Combining Gene Expression and Transcription Factor Regulation Data using Simultaneous Nonnegative Matrix Factorization

Liviu Badea
AI group
National Institute for Research and Development in Informatics
Bucharest, Romania

Abstract - *Applied to microarray data, Nonnegative Matrix Factorization (NMF) can be viewed as a generalized clustering algorithm allowing for gene overlaps – an important feature in this domain where genes can be involved in several biological processes. In this paper we present siNMF, a generalization of NMF that can simultaneously factorize a gene expression matrix and a matrix of transcription regulatory influences. Thus, siNMF constructs gene clusters taking into account not just expression information, but also background knowledge on potential regulatory factors of the clusters. A preliminary application of the algorithm to a real-life pancreatic cancer dataset shows the feasibility of our approach.*

Keywords: Gene expression data / microarray analysis, clustering with transcription factor regulatory information.

1 Introduction and motivation

The introduction of high throughput technologies for measuring gene expression, such as microarrays, has allowed a revolutionary transition from the exploration of the expression of a handful of genes to that of entire genomes. However, despite its enormous potential, microarray data has proved difficult to analyze, partly due to the significant amount of noise, but also due to the large number of factors that influence gene expression (many of which are *not* at the mRNA/transcriptome level) as well as the complexity of their interactions.

One of the most successful microarray data analysis methods has proved to be *clustering* (of genes and/or samples), and a large variety of such methods have been proposed and applied to real-life biological data. This large body of work, impossible to extensively review here, has emphasized important limitations of existing clustering algorithms:

(1) Most clustering methods produce *non-overlapping* clusters. However, since genes are typically involved in several biological processes, "non-overlapping" clustering methods, such as hierarchical clustering (HC) [2], self-organizing maps (SOM) [9], k-means clustering, etc., tend to be unstable, producing different gene clusters for only slightly different input samples (e.g. in the case of HC), or depending on the choice of initial conditions (as in the case of SOM, or k-means). Algorithms allowing for overlapping clusters, such as fuzzy k-means [4] achieved significant improvements w.r.t. "non-overlapping" clustering, but they still have the problems discussed below.

(2) Most algorithms perform clustering along a single dimension comparing e.g. genes w.r.t. *all* the available samples, whereas in reality genes have coordinated expression levels only for certain subsets of conditions. Algorithms dealing with this problem, such as biclustering [10], coupled-two way clustering (CTWC) [3], ISA (iterative signature) [1] have other problems mostly related to the control of overlap between biclusters.

(3) Although genes are subject to both positive and negative influences from other genes, the *robustness* of biological systems requires that an observed change in the expression level of a given gene is the result of *either* a positive *or* a negative influence rather than a complex combination of positive and negative influences that partly cancel out each other (as in the case of Principal Component Analysis).

Nonnegative Matrix Factorization (NMF) [7] deals with this problem by searching for *nonnegative* decompositions of (nonnegative) data. The observed localized nature of the decompositions seems to be a byproduct of the nonnegativity constraints [7].

Recently, Brunet at al [5] applied NMF for clustering *samples* in a *non-overlapping* mode for three cancer datasets. On the other hand, Kim and Tidor [6] used NMF to cluster *genes* in the context of a large dataset of yeast perturbation experiments (spotted arrays). Although NMF has the tendency of producing sparse representations, the factorizations obtained were subjected to thresholding and subsequent reoptimization to obtain sufficiently sparse clusters.

Unfortunately however, microarray data are noisy and it might be useful to be able to take into account any background knowledge that may be available. For example, data about *transcription factor binding*[1] or curated databases about transcription regulation could be used not only to *interpret* the resulting clusters (in terms of the potential regulatory modules that may be driving the coordinated expression of the cluster genes), but also to *guide* the clustering process itself.

In this paper we show how Nonnegative Matrix Factorization (NMF) can be generalized to take into account data about transcription regulation when constructing clusters. Our method is called "*simultaneous Nonnegative Matrix Factorization*" (siNMF) since it factorizes gene expression and transcription regulation data simultaneously.

Using the Transcription Regulatory Element Database TRED as background knowledge, we applied our factorization algorithm to a pancreatic cancer dataset and show that it was able not only to correctly separate the tumor samples from the normal ones (without being provided with class information), but also to recover the associated regulatory factors, which may be driving the expression of the genes responsible for this disease.

2 Simultaneous Nonnegative Matrix Factorization (siNMF)

Given a gene expression matrix X_{sg} (the index s denotes samples, while g stands for genes) and a transcription factor regulatory matrix B_{fg} (which is 1 whenever transcription factor f regulates gene g), *siNMF* simultaneously factorizes the two matrices as follows:

$$X_{sg} \approx \sum_c A_{sc} \cdot S_{cg} \qquad (1)$$

$$B_{fg} \approx \sum_c C_{fc} \cdot S_{cg} \qquad (2)$$

with the additional nonnegativity constraints:

$$A_{sc} \geq 0,\ \ S_{cg} \geq 0,\ \ \mathrm{C}_{fc} \geq 0. \qquad (3)$$

where X_{sg} is the expression level of gene g in data sample s, A_{sc} the expression level of the biological process (cluster) c in sample s, S_{cg} the membership degree of gene g in c and C_{fc} the involvement of transcription factor f in the regulation of cluster c.

Note that the gene cluster membership matrix S is common to the two factorizations, as it is influenced both by the gene expression data X, as well as by the regulatory data B. The nonnegativity constraints (3) express the obvious fact that expression levels, membership degrees and regulatory factor involvements cannot be negative.

More formally, the factorization (1-3) can be cast as a constrained optimization problem:

$$\min C(A,S) = \frac{1}{2} \| X - A \cdot S \|_F^2 + \frac{\beta}{2} \| B - C \cdot S \|_F^2 \qquad (4)$$

subject to the nonnegativity constraints (3) ($\| \cdot \|_F$ is the Frobenius norm of a matrix).

The weight β ensures a proper balance between the two error terms and was taken in the following experiments to be $\beta = \beta_0 \frac{\| X \|}{\| B \|}$ with β_0=1.

The optimization problem (4) can be solved using *multiplicative update rules* in a manner similar to Lee and Seung's seminal *Nonnegative Matrix Factorization* (*NMF*) algorithm [8] (ε is a small regularization parameter):

```
siNMF(X, A_0, S_0) → (A,S)
$A \leftarrow A_0,\ S \leftarrow S_0,\ C \leftarrow C_0$ (typically $A_0, S_0, C_0$ are initialized randomly)
loop
    $S_{cg} \leftarrow S_{cg} \frac{(A^T \cdot X + \beta C^T \cdot B)_{cg}}{((A^T \cdot A + \beta C^T \cdot C) \cdot S)_{cg} + \varepsilon}$
    $A_{sc} \leftarrow A_{sc} \frac{(X \cdot S^T)_{sc}}{(A \cdot S \cdot S^T)_{sc} + \varepsilon}$
    $C_{fc} \leftarrow C_{fc} \frac{(B \cdot S^T)_{fc}}{(C \cdot S \cdot S^T)_{fc} + \varepsilon}$
until convergence
normalize the rows of S to unit norm by taking advantage of the scaling invariance of the factorization: $S \leftarrow D^{-1} \cdot S$, $A \leftarrow A \cdot D$, $C \leftarrow C \cdot D$, where $D = diag\left(\sqrt{\sum_g S_{cg}^2}\right)$.
```

Note that such a factorization can be viewed as a "soft" clustering algorithm allowing for *overlapping gene clusters*, since we may have several significant S_{cg} entries on a given column g of S (so a gene g may "belong" to several clusters c). The final normalization of the rows of S renders the resulting clusters comparable to each other.

The algorithm above constructs the gene clusters S guided by both gene expression data and regulatory data B. If we would not want the construction of the gene clusters to be influenced by the regulator binding data B, we would have to use the following update rule for S:

$$S_{cg} \leftarrow S_{cg} \frac{(A^T \cdot X)_{cg}}{(A^T \cdot A \cdot S)_{cg} + \varepsilon}.$$

In the following, we assume that the gene expression data is given in an $n_s \times n_g$ matrix X, where n_s and n_g are the numbers of samples and genes respectively, so that X_{sg} represents the expression level of gene g in sample s.

[1] For example, location analysis data resulting from *ChIP* on chip experiments (*chromatin immunoprecipitation*).

A sparse factorization $X \approx A{\cdot}S,\ B \approx C{\cdot}S$ will be interpreted as a *simultaneous clustering of the genes* into overlapping clusters c corresponding to the rows S_{cg} of S

- w.r.t. the expression data X, and
- w.r.t. the transcription regulation data B, resulting in gene regulatory programs C_{fc} that may explain the concerted expression of the genes in the cluster c.

Note that clusters and regulatory programs can be overlapping, since the columns of S and C may have several significant entries. However, although overlaps are allowed, the algorithm will not produce highly overlapping clusters, due to the sparseness constraints. This is unlike many other clustering algorithms that allow clusters to overlap, which have to resort to several parameters to keep excessive cluster overlap under control.

3 Extracting the main regulatory programs in pancreatic cancer

Although the main goal of this paper is the presentation of a new clustering algorithm able to deal with transcription regulatory data as background knowledge rather than obtaining new biological insights, we also briefly discuss our initial attempts at applying our algorithm to a pancreatic cancer microarray dataset.

Despite the enormous recent progress in understanding cancer at a molecular level, the precise details are still elusive for many types of carcinomas. Pancreatic cancer is a particularly aggressive disease, with a very poor prognosis, requiring a more precise understanding of its molecular pathogenesis. The technological progress initiated by the introduction of gene expression microarrays about a decade ago has enabled large scale whole genome studies with the aim of identifying disease-specific genes. Unfortunately, the gene sets obtained as a result of clustering or differential expression analysis are hard to interpret, and the transcription regulatory programs controlling them are difficult to determine.

In the following, we describe the application of our simultaneous factorization algorithm to a pancreatic ductal adenocarcinoma (PDAC) dataset produced in the framework of the GENOPACT project.

The dataset contains microarray measurements (using Affymetrix U133 Plus 2.0 whole genome chips) for 38 pairs of PDAC and respectively normal samples (76 samples in total). After filtering out the probe-sets (genes) with relatively low expression as well as those with a nearly constant expression value^{2}, we were left with 12209 probe-sets. The gene expression matrix was subsequently logarithmized (since typical gene expression values are log-normally distributed) and given to the *siNMF* algorithm together with transcription regulatory data retrieved from the *Transcription Regulatory Element Database TRED* [11]. (TRED is expert curated and thus highly reliable. However, its main limitation is the relatively low coverage of transcription factor regulation, as it includes data on only about 154 factors (including families) and 2969 target genes.)

An important parameter of the factorization is its *internal dimensionality* (the number of clusters n_c). To avoid overfitting, we estimated the number of clusters n_c as the number of dimensions around which the change in relative error $d\varepsilon\ /dn_c$ of the factorization of the real data "*reaches from above*" the change in relative error obtained for a randomized dataset (similar to [6]) – see Figure 1. This analysis estimated the internal dimensionality of the dataset to be around n_c=5.

Running *siNMF* with β_0=1 and n_c=5 on the complete set of 12209 produced the sample clusters depicted in Figure 2. Note that cluster 2 recovers relatively well^{3} the distinction between normal and tumor samples genes, although the algorithm was never provided with class information related to the samples. The associated regulatory programs C are shown in the Annex.

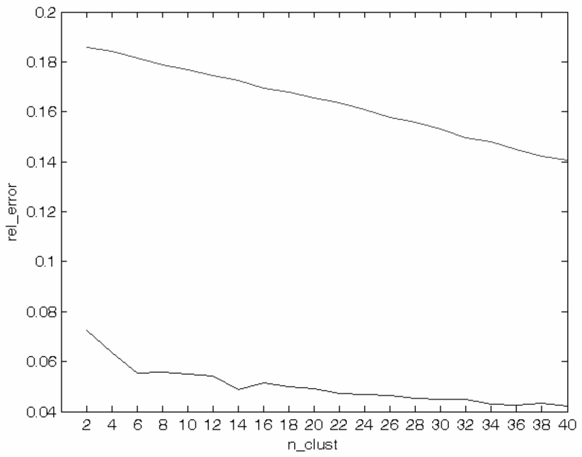


Figure 1. Determining the internal dimensionality of the pancreatic cancer dataset

The largest regulatory program also belongs to cluster 2. The top transcription factors presumably controlling this cluster are given in the Table below. Many of these transcription regulators have been mentioned in the literature in relation with pancreatic cancer. For example, SP1 is known to be involved in pancreatic cancer [12], as is the critical tumor suppressor p53 [13]. JUN, a component of the AP1 transcription complex is a well-known oncogene and has been frequently linked to pancreatic cancer (e.g. [14]).

[2] Only genes with an average expression value over 100 and with a standard deviation above 50 were retained.

[3] Certain "normal" samples (such as N30308 and N40726) which in our analysis are "closer" to the tumor samples than to the other normal ones were later reanalyzed histologically and found to be highly fibrotic (pancreatic tumor tissue is typically very fibrotic and the respective normal samples were possibly collected from a site too close to the tumoral tissue).

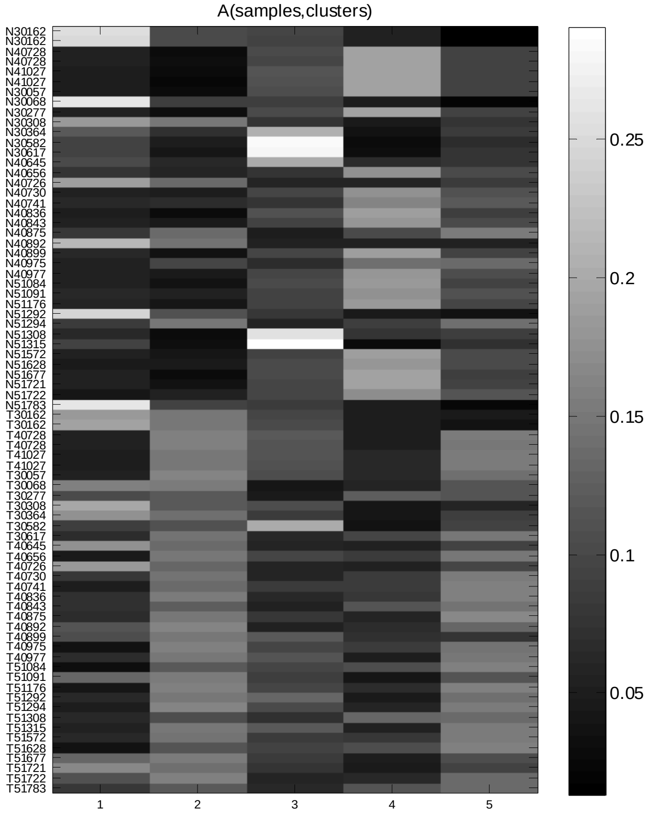
Transcription factor	Coefficient in C
SP	4.95463
SP1	4.442609
AP2	4.075412
TFAP2A	4.060308
p53	3.8224
TP53	3.797026
NFKB	3.39351
AP1	3.168997
JUN	3.168997
NFKB1	3.098774
AR	2.537124
CREB	2.4364
CREB1	2.411284
ETS1	1.842133
EGR1	1.74613
CEBPA	1.707971
CEBP	1.681746
MYB	1.548494
ER	1.542522
RAR	1.500031
NFI	1.458183
RELA	1.448307
NFIC	1.441207
RARA	1.435847
SMAD	1.373663
STAT	1.273818
SP3	1.214316
HIF	1.192895
SPI1	1.180106
OCT	1.177372
USF1	1.127539
HIF1A	1.078189
PPAR	1.052136
ATF1	1.048764
POU2F1	1.034346
ESR1	0.995088
STAT1	0.97143
STAT3	0.90882

4 Conclusions

Although widely used in microarray data analysis, existing clustering algorithms have serious problems, the most important one being related to the fact that biological processes are overlapping rather than isolated. Microarray measurements are also very noisy, which can only be compensated by additional background knowledge, such as regulatory influences.

In this paper we have introduced a clustering algorithm capable of taking into account not just expression information, but also background knowledge on potential transcription regulatory factors of the clusters. A key ingredient of this algorithm is the nonnegativity constraint, which ensures the sparseness of the factorizations. Our preliminary results on a real-life pancreatic cancer dataset are encouraging, but a more detailed biological analysis will be the focus of subsequent work. (Larger versions of the Figures can also be found at http://www.ai.ici.ro/biocomp07/.)

Figure 2. The sample clusters (matrix A) in the pancreatic cancer dataset



Acknowledgements

This work was partially supported by the BIOINFO and GENOPACT projects (CEEX 32/2005, 56/2005). I am particularly grateful to Prof. Irinel Popescu for a wonderful collaboration in these projects.

5 References

[1] Bergmann S., J. Ihmels, N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review E 67, 031902 (2003).

[2] Eisen M.B., P.T. Spellman, P.O. Brown, D. Botstein. Cluster analysis and display of genome-wide expression patterns, PNAS Vol.95, 14863-8, Dec.1998.

[3] Getz G., Levine E., Domany E. Coupled two-way clustering analysis of gene microarray data, PNAS Vol. 97, 12079-84, 2000.

[4] Gasch A.P, Eisen M.B. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biol. 2002, Oct 10;3(11).

[5] Brunet J.P., Tamayo P., Golub T.R., Mesirov J.P. Metagenes and molecular pattern discovery using matrix factorization. PNAS 101(12):4164-9, 2004, Mar 23.

[6] Kim P.M., Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. 2003 Jul;13(7):1706-18.

[7] Lee D.D., H.S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, vol. 401, no. 6755, pp. 788-791, 1999.

[8] Lee D.D., H.S. Seung. Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing 13 (Proc. NIPS*2000), MIT Press, 2001.

[9] Tamayo P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, PNAS 96(6):2907-12 (1999).

[10] Cheng Y, Church GM. Biclustering of expression data. Proc. ISMB 2000; 8:93-103.

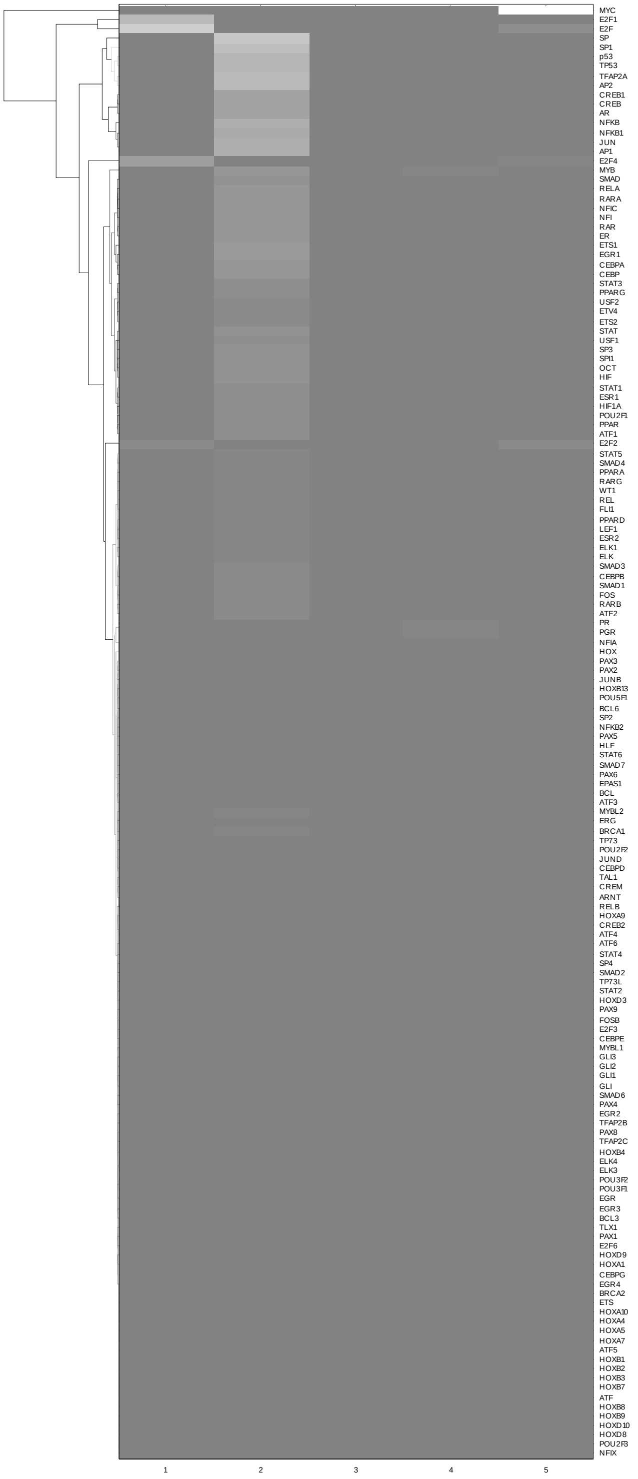
[11] TRED. Transcriptional Regulatory Element Database. http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home.

[12] Jungert K, Buck A, Wichert G, Adler G, Konig A, Buchholz M, Gress TM, Ellenrieder. Sp1 Is Required for Transforming Growth Factor-beta-Induced Mesenchymal Transition and Migration in Pancreatic Cancer Cells. Cancer Res. 2007 Feb 15;67(4):1563-70.

[13] Dong M, Dong Q, Zhang H, Zhou J, Tian Y, Dong Y. Expression of Gadd45a and p53 proteins in human pancreatic cancer: Potential effects on clinical outcomes. J Surg Oncol. 2007 Jan 17.

[14] Okutomi Y, Shino Y, Komoda F, Hirano T, Ishihara T, Yamaguchi T, Saisho H, Shirasawa H. Survival regulation in pancreatic cancer cells by c-Jun. Int J Oncol. 2003 Oct;23(4):1127-34.

Annex. The regulatory programs controlling the clusters (matrix C)


MYC
E2F1
E2F
SP
SP1
p53
TP53
TFAP2A
AP2
CREB1
CREB
AR
NFKB
NFKB1
JUN
AP1
E2F4
MYB
SMAD
RELA
RARA
NFIC
NFI
RAR
ER
ETS1
EGR1
CEBPA
CEBP
STAT3
PPARG
USF2
ETV4
ETS2
STAT
USF1
SP3
SPI1
OCT
HIF
STAT1
ESR1
HIF1A
POU2F1
PPAR
ATF1
E2F2
STAT5
SMAD4
PPARA
RARG
WT1
REL
FLI1
PPARD
LEF1
ESR2
ELK1
ELK
SMAD3
CEBPB
SMAD1
FOS
RARB
ATF2
PR
PGR
NFIA
HOX
PAX3
PAX2
JUNB
HOXB13
POU5F1
BCL6
SP2
NFKB2
PAX5
HLF
STAT6
SMAD7
PAX6
EPAS1
BCL
ATF3
MYBL2
ERG
BRCA1
TP73
POU2F2
JUND
CEBPD
TAL1
CREM
ARNT
RELB
HOXA9
CREB2
ATF4
ATF6
STAT4
SP4
SMAD2
TP73L
STAT2
HOXD3
PAX9
FOSB
E2F3
CEBPE
MYBL1
GLI3
GLI2
GLI1
GLI
SMAD6
PAX4
EGR2
TFAP2B
PAX8
TFAP2C
HOXB4
ELK4
ELK3
POU3F2
POU3F1
EGR
EGR3
BCL3
TLX1
PAX1
E2F6
HOXD9
HOXA1
CEBPG
EGR4
BRCA2
ETS
HOXA10
HOXA4
HOXA5
HOXA7
ATF5
HOXB1
HOXB2
HOXB3
HOXB7
ATF
HOXB8
HOXB9
HOXD10
HOXD8
POU2F3
NFIX