Pacific Symposium on Biocomputing 13:279-290(2008)

EXTRACTING GENE EXPRESSION PROFILES COMMON TO COLON AND PANCREATIC ADENOCARCINOMA USING SIMULTANEOUS NONNEGATIVE MATRIX FACTORIZATION

LIVIU BADEA
AI Lab, National Institute for Research and Development in Informatics
8-10 Averescu Blvd., Bucharest, Romania, badea@ici.ro

In this paper we introduce a clustering algorithm capable of simultaneously factorizing two distinct gene expression datasets with the aim of uncovering gene regulatory programs that are common to the two phenotypes. The *siNMF* algorithm simultaneously searches for two factorizations that share the same gene expression profiles. The two key ingredients of this algorithm are the nonnegativity constraint and the offset variables, which together ensure the sparseness of the factorizations.

While cancer is a very heterogeneous disease, there is overwhelming recent evidence that the differences between cancer subtypes implicate entire pathways and biological processes involving large numbers of genes, rather than changes in single genes. We have applied our simultaneous factorization algorithm looking for gene expression profiles that are common between the more homogeneous pancreatic ductal adenocarcinoma (PDAC) and the more heterogeneous colon adenocarcinoma. The fact that the PDAC signature is active in a large fraction of colon adeocarcinoma suggests that the oncogenic mechanisms involved may be similar to those in PDAC, at least in this subset of colon samples.

There are many approaches to uncovering common mechanisms involved in different phenotypes, but most are based on comparing gene lists. The approach presented in this paper additionally takes gene expression data into account and can thus be more sensitive.

1 Introduction and motivation

Understanding cancer at the molecular level is a daunting task due to the enormous *heterogeneity* of this disease, depending not only on tissue and cell type, the progenitor cells involved, but also on the stochastic nature of genomic mutations as well as the associated local evolutionary processes.

However, not all cancers are equally heterogeneous. An ongoing microarray study of pancreatic ductal adenocarcinoma (PDAC) [6] involving 76 samples (i.e. 38 normal-tumor pairs) has revealed a surprising homogeneity of this particularly deadly type of cancer, characterized by a strong so-called “desmoplastic reaction” (fibrosis), as well as by a very high metastatic potential.

A preliminary analysis of the genes differentially expressed between tumor and control samples emphasized the essential role of the *TGF-beta pathway* in PDAC. Remarkably, the TGF-beta pathway links the two observed phenotypes: fibrosis/ extracellular matrix proliferation and the aggressive metastatic potential of PDAC, the latter being due to the fact that TGF-beta controls the so-called epithelial-mesenchymal transition (EMT).

As opposed to PDAC, sporadic colon adenocarcinoma are very heterogeneous and their best current classification based on the presence or absence of microsatellite instabilities (MSI-L, MSI-H and MSS) [1] is far from ideal from the point of view of gene expression.

To obtain a better subclassification of sporadic colon adenocarcinomas, we have applied various unsupervised clustering algorithms to a large colon cancer dataset (204 samples). Interestingly, a large colon adenocarcinoma subclass expressed a set of genes very similar to the genes differentially expressed in pancreatic ductal adenocarcinoma.

This immediately leads to the question of whether the TGF-beta related mechanism involved in PDAC is also at work in at least a subset of colon adenocarcinoma. An ad-hoc approach (like the one mentioned above) based on overlaps of gene lists is however far from satisfactory, since it entirely ignores the quantitative gene expression data available.

This paper presents a more sophisticated method of extracting the gene expression profiles common to a pair of distinct phenotypes (e.g. diseases) for which microarray studies are available. The method involves a generalization of Nonnegative Matrix Factorization (NMF) and is called "*simultaneous NMF*" (*siNMF*), since it factorizes two gene expression datasets simultaneously. More precisely, the *siNMF* algorithm searches for two factorizations (of the two gene expression datasets) *sharing the same gene expression profiles*. This allows us to discover the gene expression profiles that are common to pairs of subclasses in the two datasets.

In the special case of PDAC and sporadic colon adenocarcinoma, we found a gene expression profile highly enriched in target genes of the TGF-beta pathway that is involved in the majority of PDAC cases as well as a large subclass of colon cancers.

2 The datasets

For the present study we have used two large PDAC and sporadic colon adenocarcinoma microarray datasets, which we briefly describe below.

2.1 The pancreatic ductal adenocarcinoma dataset

The pancreatic ductal adenocarcinoma (PDAC) dataset was produced in the framework of our GENOPACT project [6]. The dataset contains microarray measurements produced with Affymetrix U133 Plus 2.0 whole genome chips for 38 pairs of PDAC and respectively control samples (76 samples in total). [1] The raw

[1] As far as we know, the sample size of our study is significantly larger than all published microarray studies of pancreatic ductal adenocarcinoma.

scanning data was preprocessed with the RMA normalization and summarization algorithm from the R package. (The logarithmized form of the gene expression matrix was subsequently used, since typical gene expression values are log-normally distributed.) 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 7232 probe-sets. Finally, the Euclidean norms of the expression levels for the individual genes were normalized to 1 to disallow genes with higher absolute expression values to overshadow the other genes in the factorization.

2.2 The sporadic colon adenocarcinoma dataset

Because of the known heterogeneity of sporadic colon adenocarcinoma, a dataset much larger than the pancreatic dataset described above was needed. We combined 182 colon adenocarcinoma samples from the expO database [7] with 22 control samples from [8] to obtain a 204 sample dataset. (All of these had been measured on Affymetrix U133 Plus 2.0 chips.) After applying the same filtering step as the one used in the PDAC dataset (average expression > 100 and standard deviation > 100), we obtained a smaller set of 5617 probe-sets. The resulting gene expression matrix was also logarithmized before factorization and the Euclidean norms of the individual genes were normalized to 1.

In the following we describe the factorization algorithm in more detail before presenting its application to the two datasets.

3 Simultaneous Nonnegative Matrix Factorization with offset

SiNMF simultaneously factorizes two (non-negative) gene expression matrices $X_{sg}^{(1)}$ and $X_{sg}^{(2)}$ (the index *s* denotes samples, while *g* stands for genes) as follows:

$$X_{sg}^{(1)} \approx \sum_c A_{sc}^{(1)} \cdot S_{cg} + So_g^{(1)} \quad (1)$$

$$X_{sg}^{(2)} \approx \sum_c A_{sc}^{(2)} \cdot S_{cg} + So_g^{(2)} \quad (2)$$

with the additional nonnegativity constraints:

$$A_{sc}^{(1)} \geq 0,\ A_{sc}^{(2)} \geq 0,\ S_{cg} \geq 0,\ So_g^{(1)} \geq 0,\ So_g^{(2)} \geq 0 \quad (3)$$

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 So_g the expression offset of gene *g*.

Note that the gene cluster membership matrix *S* is common to the two factorizations, as it is influenced by both gene expression datasets $X^{(i)}$. The

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

nonnegativity constraints (3) express the obvious fact that expression levels, membership degrees and expression offsets cannot be negative.

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

$$\min C(A^{(i)}, S, So^{(i)}) = \frac{1}{2} \| X^{(1)} - A^{(1)} S - e^{(1)} So^{(1)} \|_F^2 + \frac{\beta}{2} \| X^{(2)} - A^{(2)} S - e^{(2)} So^{(2)} \|_F^2 \quad (4)$$

subject to the nonnegativity constraints (3) ($\| \cdot \|_F$ is the Frobenius norm of a matrix, while $e^{(i)}$ is a column of 1 of size equal to the number of samples of $X^{(i)}$).

The weight β ensures a proper balance between the two error terms and was taken in the following experiments to be $\beta = \beta_0 \frac{\| X^{(1)} \|^2}{\| X^{(2)} \|^2}$ with β_0=1.

The optimization problem (4) can be solved using *multiplicative update rules*[3] in a manner similar to Lee and Seung's seminal *Nonnegative Matrix Factorization* (*NMF*) algorithm [5] (ε is a small regularization parameter):

siNMF($X^{(1)}, X^{(2)}, A_0^{(1)}, A_0^{(2)}, S_0, So_0^{(1)}, So_0^{(2)}$) $\to$ ($A^{(1)}, A^{(2)}, S$)
$A^{(1)} \leftarrow A_0^{(1)}$, $A^{(2)} \leftarrow A_0^{(2)}$, $S \leftarrow S_0$, $So^{(1)} \leftarrow So_0^{(1)}$, $So^{(2)} \leftarrow So_0^{(2)}$ (typically $A_0^{(1)}$, $A_0^{(2)}$, S_0, $So_0^{(1)}$, $So_0^{(2)}$ are initialized randomly)
loop

$$S_{cg} \leftarrow S_{cg} \frac{(A^{(1)T} \cdot X^{(1)} + \beta A^{(2)T} \cdot X^{(2)})_{cg}}{(A^{(1)T} \cdot (A^{(1)} \cdot S + e^{(1)} \cdot So^{(1)}) + \beta A^{(2)T} \cdot (A^{(2)} \cdot S + e^{(2)} \cdot So^{(2)}))_{cg} + \varepsilon}$$

$$A_{sc}^{(i)} \leftarrow A_{sc}^{(i)} \frac{(X^{(i)} \cdot S^T)_{sc}}{((A^{(i)} \cdot S + e^{(i)} \cdot So^{(i)}) \cdot S^T)_{sc} + \varepsilon} \qquad \text{for } i \in \{1,2\}$$

$$So_g^{(i)} \leftarrow So_g^{(i)} \frac{(e^{(i)T} \cdot X^{(i)})_g}{(e^{(i)T} \cdot (A^{(i)} \cdot S + e^{(i)} \cdot So^{(i)}))_g + \varepsilon} \qquad \text{for } i \in \{1,2\}$$

until convergence
normalize the rows of *S* to unit norm by taking advantage of the scaling invariance of the factorization: $S \leftarrow D^{-1} \cdot S$, $A^{(1)} \leftarrow A^{(1)} \cdot D$, $A^{(2)} \leftarrow A^{(2)} \cdot D$, where $D = diag\left(\sqrt{\sum_g S_{cg}^2}\right)$.

The final normalization of the rows of *S* renders the resulting clusters comparable to each other.

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*). However, although overlaps are allowed, the algorithm will not produce highly

[3] The derivation of the above rules is very similar to the derivation of the original Lee and Seung update rules and is not reproduced here for lack of space.

overlapping clusters, due to the nonnegativity constraints and to the offset variables. 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.

4 Nonnegative Matrix Factorizations with offset

Before discussing in more detail the application of *siNMF* to the adenocarcinoma datasets mentioned in the Introduction, we explain in more detail the role of the offset terms *So* in the factorizations (1-2) above.

To make things simpler, we consider a single NMF factorization with offset rather than the simultaneous one from (1-2):

$$X_{sg} \approx \sum_c A_{sc} \cdot S_{cg} + So_g \qquad (5)$$

with the additional nonnegativity constraints:

$$A_{sc} \geq 0,\ S_{cg} \geq 0,\ So_g \geq 0. \qquad (6)$$

The main role of the “offset” *So* is to absorb the constant expression levels of genes, thereby making the cluster samples S_{cg} “cleaner”.

The associated multiplicative update rules can be easily derived using the method of Lee and Seung [5]:

$$A_{sc} \leftarrow A_{sc} \frac{\left(XS^T\right)_{sc}}{\left(\left(eS_0 + AS\right)S^T\right)_{sc} + \varepsilon}$$

$$S_{cg} \leftarrow S_{cg} \frac{\left(A^T X\right)_{cg}}{\left(A^T\left(eSo + AS\right)\right)_{cg} + \varepsilon}$$

$$So_g = So_g \frac{\left(e^T X\right)_g}{\left(e^T\left(eSo + AS\right)\right)_g + \varepsilon}$$

Figure 1 below presents a comparison between the factorizations produced by the standard NMF algorithm and its improvement NMF_{offset} on a synthetic dataset in which columns 36 to 85 are constant “genes”. As can be easily seen in the Figure, these “genes” are reconstructed by the standard NMF algorithm from combinations of clusters, while NMF_{offset} uses the additional degrees of freedom *So* to produce null cluster membership degrees S_{cg} for the constant genes. Moreover, NMF_{offset} recovers with much more accuracy than standard *NMF* the original sample clusters, the standard NMF algorithm being confused by the cluster overlaps. This improvement in recovery of the original clusters is very important in our application, where we aim at a correct sub-classification of samples.

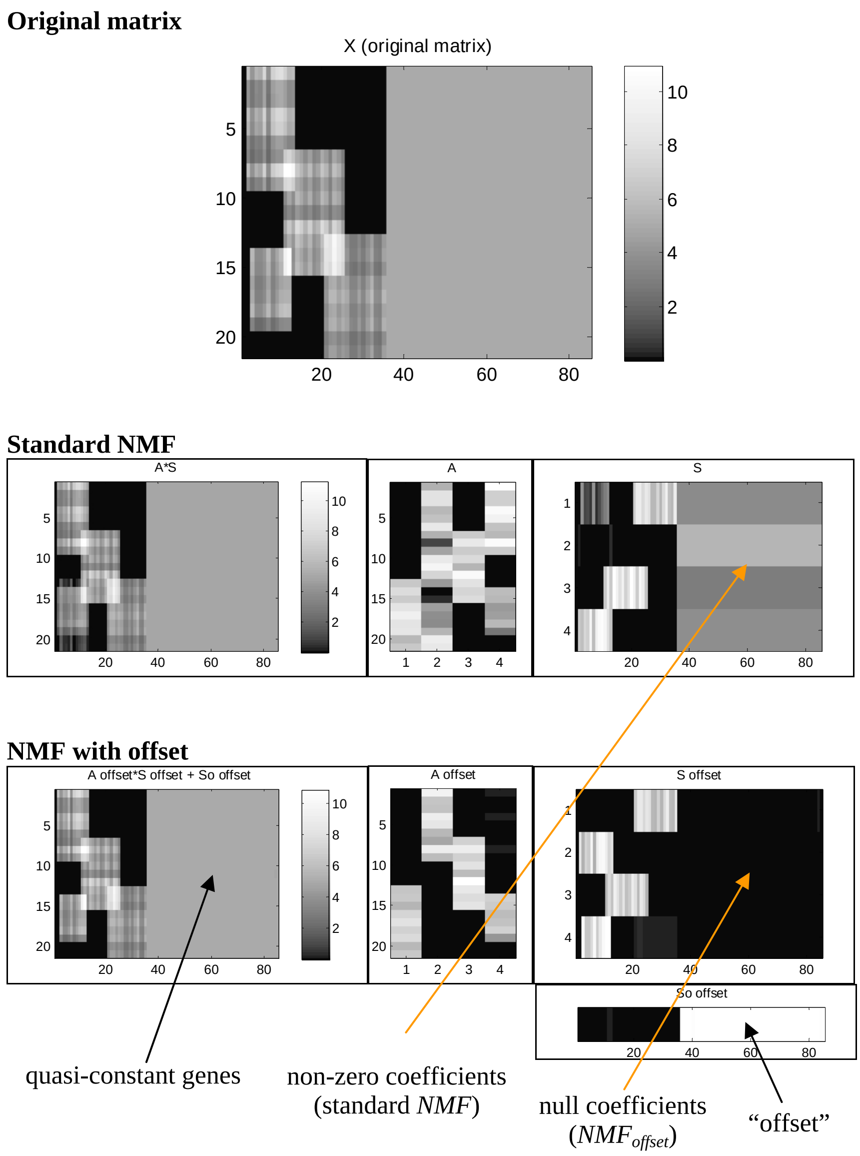

Figure 1. Comparing standard *NMF* with NMF_{offset}

5 Simultaneous factorization of the PDAC and colon adenocarcinoma datasets

In the following we describe the results obtained by applying *siNMF* to the PDAC and sporadic colon adenocarconoma datasets.

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 largest number of dimensions around which the change in relative error $\frac{d\varepsilon}{dn_c}$ of the factorization of the real data is still significantly larger than the change in relative error obtained for a randomized dataset [4] (similar to [9]) – see also Figure 2 below. Using this analysis we estimated the internal dimensionality of the dataset to be between 5 and 7. In the following, we used the conservative value n_c=5.

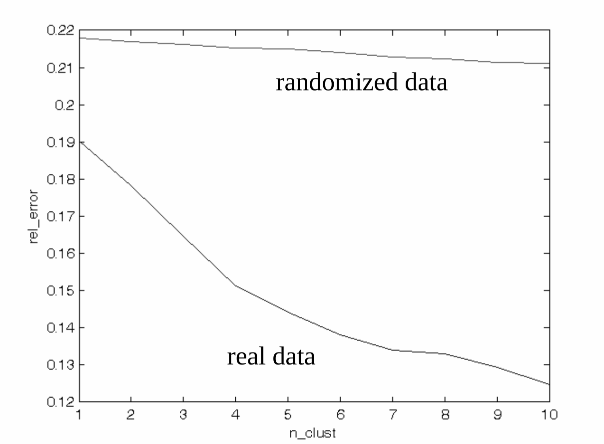

Figure 2. Determining the internal dimensionality of the datasets

We then ran the *siNMF* algorithm with n_c=5 and β_0=1 on the two datasets described previously restricted to the set of common probe-sets (4677 probe-sets).

Since the pancreatic ductal adenocarcinoma dataset is more homogeneous, we first inspected the sample cluster matrix $A^{(1)}$ to determine the cluster that best discriminates between tumor and control samples (see Figure 3 below).

[4] The randomized dataset was obtained by randomly permuting for each gene its expression levels in the various samples. The original distribution of the gene expression levels is thereby preserved.

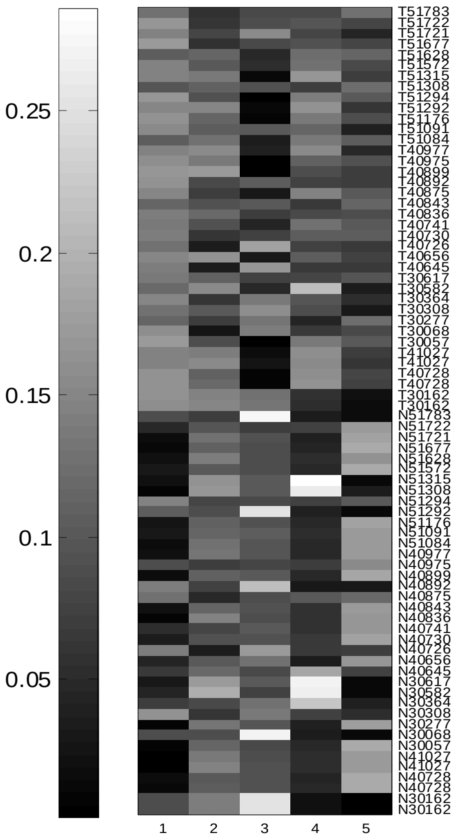

Figure 3. The normalized sample cluster matrix $A^{(1)}$ for the PDAC dataset

Note that cluster 1 recovers relatively well[5] the distinction between tumor and control samples in PDAC, although the algorithm was never provided with class information related to the samples. Similarly, cluster 5 is also significantly well correlated with the tumor-control distinction. In fact, while cluster 1 contains genes that are overexpressed in tumors, cluster 5 comprises mainly downregulated genes. The supplementary material online at www.ai.ici.ro/psb08/ contains the complete

[5] A number of 5 "control" samples (N51294, N40892, N40875, N40726 and N30308) which in our analysis are "closer" to the tumor samples than to the other control ones were later reanalyzed histologically and found to be highly fibrotic (pancreatic tumor tissue is typically very fibrotic and the respective control samples were possibly collected from a site too close to the tumoral tissue).

lists of genes for these clusters. (The threshold used for extracting gene clusters from the *S* matrix was $\sqrt{2} / \sqrt{n_g} = 0.0207$.)

For a more comprehensive biological interpretation of these sets of genes, we then looked for enrichment in known biological annotations using the *L2L Microarray Analysis Tool* [10]. As previously observed in the isolated analysis of PDAC, cluster 1 was enriched in TGF-beta target genes ("tgfbeta_all_up" with p-value 3.75e-29, "tgfbeta_early_up" with p-value 2.94e-25), as well as in the following Gene Ontology [11] Biological Process annotations:

GO Biological Process	**p-value**
response to wounding	9.44e-28
inflammatory response	2.43e-26
response to external stimulus	6.26e-26
defense response	1.74e-21
cell adhesion	8.14e-20
immune response	1.05e-18
organ development	5.13e-18
response to stress	1.10e-14
Chemotaxis	1.31e-14

The following L2L cancer gene expression modules were significantly affected: "ECM and collagens" with p-value 5.25e-82 and "Immune (humoral) and inflammatory response" with p-value 6.17e-65.

All of this is in line with the observed phenotype of PDAC, which involves an over-proliferation of the extracellular matrix (fibrosis, "desmoplastic reaction") and inflammation, supporting the view of cancer as an abnormal response to wounding.

It is impossible to present here a complete analysis of the cluster 1 genes. Some of the most significant ones are:

- INHBA (activin/inhibin betaA) – a ligand for the activin receptor (which triggers a TGF-beta-like pathway),
- POSTN (periostin), which is known to have an active role in the epithelial-mesenchymal transformation and metastasis [12] and whose over-expression promotes metastatic growth of colon cancer by augmenting cell survival via the Akt/PKB pathway [13], as well as enhancing invasion and angiogenesis,
- SULF1, which is known to regulate growth and invasion of pancreatic cancer cells by Interfering with Heparin-binding Growth Factor Signalling [14], etc.

Using the Transcriptional Regulatory Element Database TRED, we found that many of the genes in cluster 1 are controlled by the following transcription factors or a combination thereof: SP1, AP1, AP2, NF-kB, p53, ER, ETS1, SMAD family, CEBPA, etc. (Many direct and indirect TGF-beta targets are controlled by combinations of these factors.)

After having characterized the gene cluster 1 as being the main discriminator between tumor and control PDAC samples, we investigated its function in the colon adenocarcinoma dataset. Figure 4 below presents the normalized [6] sample cluster matrix $A^{(2)}$ for the colon dataset sorted with respect to the first column (cluster). The first sample cluster thus contains 91 tumor samples – half of the total number of 182 colon tumor samples! In other words, *the PDAC gene expression program that distinguishes tumors from controls is active in half of the colon adenocarcinoma we investigated and is highly expressed* [7] *in about* 12%.

On the other hand, cluster 5 is not significantly overexpressed in the normal colon samples (as it was in the PDAC) samples – this may be due either to the small number of normal colon samples, or to the differences in gene expression programs of these tissues (pancreas vs. colon).

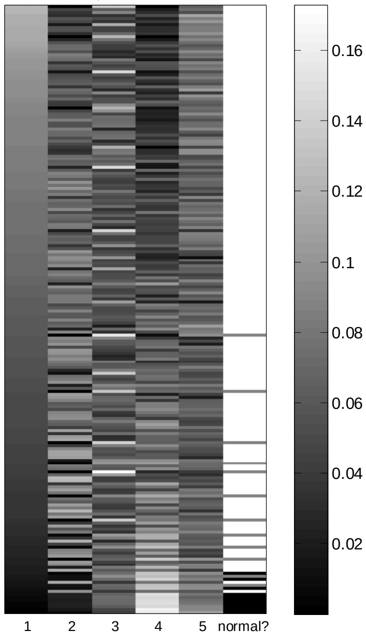

Figure 4. The normalized sample cluster matrix $A^{(2)}$ for the colon dataset (last column shows sample class: black:normal, gray:cancer susceptibility, white:tumor)

6 Conclusions and related work

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. In this paper we have

[6] i.e. $A_c^{(2)} / ||A_c^{(2)}||$. We use a sample threshold $1/\sqrt{n_s} = 0.07$.

[7] over the threshold $\sqrt{2}/\sqrt{n_s} = 0.099$.

introduced a clustering algorithm capable of simultaneously clustering two distinct gene expression datasets with the aim of uncovering gene regulatory programs that are common to the two phenotypes. The two key ingredients of this algorithm are the nonnegativity constraint and the offset variables, which together ensure the sparseness of the factorizations.

Most unsupervised gene expression data analysis methods require a careful selection of genes that are “significant” for subsequent sub-class discovery. But class discovery and “significant gene” selection are tightly inter-connected and cannot be easily separated. Thus, another very important advantage of nonnegative factorization approaches with respect to other methods consists in the fact that they eliminate the need for such an explicit gene selection step prior to classification.

While cancer is a very heterogeneous disease, there is overwhelming recent evidence that the differences between cancer subtypes implicate entire pathways and biological processes involving large numbers of genes, rather than changes in single genes. This has lead us to the following strategy for discovering these processes. We have started with a relatively homogeneous cancer subtype, namely pancreatic ductal adenocarcinoma, for which we have determined the gene group that best distinguishes tumors from controls thereby verifying the homogeneity of this subtype. Then we have applied our simultaneous factorization algorithm looking for gene expression profiles that are common between the more homogeneous PDAC and the more heterogeneous colon adenocarcinoma. The fact that the PDAC signature is active in a large fraction of colon adeocarcinoma suggests that the oncogenic mechanisms involved may be similar to those in PDAC, at least in this subset of colon samples.

The *simultaneous Nonnegative Matrix Factorization* algorithm presented in this paper generalizes the simpler version introduced in [2] by estimating “offsets” for the individual genes, which produces much cleaner gene clusters. Moreover, in [2] we used siNMF to guide the factorization of gene expression data by transcription regulation data, while in this paper we are concerned with finding common mechanisms in different types of adenocarcinoma.

The *siNMF* algorithm is also related in spirit with the generalized SVD algorithm (GSVD) [3], which was applied by Alter et al. for comparing two cell cycle datasets. There, the “common part” of the decomposition is represented by the samples (rather than the genes, as in our approach).

There are many approaches to uncovering common mechanisms involved in different phenotypes, but most are based on comparing gene lists. The approach presented in this paper additionally takes gene expression data into account and can thus be more sensitive.

Of course, this work represents just a first step towards a molecular-level classification of sporadic colon adenocarcinoma, going beyond the simpler one based on microsatellite instability status [1].

Acknowledgments. This work was partially supported by the BIOINFO and GENOPACT projects. I am particularly grateful to Prof. I. Popescu for the collaboration in these projects and to the reviewers for some very useful suggestions for improving this work.

References

1. Jass JR, Biden KG, Cummings MC, Simms LA, Walsh M, Schoch E, Meltzer SJ, Wright C, Searle J, Young J and Leggett BA. Characterisation of a subtype of colorectal cancer combining features of the suppressor and mild mutator pathways. J.Clin.Pathol. 52: 455-460, 1999.
2. Badea L. Combining Gene Expression and Transcription Factor Regulation Data using Simultaneous Nonnegative Matrix Factorization. Proc. BIOCOMP'07, CSREA Press, 2007.
3. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci U S A. 2003 Mar 18;100(6):3351-6.
4. 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.
5. 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.
6. GENOPACT project (CEEX 56/2005).
7. expO. Expression Project for Oncology http://expo.intgen.org/expo/geo/goHome.do
8. Hong Y, Ho KS, Eu KW, Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. Clin Cancer Res. 2007 Feb 15;13(4):1107-14.
9. Kim P.M., Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. 2003 Jul;13(7):1706-18.
10. L2L. L2L Microarray Analysis Tool http://depts.washington.edu/l2l/
11. Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. 25:25-29, 2000.
12. Yan W, Shao R. Transduction of a mesenchyme-specific gene periostin into 293T cells induces cell invasive activity through epithelial-mesenchymal transformation. J Biol Chem. 2006 Jul 14;281(28):19700-8.
13. Bao S, Ouyang G, Bai X, Huang Z, Ma C, Liu M, Shao R, Anderson RM, Rich JN, Wang XF. Periostin potently promotes metastatic growth of colon cancer by augmenting cell survival via the Akt/PKB pathway. Cancer Cell. 2004 Apr;5(4):329-39.
14. Abiatari I., J. Kleeff, J. Li, K. Felix, N.A. Giese, M.W. Büchler, H. Friess. Hsulf-1 Regulates Growth and Invasion of Pancreatic Cancer Cells by Interfering with Heparin-binding Growth Factor Signalling. J Clin Pathol. 2006 Oct;59(10):1052-8.
15. TRED. Transcriptional Regulatory Element Database. http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home.
16. 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.
17. Cheng Y, Church GM. Biclustering of expression data. Proc. ISMB 2000; 8:93-103.