Semantic Multi-Classifier Systems for the Analysis of Gene Expression Profiles

Ludwig Lausser*, Florian Schmid*, Matthias Platzer, Mikko J. Sillanpää, and Hans A. Kestler**

Abstract The analysis of biomolecular data from high-throughput screens is typically characterized by the high dimensionality of the measured profiles. Development of diagnostic tools for this kind of data, such as gene expression profiles, is often coupled to an interest of users in obtaining interpretable and low-dimensional classification models; as this facilitates the generation of biological hypotheses on possible causes of a categorization. Purely data driven classification models are limited in this regard. These models only allow for interpreting the data in terms of marker combinations, often gene expression levels, and rarely bridge the gap to higher-level explanations such as molecular signaling pathways.

Here, we incorporate into the classification process, additionally to the expression profile data, different data sources that functionally organize these individual gene expression measurements into groups. The members of such

Ludwig Lausser · Matthias Platzer · Hans A. Kestler
Leibniz Institute on Aging, Jena, Germany
✉ [ludwig.lausser,matthias.platzer,hans.kestler]@leibniz-fli.de

Mikko J. Sillanpää
University of Oulu, Finland
✉ mikko.sillanpaa@oulu.fi

Ludwig Lausser, Florian Schmid, Hans A. Kestler
Medical Systems Biology, Ulm University, Germany
✉ [ludwig.lausser,florian-1.schmid,hans.kestler]@uni-ulm.de
* contributed equally
** corresponding author

ARCHIVES OF DATA SCIENCE, SERIES A
KIT SCIENTIFIC PUBLISHING
Vol. 1, No. 1, S. 157–176, 2016

DOI 10.5445/KSP/1000058747/09
ISSN 2363-9881

a group of measurements share a common property or characterize a more abstract biological concept. These feature subgroups are then used for the generation of individual classifiers. From the set of these classifiers, subsets are combined to a multi-classifier system. Analysing which individual classifiers, and thus which biological concepts such as pathways or ontology terms, are important for classification, make it possible to generate hypotheses about the distinguishing characteristics of the classes on a functional level.

1 Introduction

The high dimensionality of biomolecular data is one of the major challenges for machine learning algorithms in the field of bioinformatics. The enormous amount of measurements (e.g. gene expression levels) complicates the development of reliable and interpretable models. Initial feature selection can improve the performance of a trained model. This type of model reduction can aid in identifying causes for the predictive ability of the model, which can then further be validated in other experiments. However, feature sets derived in purely data driven or model driven feature selection processes rarely allow a functional interpretation. Measurements are typically selected according to a mathematical performance measure and without respect to known relationships or dependencies. Therefore, these feature sets can rather be regarded as a collection of diverse fragments then as a description of biological processes such as molecular signaling cascades or pathways.

Functional relationships and dependencies can rarely be inferred from a single dataset. Additional knowledge in the form of meta information, i.e. information about information, is needed for grouping or selecting the measurements in an interpretable way. This information can be extracted from a large corpus of biological literature and databases, see e.g. Galperin et al (2015) for an overview of current molecular databases. It aids in focusing on the construction of dedicated feature sets for a single biological process or a small set of biological processes.

The idea of incorporating meta information in the training of predictive models is not new. An overview on recent approaches is given by Porzelius et al (2011). They can mainly be divided into two categories. The first one consists of algorithms that try to guide traditional feature selection processes. For example, Binder and Schumacher (2009) incorporate knowledge on signaling pathways

into a boosting model by penalizing the score of the single base learners. Johannes et al (2010) developed a version of recursive feature elimination that is guided by the structure of a protein-protein interaction network. The second category enforces the usage of the given meta information more directly. Abraham et al (2010) construct an intermediate representation of the original measurements. The measurements of one category are replaced by a single feature. Lottaz and Spang (2005) developed an hierarchical classifier system that follows the structure of the gene ontology.

In this work, we propose a knowledge based feature selection algorithm that operates on a predefined vocabulary, i.e. a set of interpretable terms taken form molecular signaling pathways, gene ontology, etc. These verbal phrases are assumed to be reflected in the dataset by a known subset of gene expression measurements. A sparse set of these terms will then be selected and combined in the training of a multi-classifier system.

2 Methods

Classification is the task of predicting the class label $y \in \mathscr{Y}$ of an object on the basis of a vector of measurements, often termed features, $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^T \in \mathscr{X} \subseteq \mathbb{R}^n$. The underlying decision criterion is typically formalized as a decision function (a classifier) $c : \mathbb{R}^n \to \mathscr{Y}$. A classifier $c \in \mathscr{C}$ is initially selected according to a set of m labeled training examples $\mathscr{L} = \{(\mathbf{x}_j, y_j)\}_{j=1}^m$ and denoted by $c_{\mathscr{L}}$ if the chosen training set is relevant:

$$\mathscr{C} \times \mathscr{L} \xrightarrow{\text{train}} \mathscr{C}. \tag{1}$$

An important property of a trained classifier is its risk in misclassifying new, unseen samples

$$\mathscr{R}(c) = \int \mathbb{I}_{[c(\mathbf{x}) \neq y]} dP(x, y). \tag{2}$$

Here $\mathbb{I}_{[]}$ denotes the indicator function.

The risk of a classifier is typically estimated in a resampling experiment as the $r \times f$ cross-validation (Japkowicz and Shah, 2011). Here, the available data $\mathscr{S}$ is split into f folds of approximately equal size. A number of f experiments are performed in which each fold of samples is tested by a classifier trained on the remaining samples. This procedure is repeated for r permutations of $\mathscr{S}$ in order to make the cross-validation error independent from particular data

partitions. Let $\mathscr{L}_{ij}$ and $\mathscr{T}_{ij}$ denote the training and test sets of the ith run and the jth split. The error estimation of $r \times f$ cross-validation is then given by

$$R_{r\times f} = \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{f}\frac{1}{|\mathscr{T}_{ij}|}\sum_{(\mathbf{x},y)\in\mathscr{T}_{ij}}\mathbb{I}\left[c_{\mathscr{L}_{ij}}(\mathbf{x})\neq y\right]. \tag{3}$$

A second important characteristic of a trained classifier is its interpretability. It can be seen as the classifiers ability of giving insights into the properties of a dataset (e.g. identifying important components or dependencies). The interpretability of a trained classifier depends on two distinct properties, syntactical and semantic interpretability.

Syntactical or structural interpretability

The interpretability of a decision function is dependent on its structural properties. The higher the complexity of a decision boundary the lower is its interpretability. The syntactic properties of a classifier can mainly be derived from its concept class $\mathscr{C}$. Possible notions of structural complexity are the number of parameters (Hastie et al, 2001) or the VC-Dimension (Vapnik, 1998).

Semantic interpretability

The interpretability of a classifier is also dependent on the set of measurements that is utilized for a prediction. For instance a selected measurement seems to influence a classification result while a deselected one does not or should not. Other more abstract semantic explanations can be revealed by analyzing the selected feature combinations or structures developed by the trained classifier. Analyses of this type are for example the (gene set) enrichment analysis for the analysis of feature sets (Hung et al, 2012) or principal component analysis (Jolliffe, 2002). The abstract terms that can be detected by these methods are typically strongly affected by noise and should be regarded as fuzzy concepts.

2.1 Feature selection

A common step in the training process of classification models is the selection of informative features (Guyon et al, 2006)

$$\mathscr{C} \times \mathscr{L} \xrightarrow{\text{select}} \mathscr{I} = \{\mathbf{i} \in \mathbb{N}^{\hat{n} \leq n} | i_k < i_{k+1}, 1 \leq i_k \leq n\}. \tag{4}$$

Here $\mathscr{I}$ indicates the set of all sorted and repetition free index vectors of maximal length n. A single element $\mathbf{i} \in \mathscr{I}$ is called a *signature*. It will be denoted by $\mathbf{i} = (i_1, \ldots, i_{\hat{n}(\mathbf{i})})^T$, where $\hat{n}(\mathbf{i}) \leq n$ is the size of $\mathbf{i}$. The elements of a signature indicate the selection of measurements $\mathbf{x}^{(\mathbf{i})} = (x^{(i_1)}, \ldots, x^{(i_{\hat{n}(\mathbf{i})})})^T$ that will be considered in the learning phase of the classifier and for predicting the class label of new unseen samples. It will be called a *feature set* or *feature vector* in the following.

Feature selection is typically a data driven process. That is, a feature set is chosen according to some kind of quality criterion that measures the "informativeness" of the single measurements (univariate feature selection) or a combination thereof (multivariate feature selection). If it can be applied without any knowledge of any other parts of the training algorithm, it can be seen as a preprocessing *filter*.

Feature selection becomes model driven, if knowledge about the concept class $\mathscr{C}$ is incorporated into the selection process. Here, an evaluation criterion is based on the performance (e.g. accuracy) of the classification model $c \in \mathscr{C}$ trained on the current feature combination. The category of model driven feature selection methods comprises the category of *wrappers*, which evaluates general performance measures, and *embedded feature selectors*, which evaluates model specific characteristics.

Data driven and model driven feature selectors share a common search space of $2^n - 1$ feature combinations. It can hardly be analyzed exhaustively due to its exponential growth in n. Most feature selectors are based on heuristic or stochastic search strategies. They usually do not guarantee to find a global optimal solution.

Although data or model driven feature selection clearly reduces the measurements that are involved in generating a decision boundary, it is often questionable if it really simplifies the semantic interpretability of a classifier. Measurements selected according to some kind of performance criterion rarely can be summarized under some interpretable term $\mathbf{v}$. The reason for this is the lack of knowledge about local, temporal or functional dependencies among the measurements. In a purely data or model driven setting, these relationships have to be learned from scratch and often remain undetected.

2.2 Knowledge-based feature selection

In this work we propose a knowledge based feature selection algorithm that allows for incorporating an experimenter's domain knowledge into the feature selection process. A domain expert often possesses knowledge about the experimental setup which typically can not be utilized by the machine learning algorithm. For example, a possible functional grouping of features / measurements is typically known to an experimenter but unknown to the algorithm. The domain expert may also have knowledge about the subject of an investigation, the corresponding measurements and their interactions, etc. For example, an expert in molecular biology has some a-priori knowledge about the molecules that are involved in a certain type of cellular process.

The interactions and relationships described above can typically be summarized by a short verbal phrase that conveys some semantic knowledge to the domain expert (e.g. video-signal, citrate cycle, insulin-secretion). We will call such a phrase an abstract *term* or *word* $\mathbf{v}$. A set of words will be called a vocabulary $\mathscr{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{|\mathscr{V}|}\}$. It reflects the external domain knowledge that should be incorporated into an experiment. We will use a word or term $\mathbf{v}$ synonymously with its associated signature $\mathbf{i}$. That is a vocabulary can be seen as a subset $\mathscr{V} \subseteq \mathscr{I}$.

In contrast to a purely model or data driven feature selection, our method constructs feature sets that can be seen as a union of the elements of a subset of the vocabulary $\mathscr{V}$

$$\mathscr{C} \times \mathscr{L} \times \mathscr{V} \xrightarrow{\text{select}} \bigcup_{\mathbf{v} \in \mathscr{V}'} \mathbf{v}, \mathscr{V}' \subseteq \mathscr{V}. \tag{5}$$

That is, the final feature set will include all measurements that are associated to the selected words $\mathscr{V}'$. Without loss of generality, we assume that a typical vocabulary will result in $|\mathscr{V}| \ll |\mathscr{I}|$ and $\forall \mathbf{v} \in \mathscr{V} : \hat{n}(\mathbf{v}) > 1$. In this case a knowledge based feature selection will lead to a reduction of the search space complexity from $2^n - 1$ to $2^{|\mathscr{V}|} - 1$.

Although the final set of features is constructed by selecting a set of words, it is questionable, if the corresponding union of feature sets really reflects the chosen terms. These sets can be overlapping. Their union can implicitly include signatures of additional terms. In order to keep the interpretability of the final signature, we have chosen to couple our knowledge-based feature selection to a multi-classifier system that evaluates each term independently.

2.2.1 Semantic base classifiers (SBC)

Our multi-classifier system is constructed of semantic base classifiers of type

$$c_{\mathbf{v}} : \mathbf{x}^{(\mathbf{v})} \mapsto y. \tag{6}$$

Here $c_{\mathbf{v}}$ denotes a classifier that is restricted to the signature of $\mathbf{v}$ and is therefore associated to this term. The suitability of a term $\mathbf{v}$ is estimated in a 3×3 cross-validation experiment on the learning set $\mathscr{L}$. The signature is therefore evaluated by a multivariate criterion.

A single term $\mathbf{v}^* \in \mathscr{V}$ can be chosen by ranking all terms in $\mathscr{V}$ according to their achieved cross-validation errors

$$\mathbf{v}^* = \arg\min_{\mathbf{v} \in \mathscr{V}} R_{3\times 3}(\mathscr{C}_{\mathbf{v}}, \mathscr{L}), \tag{7}$$

where $\mathscr{C}_{\mathbf{v}}$ denotes a restriction of the chosen concept class $\mathscr{C}$ to the selected term $\mathbf{v}$. The final base classifier $c_{\mathbf{v}^*} \in \mathscr{C}_{\mathbf{v}^*}$ will be trained on all samples in $\mathscr{L}$ and will be seen as an expert in interpreting $\mathbf{v}^*$. In principal, each training algorithm and concept class can be chosen for the underlying training of a semantic base classifier. For our experiments, we have chosen the nearest neighbor classifier (NNC) proposed by Fix and Hodges (1951).

2.2.2 Semantic multi-classifier systems (SMCS)

The multi-classifier system itself can be seen as a decomposable decision rule that is based on an ensemble of semantic base classifiers $\mathscr{E} = \{c_i\}_{i=1}^{|\mathscr{E}|}$, $c_i \in \mathscr{C}$. The final decision rule will be denoted by $h_{\mathscr{E}}$. The training of $h_{\mathscr{E}}$ corresponds to a selection process in which the most suitable set of experts is constructed.

We have chosen an unweighted majority vote h_{maj} as a fusion architecture. It returns the most frequent prediction of the base classifiers as its own prediction and therefore allows a direct interpretation

$$h_{maj}(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathscr{Y}} |\{c(\mathbf{x}) = y \,|\, c \in \mathscr{E}\}|. \tag{8}$$

The fusion on a symbolic level prohibits interactions on a feature level and conserves the interpretability of the final signature.

The ensemble members are selected in an iterative way. Similar to Equation 7 in each iteration t, a term $\mathbf{v}_t$ is chosen that minimizes the error estimate in a

3×3 cross-validation experiment on the samples of $\mathscr{L}$. The selection of the current term is restricted to those terms that were not selected before. Formally

$$\mathbf{v}_t = \arg\min_{\mathbf{v} \in \mathscr{V}_t} R_{3\times 3}(\mathscr{C}_{\mathbf{v}}, \mathscr{L}), \tag{9}$$

with $\mathscr{V}_t = \mathscr{V}_{t-1} \setminus \{\mathbf{v}_{t-1}\}$ and $\mathscr{V}_1 = \mathscr{V}$. The corresponding base classifiers $c_{\mathbf{v}_t}$ are again trained on all samples in $\mathscr{L}$.

3 Experimental setup and results

3.1 Basic setup

The proposed semantic multi-classifier systems are evaluated in the setting of classifying gene expression profiles. We conduct nested cross-validation experiments to assess their performance (Varma and Simon, 2006) on six different microarray data sets. For the outer cross-validation experiment a 10×10 cross-validation is chosen. The training data of every split is used to select a suitable set of features (signatures) and to train the classifier model. The model selection process for this classifier is based on an internal 3×3 cross-validation as discussed in Sect. 2.2.2. For all experiments, the nearest neighbour classifier (NNC) was chosen as single or base classifier. The semantic classifier systems (SBC and SMCS) were compared to NNCs that use all features and those that incorporate a purely data-driven feature selection process, i.e. the top k features with the highest absolute Pearson correlation to the class label were chosen. The number of features k was predetermined with regard to the chosen vocabulary (k = mean signature size, see Table 2) . All experiments were conducted with the TunePareto-Software for classifier evaluation (Müssel et al, 2012).

Datasets

The experiments are conducted on different two class diagnostic classification tasks. All are related to ageing associated diseases. The data sets are obtained from high-throughput microarray experiments from different technological platforms. All data sets are publicly available from the Gene Expression Omnibus

Table 1 Basic characteristics of the analysed data sets with citation, Gene Expression Omnibus ID (GEOid), feature number (Feat.), sample number (Samp.), and class distribution (Cl.0 and Cl.1).

Dataset	Citation	GEOid	Feat.	Samp.	Cl.0	Cl.1
Alzheimer's disease	Liang et al (2008)	GSE5281	54613	161	74	87
Leukemia	Alcalay et al (2005)	GSE34860	22215	78	21	57
Thyroid cancer	Maenhaut et al (2011)	GSE29265	54613	49	20	29
Lung cancer	Hou et al (2010)	GSE19188	54613	156	65	91
Melanoma	Xu et al (2008)	GSE8401	22215	83	31	52
Pancreatic cancer	Zhang et al (2013)	GSE28735	32321	90	45	45

Table 2 Characteristics of the vocabularies used from the MSigDB (Subramanian et al, 2005) (KEGG, CHROM) and Gene Ontology (Ashburner et al, 2000) (GO), with the number of terms, the number of elements associated to one term (signature) and the total number of covered genes in the database.

	number of terms	minimal signature size	median signature size	mean signature size	maximal signature size	total number of covered genes
KEGG	186	10	53	69	389	5267
GO	3125	10	20	40	492	15992
CHROM	326	5	65.5	91	948	30010

(`http://www.ncbi.nlm.nih.gov/geo/`) database. A brief summary of the data is given in Table 1.

Vocabularies

In our experiments we have used three different sources of meta information:

1. KEGG – Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000) is a collection of molecular signaling pathways,
2. GO – Gene Ontology (Ashburner et al, 2000) is a standardized terminology for the categorization of gene products, here we limited our terms to those that have a set size in the interval from 10 to 500, and
3. CHROM – Chromosomal Location is the position of the corresponding gene within the human genome.

An overview on their key characteristics is given in Table 2. The signatures are extracted from MSigDB (Subramanian et al, 2005) and Gene Ontology (Ashburner et al, 2000). All identifiers have been mapped to gene names. They can be regarded as knowledge of domain experts in molecular biology.

Table 3 Results of the 10×10 fold cross-validation experiments with the KEGG pathways vocabulary. Mean error rates in % $\pm$ standard deviations are given. Feature numbers are given (*All features*), predetermined (*Feature selection*), or averages (*SBC* and *SMCS*).

	Alzheimer's disease (Liang et al, 2008)		Leukaemia (Alcalay et al, 2005)	
	cv-error	features	cv-error	features
All features	9.32 ± 0.72	54613	13.59 ± 0.90	22215
Feature selection	10.31 ± 1.53	69	4.10 ± 0.81	69
SBC (KEGG)	7.37 ± 1.34	325.41	9.23 ± 1.99	174.16
SMCS (KEGG)	7.02 ± 1.02	281.8	8.33 ± 1.51	173.15

3.2 Experimental results

In the following we exemplify our method of semantic multi-classifier systems on selected combinations of vocabularies and data sets. Due to size limitations we do not show all 18 combinations. The selected classification approaches are by no means biased in terms of accuracy, etc., but rather give an arbitrary assignment of data sets and used domain knowledge. In the following each of the tested vocabularies is introduced by a short description first and then validated on two datasets.

3.2.1 Kyoto Encyclopedia of Genes and Genomes (KEGG):

The Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000) is a manually curated collection of molecular signaling and metabolic pathways that regulate different processes in or between cells. A single term from this vocabulary reflects the molecules (more precisely the gene products) that are involved in the signaling process. An example for a KEGG pathway is the *insulin signaling pathway*. It provides the list of molecules that are affected by the binding of hormone insulin to the corresponding receptor of a cell. We tested two datasets using the KEGG-pathways as meta-information. A summary can be found in Table 3.

Alzheimer's disease data set

The Alzheimer dataset was collected by Liang et al (2008) and is available in the Gene Expression Omnibus (GEO) under GSE5281. It comprises brain tissue samples taken post mortem from subjects suffering from Alzheimer's disease (74 samples) and controls (87 samples). Each gene expression profile consists of 54613 probe sets. Applied to all measurements the NNC achieves an cv-error of $9.32\% \pm 0.72$. With feature selection the cv-error is $10.31\% \pm 1.53$. Lower errors are achieved when meta information is used. Coupled to the vocabulary of KEGG pathways a single semantic base classifier achieves an cv-error of $7.37\% \pm 1.34$. A semantic ensemble of three base classifiers achieves an cv-error of $7.02\% \pm 1.02$. Fig. 1a) shows the frequency of the KEGG pathways that are selected in the 10×10 cross-validation. The *insulin signaling pathway* is selected in 91% of the cross-validation splits. It is known that this pathway is impaired in Alzheimer patients (Candeias et al, 2012).

Leukaemia data set

The Leukaemia dataset collected by Alcalay et al (2005) consists of 57 samples of acute myeloid leukaemia with aberrant cytoplasmic localization of nucleophosmin following mutations in the NPM putative nucleolar localization signal and 21 samples without this specific mutation (GSE34860). Each gene expression profile consists of 22215 probe sets. Using all features leads to the lowest performance ($13.59\% \pm 0.90$). With feature selection obtains the best cv-errors ($4.10\% \pm 0.81$). Utilizing the vocabulary of KEGG pathways the best semantic base classifier improves the cv-error (compared to all features) by 4% to $9.23\% \pm 1.99$. The semantic ensemble is able to lower the error rate by another percent ($8.33\% \pm 1.51$). In this case the KEGG pathways *hematopoietic cell lineage* and *cell adhesion molecules cams* are selected most frequently (Fig. 1b). Both terms have been reported in the context of leukaemia (Bonnet and Dick, 1997; Noto et al, 1994).

3.2.2 Gene Ontology (GO):

The Gene Ontology (Ashburner et al, 2000) is currently one of the most prominent attempts of constructing an organized and standardized terminology for

Table 4 Results of the 10×10 fold cross-validation experiments with the GO terms vocabulary. Mean error rates in % $\pm$ standard deviations are given. Feature numbers are given (*All features*), predetermined (*Feature selection*), or averages (*SBC* and *SMCS*).

	Thyroid cancer (Maenhaut et al, 2011)		Lung cancer (Hou et al, 2010)	
	cv-error	features	cv-error	features
All features	11.22 ± 1.73	54613	8.14 ± 0.61	54613
Feature selection	12.45 ± 2.25	40	6.22 ± 1.51	40
SBC (GO)	11.63 ± 3.34	159.93	4.49 ± 0.74	48.83
SMCS (GO)	6.73 ± 2.16	208.67	4.74 ± 1.14	92.77

the categorization of gene products. It provides an hierarchical ontology of terms that covers three different fields: biological processes, associated cellular components and molecular functions. Most of these terms are linked to manually curated gene lists. The Gene Ontology provides for example the term *cell aging*, which is linked to the list of genes that are known to influence the aging process of cells. The vocabulary of GO terms was tested in two different scenarios (Table 4).

Thyroid cancer

The Thyroid cancer dataset was collected by Maenhaut et al (2011) (GSE29265). Its 49 thyroid samples have been categorised into non-tumour control (20 samples) and thyroid carcinoma (29 samples). The dimensionality of the dataset is 54613. Compared to the experiments with all measurements and data driven feature selection (error rates $11.22\% \pm 1.73$ and $12.45\% \pm 2.25$) the knowledge-based ensemble clearly improves the result. The error rate for the semantic ensemble is $6.73\% \pm 2.16$. A single base classifier is not able to reach this performance ($11.63\% \pm 3.34$). Looking at the selected categories in the cross-validation experiment (Fig. 1c) of the ensemble, we find the *chondroitin sulfate metabolic process* term as the one which is most frequently selected (Infanger et al, 2006).

Table 5 Results of the 10×10 fold cross-validation experiments with the chromosomal locations vocabulary. Mean error rates in % $\pm$ standard deviations are given. Feature numbers are given (*All features*), predetermined (*Feature selection*), or averages (*SBC* and *SMCS*).

	Melanoma (Xu et al, 2008)		Pancreatic ductal adenocarcinoma (Zhang et al, 2013)	
	cv-error	features	cv-error	features
All features	8.19 ± 1.11	22215	24.67 ± 1.95	32321
Feature selection	8.92 ± 2.39	91	23.56 ± 3.28	91
SBC (CHROM)	7.59 ± 1.14	165.02	22.89 ± 3.24	69.13
SMCS (CHROM)	6.51 ± 1.63	148.39	19.78 ± 3.00	65.79

Lung cancer

The Lung cancer dataset (GSE19188) collected by (Hou et al, 2010) comprises samples of non-small cell lung cancer (91 samples) and adjacent normal tissue (65 samples). Each profile consists of 54613 probe sets. The mean cv-error achieved by the NNC on all features is $8.14\% \pm 0.61$. By using data driven feature selection this result can be improved to $6.22\% \pm 1.51$. On this dataset a single semantic base classifier achieves a slightly better classification performance than the ensemble. The cv-errors are $4.49\% \pm 0.74$ for the base classifiers and $4.74\% \pm 1.14$ for the ensemble. The most frequently selected term is related to ascorbic acid (vitamin C) metabolism (Fig. 1d). Ascorbic acid has been reported to have the ability to kill cancer cells under certain conditions (Chen et al, 2005).

3.2.3 Chromosomal location:

The vocabulary of chromosomal locations (CHROM) can also be used to organize the set of gene expression levels. Here, we restrict ourselves to the human genome. It is organized in 22 pairs of autosome chromosomes and one pair sex chromosomes. Each of the chromosomes can be divided into several cytobands. They can be used to indicate local aberrations. A single term out of this vocabulary gives the index of the chromosome, the chromosome arm (p = short arm, q = long arm), and the cytogenetic bands position on the chromosome arm. For example, *chr17p12* denotes band 1, subband 2 on the short arm of the 17th chromosome. Our experiments with the vocabulary of chromosomal locations are summarized in Table 5.

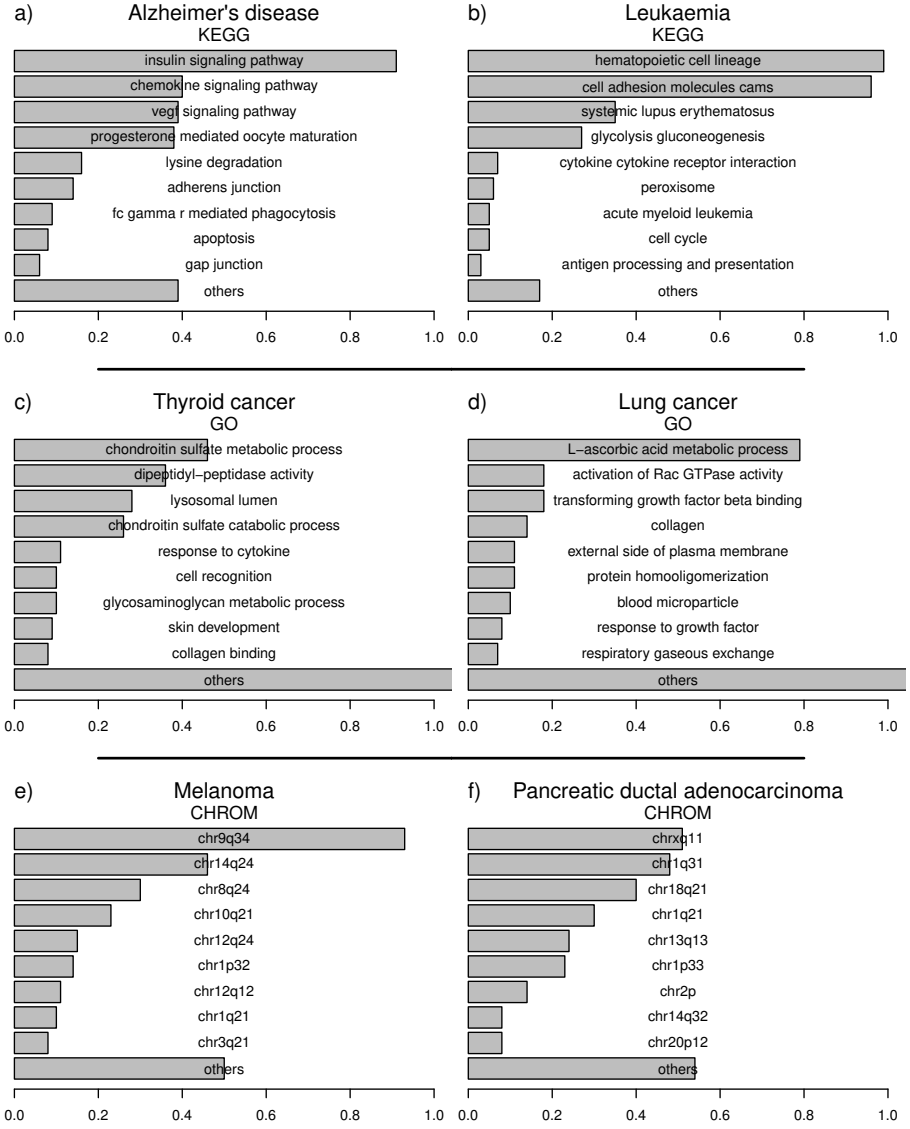

Fig. 1 Frequencies of the terms selected by the semantic multi-classifier system (SMCS) in the 10×10 cross-validation experiments. In total the frequency of 300 terms ($= 10 \times 10 \times 3$) is depicted in each diagram (a to f), normalized to the 100 experiments conducted each. The top nine selected terms are shown. The tenth bar "others" summarizes all categories that are selected less frequent.

Melanoma

The Melanoma dataset (Xu et al, 2008) was collected with the purpose to distinguish between primary melanomas and melanoma metastasis (GSE8401). Both classes are represented by 31 and 52 samples, respectively. The dimensionality of the corresponding gene expression profiles is 22215. For this dataset the data driven feature selection (cv-error: $8.92\% \pm 2.39$) performs worse than using all measurements (cv-error: $8.19\% \pm 1.11$). Using the vocabulary of chromosomal locations (CHROM) as meta information allows to improve the performance. A single semantic base classifier achieves an cv-error of $7.59\% \pm 1.14$. The semantic ensemble improves the cv-error to $6.51\% \pm 1.63$. The most frequently selected chromosomal band is *9q34* (Fig. 1e). It contains the ASS gene which is known to play a role in the cell death in melanomas (Savaraj et al, 2007).

Pancreatic ductal adenocarcinomas

The second dataset tested with chromosomal locations was collected by Zhang et al (2013) (GSE28735). Gene expression values of 45 pancreatic ductal adenocarcinomas and 45 adjacent non-tumour tissues have been measured in profiles of 32321 probe sets. The best results ($19.78\% \pm 3.00$) are achieved by ensembles using the vocabulary of chromosomal locations as meta information. Semantic base classifiers are able to achieve an cv-error of $22.89\% \pm 3.24$. Using all features or data driven feature selection leads to $24.67\% \pm 1.95$ and $23.56\% \pm 3.28$ cv-error, respectively. For this dataset three chromosomal bands are selected with comparable frequencies in the cross-validation experiment (Fig. 1f). The ensemble selects *Xq11*, *1q31* and *18q21* in most of the cases. For *1q31* and *18q21* an association to pancreatic cancer has been reported (Chen et al, 2003; Hahn et al, 1995).

4 Conclusion

We present a knowledge based approach for the design of classifier systems that are interpretable in abstract terms. The basic algorithm incorporates meta information in the form of a vocabulary of signatures (terms) that can be used for constructing a decision rule. The design of the algorithm ensures a high-level interpretability and eliminates the need for revealing an interpretation

via reconstruction methods. Our experiments suggest that knowledge based classifiers can be applied beneficially in the field of analyzing gene expression profiles. The constructed models fit into the biomedical context of the analysed diseases. The classification results indicate that selecting only a single term out of a vocabulary neither leads to optimal classification performance nor results in a highly stable selection. Combining a small set of terms improves the classification performance in almost all experiments.

Compared to other approaches the proposed multi-classifier systems excel other approaches by their superior interpretability. Yet, there might be more sophisticated classifier systems that outperform the proposed methods in terms of prediction accuracy. Subsequent work will be focused on the design of classifier systems and other model types that also use continuous outcomes that allow a suitable tradeoff between interpretability and prediction accuracy. For genetic data the presence of close genetic relationships among collected individuals may also bias the results (Habier et al, 2007; Dekkers, 2010), for expression data this is unclear. Integrating meta information in the form of these vocabularies might also be useful for guiding the selection of causal models (Mayo, 1996; Pearl, 2009).

The experiments of this investigation reveal an additional question for the design of a knowledge based classifier system. Although the selected kind of meta information will mainly be determined by the design of a medical/biological study, there may be some a-priori hints on the suitability of a vocabulary of signatures. These hints might be given in the structural properties of a vocabulary (e.g. overlap between signatures) but also in their semantic interpretation (e.g. local information vs functional information). This question can be addressed in more detailed analyses on available sources of meta information for gene expression profiles.

Acknowledgements The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n°602783 (to HAK), the German Research Foundation (DFG, SFB 1074 project Z1 to HAK), and the Federal Ministry of Education and Research (BMBF, Gerontosys II, Forschungskern SySt aR, project ID 0315894A to HAK).

References

Abraham G, Kowalczyk A, Loi S, Haviv I, Zobel J (2010) Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. BMC Bioinformatics 11(1):277–291, DOI 10.1186/1471-2105-11-277

Alcalay M, Tiacci E, Bergomas R, Bigerna B, Venturini E, Minardi SP, Meani N, Diverio D, Bernard L, Tizzoni L, Volorio S, Luzi L, Colombo E, Lo Coco F, Mecucci C, Falini B, Pelicci PG (2005) Acute myeloid leukemia bearing cytoplasmic nucleophosmin (NPMc+ AML) shows a distinct gene expression profile characterized by up-regulation of genes involved in stem-cell maintenance. Blood 106(3):899–902, DOI 10.1182/blood-2005-02-0560

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25(1):25–29, DOI 710.1038/75556

Binder H, Schumacher M (2009) Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. BMC Bioinformatics 10(1):18–28, DOI 10.1186/1471-2105-10-18

Bonnet D, Dick JE (1997) Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nature Medicine 3(7):730–737, DOI 10.1038/nm0797-730

Candeias E, Duarte AI, Carvalho C, Correia SC, Cardoso S, Santos RX, Plácido AI, Perry G, Moreira PI (2012) The impairment of insulin signaling in alzheimer's disease. IUBMB Life 64(12):951–957, DOI 10.1002/iub.1098

Chen Q, Espey MG, Krishna MC, Mitchell JB, Corpe CP, Buettner GR, Shacter E, Levine M (2005) Pharmacologic ascorbic acid concentrations selectively kill cancer cells: Action as a pro-drug to deliver hydrogen peroxide to tissues. Proceedings of the National Academy of Sciences of the United States of America 102(38):13,604–13,609, DOI 10.1073/pnas.0506390102

Chen YJ, Vortmeyer A, Zhuang Z, Huang S, Jensen RT (2003) Loss of heterozygosity of chromosome 1q in gastrinomas: Occurrence and prognostic significance. Cancer Research 63(4):817–823

Dekkers JCM (2010) Use of high-density marker genotyping for genetic improvement of livestock by genomic selection. CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources 5:1–13

Fix E, Hodges JL (1951) Discriminatory analysis: Nonparametric discrimination: Consistency properties. Tech. Rep. Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolf Field, Texas

Galperin MY, Rigden DJ, Fernández-Suárez XM (2015) The 2015 nucleic acids research database issue and molecular biology database collection. Nucleic Acids Research 43(D1):D1–D5, DOI 10.1093/nar/gku1241

Guyon I, Gunn S, Nikravesh M, Zadeh LA (2006) Feature Extraction: Foundations and Applications. Springer, Heidelberg

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397, DOI 10.1534/genetics.107.081190

Hahn SA, Seymour AB, Hoque ATMS, Schutte M, da Costa LT, Redston MS, Caldas C, Weinstein CL, Fischer A, Yeo CJ, Hruban RH, Kern SE (1995) Allelotype of pancreatic adenocarcinoma using xenograft enrichment. Cancer Research 55(20):4670–4675

Hastie T, Tibshirani R, Friedman JH (2001) The Elements of Statistical Learning. Springer, New York

Hou J, Aerts J, den Hamer B, van IJcken W, den Bakker M, Riegman P, van der Leest C, van der Spek P, Foekens JA, Hoogsteden HC, Grosveld F, Philipsen S (2010) Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS ONE 5(4):1–12, DOI 10.1371/journal.pone.0010312

Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C (2012) Gene set enrichment analysis: performance evaluation and usage guidelines. Briefings in Bioinformatics 13(3):281–291, DOI 10.1093/bib/bbr049

Infanger M, Kossmehl P, Shakibaei M, Bauer J, Kossmehl-Zorn S, Cogoli A, Curcio F, Oksche A, Wehland M, Kreutz R, Paul M, Grimm D (2006) Simulated weightlessness changes the cytoskeleton and extracellular matrix proteins in papillary thyroid carcinoma cells. Cell and Tissue Research 324(2):267–277, DOI 10.1007/s00441-005-0142-8

Japkowicz N, Shah M (2011) Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, Cambridge

Johannes M, Brase JC, Fröhlich H, Gade S, Gehrmann M, Fälth M, Sültmann H, Beißbarth T (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics 26(17):2136–2144, DOI 10.1093/bioinformatics/btq345

Jolliffe IT (2002) Principal Component Analysis. Springer, Heidelberg

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28(1):27–30

Liang WS, Reiman EM, Valla J, Dunckley T, Beach TG, Grover A, Niedzielko TL, Schneider LE, Mastroeni D, Caselli R, Kukull W, Morris JC, Hulette CM, Schmechel D, Rogers J, Stephan DA (2008) Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. Proceedings of the National Academy of Sciences 105(11):4441–4446, DOI 10.1073/pnas.0709259105

Lottaz C, Spang R (2005) Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. Bioinformatics 21(9):1971–1978, DOI 10.1093/bioinformatics/bti292

Maenhaut C, Detours V, Dom G, Handkiewicz-Junak D, Oczko-Wojciechowska M, Jarzab B (2011) Gene expression profiles for radiation-induced thyroid cancer. Clinical Oncology 23(4):282–288, DOI 10.1016/j.clon.2011.01.509

Mayo DG (1996) Error and the Growth of Experimental Knowledge. University of Chicago Press, Chicago

Müssel C, Lausser L, Maucher M, Kestler H (2012) Multi-objective parameter selection for classifiers. Journal of Statistical Software 46(1):1–27, DOI 10.18637/jss.v046.i05

Noto RD, Schiavone EM, Ferrara F, Manzo C, Pardo CL, Vecchio LD (1994) All-trans retinoic acid promotes a differential regulation of adhesion molecules on acute myeloid leukaemia blast cells. British Journal of Haematology 88(2):247–255, DOI 10.1111/j.1365-2141.1994.tb05014.x

Pearl J (2009) Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge University Press, New York

Porzelius C, Johannes M, Binder H, Beißbarth T (2011) Leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients. Biometrical Journal 53(2):190–201, DOI 10.1002/bimj.201000155

Savaraj N, Wu C, Kuo MT, You M, Wangpaichitr M, Robles C, Spector S, Feun L (2007) The relationship of arginine deprivation, argininosuccinate synthetase and cell death in melanoma. Drug Target Insights 2:119–128, DOI 10.4137/DTI.S0

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102(43):15,545–15,550, DOI 10.1073/pnas.0506580102

Vapnik V (1998) Statistical Learning Theory. Wiley, New York

Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7(1):91–98, DOI 10.1186/1471-2105-7-91

Xu L, Shen SS, Hoshida Y, Subramanian A, Ross K, Brunet JP, Wagner SN, Ramaswamy S, Mesirov JP, Hynes RO (2008) Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases. Molecular Cancer Research 6(5):760–769, DOI 10.1158/1541-7786.MCR-07-0344

Zhang G, He P, Tan H, Budhu A, Gaedcke J, Ghadimi BM, Ried T, Yfantis HG, Lee DH, Maitra A, Hanna N, Alexander HR, Hussain SP (2013) Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. Clinical Cancer Research 19(18):4983–4993, DOI 10.1158/1078-0432.CCR-13-0209