Simultaneous Feature Selection and Clustering for Gene Expression Data Using Nonnegative Matrix Factorizations with Offset

Liviu Badea

AI Lab, National Institute for Research and Development in Informatics (ICI)

8-10 Averescu Blvd., Bucharest, Romania

badea@ici.ro

Abstract: In this paper we show that adding offset terms to standard Nonnegative Matrix Factorization can improve clustering even without an explicit feature (gene) selection step.

Given that most cancer subtypes are very heterogeneous diseases, we apply our algorithm to a large public colon cancer gene expression dataset to differentiate the main genomic-level subtypes of the disease.

Liviu Badea, a senior researcher at the National Institute for Research and Development in Informatics, Bucharest, graduated with honors in Computer Science from "Politehnica" University Bucharest in 1990. In 1996 he obtained his PhD from the same university under the supervision of Prof. Cristian Giumale with a topic in Artificial Intelligence. Dr. Badea's current research interests are in the fields of Bioinformatics, Artificial Intelligence and the Semantic Web.

Keywords: bioinformatics, gene expression data analysis

1. Introduction and Motivation

Understanding cancer at the genomic 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. Moreover, 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. For example, 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. It is therefore essential to make use of existing gene expression data for obtaining a better classification of the disease subtypes, which would enable different treatments specifically targeted to the particular subtypes.

Clustering methods for gene expression data are essential for determining the main subclasses of these diseases. Unfortunately, most existing clustering methods are very sensitive w.r.t. the set of features (genes) considered in the clustering process. Therefore, a careful selection of the set of *relevant* features (genes) is needed.

For example, we could consider selecting the genes with a large standard deviation, but this would favor genes with large expression levels, while excluding possible causal factors with low expression (such as key transcription factors).[1] This problem could be corrected by considering genes with large *relative* standard deviations $\sigma(g) / \mu(g)$ [2], but this would again favor only a specific subset of genes, namely those with large isolated spikes, as in Figure 1, as opposed to genes that have consistently high or consistently low expression levels in larger subsets of samples.

[1] Since gene expression data tend to be log-normally distributed, the standard deviations tend to be proportional to the mean expression levels.

[2] $\sigma(g)$ is the standard devition of gene g, while $\mu(g)$ is its mean.

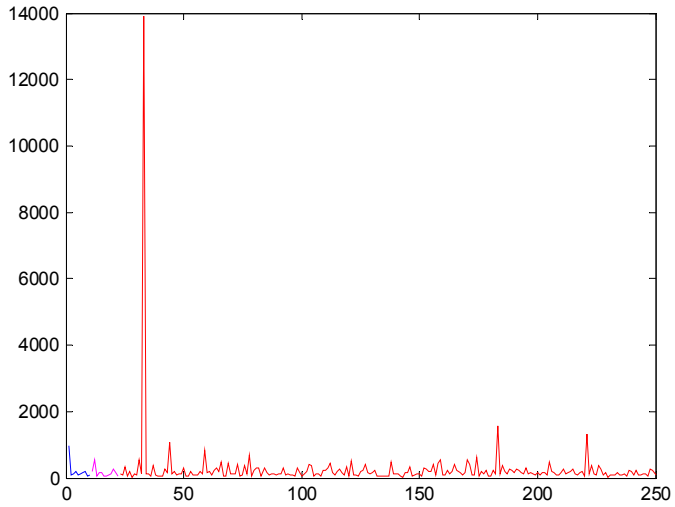

Figure 1. Relative standard deviations tend to favor genes with large spikes (e.g. IGLJ3)

Since the disease subclasses are unknown, we could try selecting the genes with clear bi- or multi-modal distributions, as in Figure 2. Such genes have very different expression values in at least two distinct subsets of samples, but it turns out that *these subsets of samples are completely different for each such bimodal gene* (reflecting perhaps the normal genetic diversity between individuals, as in Figure 2, where the bimodal gene is a member of the HLA complex). Therefore, selecting the genes based only on the bimodality of their distributions will not allow the grouping of samples into well-defined subclasses. We need to select a set of genes which all have similar expression levels in unknown but stable subsets of samples. Thus it turns out that we are in a circular situation in which we need a clustering method for selecting the features (genes) to be used by another clustering method.

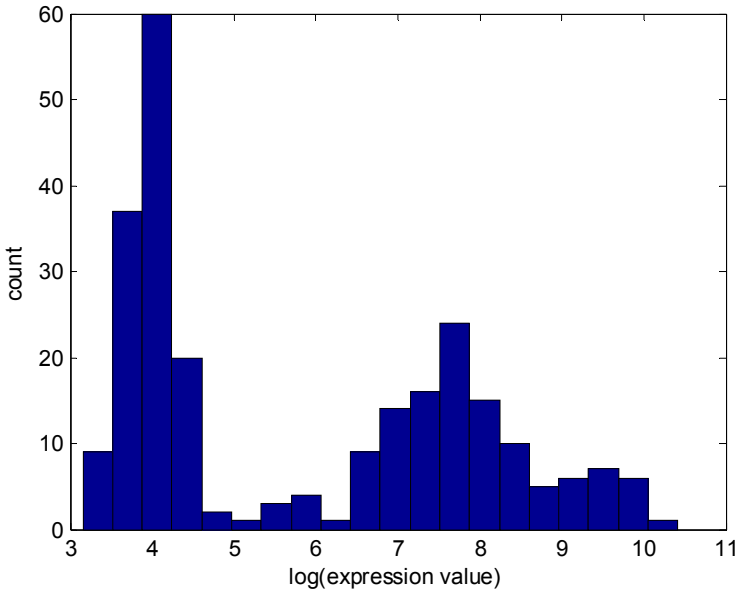

Figura 2. A gene with a bimodal distribution (HLA-DQA1) in the sporadic colon cancer dataset

Our previous results on clustering lung cancer data [**Error! Reference source not found.**] showed that Nonnegative Matrix Factorizations (NMF) [6] are quite well suited for clustering gene expression data without the need of previous feature selection. This is unlike most currently used clustering algorithms

(such as hierarchical clustering [**Error! Reference source not found.**]), which are sensitive to the sets of selected features (genes).

Still, the standard NMF algorithm [6,7] is in certain ways imperfect in domains such as microarray data clustering, where we have large numbers of features (genes), most of which are irrelevant *to varying degrees*. Assuming that the relevant genes lead to well defined gene and sample clusters, we need to explain how NMF succeeds in approximating the irrelevant genes (if NMF did not succeed in approximating the irrelevant genes, the error term associated to these genes would dominate the total error, thereby "drowning" the error term associated to the relevant genes and thus disallowing their grouping). Interestingly, we have observed that NMF reconstructs the irrelevant, nearly constant genes out of the significant gene clusters. Still, although this allows it to deal with large numbers of irrelevant genes, the sample clusters will be affected by such irrelevant genes.

In this paper we should how the standard NMF algorithm can be adapted to deal with quasi-constant irrelevant genes without affecting the sample clusters.

To obtain a better subclassification of sporadic colon adenocarcinomas, we have applied this modified unsupervised clustering algorithm 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 [5].

2. Nonnegative Matrix Factorizations with Offset (NMF_{offset})

As mentioned in the Introduction, standard NMF will reconstruct quasi-constant genes g as *superpositions of the nontrivial clusters* of the factorization, so that the genes g will have significant membership degrees S_{cg} for most clusters (see Figure 3).

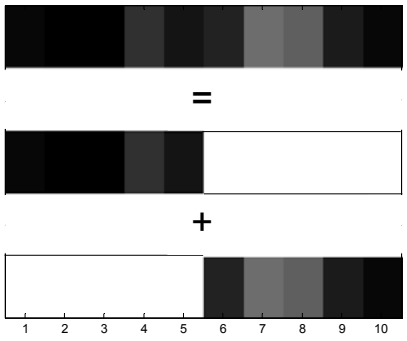

Figure 3. Quasi-constant genes are reconstructed as superpositions of nontrivial clusters

This unnatural reconstruction of quasi-constant genes can be avoided by adding a supplementary degree of freedom So_g for each gene g, representing the „offset" of g. More precisely, a *nonnegative factorization with offset* of the $n_s \times n_g$ (samples × genes) gene expression matrix X as a product of an $n_s \times n_c$ (samples × clusters) matrix A and an $n_c \times n_g$ (clusters × genes) matrix S takes the form:

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

with the additional nonnegativity constraints:

$$A_{sc} \geq 0, S_{cg} \geq 0, So_g \geq 0 \qquad (2)$$

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, So_g the expression offset of gene g and $e = (1\ 1\ ...\ 1)^T$. The nonnegativity constraints (2) express the obvious fact that expression levels and membership degrees cannot be negative.

More formally, this can be cast as a constrained optimization problem:

$$\min f(A,S,So) = \frac{1}{2} \| X - A \cdot S - e \cdot So \|_F^2 = \frac{1}{2} \sum_{s,g} (X - A \cdot S - e \cdot So)_{sg}^2 \qquad (3)$$

subject to the nonnegativity constraints (2).

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

The associated multiplicative update rules can be easily derived using the method of Lee and Seung [7]. The gradient of the function *f* is:

$$\frac{\partial f}{\partial A} = -(X - eSo - AS)S^T$$

$$\frac{\partial f}{\partial S} = -A^T (X - eSo - AS)$$

$$\frac{\partial f}{\partial So} = -e^T (X - eSo - AS) .$$

The additive update rules for the matrices *A, S* and *So* are:

$$A \leftarrow A - \mu_A \frac{\partial f}{\partial A} = A + \mu_A (X - eSo - AS)S^T \qquad (4)$$

$$S \leftarrow S - \mu_S \frac{\partial f}{\partial S} = S + \mu_S A^T (X - eSo - AS) \qquad (5)$$

$$So \leftarrow So - \mu_{So} \frac{\partial f}{\partial So} = So + \mu_{So} e^T (X - eSo - AS) \qquad (6)$$

For obtaining multiplicative update rules, we choose the factors μ to be matrices as follows:

$$\mu_A = \frac{A}{(eSo + AS)S^T}$$

$$\mu_S = \frac{S}{A^T (eSo + AS)}$$

$$\mu_{S_0} = \frac{So}{e^T (eSo + AS)} .$$

Replacing these factors in formulas (4-6) above, we obtain the following algorithm using *multiplicative* update rules:[3]

$\underline{\text{NMF}_{\text{offset}}(X, A_0, S_0, So_0) \rightarrow (A, S)}$

$A \leftarrow A_0$, $S \leftarrow S_0$, $So \leftarrow So_0$ (typically A_0, S_0 and So_0 are initialized randomly)

loop

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

[3] The convergence proofs are very similar to the one for NMF [7].

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

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

until convergence

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

Figure 4 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.

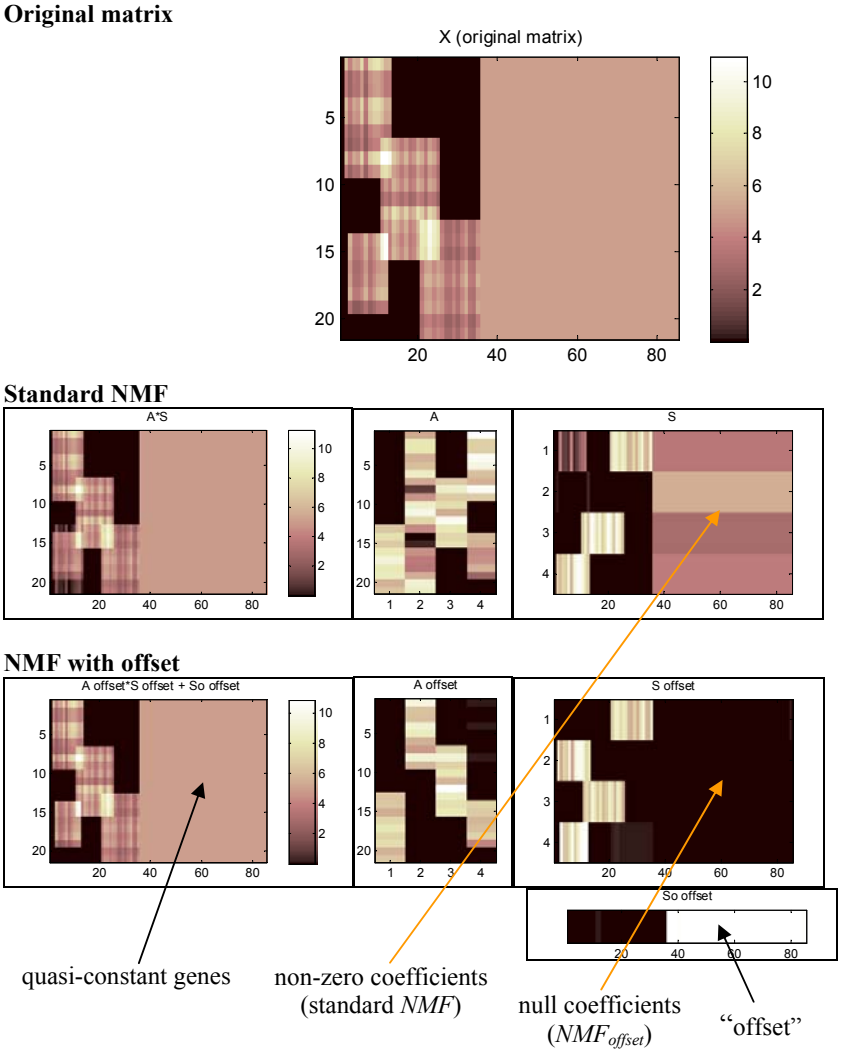

Figure 4. Comparing standard *NMF* with NMF_{offset}

3. Clustering the Sporadic Colon Adenocarcinoma Dataset

Because of the known heterogeneity of sporadic colon adenocarcinoma, an as large as possible dataset was needed. For this, we combined 182 colon adenocarcinoma samples from the expO database [8] with 22 control samples from [9] to obtain a 204 sample dataset. (All of these had been measured on Affymetrix U133 Plus 2.0 chips.) The raw 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[4], we were left with 5617 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.

In the following we briefly describe the application of NMFoffset to the sporadic colon adenocarconoma dataset.

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 [5] (similar to [10]) – see also Figure 5 below. Using this analysis we estimated the internal dimensionality of the dataset to be between 3 and 7.

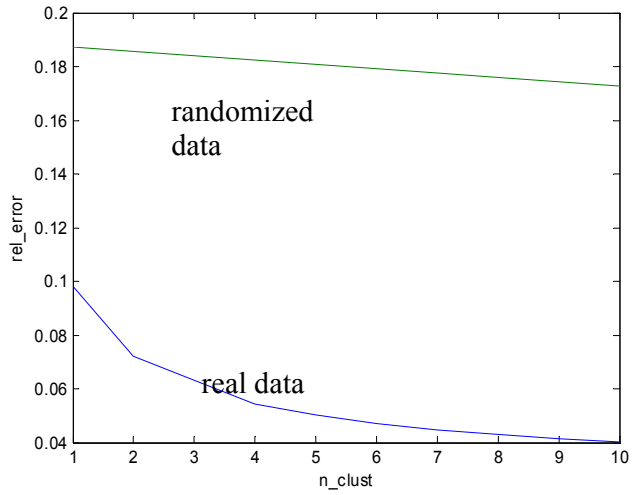

Figure 5. Determining the internal dimensionality of the datasets

Using NMF_{offset} with a conservative n_c=3 on the colon cancer gene expression data, we obtained the two-way clustering diagram from Figure 6. (The dendrograms on the Figure were obtained by hierarchical clustering of the rows of A and of the columns of S respectively.)

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

[5] 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.

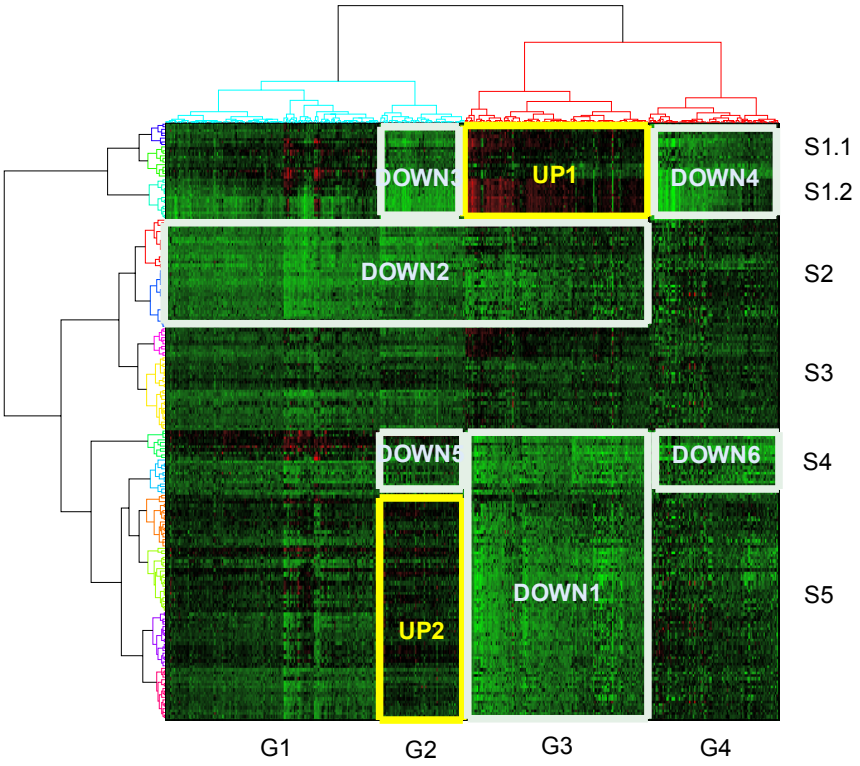

Figure 6. Two-way clustering of the colon cancer dataset

Notice the significant heterogeneity of the disease: there are at least 4 larger gene clusters (*Gi*) and 5-6 sample clusters (*Si*), the normal samples being included in cluster *S1*. In order to characterize the clusters of samples, we searched for gene clusters discriminating between them. The following Table presents the discriminating gene sets for all sample cluster pairs:

	S1	S2	S3	S4	S5
S1		G3, (G4)	G3, G4	G3	G3, G4, G2
S2			G3	G4	G2
S3				G3, G4	G2, G3
S4					G2
S5					

Next, using various annotations (Gene Ontology, GENMAPP and KEGG pathways, protein domains from INTERPRO, etc.) and literature references, we tried to find a biological interpretation of the gene clusters.

Cluster *G2* contains genes that are upregulated (w.r.t. normal tissue) in certain tumors (especially from *S5*). These genes are involved in TGF-beta signalling and inflammatory response. The TGF-beta pathway is very important in several cancer subtypes, as it plays a dual role: tumor suppressor in the initial phases and pro-metastatic in more advanced stages (the relevant genes are INHBA, SPP1 and THBS1). The cluster also contains many genes involved in cell adhesion, many controlled by the TGF-beta pathway.

Cluster *G3* contains genes overexpressed in normal tissue, especially w.r.t. tumor samples from *S2* and *S5*. They are mainly involved in metabolic processes, cellular transport and signalling. There is a marginal overlap with the gene sets reported by Watanabe (2006) to differentiate between MSI and MSS: ACE2, DUOX2, FABP1, HMGCS2, KRT20, LOC63928, PIGR, SLC26A2, SLC26A3. Also, there is a small overlap (ABP1, CDKN2B, ENTPD, FABP1, FCGBP, IGJ, MUCDHL, PDE9A, SLC4A4, XDH) with gene set SANSOM_APC_5_DN from MsigDB, which contains genes underexpressed in a cell line with an inactivated APC gene (these are probably target genes of the Wnt pathway).

Cluster *G4* contains genes from the Wnt pathway (AXIN2, MMP7, PLCB4), 3 metalloproteinases

(MMP-1,3,7) as well as genes involved in cell signalling, transcription control and transport. It overlaps with the gene set differentiating between MSI and MSS (CEL, IQGAP3, KRT23, LY6G6D, REG4, SPINK1, TDGF1, VAV3), the SANSOM_APC_LOSS4_UP gene set (ASCL2, AXIN2, CDCA7, ECT2, EHF, GGH, PSAT1), as well as with a gene set differentiating between tumors with KRAS mutations and tumors with BRAF mutations (IL8, CXCL5 and MMP1 - overexpressed in tumors with BRAF mutations, CEL, DACH1, NOX1, TDGF1 - overexpressed in tumors with KRAS mutations), suggesting that the first subcluster in *G4* may contain genes involved in tumors with BRAF mutations.

4. Conclusions and Related Work

In this paper we have shown that adding offset terms to standard Nonnegative Matrix Factorization can improve clustering even without an explicit feature (gene) selection step. Thus, NMF_{offset} performs simultaneous feature selection *and* clustering.

The molecular profiles of the samples from this colon adenocarcinoma dataset show the heterogeneity of colon cancer at the genomic level, which overlaps only partially with the currently known subtypes (MSI-H, MSI-L, MSS), probably because of the activation of different pathways, which interact in complex ways. This heterogeneity shows that most existing microarray studies of colon cancer are probably too small (in terms of number of samples) to produce statistical relevant results even for the *frequent* subtypes. The present study is, as far as we know, the largest published analysis of colon cancer.

Acknowledgments

This work is dedicated to Florin Filip for his 60th anniversary as well as for his constant support throughout the years for ICI and for *the Studies in Informatics and Control* journal, which he founded 16 years ago. I am particularly grateful to Acad. Filip for his encouragement of our research in the area of bioinformatics. It is therefore appropriate that this paper should appear in this journal.

This work was partially supported by the BIOINFO project.

REFERENCES

1. JASS, JR, BIDEN KG, CUMMINGS MC, SIMMS LA, WALSH M, SCHOCH E, MELTZER SJ, WRIGHT C, SEARLE J, YOUNG J., 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. EISEN, M.B., P.T. SPELLMAN, P.O. BROWN, D. BOTSTEIN, **Cluster Analysis and Display of Genome-Wide Expression Patterns**, PNAS 95, 14863-8.
3. BADEA, L., **Combining Gene Expression and Transcription Factor Regulation Data using Simultaneous Nonnegative Matrix Factorization**, Proc. BIOCOMP'07, CSREA Press, 2007, pp. 127-131.
4. BADEA, L., D. TILIVEA, **Stable Biclustering of Gene Expression Data with Nonnegative Matrix Factorizations**, Proc. IJCAI-07, Hyderabad, India, pp. 2651-2656.
5. BADEA, L., **Extracting Gene Expression Profiles Common to Colon and Pancreatic Adenocarcinoma Using Simultaneous nonNegative Matrix Factorization**, Proc. PSB-08, to appear.
6. LEE, D.D., H.S. SEUNG, **Learning the Parts of Objects by Non-negative Matrix Factorization**, Nature, Vol. 401, No. 6755, 1999, pp. 788-791.
7. 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.
8. **expO.** Expression Project for Oncology http://expo.intgen.org/expo/geo/goHome.do
9. 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.
10. KIM, P.M., TIDOR B**., Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression data**, Genome Res. 2003 Jul;13(7):1706-18.
11. 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.
12. CHENG, Y, CHURCH GM., **Biclustering of expression data**, Proc. ISMB 2000; 8:93-103.