Integrating mRNA and miRNA expression with interaction knowledge to differentiate myelodysplastic syndrome

Michael Anděl[1], Jiří Kléma[1], and Zdeněk Krejčík[2]

[1] Department of Computer Science and Engineering, Prague, Czech Republic,
{andelmi2,klema}@fel.cvut.cz,
[2] Institute of Hematology and Blood Transfusion, Prague, Czech Republic,
zdenek.krejcik@uhkt.cz

Abstract: Onset and progression of a genetically conditioned disease depend not only on genes themselves, but mainly on their expression during transcriptional and proteosynthetic process. Monitoring gene expression merely at its transcription level often proves insufficient for an automated disease understanding and prediction. An integration of diverse high-throughput measurements and prior knowledge is needed to capture gene expression in a holistic way. In this paper, we apply a recent matrix factorization integration method to build a plausible and comprehensive predictive model of an outcome or progress of myelodysplastic syndrome, a blood production disease often progressing to leukemia. We propose an efficient learning methodology that enables to maximize predictive performance and keep the main assets of the original method. The resulting model shows a comparable predictive accuracy with a straightforward data integration method while being more understandable and compact. The identified gene expression regulatory units with the best predictive performance will be subject of further biological analysis.

1 Introduction

A lot of severe diseases are genetically conditioned. The outcome or progress of such a disease depends not only on the patient's genome, but also on the manifestation of certain genes. The overall gene activity during the transcriptional-translational process is called *gene expression* (GE). It is the process through which genes synthesize their products and afflict the phenotype. It is possible to sense the activity of a gene as the measured amount of gene transcripts during its expression process. Current technological progress enables to measure the activity of thousands of genes simultaneously in one tissue sample. One may so feel being capable of predicting the disease outcome, progress or related issues based on acquired *gene expression data* [7]. In other words, to build a molecular classifier with particular genes as features, gene expression levels as feature values and phenotype as target variable.

Nevertheless, recent studies suggest that such a molecular classifier based solely on GE data is often not sufficiently accurate nor understandable [5]. The lack of accuracy can be caused by technical difficulties such as the noise in data as well as immense number of features. Too many features may lead to overfitting, not to mention the features are often redundant, irrelevant or highly dependent. As a reaction to this observation, the prevailing trend in GE data classification is focused on considering entire sets of genes rather than particular genes as the features [1, 8, 11, 15]. The gene sets are related to known or yet unknown biological processes as gene transcription regulation or metabolic pathways. Current effort is to reformulate GE features to gene sets and build models upon whole biological processes. Resulting models should be more precise, robust and, of course, biologically meaningful and more understandable for the experts.

But still, the results of set-level models may turn out disappointing. Gene expression is a complex process with multiple phases and components, which makes measurement of gene activity non-trivial. The acquired data are often confusing in their nature and difficult to interpret and apply. Current molecular biology addresses this difficulty through monitoring the activity of gene expression within multiple components at more stages of the process. The multilevel measurement of GE results in potentially more informative, but much larger data. Therefore another challenge for data analysis raises. A meaningful and comprehensive integration of multiple measurements or multiple data sources is desirable.

In this work we propose a robust classification framework for knowledge-based integration of molecular expression data. Currently, our quantitative measurements cover two fundamental transcript types: messenger RNA (mRNA) and microRNA (miRNA), both the crucial components of the overall gene expression process. mRNA serves as a carrier of genetic information from DNA to proteins. miRNA is a small non-coding molecule acting in transcriptional and post-transcriptional regulation, often hastening mRNA degradation and inhibiting translation of a complementary mRNA into protein. Our challenge is to integrate the measurements of these different types of ribonucleic acids in a biologically meaningful way with great regard to the predictive accuracy of resulting models. The integration is driven by the existing knowledge on miRNA targets and gene-gene interactions. The ultimate goal is a valid and robust decision support tool for immediate use in clinical or experimental practice.

The framework is tested on a particular domain of *myelodysplastic syndrome* (MDS) [23]. The data were

provided by the Institute of Hematology and Blood Transfusion in Prague. MDS obstructs blood stem cells in bone marrow from maturation, resulting in shortage of healthy blood cells. Consequent symptoms are anemia, increased susceptibility to bleeding and infection. What is more, great deal of MDS patients progress to treatment resistant acute myeloid leukaemia. Although many patients are asymptotic, the leukaemia may come out, though. Another issue, reflected in the data is the *chromosome 5q deletion syndrome* (*del(5q)*). del(5q) has similar symptoms as MDS, but mostly does not result in leukaemia. Therefore the del(5q) patients without MDS require different treatment than those with MDS. If this is not confusing enough, del(5q) may progress in MDS. It is evident, that sharp discrimination between healthy and afflicted patients and between the above-mentioned syndromes is needed.

2 Domain description and formalization

Gene expression is the overall process of transferring information from the genome towards the tangible signs of the individual, which are generally called *phenotype*. During the process, the gene is firstly transcribed into the molecule of *messenger* RNA (mRNA), which subsequently migrates towards the ribosomes, where it is translated to a protein. The protein levels determine the final phenotype. GE is most often monitored in its transcriptional phase since the transcript level is easiest to measure. The phenotype prediction stems from the basic assumption that a higher amount of detected mRNA implies a higher amount of translated protein, and therefore higher manifestation of respective gene. Currently the most popular methods for measuring expression level of the genes are the *microarray* and *RNA-Seq* technologies, which enable measuring the activity of thousands of genes in parallel.

As mentioned before, cellular pathology is still not well explored, and therefore it is often unclear which of the thousands of genes are disease related. Analyzing them all may lead to overfitting. What is more, the phenotype is not afflicted by the genes separately, but there is a complex synergy of involved genes. The expression activities of particular genes are often linked together, while *transcription factor* (proteomic functional product), synthesized according to one gene, may control, i.e. *upregulate* or *downregulate*, the transcription of several other genes. That is why one aims at analyzing GE data in terms of *gene sets* or *functional units*, based on gene regulatory networks. The gene-gene interaction networks are *partially* discovered and stored in genomic knowledge bases. The GE data are reformulated in new features, corresponding to the gene sets or heterogeneous biological process units, with the aid of certain genes function already known. The prior (background) knowledge is utilized to control or validate the discovery of novel knowledge. Its application results in more accurate, robust and biologically plausible predictive and descriptive models.

However, correlation between the amount of mRNA as the gene transcription product and the amount of protein as the gene translation product is often much weaker than expected [17, 25]. For that reason the attempt to improve model accuracy through gene set features may often fail. Gene expression process is subdued to more regulatory mechanisms than protein-gene *pre-transcriptional* regulation mentioned above. One of the essentials is gene *post-transcriptional* inhibition through *miRNA*. miRNA regulators are short (22-nt long RNA) sequences of noncoding RNA with crucial role in GE process. Despite its undeniable impact, miRNA was discovered not long ago [16], hence it is subject of intensive biological interest. The molecule of miRNA binds to mRNA molecule, suppressing its further functions. The amount of transcribed mRNA is thus reduced, and the expression of corresponding gene is put down. Malfunction of even one miRNA sequence regulator may cause a severe disease [19]. It is not an easy quest for molecular biologists [12] and bioinformaticians [26] to determine which mRNA sequences target a particular gene. This research is referred to as *target prediction*, an increasing number of the validated and predicted miRNA-gene interactions is available in public target databases such as [6, 24]. The amount of miRNA is measured by miRNA microarrays working in the analogous way as mRNA microarrays. Still more labs issue the miRNA measurements along with the common mRNA profiles in order to capture GE process at more levels and in a broader systematic view [18].

Nevertheless, in order to exhaustively utilize all the advantages contained in simultaneous measurement of mRNA and miRNA expression levels (features) on the same set of samples, one needs to engage the prior knowledge about the *interaction* between particular miRNA and mRNA molecules respectively. One miRNA sequence can target a mRNA code associated with more genes and contrariwise, one gene can be regulated by more miRNAs. Henceforth, the analysis of GE data must be led through entire gene-miRNA *modules* (regulatory units). The integration of mRNA and miRNA features is a non-trivial task due to several reasons: a) the relationship between miRNAs and genes is many-to-many, so the brute force search in known or possible interactions would lead to combinatorial explosion, b) many miRNA-gene interactions are false positive, all the miRNA sequences have not even been discovered, c) little is known about the shape, role and occurrence of modules in the miRNA-gene regulation system. Accordingly, an intelligent method of miRNA and mRNA feature integration should consider the known miRNA-gene interactions and confirm them based on measured data. Finally, based on relevant interactions it would identify present GE regulatory modules. As for the purpose of classification, the last but not least task is to reformulate the data samples in terms of learned modules.

Our challenge is to provide such an integration method for the myelodysplastic syndrome data, acquired through mRNA and miRNA microchips. The method should take

into account the recent knowledge and model the regulatory function units with subsequent use in diagnostic or treatment classification tasks. Let $\mathscr{G} = \{g_1, ..., g_{M^g}\}$ be a the genes, whose expression activity is sensed through the mRNA microarray platform, $\mathscr{R} = \{r_1, ..., r_{M^\mu}\}$ be known miRNA sequences detected through the miRNA platform, $\mathscr{S} = \{s_1, ..., s_N\}$ be the interrogated samples (tissues, patients) and $\mathscr{U} = \{u_1, ..., u_K\}$ be GE regulatory units or biological processes. Then $x^g : \mathscr{G} \times \mathscr{S} \to \mathbb{R}$ is the activity of measured genes within particular samples *in terms of mRNA*, $x^\mu : \mathscr{R} \times \mathscr{S} \to \mathbb{R}$ is the activity of measured miRNA regulators within the samples. $\mathscr{I} : \mathscr{G} \times \mathscr{G} \to \mathbb{B}$ represents known protein-gene regulatory network. The network can be seen as a graph with the genes as vertices and the *known* interactions as edges. $\mathscr{C} : \mathscr{R} \times \mathscr{G} \to \mathbb{B}$ represents the known miRNA-gene control system. It can be interpreted as a bipartite graph, with the genes and miRNAs as vertices and interactions between miRNA regulators and targeted genes. The integration method should take into account these four data sources and knowledge inputs respectively and provide an output in the form $z : \mathscr{R} \times \mathscr{G} \times \mathscr{S} \times \mathscr{U} \to \mathbb{R}$, i.e. the *virtual expression* of the entire set of miRNA-gene regulatory modules.

3 Related work

The most straightforward and intuitive way to integrate the data from mRNA and miRNA platforms, measured on the *same* set of samples, is a mere concatenation of these two sets of RNA profiles for each sample [13]. The miRNA measurements are viewed as just another kind of features besides the mRNA profiles. But the data integrated in such a *blind* way are unsurprisingly larger than simple mRNA data sets, and thus liable to overfitting or noise as mentioned above, not to mention poor interpretability of the resulting model. Additionally, certain gene features (mRNA) and miRNA features are highly associated as miRNA performs the regulation of gene expression. But these relations may not be visible in the data, as miRNA inhibition of a gene displays more in the amount of the synthesized protein, rather than in the momentary concentration of its transcript (mRNA). Therefore the utilization of the known miRNA-gene interactions is advisable.

[9] presents an interesting tool for inferring a disease specific miRNA-gene regulatory network, based on prior knowledge and user data (miRNA and mRNA profiles). However, this method does not address the way to break down the large inferred network into smaller regulatory units, which are essential for subsequent classification. The method of *data specific* identification of miRNA-gene regulatory modules is proposed in [20] and [22], where the modules are searched as maximal bi-cliques or induced as decision rules respectively. But none of these methods gives an intuitive way to *express* the identified modules within the sample set. Contrariwise, [10] provides a black box integration procedure for several data sources as mRNAs, miRNAs, methylation data etc., with an immediate classification output. Nevertheless, this method has no natural interpretation of the learned predictive models, which is unsuitable for an expert decision-making tool. [27] presents a computational framework for integration of multiple types of genomic data to identify miRNA-gene regulatory units by the means of multiple *nonnegative matrix factorization* (NMF). Unlike the other above-mentioned methods, the multiple NMF-based framework utilizes the gene-gene interaction knowledge as well as the miRNA-gene interactions. The identified GE regulatory units thus consist both of the miRNA-gene regulatory module and the gene-gene regulatory module. The authors evaluate their resulting *co-modules* in terms of biological relevance, enrichment analysis but not as to the predictive accuracy. However, NMF is an intuitive method of data modeling with a direct sample transformation to the new feature space.

4 Materials and methods

For the reasons mentioned above, the first step in seeking a way for integration and classification of MDS data will be *sparse network regularized multiple nonnegative matrix factorization* (SNMNMF) from [27]. This section briefly describes the family of NMF methods in general and specifies the applied SNMNMF method. Finally, the SNMNMF application in classification is explained.

4.1 NMF

NMF [14] is a class of methods for data modeling and approximation widely used in other machine learning applications such as computer vision, text mining. Let $\mathbf{X}^g \in \mathbb{R}^{N \times M^g}$ be a data matrix of gene expression, measured as amount corresponding mRNAs, with N samples and M^g features (genes), x_{ij} be expression of gene g_j in sample s_i. NMF then approximates the data as a linear combination of K feature subsets $\mathbf{X}^g \approx \mathbf{WH}$, with $\mathbf{H} \in \mathbb{R}^{K \times M^g}$ a soft membership assigning the features into K feature subsets or modules and $\mathbf{W} \in \mathbb{R}^{N \times K}$ the weight matrix assigning a weight w_{ij} to each j-th feature subset within i-th sample. The of $\mathbf{W}$ are commonly understood as the data samples in the new feature (module) representation [21].

Computation of the matrices W and H is formulated as an optimization problem. The objective is some kind of metric between the original data matrix $\mathbf{X}$ and its approximation $\mathbf{WH}$. The basic constraint is the nonnegativity $\mathbf{W}, \mathbf{H} \geq 0$. Due to such a vague definition of NMF there is really huge amount of factorization methods and appropriate optimization algorithms.

4.2 SNMNMF

Let $\mathbf{X}^g \in \mathbb{R}^{N \times M^g}$ be a data matrix of gene (mRNA) expression and $\mathbf{X}^\mu \in \mathbb{R}^{N \times M^\mu}$ be a data matrix of miRNA activity, with N samples and M^g genes and M^μ miRNA regulators. The multiple matrix factorization models these data

as a linear combination of K gene-gene regulatory modules $\mathbf{H}^g \in \mathbb{R}^{K \times M^g}$ and K miRNA-gene regulatory modules $\mathbf{H}^\mu \in \mathbb{R}^{K \times M^\mu}$ [27], i.e. $\mathbf{X}^\mu = \mathbf{W}\mathbf{H}^\mathbf{g}$ and $\mathbf{X}^\mu = \mathbf{W}\mathbf{H}^\mu$ respectively. The unification of k-th gene-gene module and k-th miRNA-gene module constitutes a miRNA-gene regulatory *comodule*. The weight of k-th comodule in n-th data sample encodes matrix $\mathbf{W} \in \mathbb{R}^{N \times K}$.

SNMNMF factorizes both data matrices in parallel, while the prior knowledge is incorporated to the factorization through network regularization constraints. The overall minimized objective function looks as follows [27]:

$$\begin{aligned} &\|\mathbf{X}^g - \mathbf{W}\mathbf{H}^g\|_F^2 + \|\mathbf{X}^\mu - \mathbf{W}\mathbf{H}^\mu\|_F^2 \\ &-\lambda_g Tr\left(\mathbf{H}^g \mathbf{A} \mathbf{H}^{gT}\right) - \lambda_\mu Tr\left(\mathbf{H}^\mu \mathbf{B} \mathbf{H}^{gT}\right) \\ &+\gamma_1 \|\mathbf{W}\|_F^2 + \gamma_2 \left(\sum \left\|h_j^\mu\right\|_F^2 + \sum \left\|h_j^g\right\|_F^2\right), \end{aligned}$$

where $\mathbf{A} \in \mathbb{B}^{M^g \times M^g}$ is the gene-gene regulatory network matrix, with $a_{ij} = 1$ if and only if the i-th gene and j-th gene interact, $\mathbf{B} \in \mathbb{B}^{M^\mu \times M^g}$ is the miRNA-gene regulatory network matrix, with $b_{ij} = 1$ if and only if i-th miRNA regulates j-th gene, h_j^μ and h_j^g are the j-th column of $\mathbf{H}^\mu$ and $\mathbf{H}^\mathbf{g}$ respectively. The third and fourth terms of the objective introduce the prior knowledge, i.e. λ_μ and λ_g encode the strength of *known* miRNA-gene ($\mathbf{B}$) and gene-gene ($\mathbf{A}$) interactions, respectively. The fifth term limits the growth of $\mathbf{W}$, while the last one encourages sparsity. The objective function is minimized by gradient descent through alternating updates of $\mathbf{W}$ and $\mathbf{H}$s [27]. λ_μ, λ_g, γ_1 and γ_2 are the unknown parameters of the model.

4.3 Classification framework

Although SNMNMF was not primarily intended as a *feature extraction* method with subsequent classification, its use in predictive modeling is intuitive. As the weight matrix $\mathbf{W}$ represents activity of the comodules in particular sample, it may be considered as a projection onto a new feature (comodule) space. Nevertheless, in order to avoid selection bias, while estimating the classification error over the transformed data, one must not incorporate the testing samples into the process of integration parametrization. In the other words, matrix factorization has to be performed on training data only, whereas the testing data are projected into the factorization just learned. Therefore a projection of testing data into the existing transformation (factorization) is needed. Such a projection, we used, is quite intuitive, though. The comodules encoded in matrices $\mathbf{H}^g$ and $\mathbf{H}^\mu$, are learned on training data through the iterative updates, alternating with updates of weight matrix $\mathbf{W}$. Subsequently, the comodule matrices learned are fixed and freshly initialized weight matrix $\mathbf{W}_{test}$ is computed by updating *only* $\mathbf{W}_{test}$ based on *test* data and fixed matrices $\mathbf{H}^g$ and $\mathbf{H}^\mu$. $\mathbf{W}_{test}$ is then considered as the test data in comodule feature space.

SNMNMF seems to be suitable for classification tasks thanks to its natural interpretability, intuitive test data projection and plausible incorporation of prior knowledge. However, its stability with regards to the random initialization of matrix factors and parameter settings remains debatable. Another challenge is the choice of proper number of comodules K, the metaparameter of the algorithm. We set K as the number of "natural" clusters in the miRNA profiles. In order to find this number, we clustered the miRNA profiles by k-means algorithm and set the number of clusters based on Hartigan heuristic, i.e. the sharpest decline of clustering homogeneity. This choice was done independently in each of 10 tasks.

5 Experiments

In this section, we describe the available MDS data and specify the experimental protocol that allows us to set the internal parameters of SNMNMF and evaluate its predictive potential in an unbiased way. Finally, the results and their possible biological interpretation will be discussed.

5.1 Data

The data provided by the Institute of Hematology and Blood Transfusion in Prague consist of microarray measurements of mRNA and miRNA profiles. The measurements were realized using Illumina chips. The mRNA dataset has 16,666 attributes representing the GE level through the amount of corresponding mRNA measured, while the miRNA dataset has 1,146 attributes representing the activity of particular miRNA regulators.

The measurements were conducted on 75 tissue samples categorized according to the several conditions: 1) tissue type: peripheral blood (PB) CD14+ monocytes vs. bone marrow (BM) CD34+ progenitor cells, 2) presence of MDS or del(5q), 3) treatment stage: before treatment (BT) vs. during treatment (DT). Henceforth the samples can be broken into 10 categories. The categories with the actual number of samples are shown in Table 1.

PB	Healthy		10
	5q-	BT	9
		DT	13
	non 5q-	BT	4
		DT	5
BM	Healthy		10
	5q-	BT	11
		DT	5
	non 5q-	BT	6
		DT	2

Table 1: The overview of MDS classes

The domain experts defined 10 binary classification tasks with a clear diagnostic and therapeutic motivation. There are 5 tasks for each tissue type, the numbers of samples are shown in parentheses:

1. **PB1**: healthy (10) vs. afflicted in PB (31),
2. **BM1**: healthy (10) vs. afflicted in BM (24),
3. **PB2**: healthy vs. untreated in PB (13),
4. **BM2**: healthy vs. untreated in BM (17),
5. **PB3**: healthy vs. untreated with del(5q) in PB (9),
6. **BM3**: healthy vs. untreated with del(5q) in BM (11),
7. **PB4**: healthy vs. treated in PB (18),
8. **BM4**: healthy vs. treated in BM (7),
9. **PB5**: afflicted with del(5q) (9) vs. afflicted without del(5q) in PB (22),
10. **BM5**: afflicted with del(5q) (8) vs. afflicted without del(5q) in BM (16).

Considering the prior knowledge, we had downloaded the interactions between genes and miRNAs from miRWalk database [6], while the knowledge about interactions between particular genes we obtained as the interactions of their corresponding proteins from [2].

5.2 Experimental procedure

We used three different classification algorithms to learn on resulting comodules: 1) naïve Bayes, 2) support vector machine (SVM) and 3) k-nearest neighbor (kNN). This selection is to avoid dependence of experimental results on a specific choice of a learning method.

For each task, SNMNMF needs to be correctly parametrized first. When the parametrization is available, the raw data can be projected onto comodules. Finally, the three learners are applied in the transformed comodule space and evaluated using 5-fold cross-validation.

The proper parameter configuration of SNMNMF was reached as follows. The parameters γ_1 and γ_2 were set to 5 as recommended by [27]. The "knowledge-strength" parameters λ_μ and λ_g, which seemed crucial for predictive accuracy, were tuned through 5-fold internal cross-validation for each particular experiment. For each parameter configuration, a factorization process was run on training subsets of the *internal* cross-validation and the predictive accuracy of learned comodules was estimated on testing subsets for each of the learners. The locally optimal parameter configuration has been validated for each of the learners by *external* 5-fold cross-validation. Eventually, this validated accuracy was considered as the final accuracy estimate reached by the optimized matrix factorization. The values of parameters were chosen from $\{5\cdot 10^{-5}, 5\cdot 10^{-4}, 10^{-3}, 0.01\}$ for λ_g and $\{10^{-3}, 0.01, 0.05, 0.1, 0.2\}$ for λ_μ. The number of iterations of SNMNMF factorization was set to 50. For each classification task the experiment was rerun from 15 initializations.

To ensure equal conditions within the course of each experiment, the matrix factors $\mathbf{H}^g$, $\mathbf{H}^\mu$ and $\mathbf{W}$ were randomly initialized only once, on the experiment beginning. The initialized matrix factors were subsequently passed to the folds of cross-validation as follows. The sample-size invariant comodule matrices $\mathbf{H}^g$ a $\mathbf{H}^\mu$ were passed unchanged. In the weight matrix $\mathbf{W}$ representing the data samples in the comodule space, only the rows that link to the particular validation fold were passed.

To sum up, $4 \times 5 = 20$ different parameter configurations were evaluated. Each parametrization was run 15 times for each of 5 internal folds. This process was repeated in each of 5 external folds. Having 10 tasks, we performed 75,000 individual factorizations. The best parametrization was found for each task, external fold and random initialization, which gives 750 λ_g and λ_μ pairs.

The SNMNMF predictive accuracy was compared with accuracy of its several straightforward alternatives. Firstly, we used only mRNA features to classify the samples. The goal was to compare SNMNMF with the most common GE classification technique. Secondly, we used only miRNA features to classify the samples in order to see direct applicability of miRNA for MDS prediction. Eventually, we evaluated the blind *merged* integration method, concatenating the mRNA and miRNA features. This reference allows us to study the asset of the advanced knowledge-driven feature extraction taken in SNMNMF. All these classification techniques were assessed through the three learners and 5-fold cross-validation.

The experiments were implemented in Python with the aid of numerical library NumPy [4] and machine learning library Orange Biolab [3].

5.3 Results

We obtained 450 predictive accuracy values (PAs) for SNMNMF integration method (10 tasks, 3 learners, 15 initializations). To compare it with the reference techniques we used the median PAs taken over the initializations. For the three reference techniques we obtained 30 PAs (10 tasks, 3 learners). The absolute accuracy of the compared classification techniques is summarized in Tables 2-4. The relative accuracy comparison of SNMNMF integration and the blind merged integration method is in Figures 1-3.

5.4 Discussion

The individual methods were evaluated for 10 classification tasks and 3 different learners, i.e., in 30 experiments. The results suggest that none of the methods shows clear dominance over the others. To obtain a global picture, each pair of the methods is mutually compared in every single experiment. The overall pairwise accuracy comparison of particular methods is graphically represented in Figure 4. To exemplify, the miRNA features dominate the mRNA features in 15 experiments, tie on 5, and surrender

task	mRNA	miRNA	merged	SNMNMF
PB1	0.85	**0.88**	**0.88**	**0.88**
BM1	0.94	**0.97**	**0.97**	**0.97**
PB2	0.75	0.76	0.70	**0.80**
BM2	0.97	**1.00**	**1.00**	**1.00**
PB3	**0.85**	0.80	**0.85**	**0.85**
BM3	**1.00**	0.95	0.95	0.95
PB4	**0.83**	0.78	0.78	0.78
BM4	**0.95**	0.87	0.87	0.93
PB5	0.93	**1.00**	0.97	0.94
BM5	0.87	**0.96**	**0.96**	0.95

Table 2: Absolute PAs for evaluated classification techniques in the view of naïve Bayes classifier

task	mRNA	miRNA	merged	SNMNMF
PB1	**0.76**	**0.76**	**0.76**	**0.76**
BM1	0.74	**0.77**	0.74	**0.77**
PB2	**0.90**	**0.77**	**0.90**	**0.90**
BM2	0.93	1.00	0.96	**1.00**
PB3	**1.00**	0.90	**1.00**	0.88
BM3	0.95	**1.00**	0.95	**1.00**
PB4	0.76	0.76	**0.80**	0.72
BM4	**0.95**	**0.95**	**0.95**	0.93
PB5	**0.71**	**0.71**	**0.71**	**0.71**
BM5	0.79	**0.87**	0.79	**0.87**

Table 3: Absolute PAs for evaluated classification techniques in the view of SVM classifier

in 10. This observation indicates that miRNA measurements have a real merit compared to the standard GE classification based on mRNA features. This may be caused by the key role of miRNA regulation in the examined disease or by the substantially smaller number of features in case of miRNA classification, which fundamentally prevents overfitting. However, the tight miRNA win confirms that mRNA information also has its worth. Concatenating miRNA and mRNA features together does not improve the accuracy, though. The blind mRNA concatenation to the auspicious miRNAs immensely increases feature space, which probably leads to overfitting.

Technically, the SNMNMF-based classification surpasses the reference techniques. But still, there are frequent cases, when namely the merged integration or miRNA model outperforms it. We conclude that the integrative SNMNMF yields predictive results clearly not worse than its counterparts. The SNMNMF results seem hopeful with respect to its biologically sound feature compression and locally-optimal parameter configuration only.

Of the different tasks and comodules given, biological relevance can be observed in several cases. miR-451 was previously reported as positive regulator of erythroid cell maturation and recently we have detected increased ex-

task	mRNA	miRNA	merged	SNMNMF
PB1	**0.98**	0.80	**0.98**	0.91
BM1	0.89	**0.97**	0.89	0.94
PB2	0.88	0.73	0.88	**0.90**
BM2	0.90	**0.97**	0.90	**0.97**
PB3	0.85	0.75	**0.95**	0.85
BM3	0.95	**1.00**	0.95	0.95
PB4	0.78	0.57	0.77	**0.81**
BM4	0.88	**0.93**	0.83	**0.93**
PB5	0.97	0.97	0.97	**1.00**
BM5	0.92	**1.00**	0.92	0.96

Table 4: Absolute PAs for evaluated classification techniques in the view of kNN classifier

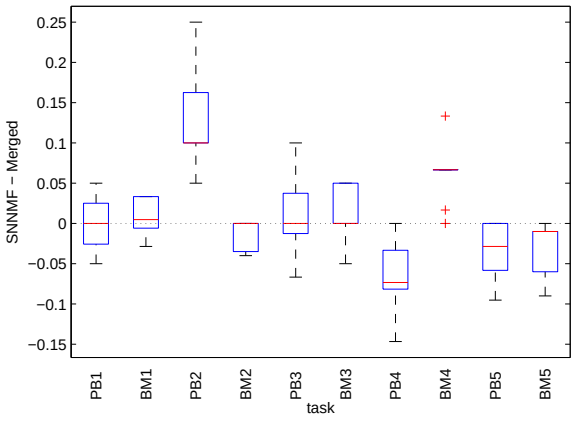

Figure 1: naïve Bayes

pression of miR-451 that was not affected by lenalidomide treatment in both BM CD34+ cells and PB monocytes of MDS patients with del(5q). In task PB1 a link of miR-451 to *hbb* (hemoglobin, beta), *hbe1* (hemoglobin, epsilon 1) and *hbq1* (hemoglobin, theta 1) genes has been found in the same comodule. miR-451 appears also in task PB3, where the interaction of jointly (in the same comodule) reported entities, namely *bax* and *cd82* (p53 signaling pathway), *rab11b* (member of RAS oncogene family), *cdkn2d* (cell cycle), *grb2* and *mapk11* (MAPK signaling pathway) and *pim1* (acute myeloid leukemia), could act in development of MDS.

miR-150 and miR-146a are known to be involved in hematopoiesis and MDS with del(5q), respectively. So called minor versions of those, miR-150* and miR-146a*, are coexpressed in the same comodule in task BM2; and downregulation of miR-150* has also been detected in BM CD34+ cells of del(5q) MDS patients before treatment compared to healthy donors. Of the genes expressed in that comodule, *bcl11a* (B-cell lymphoma/leukaemia 11A), which encodes zinc finger protein, is of importance as it functions as a myeloid and B-cell proto-oncogene and

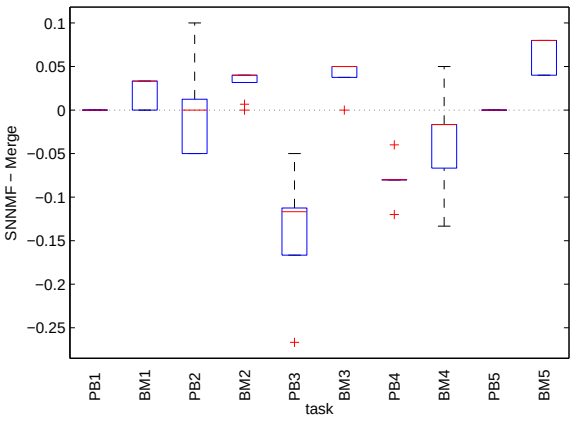

Figure 2: SVM

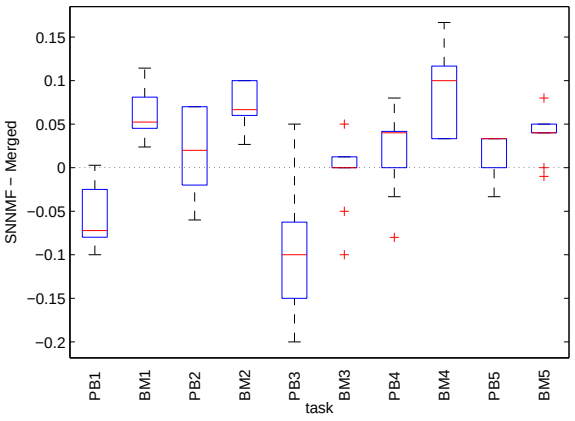

Figure 3: kNN

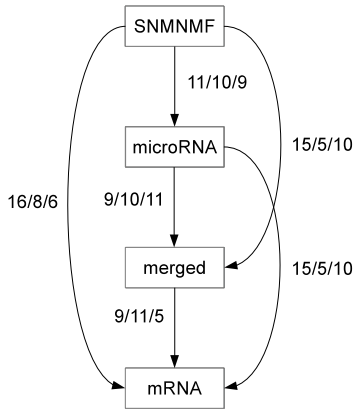

Figure 4: Pairwise accuracy comparison graph. The nodes represent particular feature sets, an edge from node a to node b, annotated as $x/y/z$ means that method a outperforms method b in x experiments, in y ties and in z looses.

therefore may play an important role in leukaemogenesis and hematopoiesis. Gene functional classification analysis of the genes of that comodule revealed some other genes (*znf319, zscan2, znf467, znf585a, znf32*) coding for yet unidentified zinc finger proteins which may be involved in transcriptional regulation, however, their role remains speculative.

Of the miRNAs jointly appeared in task BM4, miR-154 and miR-381 were significantly upregulated in BM cells of MDS del(5q) and their link to *rab23* (member of RAS oncogene family) and *wnt9b* (wingless-type MMTV integration site family, member 9B), both involved in Hedgehog signaling pathway, which has also been implicated in the growth of some cancers, is to be further explored.

6 Conclusion

The increasing amount of genomic data measured on different stages of expression process, along with the rising availability of prior knowledge about GE regulation, give us the challenging opportunity to build robust predictive models based on entire biological processes. Such models should be more comprehensible and potentially more accurate than standard GE classification based solely on one type of measurement, mostly the amount of mRNA. The integration of heterogeneous measurements and prior knowledge is non-trivial, though.

In this work we classify myelodysplastic syndrome patients. Two types of measurements are available for each sample: the amount of mRNA corresponding to gene transcription and the amount of miRNA corresponding to gene translation regulation. We investigate the possibility to utilize the biggest deal of information contained in the provided data through their integration with available prior knowledge, namely miRNA targets and protein-protein interactions. We propose the classification framework based on multiple matrix factorization. The result is a knowledge-enriched predictive model.

A large number of experiments was run to obtain an unbiased accuracy of the integrated model. The results indicate that integration of the heterogeneous measurements together with prior knowledge has its merit and prospects. The knowledge-based classification yields possibly better but clearly not worse results than simple data concatenation or omitting of one type of measurement. What is more, the integrated models are more comprehensive and interpretable. It is obvious that predictive accuracy of the SNMNMF and any other integrated model can further be increased by utilization of the most prospective raw features, in the case of MDS it would namely be the most predictive miRNAs.

But still, there is a lot of future work. The first field of improvements concerns algorithmic enhancements. Within SNMNMF it is desirable to employ an informed parameter search instead of the actual non-informed com-

plete search. Another possibility is to develop a less parameter dependent integration method. We intend to use the prior knowledge to control pseudorandom construction of weak classifiers vaguely corresponding to the individual biological processes. The weak classifiers will later be merged into an ensemble classifier.

Further, the gene regulatory network shall be extended. Currently it contains protein-protein interactions only, not considering the interactions between genes and their transcription factors. Another challenge is to employ epigenomic data, namely DNA methylation.

Acknowledgments

This research was supported by the grants NT14539, NT14377 and NT13847 of the Ministry of Health of the Czech Republic.

References

[1] G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, et al. Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC Bioinformatics*, 11:277, 2010.

[2] A. Bossi and B. Lehner. Tissue specificity and the human protein interaction network. *Molecular systems biology*, 5(1), Apr. 2009.

[3] T. Curk, J. Demšar, Q. Xu, G. Leban, et al. Microarray data mining with visual programming. *Bioinformatics*, 21:396–398, Feb. 2005.

[4] P. F. Dubois, K. Hinsen, and J. Hugunin. Numerical python. *Computers in Physics*, 10(3), May/June 1996.

[5] A. Dupuy and R. M. Simon. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *JNCI Journal of the National Cancer Institute*, 99(2):147–157, Jan. 2007.

[6] H. Dweep, C. Sticht, P. Pandey, and N. Gretz. miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of biomedical informatics*, 44(5):839–847, Oct. 2011.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct. 1999.

[8] M. Holec, J. Kléma, F. Železný, and J. Tolar. Comparative evaluation of set-level techniques in predictive classification of gene expression samples. *BMC Bioinformatics*, 13(Suppl 10):S15, 2012.

[9] G. T. Huang, C. Athanassiou, and P. V. Benos. mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic acids research*, 39(Web Server issue):W416–W423, July 2011.

[10] D. Kim, H. Shin, Y. S. Song, and J. H. Kim. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. of Biomedical Informatics*, 45(6):1191–1198, Dec. 2012.

[11] M. Krejník and J. Kléma. Empirical evidence of the applicability of functional clustering through gene expression classification. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 9(3):788–798, May 2012.

[12] M. Lagos-Quintana, R. Rauhut, Meyer, et al. New microRNAs from mouse and human. *RNA (New York)*, 9(5):175–9, 2003.

[13] G. Lanza, M. Ferracin, R. Gafà, et al. mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. *Molecular cancer*, 6:54+, Aug. 2007.

[14] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[15] E. Lee, H.-Y. Chuang, J.-W. Kim, et al. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4(11), 2008.

[16] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec. 1993.

[17] E. Lundberg, L. Fagerberg, D. Klevebring, I. Matic, et al. Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular systems biology*, 6(1), Dec. 2010.

[18] J. Nunez-Iglesias, C.-C. Liu, T. E. Morgan, et al. Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation. *PloS one*, 5(2):e8898+, Feb. 2010.

[19] K. V. Pandit, D. Corcoran, H. Yousef, M. Yarlagadda, et al. Inhibition and role of let-7d in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med*, 182(2):220–9, 2010.

[20] X. Peng, Y. Li, K. A. Walters, and E. a. o. Rosenzweig. Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*, 10(1):373+, Aug. 2009.

[21] R. Schachtner, D. Lutter, P. Knollmüller, A. M. Tomé, et al. Knowledge-based gene expression classification via matrix factorization. *Bioinformatics*, 24(15):1688–1697, 2008.

[22] D. H. Tran, K. Satou, and T. B. Ho. Finding microRNA regulatory modules in human genome using rule induction. *BMC Bioinformatics*, 9(S-12), 2008.

[23] A. Vašíková, M. Běličková, E. Budinská, and J. Čermák. A distinct expression of various gene subsets in cd34+ cells from patients with early and advanced myelodysplastic syndrome. *Leuk Res*, 34(12):1566–72, 2010.

[24] T. Vergoulis, I. S. Vlachos, P. Alexiou, G. Georgakilas, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic acids research*, 40(Database issue):D222–D229, Jan. 2012.

[25] C. Vogel and E. M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4):227–32, Apr. 2012.

[26] B. Zhang, X. Pan, Q. Wang, et al. Review: Computational identification of microRNAs and their targets. *Comput. Biol. Chem.*, 30(6):395–407, Dec. 2006.

[27] S.-H. Zhang, Q. Li, J. Liu, and X. J. Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics [ISMB/ECCB]*, 27(13):401–409, 2011.