Network-constrained Forest for Regularized Omics Data Classification

M. Anděl[a], J. Kléma[a,*], Z. Krejčík[b]

[a] *Department of Computer Science, Czech Technical University, Technická 2, Prague, Czech Republic*
[b] *Department of Molecular Genetics, Institute of Hematology and Blood Transfusion, U Nemocnice 1, Prague, Czech Republic*

Abstract

Contemporary molecular biology deals with wide and heterogeneous sets of measurements to model and understand underlying biological processes including complex diseases. Machine learning provides a frequent approach to build such models. However, the models built solely from measured data often suffer from overfitting, as the sample size is typically much smaller than the number of measured features. In this paper, we propose a random forest-based classifier that reduces this overfitting with the aid of prior knowledge in the form of a feature interaction network. We illustrate the proposed method in the task of disease classification based on measured mRNA and miRNA profiles complemented by the interaction network composed of the miRNA-mRNA target relations and mRNA-mRNA interactions corresponding to the interactions between their encoded proteins. We demonstrate that the proposed network-constrained forest employs prior knowledge to increase learning bias and consequently to improve classification accuracy, stability and comprehensibility of the resulting model. The experiments are carried out in the domain of myelodysplastic syndrome that we are concerned about in the long term. We validate our approach in the public domain of ovarian carcinoma, with the same data form. We believe that the idea of a network-constrained forest can straightforwardly be generalized to-

*Corresponding author
Email addresses: andelmi2@fel.cvut.cz (M. Anděl), klema@fel.cvut.cz (J. Kléma), zdenek.krejcik@uhkt.cz (Z. Krejčík)

wards arbitrary omics data with an available and non-trivial feature interaction network.

Keywords: omics data, microRNA, machine learning, random forest, domain knowledge

2010 MSC: 00-01, 99-00

1. Introduction

Onset and progression of heterogeneous multifactorial diseases depend on a combination of defected or altered genes, which is often too overly complex to be deciphered from an individual's genome only; instead it can be better manifested during the expression of genes [1]. *Gene expression* (GE) is the overall process by which information from a genome is transferred towards anatomical and physiological characteristics generally called *phenotype*. During the process, a gene is transcribed into the molecule of *messenger* RNA (mRNA), subjected to several transcription and translational regulatory mechanisms, and usually translated into a protein. The final protein level strongly afflicts the phenotype. Any dysfunction during the whole process may easily cause a disease.

The expression of a gene can be quantified as an abundance of gene transcript during its expression process. Current progress in high-throughput technologies such as microarrays and RNA sequencing enables affordable measurement of wide-scale gene expression on the *transcriptome* level. Therefore, the expression of thousands of genes can all be measured at once in each sample. One may thus feel capable of predicting disease outcome, progress or treatment response based on acquired GE data [2]. The phenotype prediction stems from the simplified assumption that a higher amount of detected mRNA implies a higher amount of translated protein, and therefore a higher manifestation of the respective gene. Phenotype prediction based on GE data is a natural learning task. However, many instances of this task become non-trivial within currently available GE data. The data are noisy and a small sample size together with an immense number of redundant features often leads to overfitting.

Gene expression can be seen as a complex dynamic process with many stages, components and regulatory mechanisms. A phenotype is not afflicted by particular genes separately, but there is a concert of genes involved in the expression process. The expression activities of genes are often indirectly linked together by interactions between respective proteins. The *protein-protein interactions* [3] may be involved in transporting and metabolic pathways, or in constitution of *protein complexes*. Another component of the *gene network* are the interactions between microRNAs and their target genes [4].

MicroRNAs (miRNAs) [5] serve as a component of the complex machinery which eukaryotic organisms use to tune protein synthesis. They are short ($\sim$21 nucleotides) noncoding RNA sequences which mediate post-transcriptional repression of mRNA via RNA-induced silencing complex (RISC), where miRNA serve as a template for recognizing complementary mRNA. The complementarity level of miRNA-mRNA binding initiates one of two possible mechanisms: the complete homology triggers *degradation* of target mRNA, whereas a partial complementarity leads to translational *inhibition* of target mRNA [6]. The level of miRNA expression can be measured by (e.g.) miRNA microarrays, analogically to mRNA profiling. The interactions between miRNAs and their target mRNAs, as well as interactions between proteins, are experimentally assessed in vitro or algorithmically predicted based on the structural properties of interacting molecules.

Since the journey from a genome to its phenotype manifestation is so complex and nontrivial, current trends in gene expression data analysis aim toward the integration of multiple measurement types from multiple stages of the gene expression process [7], acquired from the same set of tissues. Such an integrative analysis should provide a broader view of gene expression as a whole. This work extends our previous approaches to integrate traditional mRNA and miRNA measurements in the domain of myelodysplastic syndrome data based on non-negative matrix factorization with prior knowledge [8] and subtractive aggregation for deterministic models of the inhibition effect of miRNA [9]. In this paper, we propose a new method, based on random forest framework, which

integrates heterogeneous omics features through the knowledge of their mutual interactions. Interlinking the features by their possible interactions improves the robustness and interpretability of resulting models, and improves their empirical validity in terms of classification accuracy.

The paper is organized as follows. Section 2 reviews the recent efforts on regularization with prior knowledge in ill-posed problems with special emphasis on omics data. Section 3 firstly describes the data domain and subsequent classification tasks. Then the method itself, designed for these classification tasks is sketched, while a way of interpreting resulting models is proposed. Next, the ovarian carcinoma domain used for validation as well as the format of employed domain knowledge is described. The methodology developed and used is deeply theoretically analyzed in Sect. 4. Section 5 provides experimental results in terms of empirical validity and interpretability respectively, i.e., the predictive accuracy and examples of discovered interactions along with their biological meaning. The results are then discussed in Sect. 6. Section 7 concludes the paper.

2. Related Work

Learning from GE data is a challenging task due to its complexity and heterogeneity. On top of that, the number of variables p greatly exceeds the number of observations n, we are referring to the so-called $n \ll p$ problem that leads to overfitting [10]. However, certain learning algorithms may provide promising results even in ill posed problems like this. For example, *support vector machine* (SVM) [11] is capable of dealing with a large dimensionality with sufficient generalization. However, in GE data analysis, the model itself is often just as appreciated as its output. Henceforth, SVM is more or less a black-box model, which does not provide sufficient insight. Conversely, a decision tree is easily comprehensible, but its prediction results are often weak [12]. Since GE data have a large dimensionality with few samples, there is a great number of hypotheses, often based merely on random perturbations, which can perfectly

split the data into classes, but lack generalization. Counter-intuitively, even decision stumps (one-level decision trees) are overfitted as a consequence.

The way to address overfitting in general is *regularization* [13]. Regularization restrains the space of all hypotheses to improve generalization. In terms of machine learning, the trade off between bias and variance is tuned to deliberate a smaller structural risk. Besides initial dimensionality reduction, it may be implemented geometrically as in the case of margin classifiers [14], through certain hypothesis assumptions, complexity penalization or domain knowledge. We will focus on the last approach here, in which we promote such hypotheses that are in accord with the existing knowledge.

The prior knowledge-based regularization approaches are popular in the molecular biology domain; in particular, in omics data analysis. In the most general way, the domain knowledge is encoded as conditional probability in statistical relational learning [15, 16], or as first-order predicates in inductive logic programming [17, 18]. The advantage of these approaches is the ability to tackle the knowledge from an arbitrary domain; i.e., not only omics. However, these approaches are computationally expensive in domains with a large dimension. In omics problems where the dimension commonly exceeds 10^4, it often implies substantial problem reduction in terms of pre-processing. An alternative way is to develop a specialized learning method dedicated to a certain domain, which stems from the domain functionality and its specific assumptions and integrates them into a learning framework. As an example of dedicated method see network regularized SVM and logistic regression, [19, 20] and [21] respectively, where genes related by prior known interactions are expected to contribute *similarly* to the classification function. Among others, [22] gives an overview of recent methods for the incorporation of biological prior knowledge on molecular interactions and known cellular processes into the feature selection process to improve risk prediction of patients. [23] exemplifies a tool for the incorporation of gene network data into support vector machines. [24] proposes both supervised and unsupervised learning based on spectral decomposition of gene expression profiles with respect to the eigenfunctions of the underlying gene

network graph.

Regularization through domain knowledge is not such a frequent issue in the case of ensemble classifiers. The prior knowledge model and ensemble model are often regarded as two sides of the same coin, as both try to address the generalization problem and model enhancement. However, there is no reason not to combine both. [25] uses gene ontology terms and miRNA-mRNA target relations to create an ensemble of centroid-based weak classifiers based on tree-like modules to forecast the prognosis of breast cancer patients. [26] integrates linguistic knowledge into random forest language models originally based on n-gram counts only. The author illustrates the applicability of the ensembles in morphological language models of Arabic, prosodic language models for speech recognition and a combination of syntactic and topic information in language models. [27] proposes a random forest-based method, where the building of trees is guided by a protein network. The authors proposed a procedure for the validation of network decision modules through the forest and demonstrated that the validated modules are robust and reveal causal mechanisms of cancer development. However, their search strategy most likely does not improve the classification accuracy of resulting models. [28] iteratively builds random forests through a weighted sampling of the variables taken from modules of correlated genes. They use the OOB (out-of-bag) importance estimate of each gene involved in the forest to adapt its weight and the weight of its module for the sampling session in the next iteration. Using bootstrapping with subsequent OOB assumes a sufficiently large data sample.

3. Materials and Methods

We illustrate the proposed method in the task of disease classification based on measured mRNA and miRNA profiles complemented by the interaction network composed of the miRNA-mRNA target relations and mRNA-mRNA interactions corresponding to the interactions between their encoded proteins. Here we describe the experimental protocol, two particular domains and the resources

employed for the building of the gene regulatory network used as prior knowledge. Subsequently, the way in which the ensemble is analyzed to understand the interactions among features is presented.

3.1. Domain Description

The data, provided by our collaborative lab at the Institute of Hematology and Blood Transfusion in Prague, are related to *myelodysplastic syndrome* (MDS) [29]. 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, then the datasets were scaled to have the identical median of 1. 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 samples labeled as follows. The sample was either healthy, or afflicted. If afflicted, it was further categorized according to genotype background as a presence of partial deletion of the chromosome 5 (del(5q) or non-del(5q)), and according to the stage of lenalidomide treatment, i.e. before treatment (BT), or during treatment (DT). Together with 2 types of tissue for each of these configurations it adds up to 10 categories. Informed consent was obtained from all the subjects whose samples were used for expression profiling, and the study was approved by the Scientific Board and Ethics Committee of the Institute of Hematology and Blood Transfusion in accordance with the ethical standards of the Declaration of Helsinki.

On these categories we defined 7 binary classification tasks with a clear clinical or biological interest. The tasks were to differentiate: 1) healthy samples and afflicted samples with a particular genotype and treatment stage, 2) treatment stage of afflicted samples with del(5q), and 3) genotype background (incidence

of del(5q)) of untreated samples.

3.2. Regularized Omics Data Classification

To differentiate the data mentioned above we developed new method based on random forest (RF) framework [30]. The proposed method, *Network constrained forest* (NCF), incorporates domain knowledge in terms of prior known or predicted interaction between omics features. In the given setting, we work mainly with transcriptomic data, employing protein-protein interactions and miRNA-target interactions as a domain knowledge. This regularization by gene networks is employed to increase the stability, comprehensibility and, last but not least, the accuracy of resulting models, namely when treating the usual $n \ll p$ settings. The method learns decision trees on those features that lie *close* to the candidate genes in the feature interaction network. This selection is unlike RF, which uses randomly selected predictors *in each decision node*. Instead, the NCF firstly samples a feature as a seed, potentially the candidate for causing the phenomenon under study, then it samples the rest from a probabilistic distribution over the omics network. The distribution is parametrized to certainly prefer selecting the features lying closer the seed gene. The algorithm of NCF is thoroughly depicted in Sect. 4.2.

3.3. Understanding the model

Random forests are not by far black-box models used solely for prediction. The ensembles are frequently used for variable importance estimation, feature selection or sample proximity evaluation ([31, 32, 33]). However, in the context of NCFs, variable interactions apparently represent the most interesting piece of knowledge that can be extracted from the model. In the traditional scenario that does not involve prior knowledge, the interactions are extracted purely from the measurements [34, 35]. A pair of variables is considered to interact if a split on one variable in a tree makes a split on the other variable either systematically more possible or less possible. The interactions often serve to improve the bias in variable selection stemming from variable interaction effects.

We deal with the prior set of interactions equivalent to a feature network when building the weak classifiers. Each tree belonging to the ensemble is believed to be local in terms of this feature network, i.e., to contain features that lie close to the seed gene. For this reason, we may extract the *empirical interactions* that correspond to interactions that make edges in the shortest path connecting a pair of tree neighboring nodes in the prior gene regulatory network. By searching the whole forest, we find a large number of empirical interactions that can be employed in statistical validation of the prior interaction network under the given biological conditions. The more counts a prior interaction gets in the empirical phase, the more active it seems to be in the given context. In other words, we employ the common statistical and data mining formula "data + prior knowledge $\rightarrow$ knowledge", an interaction is extracted either if it is obvious from the measurements themselves or it is contained in the prior interaction set and not invalidated by the measurements. The prior and posterior empirical interaction sets can also be compared.

3.4. Validation Domain

As the classification tasks from MDS domain (Sect. 3.1) are mutually dependent, to validate our method we used a similar omics domain data related to another genetically determined disease from an independent source. We downloaded 17,814 mRNA (Agilent 244K Custom Gene Expression G4502A-07) and 799 miRNA (Agilent Human miRNA Microarray Rel12.0) profiles with matching samples related to ovarian carcinoma (OC) from The Cancer Genome Atlas (TCGA) repository [36]. TCGA data encompasses hundreds of heterogeneous tumor samples with different clinical backgrounds. We choose the OC, which contains a sufficient number of matching mRNA and miRNA profiles, clinically well annotated. Hence, we defined two validation data sets with regards to sample homogeneity in terms of their clinical annotation and balanced class distributions. The first data set contains 58 tumor samples of high grade (G3) and late stage (Stage IV), the latter contains 64 lower grade (G2) tumor samples. The dichotomized overall survival of respective patients was chosen as a

target attribute to be learned. The survival dichotomization threshold was set to the median value of overall survival, i.e. 25 months for high grade tumor dataset and 40 months for the lower grade. This preprocessing concludes with long-term and short-term survival classes, each with 29 samples for the high grade dataset and with 31 and 33 samples, respectively, for the latter dataset.

3.5. Available Domain Knowledge

Considering domain knowledge, in terms of gene networks, we downloaded the interactions between proteins, and genes and miRNAs, from the following publicly available databases. In vitro validated miRNA-mRNA interactions were obtained from TarBase 6.0 [37], while in silico predicted relations were downloaded from miRWalk database [4]. Protein-protein interactions were obtained from Human Protein Reference Database [38] and from [3], as to the experimentally validated and algorithmically predicted interactions, respectively. Eventually, we ended up with 9,077 genes involved in 79,288 protein-protein interactions, 463 miRNAs in 92,886 miRNA-target interactions. Regarding the validation domain, we handled 8,073 genes involved in 81,067 protein-protein interactions, 417 miRNAs in 84,332 miRNA-target interactions. The candidate causal genes, a total of 145 and 220 genes associated with MDS and OC respectively, were obtained from [39].

3.6. Experimental Protocol

To validate our method we have extended an implementation of RF in Scikit-learn [40], a Python machine learning library. Then we run number of robust experiments on our NCF. Random forest, standard classification and regression tree (CART), linear SVM and naïve Bayes (NB) classifier served as benchmark learners [13]. A simple decision tree was introduced to assess generalization of forest based algorithms. Each learning algorithm was validated in 5-times repeated 10-fold stratified cross-validation. Since all the learners should be randomly initialized, we ran each of the validation processes from 10 random

seeds. It comes out to $3 \times 10 \times 5 \times \ 10 = 1500$ learning epochs. Forest based algorithms were run with 1,024 base trees.

Since we deal with classes of different sizes, we use the Mathews correlation coefficient (MCC) as a balanced quality measure. It returns a value of between -1 and +1; +1 represents a perfect match between annotation and prediction, 0 equals random prediction and 1 indicates absolute disagreement between annotation and prediction.

4. Theory

In this section, a theoretical background and motivation for our approach is supplied. Firstly, the data and used domain knowledge is formalized to be further addressed. Since our approach is based on modification of random forest by omics network, a brief description of random forest learning framework is provided along with a biological motivation for network incorporation. The pseudocode of our NCF is also presented.

4.1. Motivation

Our approach is to modify random forest by gene network regularization. Random forest [30] is a popular tree-based ensemble model. Its general idea is to reduce the variance of single deep decision tree classification by voting. The goal is to create an ensemble of decorrelated but still accurate base tree learners. The decorrelation of the base learners is reached by limiting their feature and sample sets. Namely, the trees are built on a bootstrapped sample set, while the features are randomly subsampled in each decision node. Finally, decisions of the trees are merged to adjust variance of the prediction. The diversity among particular decisions is crucial for the generalization power of the ensemble classifier. Our intention is to use prior known interactions between omics features (simplify as genes) to encourage the diversity of base decisions. The feature subset of a tree is to be constrained not only by its size (e.g., $\sqrt{p}$), but also by the existence and type of interactions between the features. The basic hypothesis claims that

network-close entities are correlated and henceforth suitable to be grouped in the same base learner to decorrelate it from its counterparts.

This intention has an intuitive biological background. A heterogeneous multifactorial disease is caused by multiple altered loci. The effect of these *causal genes* need not be clearly detectable in mRNA abundances of corresponding genes, but may be observable in the expression of interacting genes. Similarly, the effect of translational repression by miRNA may not be observable as decreasing target mRNA abundance. In the case of translation inhibition, mRNA molecule is only partially modified while the amount of respective protein may be significantly reduced. Henceforth the effect of target inhibition gets detectable in the expression of interacting genes instead. The individual trees may vaguely correspond to the individual disease factors and their network-local manifestations. The final decision merges the base learners, that can be characterized as local, stochastic, repetitive and overlapping.

4.2. Network Constrained Forest

Let $\mathcal{G} = \{g_1, ..., g_{p_G}\}$ be the genes for which expression level (actually the transcript abundance) is measured, $\mathcal{R} = \left\{r_1, ..., r_{p_\mu}\right\}$ be miRNAs also with available expression levels. Let $\mathcal{S} = \{s_1, ..., s_n\}$, be the set of samples, where expression measurements of both $\mathcal{G}$ and $\mathcal{R}$ are available, with a binary phenotype $\mathcal{P} : \mathcal{S} \to \mathbb{B}$, Then the expression set is $\mathcal{E} : \mathcal{S} \times (\mathcal{G} \cup \mathcal{R}) \to \mathbb{R}$. Let $\mathcal{I}_{\text{PPI}} \subset \mathcal{G} \times \mathcal{G}$ represent the previously reported (curated or algorithmically predicted) binary interactions between particular genes through respective proteins. Let $\mathcal{I}_\mu \subset \mathcal{R} \times \mathcal{G}$ represent the binary interactions between particular miRNAs and their target genes. By merging these units and interactions we define an omics network $\mathcal{N} = (\mathcal{V}, \mathcal{I})$ as a directed graph with omics features $\mathcal{V} = \mathcal{G} \cup \mathcal{R}$ as the vertices and mutual interactions $\mathcal{I} = \mathcal{I}_{\text{PPI}} \cup \mathcal{I}_\mu$ as the edges. Finally, let $\mathcal{C} \subset \mathcal{G}$ be the loci (particularly genes in our study) that are believed to be associated with a disease in terms of genotype (i.e., polymorphisms, mutations).

Our approach, called Network Constrained Forest (NCF), is based on a simple idea of biasing the feature sampling process towards the genes and loci

in general, which have been previously reported as candidates for causing the phenomenon being studied (typically a disease, treatment outcome, etc.), and consequently the omics features which directly or indirectly interact with those candidate genes. The bias is guided by previously known interactions between the features in a way that those closer to the candidate causal gene in the network are more likely to be chosen as candidate features in the tree whose construction is driven by the particular causal gene as a seed. Unlike random forest, NCF does not sample the features uniformly. Instead, the features are sampled from a distribution π_c, which is certainly biased towards a potentially causal gene $c \in \mathcal{C}$. The principles of NCF are outlined as a pseudocode in Alg. 1 and 2. Reduction of NCF to RF is sketched in code comments.

Algorithm 1 Pseudocode of learning NCF as an ensemble.

Input:

Dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^{|\mathcal{G} \cup \mathcal{R}|}$ and $y_i \in \mathbb{B}$,

Gene network $\mathcal{N}$, seed genes $\mathcal{C}$, number of trees T

Output:

Ensemble $forest$

1: **for all** $c \in \mathcal{C}$ **do**

2: precompute $\pi_c : \mathcal{V} \to \mathbb{R}$ based on $\mathcal{N}$ topology

3: **for** $t \leftarrow 1 \ldots T$ **do** ▷ same in RF

4: $forest[t] \leftarrow \textsc{buildTree}(\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \{\pi_c\}_{c \in \mathcal{C}}, \mathcal{C})$

The core of NCF lies in the learning of a constrained set of (omics) features for making decisions in each node of each tree in the ensemble. We will denote it by $\mathcal{F}$. The distributions π_c which $\mathcal{F}$ to be sampled from is learned based on topological properties of network $\mathcal{N}$ (see Alg. 1.2). A distribution over the features $\mathcal{V}$ is computed for each potentially causal gene, in general locus, $c \in \mathcal{C}$. In our study it is most often the set of genes previously reported as phenotype associated in terms of e.g., nucleotide variations, chromosome deletions. In the case of missing information, it is a set sampled randomly or in future extensions, a set implicitly identified during learning of NCF. Such a distribution is required

Algorithm 2 Pseudocode of learning a base estimator of NCF.

1: **function** BUILDTREE($\mathcal{D}, \{\pi_c\}_{c\in\mathcal{C}}, \mathcal{C}$)
2: randomly select seed $c \in \mathcal{C}$
3: sample features $\mathcal{F} \subset \mathcal{V}$ from π_c ▷ in case of RF $\pi_c = \frac{\sqrt{|\mathcal{V}|}}{|\mathcal{V}|}$
4: $node.separator \leftarrow g \in \mathcal{F}$ best separating $\mathcal{D}$
5: **if** $\mathcal{D}$ are inseparable **then**
6: **return** majority class
7: $node.left \leftarrow$ BUILDTREE($\{s \in \mathcal{D}$ where g is upregulated$\}, \{\pi_c\}_{c\in\mathcal{C}}, \mathcal{C}$)
8: $node.right \leftarrow$ BUILDTREE($\{s \in \mathcal{D}$ where g is downregulated$\}, \{\pi_c\}_{c\in\mathcal{C}}, \mathcal{C}$)
9: **return** *node*

to be more dense as approaching its seed c. Henceforth, it is *implicitly* defined as a random walk of length k from the seed gene c:

$$\boldsymbol{\pi}_c^k = \boldsymbol{\pi}_c^{k-1}\mathbf{W}, \tag{1}$$

where $\pi_c^0(u) = 1$ for $u = c$ and $\pi_c^0(u) = 0$ otherwise, $\mathbf{W}$ is the probability of transition from a vertex u to a vertex v, with $W_{uv} = \{deg(u)^{-1}$, if $uv \in \mathcal{I}_{\text{PPI}}$ or $vu \in \mathcal{I}_{\text{PPI}}$ or $uv \in \mathcal{I}_{\mu}$ else $0\}$, in the feature network extended with the self transitions, i.e., loops. Note that the gene-miRNA edges are *directed*. First, miRNAs have to be accessible from the causal gene. Second, the edges cannot be treated as undirected since miRNAs often have a large number of targets and constitute network hubs; i.e., the high-degree nodes leading to the small world phenomenon. To minimize the ability of reaching distant network nodes, miRNAs represent terminal nodes in terms of random walk.

Resulting distributions π_c serve for biased feature sampling within decision nodes of base trees (see Alg. 2.3). Firstly (Alg. 2.2), a seed gene $c \in \mathcal{C}$ is randomly selected. Corresponding distribution π_c is then used for weighted sampling of $\sqrt{|\mathcal{V}|}$ features (the common RF heuristic) considered to make a decision (Alg. 2.4). Note that NCF is reducible to RF by making π_c uniform

depending only on the number of omics features $|\mathcal{V}|$. The last steps are same as in the standard RF algorithm.

4.3. Learning Assumptions of NCF

In Sect. 4.2, note Alg. 2 that describes the induction of a particular tree. Let us focus on its steps 2.2 and 2.3. Since each node uses a *different* seed gene c and consequently samples a different feature set $\mathcal{F} \subset \mathcal{G}$ from a *different* distribution π_c, the tree is forced to span its decisions over the whole network $\mathcal{N}$ and not only to fit the noise of correlated features. Let us explain it as follows.

Behold a multifactorial phenotype defined by two loci $\mathcal{L}_1$ and $\mathcal{L}_2$, where the former is more phenotype related than latter, but still imperfectly. It can manifest in such a way that $\mathcal{L}_1$ gets associated with more samples than the latter locus, while both cover whole the sample set $\mathcal{S}$. Then assume $\mathcal{L}_1$ affects p_1 measurable features which are naturally correlated with each other and with $\mathcal{L}1$. Similarly, $\mathcal{L}_2$ affects p_2 measurable features. In case $p_1 > p_2$, $\mathcal{L}_1$ related features tend to prevail in the candidate sets of any decision node when *uniform* random sampling is performed. As $\mathcal{L}_1$ is more related to the phenotype than $\mathcal{L}_2$, it eventually dominates the decision nodes of any base tree. Thus, a great portion of trees is biased towards $\mathcal{L}_1$ factor and the ensemble lacks diversity. As $\mathcal{L}_1$ is only imperfectly related to the phenotype, the overall classification performance is not the best possible. This bias will not decrease even with the addition of more base trees.

This potential pitfall is addressed by NCF. Following the assumption that genes that are close in the network are correlated in their expression [41, 42], NCF samples the features from a certain neighborhood parametrized by the walk length k. The network guided sampling thus encourages the base trees being induced to make decisions according to *decorrelated* features and to make decorrelated base predictions as a consequence.

Disagreement among predictions of base learners is the key point in ensemble learning. Most of the measures evaluating the ensemble diversity before the validation are based on base classifiers disagreement on certain training samples,

often called critical samples beyond the ensemble margin [43, 44]. When there is only few examples, the diversity might not be observable within training sample. Henceforth, NCF manages the diversity *implicitly*.

4.4. Heuristic Handling of Hyperparameter

One of the key issues is finding the optimal walk length k^*. Short walks tend to generate small neighborhoods; the individual trees grow larger to fit the training data. Long walks get closer to the traditional RFs as the constrained feature sets become less dependent on the seed and the network topology. Note that NCF does not fully converge to RF for any walk length as the whole path starting in the seed is always taken. Instead, it converges to the stationary distribution of random walk, i.e. is $\pi_c^{\infty}(v) = deg(v)/|\mathcal{I}|$. This means that such a *degenerated* NCF samples the features regarding their importance in terms of their degree, i.e., the number of interconnecting features, without respecting nature of their interactions including putative correlations.

In this paper we propose a heuristic to set k based on training data only with no need of e.g. nested cross-validation. The heuristic is based on the theory of bias-variance trade-off, namely on decreasing *incidence of underfitted trees* in the forest with growing walk length. In our case, the walk is stopped one step before the empirical incidence of underfitted trees I (the percentage of trees that do not fit the training data perfectly) ceases to decrease or approaches zero:

$$k^* = \underset{k=1\ldots10}{\arg\min}(I(k+1) < \epsilon \vee I(k) - I(k+1) < \epsilon) \tag{2}$$

At this length, the neighborhood is large enough to provide sufficient accuracy, yet still small enough not to overfit the training data. The tree depth is limited to 2 here to avoid perfect fit regardless of the walk length. The value of ϵ was set to 1%. The relationship between the walk length and the underfitted tree incidence is shown in Sect. 5 and Fig. 1. The value of k adjusts the learning bias to the given problem. Usually, NCF uses fewer features than RF while reaching equal or better predictive results.

5. Results

The results illustrate an empirical evaluation of our method within 9 classification tasks. The evaluation is performed in terms of the classification accuracy plotted in Fig. 1. The graphs depict the progress of an empirical estimate of NCF classification accuracy as a function of walk length. Since the walk length k is a key learning parameter of NCF, we also plot the incidence of underfitted trees, which serves as a heuristic to assess a proper value of k^*. The MDS and ovarian cancer class definitions reached through dichotomization are available in the subfigure legends of Fig. 1. The classes in the individual tasks are encoded according to the nomenclature in Sect. 3.1 and Sect. 3.4, respectively.

The accuracy reached with the proper heuristic setting of k^* (note that it is not the walk length with the optimistically biased maximum accuracy) is compared with the other learning algorithms in Tab. 1. The overall picture can be seen in terms of an average over 9 tasks. As the MCC values reached in different tasks can be seen as incomparable, the classification algorithms are also evaluated in terms of their average ranking. The algorithms are sorted and ranked according to their MCC scores in each of the tasks (from 1st to 4th) first, then the average of their ranks is calculated. The lower the rank, the better the algorithm. Both in Fig. 1 and Tab. 1 only the median of 10 randomly seeded runs representing a robust accuracy estimate for each classification task are shown.

The table illustrates a shift in predictive performance among several learning methods, starting from the comprehensible but empirically invalid decision tree through random forest to the accurate though black-box SVM. The results suggest that our network-enriched RF provides a good trade off between these two extremes. NCF shows good classification accuracy, while being more comprehensible than black-box models (see Sect. 5.1). In most of the cases, NCF has a better or equal predictive power than the state-of-the-art RF and as a whole, in terms of classification accuracy, is even competitive with the black-box SVM.

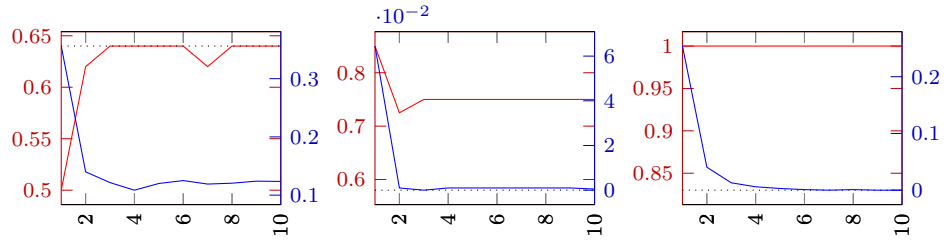
Table 1: The comparison of NCF with other classifiers. Median MCC values for 7 MDS classification tasks and 2 OC tasks. The k^* values set according to Equation 2 were applied for NCF. The other classifiers worked with the default settings.

task #	**Class Ratio**	**NCF**	**RF**	**CART**	**SVM**	**NB**
MDS1	5:10	0.64	0.64	0.45	0.60	**0.80**
MDS2	4:10	0.85	0.58	0.40	**1**	0.62
MDS3	6:10	**1**	0.83	0.58	**1**	0.83
MDS4	4:9	**1**	0.88	0.16	**1**	0.49
MDS5	6:11	0.93	60	0.44	**1**	0.57
MDS6	13:9	0.46	0.63	0.07	**0.64**	0.23
MDS7	5:11	**0.74**	0.23	0.13	0.30	0.26
OC1	29:29	**0.32**	0.24	0.06	0.28	0.20
OC2	31:33	**0.49**	0.37	0.25	0.36	0.46
Average accuracy		**0.71**	0.56	0.37	0.69	0.42
Average ranking		**1.6**	3	5	**1.5**	3.2

5.1. NCF interaction exploitation

To illustrate the process of NCF understanding, the set of 10 differently initialized forests constructed for MDS7 was taken. The task is to assess the impact of treatment in the group of BM del(5q) patients. 107 feature interaction pairs that appeared 10 and more times were extracted. The core of underlying subnetwork is shown in Fig. 2.

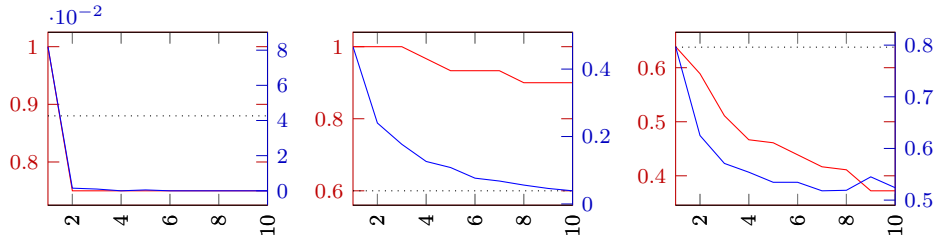
For several interaction pairs, their hypothetically calculated relationship or involvement in MDS and/or leukemia have a solid experimental support in reality. One example with a high score is an interaction pair found in the case of EGFR–CBL. Whereas CBL, the E3 ubiquitin-protein ligase involved in cell signalling and protein ubiquitination, is already known to control the fate of



(a) MDS1 – BM: H vs. A DT non-del(5q)

(b) MDS2 – PB: H vs. A BT non-del(5q)

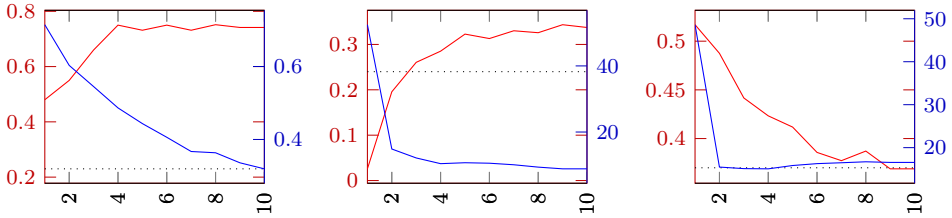
(c) MDS3 – BM: H vs. A BT del(5q)



(d) MDS4 – PB BT: non-del(5q) vs. del(5q)

(e) MDS5 – BM BT: non-del(5q) vs. del(5q)

(f) MDS6 – PB del(5q): BT vs. DT



(g) MDS7 – BM del(5q): BT vs. DT

(h) OC1 – G3 Stage IV: long- vs. short-term

(i) OC2 – G2: long- vs. short-term

Figure 1: Experimental results for 7 MDS classification tasks (graph 1a – 1g) and 2 validation tasks related to the overall survival of ovarian cancer (graph 1h – 1i). The development of Mathews correlation coefficient (MCC) (in red, the left y-axis) and the incidence of underfitted trees (in blue, the right y-axis) with increasing walk length (the x-axis). The MCC values of benchmarking RFs (that do not work with the walk length) are shown in dotted lines.

EGRF (epidermal growth factor receptor) ([45]), mutations in *Cbl gene* have been related to MDS and acute myeloid leukemia (AML) ([46, 47]). Additionally, CBL forms other interesting pairs, e.g., the interaction between CBL and

ABL1 (Abelson murine leukemia viral oncogene homolog 1) is also well documented (e.g., [48]). *Abl1* is a proto-oncogene that, activated by t(9;22) translocation, creates a new fusion gene, BCR-ABL which is typically associated not only with chronic myelogenous leukemia (CML) but also in some cases with acute lymphoblastic leukemia (ALL) and occasionally with AML ([49]). Still more pairs contain CBL; from those, at least interactions with CRK, CD2AP and AXL were previously reported ([50, 51], respectively), although up to now they have not been seen as relevant factors in MDS or leukemia.

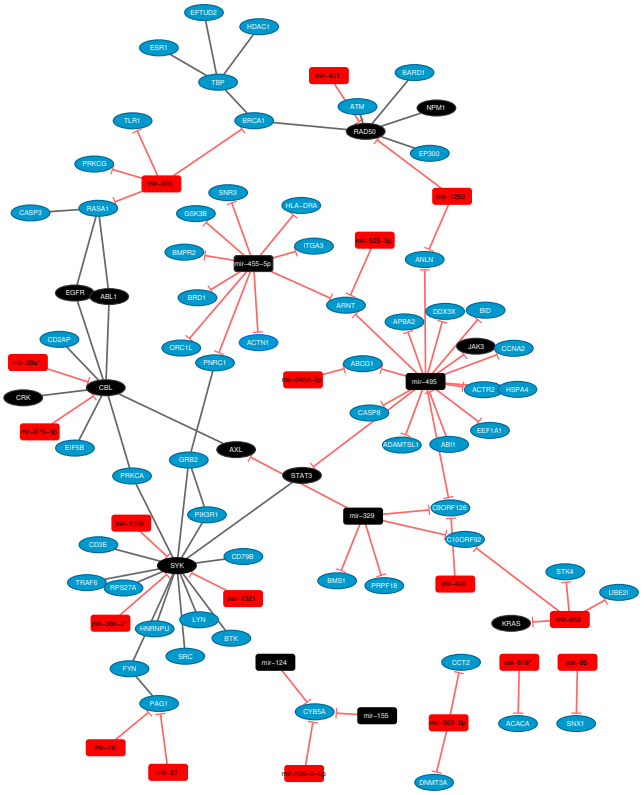


Figure 2: Discovered interaction subnetwork visualized in Cytoscape [52]. Oval entities correspond to protein coding genes, the rectangles to miRNA-genes. Black edges correspond to the interactions of respective proteins and red inhibitory edges to the interactions between miRNA and mRNA of inhibited gene. The black entities refer to genes and miRNAs respectively, reported in the text.

Another important high-score hit is NPM1–RAD50. Besides functioning as DNA-repair proteins found to be deregulated together in ovarian cancer ([53]), NPM1 (nucleophosmin) is known to be involved in AML, MDS, and acute promyelocytic leukemia ([54]). Interestingly, RAD50, as a DNA-repair protein, is also a potential factor in the etiology of AML and possibly MDS ([55]).

Besides the aforementioned genes, interaction pairs involve other genes, e.g., KRAS, JAK3, STAT3, and SYK, whose occurrence in MDS is known and under further investigation (for an overview, see [39]). Somewhat surprisingly, within the interactions there are only several hits for miRNA-coding genes previously reported as deregulated in MDS (e.g. miR-124, miR-329, miR-355 and miR-155) (reviewed in [56]). The other miRNA listed in interactions could then possibly represent new targets to which we should turn our attention considering MDS, i.e. miR-495 which has previously been associated with acute myeloid leukemia with mixed lineage leukemia rearrangements [57] but not with MDS.

To sum up, 36 out of 107 (33 %) high-scoring interaction pairs contain a gene (or the whole pair) whose (more or less significant) relationship to MDS has already been reported [39]. Such an occurrence is far from random and is obviously the result of a successful computational approach. Therefore, some of the interactions (or interacting counterparts) could turn out to be promising candidates for further investigation for their role in MDS and/or leukemia through both a literature search as well as laboratory experiments.

6. Discussion

In order to gain a deeper insight into the mechanism of NCF with respect to its predictive performance, in Fig. 1 the relationship between the empirical estimate of generalization error, walk length as a regularization parameter, and incidence of underfitted trees used to set a heuristic value of the parameter can be observed. This relationship shows a few consistent patterns. In MDS1 (Fig. 1a) a prematurely converging heuristic can be observed. This most

probably leads to underfitting of the model, which leads to the stagnation of predictive accuracy close to the baseline represented as a performance of RF. The mutually related tasks MDS2 and MDS4 (Fig. 1b and 1d), which share the definition of one the classification classes, show the heuristic falling to zero; at the same time, the initially promising value of MCC is reduced. Such a trend suggests overfitting, most probably due to the small sample size of the shared class, which contains only 4 samples (see Tab. 1). Conversely, stable good classification performance is manifested within another two mutually related tasks MDS3 and MDS5 (Fig. 1c and 1e). Very interesting results are shown in task MDS7, which manifest a slow decrease of the heuristic and attendant growth of the accuracy far above the baseline. As to the OC tasks which are obviously more complex due to the larger number of samples (see Tab. 1), NCF, under the proper parametrization, beats standard RF and even the black-box SVM. As to the heuristic settings of the walk length parameter k, it can be stated that the heuristic finds the values of generalization error very close to their optima.

To understand how particular types of domain knowledge influence the NCF predictive performance, we ran several additional experiments. The experiments were run under the same protocol as defined in Sect. 3.6 and the results were aggregated in Tab. 2. Particularly, the NCF was run with seed genes randomly drawn to study the direct influence of the candidate causal genes (see Tab. 2, *NCF-Complete* vs. *NCF-RandSeed*). Next, protein-protein interactions only were submitted to NCF to assess the influence of miRNA-target interactions and miRNA inhibitory mechanisms in general (*NCF-Complete* vs. NCF-PPI in Tab. 2). The results suggest that prior candidate gene selection itself slightly improves the NCF performance. The candidate causal genes help to focus towards the prospective regions of the feature network in tasks with the small sample mRNA and miRNA profiles. The knowledge of miRNA profiles and miRNA-target interactions have indisputable positive effect on the NCF performance. The latter observation may suggest that intelligent integration of miRNA features to the model has a positive impact on its validity. Similarly to the performance relationship between NCF and RF discussed in Sect. 5, NCF

based on mRNA profiles and protein-protein interactions only outperforms its RF counterpart based on mRNA profiles only (*NCF-PPI* vs. RF-mRNA in Tab. 2).

Table 2: The influence of the individual ingredients of NCF on its predictive performance. Median MCC values for 7 MDS classification tasks and 2 OC tasks. The influence of seed genes and miRNA-target interactions, respectively, on the predictive validity under the heuristic settings of k^*.

Task Name	**NCF**			**RF**	
	Complete	**RandSeed**	**PPI**	**Merged**	**mRNA**
MDS1	**0.64**	0.60	0.58	**0.64**	0.40
MDS2	0.85	0.7	**0.90**	0.58	0.43
MDS3	**1**	**1**	0.83	0.83	0.37
MDS4	**1**	**1**	0.90	0.88	0.42
MDS5	0.93	**0.95**	0.63	0.60	0.49
MDS6	0.46	0.44	**0.67**	0.63	0.51
MDS7	**0.74**	0.72	0.31	0.23	0.45
OC1	**0.32**	0.30	0.29	0.24	0.22
OC2	**0.49**	0.34	0.47	0.37	0.35
Average	**0.71**	0.67	0.62	0.56	0.40
Rank	**1.5**	2.5	2.5	3.2	4.4

7. Conclusion and Future Work

We propose a general parameter-free method for learning from high-dimensional and low-sample size data complemented by a feature interaction network. The method of network-constrained forest stems from the well-known random forests.

The main difference is that decorrelation of the individual weak classifiers is not reached through bootstrap sampling and random subsetting of the features, but pseudorandom subsetting driven by the feature interaction network. The individual trees deal with feature sets sampled from different areas of a feature network, the curated feature interactions thus tend to be promoted. Still, they are not strictly imposed; the method remains stochastic in its nature and an arbitrary feature relationship may appear in a tree. The probability of its occurrence increases with the decreasing path length between the pair of features in the network and increasing interaction observed in measurements. Unlike our previous efforts, we do not rely on feature extraction based on prior modules such as pathways or simple interaction subgraphs [58, 59, 9].

The method was applied to improve classification accuracy and comprehensibility of gene expression-based disease models. The obtained results suggest that introducing domain knowledge improves the accuracy of the forest and increases its compliance with the current knowledge. We believe that the method is able to benefit from stochastic identification of the subset of earlier reported general interactions; this subset manifests in the given context.

In future work, we will aim at truly omics experiments. We will employ more data types such as epigenetic data, namely DNA methylation arrays, and further extend utilized prior knowledge; for example, with information about transcription factor interactions and problem related pathways. At the moment, the main limitation lies in the simplifying assumption of measurement completeness; all the measurements must be available for all the samples, which is often not the case. Some patient mRNA profiles may missing, even though protein levels might be available for them, etc. The authors of [60, 61] demonstrate that feature networks represent a suitable regularization tool in other domains, as well; such as document topic prediction and click prediction. Eventually, we plan to proceed further beyond classification, namely to analyze the resulting forest. To be more precise, an analysis of successful trees in terms of gene ontology terms could be provided. Dealing with artificially generated data should answer the general applicability of the given method. The analysis

should also work with different ratios of n and p (as the number of samples grows the methods such as sparse SVM seem to be natural competitors [62], at least in terms of accuracy) and feature network sizes and topologies as well as feature interaction strengths (the stronger the curated interactions manifest in the measurements, the more prior knowledge applies to the given domain but it is easier to be identified from the measurements themselves).

Acknowledgment

This work was supported by the grant NT14539 of the Ministry of Health of the Czech Republic. We also thank to M. Beličková for providing expression profiling data and biological material.

References

[1] C. Giallourakis, C. Henson, M. Reich, et al., DISEASE GENE DISCOVERY THROUGH INTEGRATIVE GENOMICS 6 (1) (2005) 381–406. doi:10.1146/annurev.genom.6.080604.162234.
URL http://dx.doi.org/10.1146/annurev.genom.6.080604.162234

[2] T. R. Golub, D. K. Slonim, P. Tamayo, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science 286 (5439) (1999) 531–537. arXiv:http://www.sciencemag.org/content/286/5439/531.full.pdf.
URL http://www.sciencemag.org/content/286/5439/531.abstract

[3] A. Bossi, B. Lehner, Tissue specificity and the human protein interaction network. 5 (1). doi:10.1038/msb.2009.17.
URL http://dx.doi.org/10.1038/msb.2009.17

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

[5] R. C. Lee, R. L. Feinbaum, V. Ambros, The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14., Cell 75 (5) (1993) 843–54.
URL http://view.ncbi.nlm.nih.gov/pubmed/8252621

[6] M. R. Fabian, N. Sonenberg, The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC, Nat Struct Mol Biol 19 (6) (2012) 586–93. doi:10.1038/nsmb.2296.

[7] S. Sass, F. Buettner, N. Mueller, et al., A modular framework for gene set analysis integrating multilevel omics data 41 (21) (2013) 9622–9633. doi:10.1093/nar/gkt752.
URL http://dx.doi.org/10.1093/nar/gkt752

[8] M. Anděl, J. Kléma, Z. Krejčík, Integrating mRNA and miRNA expressions with interaction knowledge to predict myelodysplastic syndrome, in: ITAT 2013: Workshop on Bioinformatics in Genomics and Proteomics, 2013, pp. 48–55.

[9] J. Kléma, J. Zahálka, M. Anděl, et al., Knowledge-based Subtractive Integration of mRNA and miRNA Expression Profiles to Differentiate Myelodysplastic Syndrome, in: Proceedings of Int. Conf. on Bioinformatics Models, Methods and Algorithms, 2014, pp. 31–39.

[10] J. Kalina, Classification methods for high-dimensional genetic data, Biocybernetics and Biomedical Engineering 34 (1) (2014) 10–18.

[11] I. Guyon, J. Weston, S. Barnhill, et al., Gene Selection for Cancer Classification using Support Vector Machines 46 (2002) 389–422.

[12] M. H. Asyali, D. Colak, O. Demirkaya, et al., Gene Expression Profile Classification: A Review, Current Bioinformatics 1 (2006) 55–73.

[13] T. Hastie, R. Tibshirani, J. Friedman, et al., The elements of statistical learning, Vol. 2, Springer, 2009.

[14] A. J. Smola, P. J. Bartlett (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, USA, 2000.

[15] Z. Wei, H. Li, A Markov random field model for network-based analysis of genomic data, Bioinformatics 23 (12) (2007) 1537–1544.

[16] S. Imoto, T. Higuchi, T. Goto, et al., Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, Journal of Bioinformatics and Computational Biology 2 (01) (2004) 77–98.

[17] E. Ryeng, B. K. Alsberg, Microarray data classification using inductive logic programming and gene ontology background information, Journal of Chemometrics 24 (5) (2010) 231–240.

[18] I. Trajkovski, F. Zelezny, N. Lavrac, et al., Learning relational descriptions of differentially expressed gene groups, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 38 (1) (2008) 16–25.

[19] Y. Zhu, X. Shen, W. Pan, Network-based support vector machine for classification of microarray samples, BMC Bioinformatics (2009) 1–21.

[20] O. Lavi, G. Dror, R. Shamir, Network-induced classification kernels for gene expression profile analysis, Journal of Computational Biology 19 (6) (2012) 694–709.

[21] C. Li, H. Li, Network-constrained regularization and variable selection for analysis of genomic data, Bioinformatics 24 (9) (2008) 1175–1182.

[22] C. Porzelius, M. Johannes, H. Binder, et al., Supporting information leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients., Biometrical Journal 53 (2) (2011) 190–201. doi:10.1002/bimj.201000155.

[23] M. Johannes, H. Fröhlich, H. Sültmann, et al., pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery., Bioinformatics 27 (10) (2011) 1442–1443.

[24] F. Rapaport, A. Zinovyev, M. Dutreix, et al., Classification of microarray data using gene networks., BMC Bioinformatics 8.

[25] X. Zhou, J. Liu, X. Ye, et al., Ensemble classifier based on context specific miRNA regulation modules: a new method for cancer outcome prediction., BMC Bioinformatics 14 (S-12) (2013) S6.
URL http://dblp.uni-trier.de/db/journals/bmcbi/bmcbi14S.html#ZhouLYWX13

[26] Y. Su, Knowledge Integration Into Language Models: A Random Forest Approach, Ph.D. thesis, The Johns Hopkins University, Baltimore, Maryland, 90 p. (4 2009).

[27] J. Dutkowski, T. Ideker, Protein Networks as Logic Functions in Development and Cancer., PLoS Comput Biol 7 (9).

[28] Z. Chen, Z. Weixiong, Integrative Analysis Using Module-Guided Random Forests Reveals Correlated Genetic Factors Related to Mouse Weight 9 (3) (2013) e1002956. doi:10.1371/journal.pcbi.1002956.
URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1002956

[29] A. Vašíková, M. Beličková, E. Budinská, et al., A distinct expression of various gene subsets in CD34+ cells from patients with early and advanced myelodysplastic syndrome., Leuk Res 34 (12) (2010) 1566–72.
URL http://www.biomedsearch.com/nih/distinct-expression-various-gene-subsets/20303173.html

[30] L. Breiman, Random Forests, Machine Learning (2001) 5–32.

[31] A. Verikas, A. Gelzinis, M. Bacauskiene, Mining data with random forests: A survey and results of new tests 44 (2) (2011) 330–349. doi:10.1016/j.patcog.2010.08.011.
URL http://www.sciencedirect.com/science/article/pii/S0031320310003973

[32] H. Deng, G. Runger, Gene selection with guided regularized random forest 46 (12) (2013) 3483–3489. doi:10.1016/j.patcog.2013.05.018.
URL http://www.sciencedirect.com/science/article/pii/S0031320313002422

[33] M. B. Kursa, Robustness of Random Forest-based gene selection methods, BMC Bioinformatics 15 (1) (2014) 8. doi:10.1186/1471-2105-15-8.
URL http://www.biomedcentral.com/1471-2105/15/8

[34] K. Lunetta, L. B. Hayward, J. Segal, et al., Screening large-scale association study data: exploiting interactions using random forests, BMC Genetics 5 (1) (2004) 32. doi:10.1186/1471-2156-5-32.
URL http://www.biomedcentral.com/1471-2156/5/32

[35] C. Strobl, A.-L. Boulesteix, T. Kneib, et al., Conditional variable importance for random forests, BMC Bioinformatics 9 (1) (2008) 307. doi:10.1186/1471-2105-9-307.
URL http://www.biomedcentral.com/1471-2105/9/307

[36] The TCGA Research Network (2014).
URL http://cancergenome.nih.gov/

[37] T. Vergoulis, I. S. Vlachos, P. Alexiou, et al., TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. 40 (2012) D222–9. doi:10.1093/nar/gkr1161.
URL http://dx.doi.org/10.1093/nar/gkr1161

[38] T. S. Prasad, R. Goel, K. Kandasamy, et al., Human Protein Reference Database - 2009 update. 37 (Database-Issue) (2009) 767–772.

[39] W. Yu, M. Clyne, M. J. Khoury, et al., Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations., Bioinformatics 26 (1) (2010) 145–146.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: Machine Learning in Python 12 (2011) 2825–2830.

[41] A. Grigoriev, A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae, Nucleic Acids Res 29 (17) (2001) 3513–9.

[42] R. Jansen, D. Greenbaum, M. Gerstein, Relating whole-genome expression data with protein-protein interactions, Genome Res 12 (2002) 37–46.

[43] R. E. Banfield, L. O. Hall, K. W. Bowyer, et al., Ensemble diversity measures and their application to thinning, Information Fusion 6 (1) (2005) 49–62.

[44] E. K. Tang, P. N. Suganthan, X. Yao, An analysis of diversity measures, Machine Learning 65 (1) (2006) 247–271.

[45] G. D. Visser Smit, T. L. Place, S. L. Cole, et al., Cbl controls EGFR fate by regulating early endosome fusion, Science signaling 2 (102) (2009) ra86.

[46] J. Rocquain, N. Carbuccia, V. Trouplin, et al., Combined mutations of ASXL1, CBL, FLT3, IDH1, IDH2, JAK2, KRAS, NPM1, NRAS, RUNX1, TET2 and WT1 genes in myelodysplastic syndromes and acute myeloid leukemias., BMC Cancer 10 (2010) 401.
URL http://www.biomedsearch.com/nih/Combined-mutations-ASXL1-CBL-FLT3/20678218.html

[47] M. Naramura, S. Nadeau, G. A. Bhopal Mohapatra, et al., Mutant Cbl proteins as oncogenic drivers in myeloproliferative disorders, Oncotarget 2 (3) (2011) 245.

[48] T. Miyoshi-Akiyama, L. M. Aleman, J. M. Smith, et al., Regulation of Cbl phosphorylation by the Abl tyrosine kinase and the Nck SH2/SH3 adaptor, Oncogene 20 (30) (2001) 4058–4069.

[49] M. Talpaz, N. P. Shah, H. Kantarjian, et al., Dasatinib in imatinib-resistant Philadelphia chromosome–positive leukemias, New England Journal of Medicine 354 (24) (2006) 2531–2541.

[50] M. Garcia-Guzman, E. Larsen, K. Vuori, The proto-oncogene c-Cbl is a positive regulator of Met-induced MAP kinase activation: a role for Crk adaptor., Oncogene 19 (35) (2000) 4058–4065.

[51] P. Valverde, Effects of Gas6 and hydrogen peroxide in Axl ubiquitination and downregulation 333 (1) (2005) 180–185.

[52] M. E. Smoot, K. Ono, J. Ruscheinski, et al., Cytoscape 2.8: new features for data integration and network visualization, Bioinformatics 27 (3) (2011) 431–432.

[53] R. S. Kalra, S. A. Bapat, Enhanced levels of double-strand DNA break repair proteins protect ovarian cancer cells against genotoxic stress-induced apoptosis 6 (1) (2013) 66.

[54] B. Falini, I. Nicoletti, N. Bolli, et al., Translocations and mutations involving the nucleophosmin (NPM1) gene in lymphomas and leukemias, Haematologica 92 (4) (2007) 519–532.

[55] J.-Y. Shi, Z.-H. Ren, B. Jiao, et al., Genetic variations of DNA repair genes and their prognostic significance in patients with acute myeloid leukemia 128 (1) (2011) 233–238.

[56] G. Rhyasen, D. Starczynowski, Deregulation of microRNAs in myelodysplastic syndrome, Leukemia 26 (1) (2012) 13–22.

[57] X. Jiang, H. Huang, Z. Li, et al., miR-495 is a tumor-suppressor microRNA down-regulated in MLL-rearranged leukemia, Proceedings of the National Academy of Sciences 109 (47) (2012) 19397–19402.

[58] M. Holec, J. Kléma, F. Železnỳ, et al., Comparative evaluation of set-level techniques in predictive classification of gene expression samples, BMC Bioinformatics 13 (Suppl 10) (2012) S15.

[59] M. Krejník, J. Kléma, Empirical evidence of the applicability of functional clustering through gene expression classification, IEEE/ACM Trans. Comput. Biol. Bioinformatics 9 (3) (2012) 788–798. doi:10.1109/TCBB.2012.

23.
URL http://dx.doi.org/10.1109/TCBB.2012.23

[60] T. Sandler, J. Blitzer, P. P. Talukdar, et al., Regularized Learning with Networks of Features, in: D. Koller, D. Schuurmans, Y. Bengio, et al. (Eds.), Advances in Neural Information Processing Systems 21, Curran Associates, Inc., 2009, pp. 1401–1408.
URL http://papers.nips.cc/paper/3427-regularized-learning-with-networks-of-features.pdf

[61] D. Chakrabarti, R. Herbrich, Speeding Up Large-scale Learning with a Social Prior, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, ACM, New York, NY, USA, 2013, pp. 650–658. doi:10.1145/2487575.2487587.
URL http://doi.acm.org/10.1145/2487575.2487587

[62] J. Bi, K. Bennett, M. Embrechts, et al., Dimensionality Reduction via Sparse Support Vector Machines 3 (2003) 1229–1243.
URL http://dl.acm.org/citation.cfm?id=944919.944971