Knowledge-based Subtractive Integration of mRNA and miRNA Expression Profiles to Differentiate Myelodysplastic Syndrome

Jiří Kléma[1], Jan Zahálka[1], Michael Anděl[1] and Zdeněk Krejčík[2]
[1]*Department of Computer Science and Engineering, Czech Technical University, Technická 2, Prague, Czech Republic*
[2]*Department of Molecular Genetics, Institute of Hematology and Blood Transfusion, U Nemocnice, Prague, Czech Republic*

Keywords: Gene Expression, Machine Learning, microRNA, Classification, Prior Knowledge, Myelodysplastic Syndrome.

Abstract: The goal of our work is to integrate conventional mRNA expression profiles with miRNA expressions using the knowledge of their validated or predicted interactions in order to improve class prediction in genetically determined diseases. The raw mRNA and miRNA expression features become enriched or replaced by new aggregated features that model the mRNA-miRNA interaction. The proposed subtractive integration method is directly motivated by the inhibition/degradation models of gene expression regulation. The method aggregates mRNA and miRNA expressions by subtracting a proportion of miRNA expression values from their respective target mRNAs. The method is used to model the outcome or development of myelodysplastic syndrome, a blood cell production disease often progressing to leukemia. The reached results demonstrate that the integration improves classification performance when dealing with mRNA and miRNA profiles of comparable predictive power.

1 INTRODUCTION

Onset and progression of *myelodysplastic syndrome*, like other genetically determined diseases, depend on the overall activity of copious genes during their expression process. Current progress in microarray technologies (Brewster et al., 2004) and RNA sequencing (Morin et al., 2008) enables affordable measurement of wide-scale gene activity, but only on the transcriptome level. Further levels of the gene expression (GE) process which prove disease, whether proteome or even phenome, are still difficult to capture. Henceforth, many natural learning tasks, such as disease diagnosis or classification, become non-trivial within current generation GE data. However, GE is a complex process with multiple phases, components, and regulatory mechanisms. Sensing GE at certain points of these phases and integrating the measurements with the aid of recent knowledge about subduing mechanisms may show the GE process in a broader, systematic view, and make the analysis comprehensible, robust and potentially more accurate.

The goal of our work is to integrate conventional GE data sources as mRNA profiles with microRNA measurements and to experimentally evaluate the merit of using the integrated data for class prediction. MicroRNAs (miRNAs) (Lee et al., 1993) serve as one component of complex machinery which eukaryotic organisms use to regulate gene expression and protein synthesis. Since their discovery, miRNA have shown to play crucial role in development and various pathologies (Croce, 2009; Sayed and Abdellatif, 2011). They are short (21 nucleotides) noncoding RNA sequences which mediate post-transcriptional repression of mRNA via multiprotein complex called RISC complex (RNA-induced silencing complex) where miRNA serve as a template for recognizing complementary mRNA. The complementarity of miRNA-mRNA binding initiates one of the two possible mechanisms: the complete homology triggers degradation of target mRNA, whereas a partial complementarity leads to inhibition of translation of target mRNA. However, despite the progress in understanding the underlying mechanisms in recent years, the effects of miRNA on gene expression is still a developing field and many important facts about mechanism of action and possible interactions remain still unclear (Fabian and Sonenberg, 2012). The level of expression of particular miRNAs can be measured by (e.g.) miRNA microarrays, analogically to well-known mRNA profiling. The resulting dataset, called the miRNA expression profile,

Kléma J., Zahálka J., Anděl M. and Krejčík Z..
Knowledge-based Subtractive Integration of mRNA and miRNA Expression Profiles to Differentiate Myelodysplastic Syndrome.
DOI: 10.5220/0004752200310039
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2014), pages 31-39
ISBN: 978-989-758-012-3

contains, similarly to mRNA profiles, tissue samples as data instances; only this time the attributes are individual miRNA sequences. Integrating mRNA and miRNA data sources may provide a better picture about the true protein amount synthesized according to respective genes, regarding the mechanisms of disease occurrence.

We propose integration stemming from the knowledge which miRNA targets which mRNA. Target prediction is a topic of active research (Tan Gana et al., 2012). The most reliable form of target prediction is experimental *in vitro* validation. Complementary *in silico* target prediction offers more miRNA targets with a higher false detection rate. The predictive algorithms either work based directly on molecular biological theory, building the relationship based on miRNA/mRNA structure and properties (Lewis et al., 2003; Dweep et al., 2011), or be data-driven; i.e., determining targets empirically using statistical or machine learning methods on as much data as possible (Wang and Naqa, 2008; Krek et al., 2005). As an example of algorithms of the first class we should mention miRanda, as an extension of the Smith-Waterman algorithm (Smith and Waterman, 1981), miRWalk (Dweep et al., 2011) and TargetScan (Lewis et al., 2003); as to those of the second class refer to miRTarget2 (Wang and Naqa, 2008) or PicTar 5 (Krek et al., 2005). Target prediction algorithms usually output a score, which for a particular mRNA and a particular miRNA quantifies the strength of the belief that the two are truly related. While there is no guarantee that the results are truly correct, employing prediction algorithms on already existing gene/miRNA expression profiles is cheap and, with the possibility of thresholding the score, one can express confidence in the results, possibly eliminating fluke results.

Despite the above-mentioned problems in target prediction, the main challenge in mRNA and miRNA data integration is different. The relationship between miRNAs and mRNAs is many-to-many, a miRNA binds to different mRNAs, while an mRNA molecule hosts binding sites for different miRNAs. Moreover, the binding can be, and often is, imperfect; with a miRNA binding only partly to its target site. One miRNA can, in addition, potentially bind to multiple locations on one mRNA. Due to all of these aspects and the fact that the mRNA-miRNA interaction itself is far from being fully understood, mRNA-miRNA data integration is a non-trivial task. Simply merging mRNA and miRNA probesets (Lanza et al., 2007) may increase current difficulties in GE classification, such as overfitting caused by the immense number of features. Hence, a smart method of reasonable integrating miRNA and mRNA features is desired.

The authors of (Huang et al., 2011) present 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 approach does not address the method of breaking 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 (Peng et al., 2009) and (Tran et al., 2008), where the modules are searched as maximal bi-cliques or induced as decision rules respectively. But none of these methods offer an intuitive way to *express* the identified modules within the sample set. Contrariwise, (Kim et al., 2012) provides a black box integration procedure for several data sources like mRNAs, miRNAs, methylation data etc., with an immediate classification output. Nevertheless, this method contains no natural interpretation of the learned predictive models, which is unsuitable for an expert decision-making tool.

In this work, we propose a novel feature extraction and data integration method for the accurate and interpretable classification of biological samples based on their mRNA and miRNA expression profiles. The main idea is to use the knowledge of miRNA targets and better approximate the actual protein amount synthesized in the sample. The raw mRNA and miRNA expression features become enriched or replaced by new aggregated features that model the mRNA-miRNA regulation instead. The sample profile presumably gets closer to the phenotype being predicted. The proposed subtractive aggregation method directly implements a simple mRNA-miRNA interaction model in which mRNA expression is modified using the expression of its targeting miRNAs. A similar approach has already been demonstrated in (Anděl et al., 2013), where we employ matrix factorization proposed in (Zhang et al., 2011) instead. In comparison to the subtractive method under study, the matrix-factorization approach leaves room for developing features corresponding to larger functional co-modules, but it could overfit training data when dealing with a small number of samples.

The method widely used for analyzing associations between two heterogeneous genome-wide measurements acquired on the same cohort is canonical correlation analysis (CCA) (Pollack et al., 2002; Stranger et al., 2007; Witten and Tibshirani, 2009). CCA is applicable for mRNA and miRNA expression integration. However, CCA is based purely on mutual correlation between distinct feature sets and disregards prior knowledge of their interaction. It rather aims to describe or simplify the underlying

data, while we focus on prediction of the decrease of respective protein level due to inhibition that does not primarily manifest in correlation. In (Li et al., 2012) the authors model heterogeneous genomic data by the means of sparse regression. The method explains mRNA matrix through decomposition into miRNA expression, copy number value and DNA methylation matrices. It follows similar descriptive goals as CCA.

Incentive for our method design comes from probe sessions performed on patients with myelodysplastic syndrome (MDS). MDS is a heterogeneous group of clonal hematological diseases characterized by ineffective hematopoiesis originating from hematopoietic stem cells (Vašíková et al., 2010). Patients with MDS usually develop severe anemia (or other cytopenias) and require frequent blood transfusion. MDS is also characterized by a high risk of transformation into secondary acute myeloid leukemia, and thus could serve as a model for the research of leukemic transformation.

Of the different cytogenetic abnormalities found in MDS, deletion of the long arm of chromosome 5 (del(5q)) is the most common aberration. MDS with isolated del(5q) exhibits a distinct clinical profile and a favorable outcome. Lenalidomide is a relatively new and potent immunomodulatory drug for the treatment of patients with transfusion-dependent MDS with del(5q). It has pleiotropic biologic effects, including a selective cytotoxic effect on del(5q) myelodysplastic clones. As miRNAs serve as key regulators of many cellular processes including hematopoiesis, a number of miRNAs have been also implicated in the pathophysiology of MDS (Rhyasen and Starczynowski, 2012; Dostalova Merkerova et al., 2011).

The paper is organized as follows. Section 2 describes the proposed subtractive method (SubAgg) including its SVD-based modification (SVDAgg) that enables different subtractive weights for different miRNAs. Section 3 describes the MDS domain, defines the learning tasks and summarizes the experimental protocol. Section 4 provides experimental results. Section 5 concludes the paper.

2 MATERIALS AND METHODS

This section covers the procedures proposed for the integration of mRNA and miRNA data. First, inputs required for correct functionality of the methods are defined in Section 2.1. Then dataset merge, a simple integration technique serving as a benchmark, is presented in Section 2.2. The new integration method, subtractive aggregation is presented in Section 2.3.

2.1 Inputs

The integration method requires two datasets; one containing mRNA measurements, and one containing miRNA measurements. Those two datasets must be matched; i.e., both must contain samples taken from the same patients and the same tissue types.

Let $\mathcal{G} = \{g_1, ..., g_n\}$ be the genes, $\mathcal{R} = \{r_1, ..., r_m\}$ be the miRNAs and $\mathcal{S} = \{S_1, ..., S_s\}$ be the interrogated samples (tissues, patients, experiments). Then $x^G : \mathcal{G} \times \mathcal{S} \to \mathbb{R}$ is the amount of respective mRNA measured by mRNA chip in particular samples, and $x^\mu : \mathcal{R} \times \mathcal{S} \to \mathbb{R}$ is the expression profile of known miRNA sequences; i.e., the amount of respective molecules measured by the miRNA chip within the samples.

For further reference, the mRNA dataset will be denoted as an $s \times n$ matrix $\mathbf{X}^G$, with s samples and n genes. Similarly, the miRNA dataset will be referred to as an $s \times m$ matrix $\mathbf{X}^\mu$, with m miRNAs. Henceforth, column vectors of the two data matrices, $\{\mathbf{x}_i^G\}_{i=1}^n$ and $\{\mathbf{x}_i^\mu\}_{i=1}^m$, represent measured expression of particular genes and miRNAs, respectively.

The integration method requires information pertaining to which miRNA targets which mRNA. The known miRNA-gene control system is represented by binary relation $\mathcal{T} \subset \mathcal{R} \times \mathcal{G}$.

2.2 Dataset Merge

The most straightforward method of obtaining integrated mRNA and miRNA data is merging the two respective datasets. This method, as mentioned above, was presented by (Lanza et al., 2007) and is included in our experimental evaluation as a benchmark. The resulting *merged* dataset simply contains column-wise concatenated mRNA and miRNA data matrices. The advantages of this integration approach are no required prior knowledge of targets and computational efficiency. Excluding prior knowledge of targets, however, means that the target relationships are to be induced empirically by the classifier itself. The question remains as to whether the classifier is capable of doing that. Also, this approach increases the already-high number of features.

2.3 Subtractive Aggregation (SubAgg)

Due to the fact, that many aspects of miRNA-mRNA interactions are not yet fully understood and remain unclear, we were forced to involve several simplifying assumptions as follows: 1) miRNA effect is strictly subtractive, 2) the measured miRNA amount is proportionally distributed among its targets, and 3) the

mRNA inhibition rate is proportional to the amount of available targeting miRNA.

The method aggregates mRNA and miRNA values by subtracting a proportion of miRNA expression values from their respective target mRNAs. At the same time, it minimizes the number of parameters needed to be learned to 1. This characteristic complies with the inconvenient sample set size and the feature set size rate.

Each gene, or rather its mRNA transcript, $g \in \mathcal{G}$ has a defined set of miRNAs which target it, $\mathcal{R}_g \subset \mathcal{R}$. Conversely, each miRNA $r \in \mathcal{R}$ has a defined set of mRNAs which it targets, $\mathcal{G}_r \subset \mathcal{G}$. Let, x_g^G be the amount of mRNA measured for respective gene g in an arbitrary tissue sample and x_r^μ be the amount of particular miRNA r measured in an arbitrary sample. Let p_r be the proportion of the amount of $r \in \mathcal{R}_g$ used to regulate the expression of gene g and σ be a coefficient representing the strength of the inhibition of mRNA by miRNAs. Since the process is considered to be strictly subtractive, the aggregated value representing the inhibited mRNA of gene g, denoted x_g^{sub} would be obtained by subtracting as follows:

$$x_g^{sub} = x_g^G - \sigma \sum_{r \in \mathcal{R}_g} p_r x_r^\mu. \quad (1)$$

This equation takes an above-mentioned simplified view of inhibition of the gene by all targeting miRNAs. Hence, proportion p_r is defined as a ratio of x_g^G to the sum of levels of all targeted mRNAs. The inhibition equation is then expanded:

$$x_g^{sub} = x_g^G - \frac{c}{|\mathcal{R}_g|} \sum_{r \in \mathcal{R}_g} \frac{x_g^G}{\sum_{t \in \mathcal{G}_r} x_t^G} x_r^\mu. \quad (2)$$

Further, the parameter σ has been expanded in Equation (2). The strength of inhibition is an unknown value, but needs to be somehow represented nonetheless. In this method, it is modeled as the product of a real parameter c and a normalizer defined as $1/|\mathcal{R}_g|$. The real parameter c represents the unknown strength of the relationship and its values are subject to experimentation. Intuitively, the larger c is, the more prominent the miRNA data are (larger c amplifies the inhibition). The c parameter can be set uniformly for all genes, or alternatively, different c values may be employed for different mRNAs. Concerning the limited sample sets and the risk of overfitting, we worked with the uniform c for all mRNAs. Its setting is further discussed in the experimental part of the paper.

It is possible to obtain the overall data matrix $\mathbf{X}^{sub}$ of inhibited mRNA by iteratively updating the submatrix $X^G_{1\ldots s,\mathcal{G}_r}$, thus calculating all x_g^{sub} in Equation (2) pertaining to one miRNA and all samples in one step. Henceforth, the implementation of Equation (2) is iterated over particular miRNAs, as there are far fewer miRNAs than mRNAs:

$$X^{sub}_{1\ldots s,\mathcal{G}_r} = X^G_{1\ldots s,\mathcal{G}_r} - c\Delta(\mathbf{u})\Delta(\mathbf{x}_r)\Delta(\mathbf{s})^{-1} X^G_{1\ldots s,\mathcal{G}_r}, \quad (3)$$

where $\Delta(\mathbf{v})$ denotes a diagonal matrix, whose (i,i)-th item is equal to the i-th value of a vector $\mathbf{v}$; $\mathbf{u}$ is a vector containing the number of targeting miRNAs for each mRNA and $\mathbf{s} = X^G_{1\ldots s,\mathcal{G}_r}\mathbf{1}_{|\mathcal{G}_r|}$ is a vector of mRNA value sums pertaining to miRNA targets in all samples.

2.4 SVD-based Aggregation (SVDAgg)

The aim of an integration method in general, is to reduce the mRNA vectors and their respective targeting miRNAs into one aggregated feature. A common known method which reduces data, preserving as much useful information contained in the non-reduced vectors as possible, is *Singular value decomposition* (SVD) (Eckart and Young, 1936).

The second method we propose, SVD-based aggregation, is based on the idea that targeting miRNAs of each gene can be represented in one dimensional basis space. So, for each gene g, the expression data submatrix $\mathbf{X}^\mu_{1\ldots s,\mathcal{R}_g}$, referring to the respective targeting miRNAs, is projected into its first singular vector:

$$\mathbf{x}^{\mu,\mathbf{svd}} = \mathbf{X}^\mu_{1\ldots s,\mathcal{R}_g}\mathbf{V}_{1\ldots|R_g|,1}, \quad (4)$$

where $\mathbf{V}$ is the singular vector matrix of targeting miRNAs. Vector $\mathbf{x}^{\mu,\mathbf{svd}}$, the new representative of targeting miRNAs, is then joined to the respective mRNA vector, and reduced into one dimensional space again:

$$\mathbf{x}_g^{G,svd} = \left[\mathbf{x}_g^G, \mathbf{x}^{\mu,svd}\right]\mathbf{U}_{1\ldots 2,1}, \quad (5)$$

where $\mathbf{U}_{1\ldots 2,1}$ is the first singular vector of the two concatenated vectors.

The new feature $\mathbf{x}_g^{G,svd}$, a virtual profile comprising the gene and its miRNAs, is computed in two steps. The reason for not aggregating the mRNA vector and respective miRNA vectors together at the same time follows. Such an alternative approach gives almost all the power to the miRNAs, and since they would constitute a majority of vectors, SVD would have a tendency to disregard the information contained in the mRNA, which is necessary to avoid. Moreover, this effect would increase with the increasing number of targeting miRNAs.

3 EXPERIMENTS

This section describes the MDS domain, defines the learning tasks, and summarizes the experimental protocol.

3.1 Datasets

The data were acquired in collaboration with the Institute of Hematology and Blood Transfusion in Prague. Illumina miRNA (Human v2 MicroRNA Expression Profiling Kit, Illumina, San Diego, USA) and mRNA (HumanRef-8 v3 and HumanHT-12 v4 Expression BeadChips, Illumina) expression profiling were used to investigate the effect of lenalidomide treatment on miRNA and mRNA expression in bone marrow (BM) CD34+ progenitor cells and peripheral blood (PB) CD14+ monocytes. Quantile normalization was performed independently for both the expression sets, the datasets were scaled to have the identical median of 1 then. 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 expression level of particular miRNAs. The measurements were conducted on 75 tissue samples categorized according to the following conditions: 1) tissue type: peripheral blood monocytes vs. bone marrow cells, 2) presence of MDS and del(5q), 3) lenalidomide treatment stage: before treatment (BT) vs. during treatment (DT). Henceforth, the samples can be broken into 10 categories. The categories, along with the actual number of samples, are shown in Table 1:

Table 1: The overview of MDS classes.

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

The domain experts defined 16 binary classification tasks with a clear diagnostic and therapeutic motivation. There are 8 tasks for each tissue type, the tissue types are encoded in the task names, while the numbers of samples are shown in parentheses. The afflicted group comprises all MDS patients regardless their treatment status.

1. **PB1**: healthy (10) × afflicted (31),
2. **BM1**: healthy (10) × afflicted (24),
3. **PB2**: healthy (10) × BT (13),
4. **BM2**: healthy (10) × BT (17),
5. **PB3**: healthy (10) × BT with del(5q) (9),
6. **BM3**: healthy (10) × BT with del(5q) (11),
7. **PB4**: healthy (10) × DT (18),
8. **BM4**: healthy (10) × DT (7),
9. **PB5**: afflicted: del(5q) (9) × non del(5q) (22),
10. **BM5**: afflicted: del(5q) (8) × non del(5q) (16).
11. **PB6**: healthy (10) × DT del(5q) (13),
12. **BM6**: healthy (10) × DT del(5q) (5),
13. **PB7**: healthy (10) × BT non del(5q) (4),
14. **BM7**: healthy (10) × BT non del(5q)(6),
15. **PB8**: del(5q): BT (9) × DT (13),
16. **BM8**: del(5q): BT (11) × DT (5).

3.2 Prior Knowledge

Considering prior knowledge, we downloaded the interactions between genes and miRNAs from publicly available databases. TarBase 6.0, strives to encompass as many miRNA-mRNA validated targeting relations scattered in literature as possible. The database, maintained by DIANA Lab, was built utilizing text-mining-assisted literature curation – literature covering the discovery of new target relationships were downloaded in XML format from MedLine, processed using text mining and the resulting candidates for addition to the database were reviewed before the actual entry by the curators (DIANA Lab personnel). Its respective target matrix, filtered so as to contain solely human data, covers 228 miRNAs, 11,996 mRNAs and 20,107 target relationships between them. When selecting only the mRNAs and miRNAs available in the actual chip probesets and carefully translating and unifying miRNA identifiers using miRBase (Kozomara and Griffiths-Jones, 2011), the TarBase covers 179 miRNAs, 8,188 mRNAs and contains 14,404 target relationships.

The miRWalk database (Dweep et al., 2011), comprises both validated and predicted targets. In our experiments, only the predicted target database is used; the entries in the validated target database are already included in TarBase 6.0. Since, according to the authors, no target prediction algorithm consistently achieves better results than the others, the predicted target database includes not only targets obtained using the eponymous miRWalk algorithm, but also targets provided by other prediction algorithms. Our experiments use five of them, which are outlined in Section 1. The predicted targets dataset used in the

experiments was obtained from miRWalk by merging the results of multiple queries on the mRNA targets of canonically-named miRNAs present in the experimental miRNA dataset. Each query consisted of up to 20 miRNAs (limit imposed by the miRWalk site), each query was restricted to targets in the 3' UTR region with p-value less or equal to 0.01. The resulting dataset obtained contains 392 miRNAs, 14,550 mRNAs and 89,402 unique predicted human miRNA-mRNA target relationships. 389 miRNAs, 12,847 mRNAs and 79,014 relationships were applicable in terms of our actual mRNA and miRNA probesets.

The merged target dataset concatenates both the above-mentioned resources. It is further referred to as the extended predicted database and contains 93,325 target relationships.

3.3 Experimental Protocol

The main aim of the experiments is to verify whether the features, extracted by prior knowledge, can improve classification quality. Since we deal with classes of different sizes, we use the Mathews correlation coefficient as a balanced quality measure. It returns a value between -1 and +1, +1 represents a perfect match between annotation and prediction. We employ three benchmarking feature sets to tackle this issue. The first contains mRNA profiles only, the second takes purely miRNA profiles, and the third concatenates them as described in Section 2.2. The knowledge-based feature sets denoted as SubAgg and SVDAgg take the merged feature set and concatenate it with the aggregated features obtained in Equations (2), (3) and Equation (5) respectively.

We used 5 times repeated stratified 5-fold cross validation to assess the performance of the proposed methods as well as their benchmarking counterparts. The whole learning workflow was implemented in R environment.

SVDAgg has no parameters, SubAgg has the inhibition strength parameter c that needs to be optimized. The most straightforward way is to set it to 1 relying purely on mRNA and miRNA expression normalization. However, the absolute mRNA and miRNA expression values can hardly be directly matched. Moreover, the relative predictive power of mRNA and miRNA feature sets varies for different tasks. That is why we tuned the optimal value of c in terms of internal cross-validation. The parameter values 10^k, k$\in\{$-2,1,0,1,2$\}$ were concerned, the best value was taken in each experimental setting and fold uniformly for all mRNAs.

In order to keep a reasonable number of features, to minimize overfitting, and maintain the constant number of features across different feature sets in terms of one learning task, we applied the well-known feature selection method SVM-RFE (Guyon et al., 2002). In each of the learning tasks, the size of the reduced feature set was established as follows. We found the number of active mRNAs and the number of active miRNAs, and took their minimum. This value served as the target feature set size for mRNA, miRNA, merged and both subtractive classifiers.

We deal with 8 binary MDS tasks defined in Section 3.1. At the same time, we have two distinct target relations (validated and extended) as described in the previous section. These target relations have different domains and ranges, the domain and range of the validated target relation make subsets of their extended counterparts. As the aggregated features concern purely the domain miRNAs and the range mRNAs we filter out the rest of mRNA and miRNA profiles from the benchmarking datasets as well. This is done in order to make the comparison of classifier performance on benchmarking datasets more relatable and better identify the potential asset of the target relationships. The absolute score is not important, the main issue is the relative comparison in terms of a single learning task. In this way, 64 different experimental settings originate (2 tissue types $\times$ 8 task definitions $\times$ 2 target relations $\times$ 2 classification algorithms). The settings are independent between tissue types, however, they deal with overlapping sample and feature sets within the same tissue type.

We employed two diverse classification algorithms to avoid the dependence of experimental results on a specific choice of learning method. Support Vector Machine (SVM) with a linear kernel and the regularization parameter $C = 1$ was taken as the first option. SVM prevails in predictive modeling of gene expression data and is usually associated with high resistance to noise in data. C setting proves robust even when learning with many relevant features (Joachims, 1998). Naïve Bayes (NB) is a simple and interpretable classifier.

4 RESULTS

The individual feature sets were tested and compared under all the experimental settings defined above. The results reached are available in Tables 2 and 3; each table summarizes the results achieved by one of the classification algorithms.

The following direct observations can be drawn from the result tables. There are settings that can be perfectly solved by either the mRNA or miRNA profiles. Then, there are settings with incomparable

score reached with the mRNA and miRNA feature set. Naturally, these settings are not suitable for any integration including the concatenation as this integration can hardly outperform the better of the raw feature sets. These settings can be a priori identified and omitted from the integration procedure, or the procedure can be parametrized in such a way that the inferior dataset has no influence on the final feature set (e.g., *c* parameter in SubAgg is set to 0).

On the other hand, when dealing with mRNA and miRNA profiles of comparable predictive power, the integration improves classification performance. In general, the knowledge-based methods outperform their concatenation benchmark. As already mentioned, we deal with dependent tasks and settings while traditional hypothesis testing asks for independence. That is why we cannot apply Wilcoxon, Friedman, or other classical tests. Instead, the methods are sorted and ranked according to their pair-wise comparison in each of the particular settings; Figure 1 provides an overall comparison graph and the last row of result tables gives the ranks averaged across all the settings. The comparison suggests that the knowledge-based feature sets dominate the rest of the feature pool.

Another useful comparison measure is the overall number of occurrences, denoted as *synergies*, in which the knowledge based features outperform both raw feature sets. The presented results show 31 and 26 synergies occurred in the case of SubAgg and SVDAgg methods respectively; only 10 synergies can be observed in the case of the benchmark integration. In the other words, when dealing with settings that cannot be perfectly solved by the original features, the knowledge based integration helps.

SVM turns out to be a better choice than naïve Bayes. Let us stress that the choice of target type (validated, extended) may seem to largely affect classification quality; however, the main reason for this difference lies in the filtering mentioned in Section 3.3. The validated and extended runs cannot be directly compared (validated clearly worse than extended). The relative comparison between the merged and the other knowledge-based methods suggests that when including the predicted targets into the aggregation, no clear improvement can be observed.

Table 2: The SVM results in terms of MCC. T stands for the tissue type, # for the task ID, R for the target relation (V means validated and E extended), mR for mRNA, miR for miRNA, mer for merged, Sub stands for SubAgg and SVD for SVDAgg. The last row gives average ranking of each feature set; the lower the rank, the better.

T	#	R	Feature set				
			mR	miR	mer	Sub	SVD
PB	1	E	0.962	0.653	0.962	1.000	1.000
PB	2	E	0.983	0.881	0.983	0.983	1.000
PB	3	E	0.979	0.812	0.979	1.000	1.000
PB	4	E	1.000	0.801	1.000	1.000	0.970
PB	5	E	0.860	0.969	0.891	0.891	0.938
PB	6	E	1.000	0.823	1.000	1.000	0.983
PB	7	E	0.755	0.861	0.719	0.826	0.791
PB	8	E	0.621	0.487	0.564	0.562	0.644
BM	1	E	0.972	0.921	0.986	0.959	0.959
BM	2	E	0.911	0.954	0.911	0.939	0.925
BM	3	E	0.944	0.981	0.944	0.944	0.944
BM	4	E	0.976	0.883	0.976	0.952	0.838
BM	5	E	0.732	0.907	0.773	0.816	0.913
BM	6	E	0.882	0.853	0.882	0.911	0.853
BM	7	E	0.947	0.974	0.974	0.947	1.000
BM	8	E	0.574	0.541	0.539	0.434	0.448
PB	1	V	0.962	0.781	0.962	1.000	0.987
PB	2	V	0.983	0.827	0.983	0.983	0.916
PB	3	V	0.938	0.771	0.979	1.000	0.959
PB	4	V	1.000	0.761	1.000	1.000	0.970
PB	5	V	0.860	0.939	0.891	0.891	0.891
PB	6	V	1.000	0.895	1.000	1.000	1.000
PB	7	V	0.719	-0.156	0.645	0.791	0.607
PB	8	V	0.621	0.506	0.620	0.523	0.586
BM	1	V	0.972	0.854	0.986	0.959	0.986
BM	2	V	0.897	0.874	0.911	0.954	0.954
BM	3	V	0.963	0.909	0.963	1.000	0.981
BM	4	V	0.976	0.856	0.976	0.952	1.000
BM	5	V	0.694	0.893	0.732	0.773	0.717
BM	6	V	0.911	0.795	0.882	0.911	0.941
BM	7	V	0.921	0.896	0.974	0.974	0.947
BM	8	V	0.574	0.607	0.501	0.541	0.574
Avg. ranking			3.08	3.81	2.88	2.56	2.67

5 CONCLUSIONS

Molecular classification of biological samples based on their expression profiles represents a natural task. However, the task proved conceptually difficult due to the inconvenient rate of the sample and feature set sizes and complexity and heterogeneity of the expression process. These characteristics often cause overfitting. Classifiers do not sufficiently generalize; instead of revealing the underlying relationships, they capture perturbations in training data. This problem can be minimized by regularization; i.e., introduction of additional knowledge. The regularized models should be more comprehensible and potentially more accurate than standard models based solely on a large amount of raw measurements.

The integration of heterogeneous measurements and prior knowledge is non-trivial, though. In this paper we proposed the subtractive method that aggregates mRNA and miRNA values by subtracting a proportion of miRNA expression values from their respective target mRNAs. The method simplifies the mRNA-miRNA interaction and minimizes the number of parameters needed to be learned to 1. We also proposed another integration method that can be perceived as an extension that enables different subtrac-

Table 3: The NB results in terms of MCC. The header symbols have the same meaning as in Table 2.

T	#	R	Feature set				
			mR	miR	mer	Sub	SVD
PB	1	E	0.840	0.732	0.840	0.840	0.799
PB	2	E	0.865	0.846	0.881	0.849	0.898
PB	3	E	0.845	0.689	0.845	0.900	0.881
PB	4	E	0.831	0.732	0.831	0.844	0.816
PB	5	E	0.794	0.984	0.858	0.875	0.860
PB	6	E	0.824	0.770	0.824	0.877	0.965
PB	7	E	0.645	0.791	0.645	0.645	0.965
PB	8	E	0.322	0.429	0.303	0.308	0.523
BM	1	E	0.929	0.915	0.944	0.930	0.943
BM	2	E	0.953	0.953	0.953	0.873	0.984
BM	3	E	0.909	0.981	0.909	0.909	0.963
BM	4	E	0.976	0.787	0.976	0.976	0.793
BM	5	E	0.593	0.871	0.732	0.713	0.753
BM	6	E	1.000	0.707	0.795	0.941	0.707
BM	7	E	0.870	0.866	0.870	0.870	0.896
BM	8	E	0.320	0.399	0.383	0.405	0.268
PB	1	V	0.853	0.551	0.853	0.880	0.880
PB	2	V	0.833	0.627	0.849	0.865	0.965
PB	3	V	0.881	0.641	0.881	1.000	1.000
PB	4	V	0.863	0.685	0.863	0.876	0.844
PB	5	V	0.842	0.937	0.842	0.875	0.860
PB	6	V	0.805	0.627	0.805	0.895	0.911
PB	7	V	0.861	0.826	0.895	0.965	0.826
PB	8	V	0.322	0.210	0.303	0.362	0.523
BM	1	V	0.886	0.887	0.915	0.930	0.929
BM	2	V	0.953	0.828	0.953	0.904	0.954
BM	3	V	0.909	0.908	0.909	0.888	0.944
BM	4	V	0.951	0.882	0.951	0.976	0.909
BM	5	V	0.578	0.815	0.694	0.733	0.694
BM	6	V	0.941	0.853	0.882	0.970	0.911
BM	7	V	0.870	0.872	0.870	0.870	0.870
BM	8	V	0.153	0.201	0.187	0.272	0.361
Avg. ranking			3.41	3.72	3.17	2.36	2.34

tive weights for different miRNAs; the weights are learned by SVD.

In this work we classified myelodysplastic syndrome patients under 64 experimental settings. We compared five types of feature sets. Two of them represented raw homogeneous expression measurements (mRNa and miRNA profiles), the third implemented their straightforward concatenation, and the last two resulted from SubAgg and SVDAgg integration. The comparison suggests that the knowledge-based feature sets dominate the rest of the feature pool, and the features resulting from the mRNA-miRNA target relation can improve classification performance.

There is still a lot of future work. More problem domains need to be considered. The prior knowledge should be extended to cover the gene regulatory network (the protein-protein interactions, interactions between genes, and their transcription factors). Another challenge is to employ epigenetic data, namely DNA methylation. Concerning the algorithmic issues, we intend to develop another parameter-free integration method where the prior knowledge controls pseudorandom construction of weak classifiers vaguely corresponding to the individual biological processes. The weak classifiers will later be merged into an ensemble classifier.

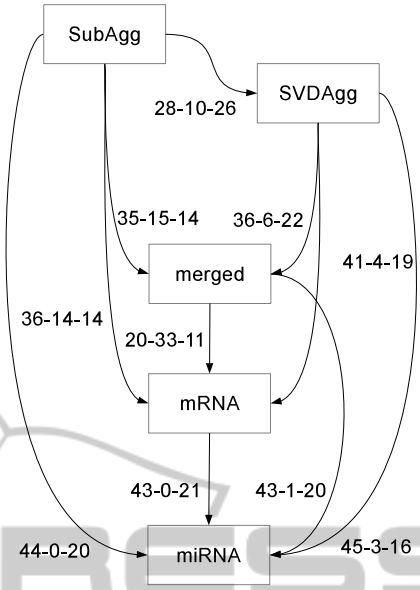

Figure 1: Pair-wise classification comparison graph. The nodes represent particular feature sets, an edge from node *a* to node *b*, annotated as *x*-*y*-*z* means that method *a* outperforms method *b* in *x* experiments, in *y* ties and in *z* losses.

ACKNOWLEDGEMENTS

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

REFERENCES

Anděl, M., Kléma, J., and Krejčík, Z. (2013). Integrating mRNA and miRNA expressions with interaction knowledge to predict myelodysplastic syndrome. In *ITAT 2013: Information Technologies – Applications and Theory, Workshop on Bioinformatics in Genomics and Proteomics*, pages 48–55.

Brewster, J. L., Beason, K. B., Eckdahl, T. T., et al. (2004). The microarray revolution: Perspectives from educators. *Biochem Mol Biol Educ*, 32(4):217–27.

Croce, C. M. (2009). Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet*, 10(10):704–14.

Dostalova Merkerova, M., Krejcik, Z., Votavova, H., Belickova, M., Vasikova, A., and Cermak, J. (2011). Distinctive microRNA expression profiles in CD34+ bone marrow cells from patients with myelodysplastic syndrome. *Eur J Hum Genet*, 19(3):313–9.

Dweep, H., Sticht, C., Pandey, P., et al. (2011). miRWalk - Database: Prediction of possible miRNA binding sites by ”walking” the genes of three genomes. *J Biomed Inform*, 44(5):839–47.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–8.

Fabian, M. R. and Sonenberg, N. (2012). The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat Struct Mol Biol*, 19(6):586–93.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422.

Huang, G. T., Athanassiou, C., and Benos, P. V. (2011). mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic acids res*, 39(Web Server issue):W416–23.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–42, Berlin. Springer.

Kim, D., Shin, H., Song, Y. S., et al. (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform*, 45(6):1191–8.

Kozomara, A. and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39(Database-Issue):152–7.

Krek, A., Grün, D., Poy, M. N., et al. (2005). Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500.

Lanza, G., Ferracin, M., Gafà, R., et al. (2007). mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. *Mol Cancer*, 6:54+.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–54.

Lewis, B. P., Shih, I.-h. H., Jones-Rhoades, M. W., et al. (2003). Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98.

Li, W., Zhang, S.-H., et al. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, 28(19):2458–66.

Morin, R., Bainbridge, M., Fejes, A., et al. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1):81–94.

Peng, X., Li, Y., Walters, K. A., et al. (2009). Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*, 10(1):373+.

Pollack, J. R., Sørlie, T., Perou, C. M., et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–8.

Rhyasen, G. W. and Starczynowski, D. T. (2012). Deregulation of microRNAs in myelodysplastic syndrome. *Leukemia*, 26(1):13–22.

Sayed, D. and Abdellatif, M. (2011). MicroRNAs in development and disease. *Physiol Rev*, 91(3):827–87.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7.

Stranger, B. E., Forrest, M. S., Dunning, M., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–53.

Tan Gana, N. H., Victoriano, A. F., and Okamoto, T. (2012). Evaluation of online miRNA resources for biomedical applications. *Genes to Cells*, 17(1):11–27.

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

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

Wang, X. and Naqa, I. M. E. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, 24(3):325–32.

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*, 8(1):28.

Zhang, S.-H., Li, Q., et al. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, 27(13):401–409.