data

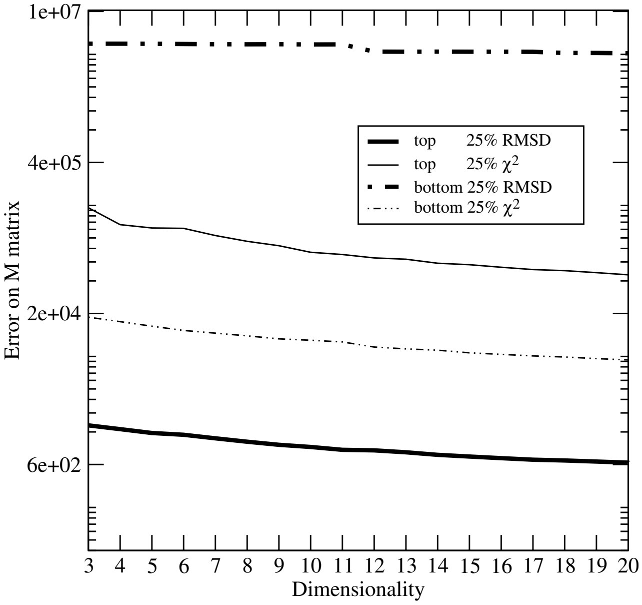

Figure 2
Misfit in M. The error between portions of the **M** and **D** matrices is shown as the dimensionality increases for the Rosetta data as fit by LS-NMF. The bottom 25% of the genes in terms of precision (i.e., highest variance) are shown together with the top 25%. Since LS-NMF relies on the variance, the algorithm fits the reliable data with improving accuracy, while ignoring poor data in terms of RMSD as the dimensionality increases.

better, which is expected as NMF does not optimize M for the known variance. More interestingly, in the LS-NMF simulation, the RMSD error appears to be independent of the dimensionality, while NMF does depend on the dimensionality. This results from the LS-NMF simulation fitting the more reliable data points tightly, while allowing the data points with high variance to be fit more loosely. This can be seen in Figure 2, where points with the highest variance (bottom 25%) do not improve in RMSD in LS-NMF, while they continue to show improvement in χ^2. The χ^2 measure takes into account the variance providing a better measure of fit in cases of varying variance.

Computational complexity

While incorporating uncertainty measurements into NMF update rules (as described in Methods) makes sense for fitting microarray data, if the modification adds too much computational complexity, it will not be useful practically. We have measured the additional computational cost when applied to the Rosetta data set, and we have found that it is a constant offset compared to the cost of the original NMF implementation, regardless of complexity. This indicates that LS-NMF has the same convergence speed as NMF after a higher set-up cost, so that LS-NMF can be used with a minimal loss of efficiency compared with the original NMF algorithm in most practical applications.

Biological insights

ROC analysis of metagenes

After normalization as described in the Methods, the contribution of each gene to a metagene is represented by a scaled *Z*-score, with a positive *Z*-score indicating that the gene is likely to be associated with that metagene, and a negative *Z*-score indicating that the gene is unlikely to be associated with the metagene. Metagenes were introduced to summarize behavior shared by many genes within an experiment that together provided the ability to classify samples [26,27]. This was similar to the concept of an eigengene from singular value decomposition [11], but with regression models using classification data driving the determination of mixing of the genes in a metagene. Metagenes have been discussed in relation to NMF analysis of microarray data previously [15], where they summarize behavior across conditions and assign fractions of the overall expression pattern for each gene to these behaviors.

Figure 3 shows that metagenes can recover known gene coregulation in the well studied yeast cell cycle data set [24]. The ROC test for the cell cycle data set relies on known sets of coregulated genes [28] and is described in Methods. Figure 3a shows the areas under the ROC curve against the number of metagenes (dimensions) for both NMF and LS-NMF. Because the cell cycle data set does not have uncertainty estimates, we use a uniform multiplicative uncertainty estimate in LS-NMF (see Equations 6–8). As expected, LS-NMF does not outperform NMF here, as there is no uncertainty information. The differences between the curves merely reflects random differences in changes generated by slightly different update rules. To verify that LS-NMF behaves like NMF in the limit, we assigned unit uncertainty to all data points, which should reduce LS-NMF to NMF, as Equations 7–8 reduce to Equations 2–3. We analyzed the cell cycle data using six dimensions in agreement with our earlier work [9] and compared the performance of LS-NMF and NMF to shrinkage-based hierarchical clustering [28]. As can be seen in Figure 3b, both NMF and LS-NMF clearly perform much better at the recovery of known biological coregulation groups (note that the curves for NMF and LS-NMF lie on top of each other). We used the Rosetta compendium [25] to explore the ability of LS-NMF to gain power from uncertainty data, with the gold standard supplied by genes that together provide a metabolic pathway (see Table 1). Figure 3c shows the performance of NMF and LS-NMF across different values of dimensionality *K* in terms of area under the ROC curve, as well as results for K-means clustering applied to scaled data and for NMF applied to scaled data. The scaled data set was generated by dividing each data point in **D** by its associated uncertainty as in the original publication [25]. For all dimensionalities, LS-NMF improves the recovery of coregulated genes by 15% over NMF and K-means. Scaling the data has an effect similar to LS-NMF (see update equations below), however this leads to problems in interpretation of the metagenes, since the amplitude of a gene in a metagene will be reduced to near zero for a gene with strongly scaled data (see Equation 1).

Prediction of functional relationships

The other way to evaluate the performance of NMF and LS-NMF is by testing their ability to predict functional relationships between genes. In the reduced Rosetta data set, 215 conditions are measurements of mRNA levels in deletion mutants of *S. cerevisiae* growing in rich media compared with mRNA levels of wildtype *S. cerevisiae* in similar conditions. Mutants showing similar changes in gene expression might be expected to have deletions of functionally related genes, allowing predictions of functional relationships between genes based on links between deletion mutants. These predictions were scored against available database information. In order to demonstrate the gain from including uncertainty measurements, predictions based on **P** matrices in Equation 1 from NMF, LS-NMF, and NMF on data scaled by the uncertainty estimates were compared at different dimensions. The dimensions chosen match previous work using estimation by Bayesian Decomposition and ClutrFree [17], where 15 dimensions were estimated to cover the data, and NMF [13], where 50 dimensions were estimated. In addition, correlations in the original data space (**D**) were also calculated as a baseline providing estimates of the ability of the data to predict functional relationships independent of any dimensionality reduction.

Predictions of functional relationships were made using pairwise Pearson correlations between experiments measured in each of the seven spaces (the original data space, and the 15 and 50 dimensional spaces with NMF, LS-NMF, and NMF on scaled data). In all spaces, only the 215 deletion mutant conditions were used for analysis. Predictions were checked against the MIPS database (see Methods), and the results are shown in Figure 4. For each of the methods, Figure 4 shows the percentage of predictions validated by MIPS as a function of the number of predictions made, which increases as the threshold for correlation is lowered. In general, the methods should exhibit the highest validation for their strongest predictions (i.e., highest thresholds, far left in Figure 4), but predictions based on the 15-dimensional NMF did not show such a trend, while 50-dimensional NMF and all LS-NMF did behave as expected. NMF applied to the scaled data and LS-NMF performed similarly at 50 dimensions, but LS-NMF performed better at 15 dimensions. Both produced better results than NMF applied to unsealed data and the unreduced original data. The better consistency of LS-NMF across changes in dimension may indicate that including uncertainty information into NMF updating rules improves the robustness of the algorithm. Note that for all methods except NMF applied to unsealed data, the methods perform at roughly the same level when only the most reliable predictions are considered (left side of figures).

Discussion

In the last several years, many analytical approaches have been used to identify groups of genes related by their similar expression profiles across different conditions, including time series, tumor samples, or different tissues. Since evolution has led to the borrowing of genes for use in multiple biological functions, the ability of NMF to estimate an expression profile as a linear combination of metagenes gains power by matching biological behavior. This power is demonstrated by the analysis of the yeast cell cycle data, where the ROC analysis shows that NMF is more powerful than hierarchical clustering at recovering coexpression groups. Nevertheless, as shown by the 63 replicated controls in the Rosetta compendium, the mRNA levels of individual genes are not equally well controlled in biological systems, leading to potentially large differences in the variance of mRNA levels between different genes. The constantly improving quality of microarrays and the ability to replicate conditions, either through repeated experimentation in model systems or through the capture of multiple related samples, provides estimates of this gene and condition specific variance. By using this valuable information in NMF update rules, the least squares non-negative matrix factorization (LS-NMF) algorithm improves the ability of this approach to recover biological knowledge as demonstrated by the analysis of the Rosetta compendium. Here, unlike in the yeast cell cycle data, individual uncertainty estimates are available at each data point. The value of this information is demonstrated in Figure 3c, where LS-NMF outperformed NMF in ROC analysis, and in Figure 4, where at both dimensionalities LS-NMF greatly increased the number of successfully recovered functional relationships.

LS-NMF may also be more stable than NMF in interpreting biological functions of genes based on the metagenes when the dimensionality is poorly estimated. In Figure 4, analysis at 15 and 50 dimensions give similar results in gene function prediction for LS-NMF, but not for NMF, where analysis at 15 dimensions failed to give meaningful results. This may result from the lack of uncertainty information in NMF, which makes each data point equally important in feedback to the update rules. A higher

Table 1: Coregulation Groups for Rosetta ROC Analysis. This table provides groups of genes that are believed to be coregulated based on the metabolic pathways of *S. cerevisiae* as summarized in the KEGG database.

Coregulation Groups		
Group	KEGG Pathway	Genes
1	glucose fermentation	ADH5 ALD5 ADH4 ADH2 ADH1 ALD4
2	phenylalanine degradation	ADH5 ARO10 ADH4 ARO9 ADH2 ADH1
3	sulfate assimilation pathway II	MET10 MET3 ECM17 MET14 MET17 MET16
4	gluconeogenesis	TDH2 PCK1 YMR323W ERR1 ERR2
5	serine-isocitrate lyase pathway	CIT2 ACO1 YMR323W ERR1 ERR2
6	TCA cycle, aerobic respiration	CIT2 KGD2 ACO1 IDH1 IDH2
7	tryptophan degradation	ADH5 ARO10 ADH4 ADH2 ADH1
8	glycolysis	TDH2 YMR323W ERR1 ERR2
9	histidine biosynthesis	HIS7 HIS4 HIS5 HIS3
10	leucine biosynthesis	LEU2 LEU1 BAT1 LEU4
11	tryptophan degradation via kynurenine	BNA4 BNA1 BNA2 BNA5

dimensionality could then yield a better factorization, while dimensionalities under some threshold would be highly influenced by the noise from mRNA levels with high variance. In LS-NMF, data points with low vairance will always influence the update rules more strongly.

LS-NMF gains its power from inclusion of uncertainty information. Such information can also be added by scaling the data by the uncertainty estimate, as was done in the original Rosetta study [25]. As can be seen in Figures 3 and 4, this gives similar though not identical results. However, the direct inclusion of uncertainty information provides both improved interpretation (15 dimensions in Figure 4) and more flexibility. The direct use of uncertainty information in LS-NMF allows extensions, such as treatment of individual data points separately based on additional information during updating (e.g., *a priori* biological knowledge linking genes), which scaling cannot include. In addition, the approach allows for straightforward addition of methods such as simulated annealing [29], which may be useful in escaping local maxima in the probability distribution.

Conclusion

We have implemented a new algorithm, LS-NMF, based on NMF, to analyze microarray data. The incorporation of uncertainty information into the analysis of mRNA transcript levels significantly improves the recovery of biological information in the form of functional links between genes. In cases where there is no variance information available, LS-NMF reduces back to NMF. LS-NMF will provide the community with a powerful new tool for analysis of high-throughput data. The implementation is straightforward, so analysis of new data types with similar variance estimates should be possible, such as mass spectrometry data. The source code and documentation is available from http://bioinformatics.fccc.edu/ by following the Open-Source link.

Methods

The LS-NMF algorithm

LS-NMF, like NMF, operates on preprocessed data from a set of expression array experiments. The data comprises estimates of mRNA transcript levels (single channel) or ratios (two channel) represented as a single matrix **D**. Each row of **D** contains the mRNA estimates for each gene in all conditions (e.g., distinct tissues, experiments, timepoints), and each column corresponds to the estimates of mRNA levels for all genes in a single condition. For a dataset comprising I genes with expression measured in J conditions, the dimensionality of matrix **D** would be $I \times J$. The goal of the NMF simulation is to find a small number of metagenes (the number of metagenes provides a dimensionality estimate), each defined as a positive linear combination of I genes. The mRNA level estimates across conditions for each gene can be approximated then as a positive linear combination of these metagenes. Mathematically, this can be expressed as an approximate factorization of matrix **D** into a pair of matrixes **A** and **P** as in Equation 1. The mock data, **M**, is the approximation of **D**, based on our estimates of **A** and **P**. The matrix ε provides for the error in the measurements in **D**. For K metagenes (i.e., K dimensions), matrix **A** is of size $I \times K$ with each of the K columns defining a metagene. The value of element A_{ik} indicates how strongly gene i is associated with metagene k. Matrix **P** is then of size $K \times J$, with each row representing the relative mRNA levels of a metagene across the conditions. The value of element P_{kj} givens the strength of metagene k in condition j.

For NMF simulation, random matrices **A** and **P** are initialized according to some scheme. For instance, they could be populated from a uniform distribution U [0,1]. The

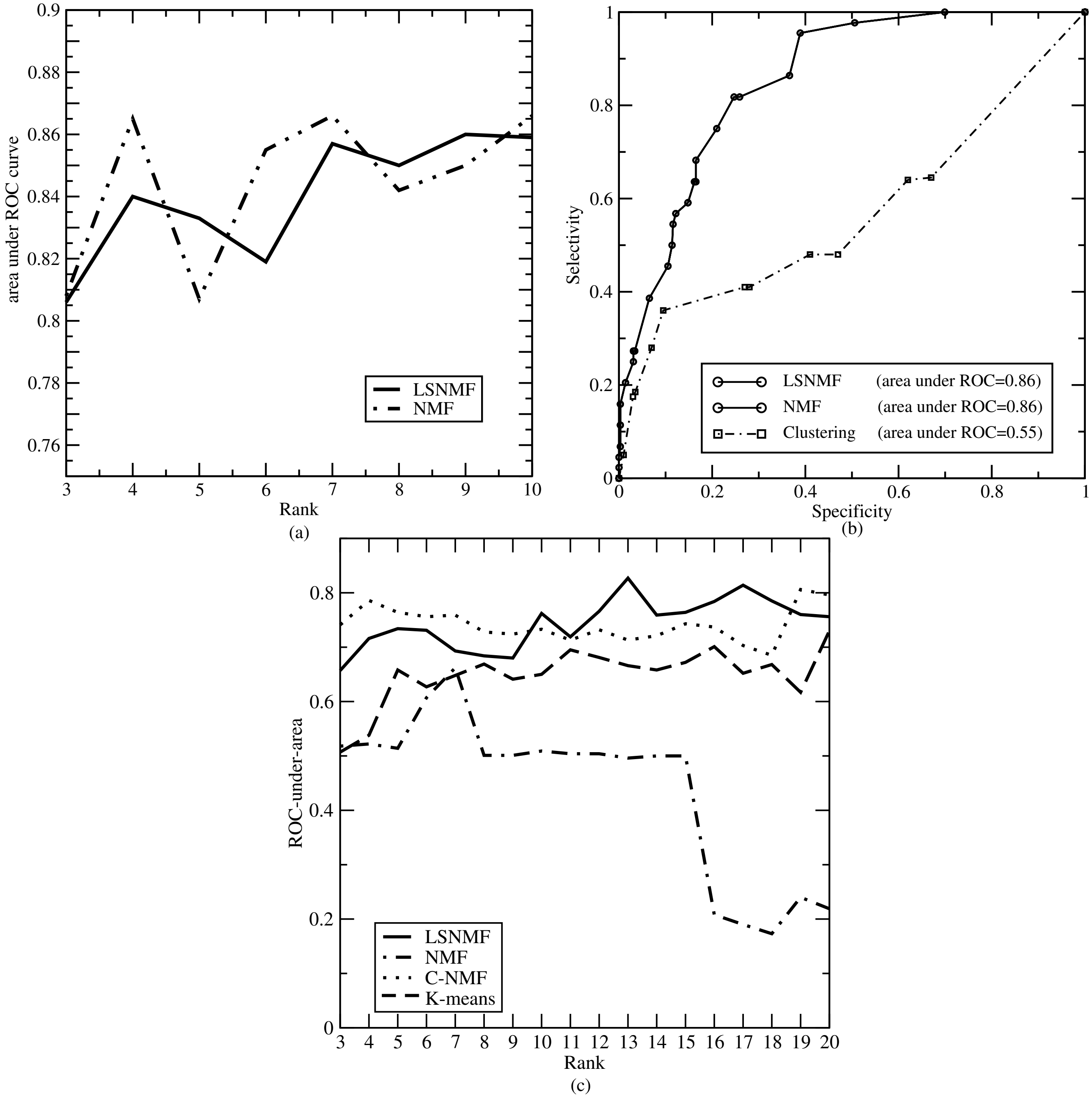

Figure 3

ROC Analysis of Yeast Cell Cycle and Rosetta Data Sets. ROC analysis for NMF and LS-NMF for the yeast cell cycle data is summarized in sections a and b, while section c shows the results for the Rosetta Compendium analysis, where variance estimates are available. In a, the total area under the ROC curves is shown for LS-NMF (solid line) and NMF (dashed line). In b, LS-NMF and NMF are compared to hierarchical clustering at a single dimensionality (the LS-NMF and NMF curves are superimposed as there is no difference). In c, the areas under the ROC curve for analysis of the Rosetta data are shown for LS-NMF (solid line), NMF (dash-dot line), K-means clustering (dashed line), and NMF on scaled data (C-NMF, dotted line). Here the variance information allows far better results to be obtained, either by scaling or by use of LS-NMF. The advantages of the approach used in LS-NMF are discussed in the text.

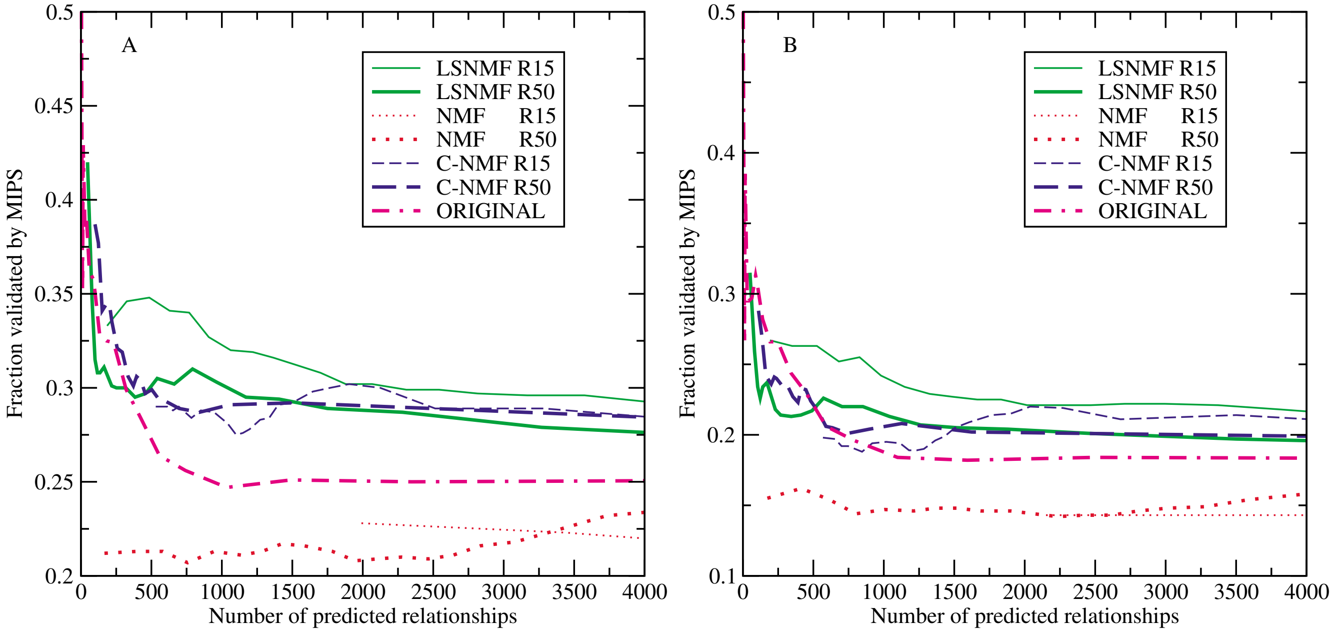

Figure 4

Comparisons of Metagenes and MIPS Functional Classes. Evaluation of metagenes for LS-NMF, NMF, and NMF on scaled data, as well as expression profiles for unreduced data, was done by determining the ability of each approach to find functionally related genes based on the MIPS functional classification. In a, the fraction of correct classifications at level 2 in the MIPS ontology are shown in terms of increasing numbers of gene pairs (i.e., decreasing threshold for correlation). In b, the same information is shown for level 3 in the MIPS ontology. The performance of NMF can be improved by utilizing a divergence based update rule, however performance is still significantly below LS-NMF, with recovery ranging from 12 to 30 times greater with LS-NMF.

two matrices are then iteratively updated using the rules [13,30],

$$P_{\alpha\mu} \leftarrow P_{\alpha\mu} \frac{\sum_i A_{i\alpha} D_{i\mu}}{\sum_i A_{i\alpha} M_{i\mu}} \qquad (2)$$

$$A_{\delta\alpha} \leftarrow A_{\delta\alpha} \frac{\sum_j D_{\delta j} P_{\alpha j}}{\sum_j M_{\delta j} P_{\alpha j}} \qquad (3)$$

$$M_{ij} = \sum_k A_{ik} P_{kj} \qquad (4)$$

which guarantees reaching a local maximum in Likelihood and minimizes

$$\|D - M\|^2 = \sum_{ij} \left(D_{ij} - M_{ij} \right)^2. \qquad (5)$$

In the original NMF algorithm, elements of **A** and **P** were updated at a pace determined by the difference between **M** and **D**. For a discussion of some approaches to updating **A** and **P**, see [31]. Since no uncertainty information enters the updating rules, every matrix element involved in the updating is weighted equally. In order to take advantage of uncertainty information, we minimize the χ^2 error,

$$x^2 = \sum_{ij} \left(\frac{\left(D_{ij} - M_{ij} \right)}{\sigma_{ij}} \right)^2 \qquad (6)$$

instead of the distance between **D** and **M**. The update rules can easily incorporate this change with

$$P_{\alpha\mu} \leftarrow P_{\alpha\mu} \frac{\sum_i A_{i\alpha} \dfrac{D_{i\mu}}{\sigma_{i\mu}}}{\sum_i A_{i\alpha} \dfrac{M_{i\mu}}{\sigma_{i\mu}}} \qquad (7)$$

$$A_{\delta\alpha} \leftarrow A_{\delta\alpha} \frac{\sum_j \dfrac{D_{\delta j}}{\sigma_{\delta j}} P_{\alpha j}}{\sum_j \dfrac{M_{\delta j}}{\sigma_{\delta j}} P_{\alpha j}} \qquad (8)$$

$$M_{ij} = \sum_k A_{ik} P_{kj} \qquad (9)$$

where σ_{ij} is the uncertainty measurement for D_{ij}. This requires the existence of a new matrix, **U**, that provides estimates of the uncertainties for all data points. One advantage of this approach is that missing data is easily handled by assigning a value of 0 (for single channel) or 1.0 (for two channel, i.e. ratio) together with a large uncertainty. This essentially allows the algorithm to ignore these data points when fitting the model.

It is straightforward to follow the procedure described by Lee and Seung [30] to verify that χ^2 is nonincreasing under the modified update rules.

Implementation of NMF and LS-NMF algorithms

The NMF algorithm was obtained from the Broad Institute as a Matlab script [15]. We converted this to a C++ version allowing modification, and we validated the C++ version by making sure we obtained the same results as for the Matlab version in a number of simulations. For LS-NMF, Equations 7–9 were implemented within C++. The code included modified update rules that take into account the additional U matrix. The code is designed for use on Beowulf clusters running Linux, and source code is available under the GNU Lesser Public License from the Fox Chase Bioinformatics web site http://bioinformatics.fccc.edu/ by following the Open-Source Software link.

Since the update rules in NMF and LS-NMF only guarantee a local minimum, the algorithms may or may not converge to the same solution on each simulation, depending on the properties of the probability distribution. To address this limitation, any simulation must be repeated multiple times (typically 20–100 individual runs) starting with different initial **A** and **P** matrices [13,15]. After the simulation, either the best [13] or average [15] factorization is selected for further analysis. We repeated NMF and LS-NMF simulations on both the cell cycle dataset and the Rosetta dataset 20 times for each dimension, *K*, between 3 and 20. Each single simulation was run for 20,000 update steps or until it converged based on a predefined χ^2 threshold (or RMSD threshold for NMF). The factorization with the lowest χ^2 error (or RMSD error for NMF) from the 20 repeated runs for each dimension was selected for further analysis. The errors were calculated as

$$x^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{\frac{1}{\sigma_{ij}^2}\left(D_{ij} - M_{ij}\right)^2\right\} \qquad (10)$$

$$RMSD = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(D_{ij} - M_{ij}\right)^2 \qquad (11)$$

Data preprocessing

Two data sets were analyzed using LS-NMF and NMF, the yeast cell cycle data set [24] and the Rosetta compendium [25]. The cell cycle data set comprises measurements of mRNA levels using Affymetrix GeneChips. Synchronization was done using a temperature sensitive mutant of cdc28, which is required for passage into the late G 1 stage of the cell cycle. Cultures were grown following temperature change to activate cdc28, and mRNA was harvested at 10 minute intervals. The data was preprocessed by the original authors to identify genes that had cell cycle periodicity, resulting in a data set with 788 genes measured at 17 time points. The 10 minute intervals beginning with *t* = 0 on release from cell cycle arrest ended at 160 minutes, providing roughly two passes through the yeast cell cycle.

The Rosetta data comprises genome-wide measurements of gene expression across 300 deletion mutants or chemical treatments using oligonucleotide microarrays. We downloaded the data from Rosetta Inpharmatics, filtered it to remove experiments where less than 2 genes underwent 3-fold changes, and finally removed genes that did not change by 3-fold across the remaining conditions, resulting in 764 gene probes and 228 conditions. The Rosetta error model, based on replicates and 63 control replications of wildtype yeast, provided the estimation of uncertainty for each data point [25]. As the data comprised log-ratios, data transformation was used to convert these measurements to positive ratios and errors were propagated from the log space to the ratio space.

Evaluating the algorithms

The most reliable criteria to evaluate algorithms applied to microarray data is the ability to recover the knowledge of biological relationships between genes, i.e., whether the suggested metagenes summarize biological knowledge. While there are few well-established benchmarks available to test the validity of metagenes, we use two well studied data sets and biological knowledge of coregulation or functional relationships to evaluate performance. For the first data set, the cell cycle data, the coregulation groups are based on biological knowledge of the yeast cell cycle and comprise 9 overlapping groups with 43 genes [28]. For the second data set, the Rosetta compendium, accuracy of the metagenes was estimated using ROC analysis based on KEGG metabolic pathways [32] and predictions of gene relationships compared to MIPS functional classification [33].

To identify membership of genes in each metagene, a threshold must be set. To do this, each row of **P** was normalized to sum to 1, and a correction factor was applied to the corresponding column of **A** to leave **M** unchanged. In order to find the metagenes in **A**, the *Z*-score was calculated for each element in each row by

$$Z_{ik} = \frac{A_{ik} - \mu_i}{\sigma_i}, \qquad (12)$$

where μ_i is the average value for gene i in **A**, and σ_i is the standard deviation. The same data transformation was done also for **P**, but it was column-based to obtain behavior across metagenes for each condition. Then we assigned gene i as a member of metagene k, if $Z_{ik}$was greater than a threshold, T. By changing the threshold value for the Z-score, we calculated an ROC curve using the procedure outlined below. It is useful to note that each gene may be assigned to multiple metagenes, which allows identification of multiple regulation of genes.

1. **Assign genes to metagenes based on the selected threshold.** Generate a Boolean connectivity matrix **C** with dimension of $I \times K$ with $C_{ik} = 1$, iff gene i was assign to metagene k.

2. **Assign metagenes to biologically verified coregulation groups.** Metagene k is assigned to represent coregulation group G_m iff the metagene maximizes $\sum_{i \in G_m} C_{ik}$ among all possible k, i.e. metagene k is the most enriched metagene for genes in group G_m.

3. **Keep a Boolean correlation matrix R with dimension of** $K \times M$. The value M is set by the number of coregulation groups, and $R_{km} = 1$ iff metagene k was assigned to represent coregulation group G_m.

4. **Calculate the sensitivity and specificity**. The true positive, true negatives, false positives, and false negatives are given by

$$\begin{aligned} TP &= \sum_{k=1}^{K}\sum_{m=1}^{M}\{R_{km}\sum_{g\in G_m} C_{gm}\} \\ TN &= \sum_{k=1}^{K}\sum_{m=1}^{M}\{R_{km}\sum_{g\notin G_m} \left(1-C_{gm}\right)\} \\ FP &= \sum_{k=1}^{K}\sum_{m=1}^{M}\{R_{km}\sum_{g\notin G_m} C_{gm}\} \\ FN &= \sum_{k=1}^{K}\sum_{m=1}^{M}\{R_{km}\sum_{g\in G_m} \left(1-C_{gm}\right)\} \end{aligned} \qquad (13)$$

yielding sensitivity and specificity given by

$$\begin{aligned} Sens &= \frac{TP}{TP+FN} \\ Spec &= \frac{TN}{TN+FP} \end{aligned}. \qquad (14)$$

5. **Plot the ROC curve**. The ROC curve is generated by plotting multiple esitmates of the $Sens$ and $1 - Spec$ based on different thresholds T. We used the trapezoidal rule to approximate the area under ROC curve.

In order to identify TN, TP, FP, FN values for Equation 13, a gold standard is required. For figures 3a and 3b, the cell cycle coregulation groups identified previously from known transcriptional response and transcription factor binding analyses in the yeast cell cycle are used [28]. For 3c, these groups are not useful, since most of these genes play a criticial role in the cell cycle and are not included in the deletion mutant set (only 5 of 43 genes are included as deletion mutants). For this data set, we instead created coreguation groups from the KEGG metabolic pathways [32]. These groups are based on the assumption that genes encoding enzymes that together provide a metabolic pathway are likely to be coexpressed. The 11 groups comprising 63 genes are provided in Table 1.

The second way that we evaluated the performance of NMF and LS-NMF was by predicting gene relationships as done by Kim and Tidor [13]. We assumed that similarity of gene expression profiles as measured by metagenes for different deletion mutants indicated a functional relationship between the deleted genes. We calculated predictions in both the original, unreduced data space (from the **D** matrix) and in the reduced dimensional spaces computed by NMF, LS-NMF, and NMF applied to scaled data (from the **P** matrix). The scaled data was produced by taking the Rosetta data and dividing by the uncertainty estimates. The Pearson correlation coefficient was calculated between all conditions (i.e. deletion mutants), and the absolute value of the correlation coefficient was used as a score for the predicted relationship. The functional relationships between deletion mutants are used as predictors of functional relationships for the deleted genes. From the Rosetta dataset, Pearson correlation coefficients were calculated for all possible pairs of 215 columns in **P** (i.e, all possible deletion mutant combinations). The scores were compared to a threshold for multiple thresholds, and successful predictions of functional relationships were determined by comparison to the MIPS Funcat annotation [33,34]. The Funcat annotations are hierarchical, comprising high level (level 1, example 41 DEVELOPMENT) to low level (level 4, example 41.01.03.03 mycelium development) categorizations. In order to have sufficient data for the analysis and to have sufficient fine resolution of function, we focused on the second and third levels of annotation (e.g., 41.01 fungal/microorganismic development, 41.01.03 tissue pattern formation). Two genes appearing in the same MIPS functional category were considered as functionally related. For all combinations of two genes predicted as functionally related by the Pearson correlation test of their deletion mutants, the total number of successful predictions were determined. The fraction of successful predictions were then plotted as the threshold was reduced and the number of predictions increased. In cases where genes do not have a MIPS functional assignment at the desired level or were classified as "UNCLASSIFIED PROTEINS", the corresponding deletion mutants were removed from the analysis. This left 168 genes at level 3, where MIPS has 441 different classifications, and 180 genes at level 2, where MIPS has 158 different classifications.

Authors' contributions

GW performed all analyses presented in this work, developed the new update equations for LS-NMF, and coded the LS-NMF algorithm. AVK analyzed the KEGG metabolic pathways and created the genes lists presented in Table 1 by comparison to the Rosetta deletion mutant data. MFO oversaw the project and suggested the use of the χ^2 measure to improve NMF.

Acknowledgements

This work was supported by NIH through the National Library of Medicine (LM008309 to mfo) and the NCI CCSG grant (CA06927 supporting mfo), and the Pennsylvania Department of Health.

References

1. Ochs MF, Godwin AK: **Microarrays in cancer: research and applications.** *Biotechniques* 2003, **34:**S4-S15.
2. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95(25):**14863-8.
3. Lukashin AV, Fuchs R: **Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters.** *Bioinformatics* 2001, **17(5):**405-14.
4. Ben-Dor A, Shamir R, Yakhini Z: **Clustering gene expression patterns.** *J Comput Biol* 1999, **6(3–4):**281-97.
5. Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes.** *Genome Res* 1999, **9(11):**1106-15.
6. Gasch AP, Eisen MB: **Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.** *Genome Biol* 2002, **3(11):**. RESEARCH0059
7. Getz G, Levine E, Domany E: **Coupled two-way clustering analysis of gene microarray data.** *Proc Natl Acad Sci USA* 2000, **97(22):**12079-84.
8. Tanay A, Sharan R, Shamir R: **Discovering statistically significant biclusters in gene expression data.** *Bioinformatics* 2002, **18(Suppl 1):**S136-44.
9. Moloshok TD, Klevecz RR, Grant JD, Manion FJ, Speier WFt, Ochs MF: **Application of Bayesian Decomposition for analysing microarray data.** *Bioinformatics* 2002, **18(4):**566-75.
10. Ochs MF: **Bayesian Decomposition.** In *The Analysis of Gene Expression Data: Methods and Software* Edited by: Parmigiani G, Garrett E, Irizarry R, Zeger S. New York: Springer Verlag; 2003.
11. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97(18):**10101-6.
12. Lee SI, Batzoglou S: **Application of independent component analysis to microarrays.** *Genome Biol* 2003, **4(11):**R76. [1465–6914 (Electronic) Journal Article]
13. Kim PM, Tidor B: **Subsystem identification through dimensionality reduction of large-scale gene expression data.** *Genome Res* 2003, **13(7):**1706-18.
14. Zhang J, Wei L, Wang Y: **Computational decomposition of molecular signatures based on blind source separation of non-negative dependent sources with NMF.** *IEEE 13th Workshop on Neural Networks for Signal Processing, 2003* 2003:409-418.
15. Brunet JP, Tamayo P, Golub TR, Mesirov JP: **Metagenes and molecular pattern discovery using matrix factorization.** *Proc Natl Acad Sci USA* 2004, **101(12):**4164-9.
16. Sibisi S, Skilling J: **Prior distributions on measure space.** *Journal of the Royal Statistical Society, B* 1997, **59:**217-235.
17. Bidaut G, Ochs MF: **ClutrFree: cluster tree visualization and interpretation.** *Bioinformatics* 2004, **20(16):**2869-71.
18. 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.
19. 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:**37-52.
20. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98(9):**5116-21.
21. Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ, Churchill GA: **Statistical analysis of a gene expression microarray experiment with replication.** *Statistica Sinica* 2002, **12:**203-218.
22. 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. [1367–4803 (Print) Evaluation Studies Journal Article]
23. Pochet N, De Smet F, Suykens JA, De Moor BL: **Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction.** *Bioinformatics* 2004, **20(17):**3185-95. [1367–4803 (Print) Evaluation Studies Journal Article Validation Studies]
24. 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:**65-73.
25. 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:**109-26.
26. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JJA, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc Nail Acad Sci USA* 2001, **98(20):**11462-7.
27. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D'Amico M, Pestell RG, West M, Nevins JR: **Gene expression phenotypic models that predict the activity of oncogenic pathways.** *Nat Genet* 2003, **34(2):**226-30.
28. Cherepinsky V, Feng J, Rejali M, Mishra B: **Shrinkage-based similarity metric for cluster analysis of microarry data.** *Proc Natl Acad Sci USA* 2003, **100(17):**9668-73.
29. Geman S, Geman D: **Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1984, **PAMI-6(6):**721-741.
30. Lee DD, Seung HS: **Learning the parts of objects by non-negative matrix factorization.** *Nature* 1999, **401(6755):**788-91.
31. Lee DD, Seung HS (Eds): *Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems 13* 2001.
32. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet.** *Nucleic Acids Res* 2002, **30:**42-6.
33. 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: **CYGD: the Comprehensive Yeast Genome Database.** *Nucleic Acids Res* 2005:D364-8.
34. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A: **MIPS: analysis and annotation of proteins from whole genomes.** *Nucleic Acids Res* 2004:D41-4.