Empirical Evidence of the Applicability of Functional Clustering through Gene Expression Classification

Miloš Krejník and Jiří Kléma

Abstract—The availability of a great range of prior biological knowledge about the roles and functions of genes and gene-gene interactions allows us to simplify the analysis of gene expression data to make it more robust, compact and interpretable. Here, we objectively analyze the applicability of functional clustering for the identification of groups of functionally related genes. The analysis is performed in terms of gene expression classification and uses predictive accuracy as an unbiased performance measure. Features of biological samples that originally corresponded to genes are replaced by features that correspond to the centroids of the gene clusters and are then used for classifier learning. Using ten benchmark datasets, we demonstrate that functional clustering significantly outperforms random clustering without biological relevance. We also show that functional clustering performs comparably to gene expression clustering, which groups genes according to the similarity of their expression profiles. Finally, the suitability of functional clustering as a feature extraction technique is evaluated and discussed.

Index Terms—Biological prior knowledge, gene expression, gene set analysis, clustering, feature extraction, classification.

1 Introduction

Currently, there is a large range of bioinformatics tools that exploit *prior knowledge* of gene function. One important way to make use of this knowledge is through *functional clustering* (FC), which aims to group genes according to their functional similarities. The notion of functional similarity is based on the assumption that genes with related functional annotation records are functionally related to each other. Various approaches for FC are available [1], [2], [3], [4], [5]. The various approaches differ in their selection, heterogeneity and amount of employed prior biological knowledge, their notion of similarity between genes and the type of clustering algorithm used. The corresponding tools vary in their availability and serviceability.

The most frequent application of FC is to simply break down a large gene list into a manageable number of functionally related groups for further efficient *interpretation*. The origin of the gene list is commonly high-throughput genomic, proteomic and bioinformatics scanning approaches (mostly expression microarrays) that enable the researcher to select interesting (typically differentially expressed) genes. Thus, the FC tools contribute to gene-annotation *enrichment analysis*. The functional gene clusters can then be used to control the subsequent experiments such that a gene cluster is given preference, e.g., if most of its gene members are associated with highly enriched annotation terms that are found in the traditional enrichment analysis of the total gene list. [6] introduced the first tool for gene ontology functional analysis, the first discussion and comparison of various statistical functional analysis models is available in [7]. The detailed overviews of enrichment tools can be found in [8], [9].

However, functional annotations can also be employed in *classification* of gene expression (GE) data to obtain more interpretable, robust and potentially accurate predictive models. Classification based on GE monitoring by DNA microarrays (often referred to as molecular classification) is a natural learning task with immediate practical uses. There have been several early success stories [10], [11], [12], followed by a large number of studies with the main goal of predicting cancer outcome (an overview is provided, e.g., in [13]). Recent surveys [14], [15] have demonstrated serious technical flaws in a large proportion of these studies, which were published in high-impact biomedical journals, and have found that most of the published results are overly optimistic. The routine application of GE classification is limited by frequent inaccuracies in the resulting classifiers and their inability to be understood by physicians. Molecular classifiers based solely on GE in most cases cannot be considered useful decision-making tools or decision-supporting tools.

Recent efforts in the field of molecular classification aim to employ additional information available for genes, proteins and tissues that are being studied. They follow the major trend that is currently prevailing in the area of general GE data analysis. The anal-

• *The authors are with the Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, Prague 6, 166 27, Czech Republic.*
E-mail: krejnmi1@fel.cvut.cz, klema@labe.felk.cvut.cz.

ysis that was formerly aimed at identifying *individual genes* that are differentially expressed across sample classes [16] now focuses on identifying entire sets of genes with significantly different expression [17], [18], [19]. The genes share a set of characteristics that are defined by prior biological knowledge. The *set-level techniques* applied to GE classification develop new features that correspond to gene sets that represent pathways, their sub-clusters or gene-ontology terms at various levels of generality [20], [21]. The authors of [22] propose a method that integrates a priori the knowledge of a gene network into a classification that results in classifiers with biological relevance, a good classification performance and an improved interpretability of the results. [23] introduced the concept of condition-responsive genes (CORGs), which are the genes with the highest discriminative power in a pathway. The activity of a pathway is defined as a vector of CORG expression activity, and markers based on CORGs have been shown to improve the predictive results when compared with the random gene subset for a pathway. In [24], the authors compute the pathway activity score and the pathway consistency score. These two scores are then used as features for classifying phenotypes. The consistency score is defined using gene interaction networks.

In this paper, we propose the use of FC as *a feature extraction* tool for subsequent classification of GE samples. The main idea is to replace the sample features that originally corresponded to genes with a lower number of more robust, more interpretable features that correspond to the gene cluster centers. The dimensionality reduction of GE data by gene clustering with subsequent classification has already been proposed in [25]. The method is referred to as the prototype gene method, and the authors suggest that more accurate (and presumably more interpretable) classifiers can be created. However, this conclusion is only drawn from a two-dataset experiment. The paper does not employ any prior knowledge regarding gene function (the authors suggest that it will be used in future works) and derives the k-mean clusters by the Euclidean distance based on the GE profiles themselves.

This paper primarily addresses *the extent to which FC is useful in the analysis of GE data.* We assert that this question can only be partially answered when FC is applied within its traditional enrichment framework. In [2] the authors note that there is not a null hypothesis test to directly compare the quality of clustering algorithms. General remarks on the challenges of assessing the capabilities of any gene-set analysis method in real experiments can be found in [26], [17]. The common difficulty is that the ground truth is never known. The clustering outcome is therefore evaluated mainly in terms of its *interpretability* and in the scope of functional annotation data. Cluster compactness and stability are the most informative indirect measures of clustering outcome quality based on this point of view. The other common way of evaluating interpretability is purely subjective. Biomedical researchers interpret particular clusters, pick the most interesting clusters (those that can be given a plausible explanation) and compare them manually with other clusters derived from other bioinformatics tools. Although the comparison is convincing and the applicability of prior biological knowledge is broadly taken for granted, this method of evaluation leaves much room for subjective analysis. The author of [27], [28] summarizes the principal reasons for the demonstrable increase in the rate of false positive findings in research in general. It is also shown that the analysis of high-dimensional molecular data is increasingly affected by the risk for false positive conclusions.

This study considers another relevant criterion of clustering quality: *performance*. The performance criterion is orthogonal to the criteria of interpretability. It evaluates the clustering outcome in a wider context of GE data that underlie the creation of the gene list that is to be interpreted. Clearly, it is important that the clusters are interpretable, but they also need to prove meaningful in the original setting. The common method of performance evaluation is as follows. First, the gene clusters that are differentially expressed among the sample classes are identified. Then, the top-ranking clusters are interpreted, and it is demonstrated that their meaning is consistent with the definition of sample classes, which typically concern diseased and non-diseased individuals or different disease variants. This method of evaluation is as subjective as the interpretability evaluation mentioned above.

We propose the employment of an indirect, but entirely objective and impartial, method based on *predictive accuracy* (PA) to assess the performance of gene clustering approaches. The PA is estimated from the classification framework. The methodology that allows us to use PA to compare the efficiency of various types of gene clustering approaches is given briefly as follows. First, the involved genes whose expression levels are measured are clustered. Second, the features of the GE samples that originally corresponded to genes are replaced by features that correspond to the centroids of the clusters. Third, the classifiers, which are prescribed by formal models to determine the class of the new, unclassified samples, are learned, and their unbiased PA is estimated. Finally, the difference in the PA achieved for the various gene-clustering approaches is statistically evaluated. Note that the first two steps correspond to the procedure called feature extraction. The last two steps implement and evaluate the classification. Here, they serve to compare the different methods of feature extraction.

We assert that, there are two necessary conditions for applying FC to the analysis of GE data. First, the

gene functional clusters need to perform better than random gene clusters (random clustering decomposes a gene list by disregarding any available information on the genes). If not, the functional clusters have no meaning for the data that created the gene list. Second, the gene functional clusters must achieve a performance comparable to that of the clusters that are based on gene expression profile similarity (the approach mentioned earlier in [25]). If not, there is a straightforward way to better cluster the genes without knowledge regarding their functionality. Consequently, the vague initial question is rephrased in terms of two technical hypotheses that compare the PA achieved by classifiers based on different types of gene clustering approaches: (1) FC leads to a better predictive performance than *random clustering* (RC) without knowledge of biological relevance; and (2) FC and *gene expression clustering* (GEC), which groups genes according to the similarity of their expression profiles, have equally predictive performances.

This study should not be taken as an effort to develop the most accurate molecular classifiers. It instead aims to provide a robust test of the hypotheses stated above regarding the applicability of prior biological knowledge for further processing and understanding of GE data. To demonstrate the direct performance of FC in feature extraction for further classification, two more comparisons are drawn. We compare the FC-based feature extraction to *feature selection* that chooses the most differentially expressed genes and to the *fundamental treatment* that learns using all original data features.

The rest of this paper is organized as follows. Section 2 gives details on the FC, RC and GEC algorithms. Section 3 describes the experimental protocol and provides and interprets the hypothesis test results. Section 4 discusses a few additional issues on the applicability of FC. Section 5 reviews the contributions of this study and outlines directions for future work.

2 METHODS

This section reviews the differences among the gene clustering approaches (FC, RC and GEC) implemented here. It also summarizes the prior biological knowledge that is used in FC.

2.1 Biological prior knowledge

In this paper, we define prior biological knowledge as any information that is not available in a GE dataset but that is related to the genes contained in the dataset. There is a rich body of knowledge available for genes including a short textual description of gene function, the cellular location, a bibliography, interaction partners and links with other genes, membership and role in pathways, referential sequences and many other pieces of information.

The way we apply the biological prior knowledge in functional clustering was mainly inspired by the popular "DAVID Gene Functional Clustering Tool" [2], which represents one of the most consistent efforts to fuse the available knowledge found in various biological annotation databases (14 annotation categories including Gene Ontology, KEGG Pathways, BioCarta Pathways, Swiss-Prot Keywords). Technically, the uniform list of annotation terms adopted from DAVID is applied to describe each gene. The background knowledge is represented as a binary gene-term matrix enable to cope with the many-to-many gene-to-term relationships that are found in functional annotation databases.

On the other hand, there are obvious limitations of such a representation. The annotation does not fuse all of the possible heterogeneous knowledge resources, and gene links or genomic sequences cannot fit this format. The binary resolution ignores variance in reliability of the individual annotation records, e.g., the Gene Ontology evidence codes (the computationally derived annotations are generally thought to be of lower quality than those inferred from direct experimental evidence [29]). Pathways are treated as gene sets, their network structure is not concerned.

Because we implemented the presented method in R, we use the annotation packages from the open source Bioconductor bioinformatics software [30]. In particular, we use two annotation packages: the Affymetrix HuGeneFL Genome Array annotation data (hu6800.db for the GPL80 platform) and the Affymetrix Human Genome U133 Set annotation data version (hgu133a.db for the GPL96 platform), which correspond to the microarray chips from the datasets used in the experiments. Last but not least, there is a technical limitation of functional clustering caused by the significant number of probes and genes without annotation. In the employed versions 2.5.0 (hu6800.db) and 2.4.5 (hgu133a.db) of the annotation packages, 23%, respectively 43% of the probes remain unannotated and thus excluded from clustering.

2.2 Gene similarity/distance

The proper distance function is a keystone of any clustering algorithm. The gene distance grows with the dissimilarity of a gene pair, and the normalized distance is a real number from $\langle 0, 1 \rangle$, where 0 is the identity and 1 indicates the maximum possible dissimilarity. The gene similarity is the complement of the distance function to 1. The simplest definition of gene distance is applied in RC, where a pair of genes is assigned a random distance value. In GEC, the Euclidean distance is used. The Euclidean distance of two genes, u and v, is defined as

$$d(u, v) = \sqrt{\sum_{i=1}^{n} (x_{iu} - x_{iv})^2}, \quad (1)$$

where n is the number of samples and x_{iu} is the expression value of the gene, u, in sample, i. In FC, the kappa similarity measure adopted from [31], [2] is used. The kappa of a gene pair is computed from the binary vectors of the annotation terms assigned to the genes (the term can be present or absent for the given gene). The kappa of two genes, u and v, is defined as

$$\kappa(u,v) = \frac{O_{uv} - A_{uv}}{1 - A_{uv}}, \tag{2}$$

where O_{uv} represents the observed co-occurrence and A_{uv} represents the chance co-occurrence. Let $\mathcal{T}$ be a set of observed annotation terms, and let C_{00} be the number of terms that occur in neither u nor v. Let C_{01} be the number of terms that occur in v, but not in u, and let C_{10} be the number of terms that occur in u, but not in v. Finally, let C_{11} be the number of terms that are observed in both u and v. Then, O_{uv} and A_{uv} are defined as

$$O_{uv} = \frac{C_{11} + C_{00}}{|\mathcal{T}|} \tag{3}$$

and

$$A_{uv} = \frac{C_{*1}C_{1*} + C_{*0}C_{0*}}{|\mathcal{T}|^2}, \tag{4}$$

where $C_{*1} = C_{01} + C_{11}$, $C_{1*} = C_{10} + C_{11}$, $C_{*0} = C_{00} + C_{10}$ and $C_{0*} = C_{00} + C_{01}$.

2.3 Clustering algorithms

Gene clusters can be found using the gene distance/similarity measures. This subsection briefly reviews the clustering algorithms used earlier in FC and GEC and explains the choice of clustering algorithms made in this study.

The contribution of gene functional annotations in GE data analysis can be most easily illustrated when an identical clustering algorithm is used for functional, random and gene expression clustering. By applying only one clustering algorithm, we can increase the reliability of the hypothesis tests, as the issue of the influence of the clustering algorithm and its parameterization on the PA can be completely omitted. Therefore, we have reviewed the clustering algorithms that were actually applied earlier in FC and GEC, studied their evaluation or reevaluated them and attempted to identify an algorithm that best fits both fields of application. The algorithm selected also needs to be computationally feasible for large, genome-wide lists. Finally, the repetitive nature of our study needs to be addressed. In GEC, clustering needs to be performed for every single cross-validation split (10,000 total runs as we deal with 10 datasets, 10 fold cross-validation, 10 numbers of clusters and 10 repetitions). In FC, only 200 runs are needed (2 platforms, 10 numbers of clusters and 10 repetitions) because the clustering is independent of the GE data. Section 3.2 discusses the experimental design in detail.

The first candidate is the heuristic fuzzy partition (HFP) clustering algorithm that was developed for the DAVID Functional Annotation Clustering Tool [2]. The authors of the tool experimentally verified that fuzzy clustering best fits the gene annotation data and the nature of functional relationships of the genes from the viewpoint of interpretability. We therefore reimplemented the HFP clustering algorithm in R [32], accelerated it and made it scalable to genome-wide experiments. However, we have found that the HFP clustering algorithm does not suit the gene profile similarities that have distributions that are unlike the kappa similarity distribution for functional annotations. The algorithm is difficult to regulate to obtain a reasonable number of reasonably sized clusters (small changes in the control parameters often result in very different clustering of the initial gene set). In addition, the HFP clustering algorithm has a higher empirical computational complexity than a crisp clustering algorithm such as k-means or k-medoids clustering, and applying it multiple times for GEC is not computationally feasible.

The second candidate for a uniform clustering algorithm is the k-means algorithm [33], which was applied for GEC in [25], [34], [35], [36], [37]. The algorithm appears to be suitable for the GEC application from the viewpoint of PA, its ease of control and its efficiency for repetitive execution. Although the algorithm cannot be immediately applied to FC because it deals with cluster centroids whose functional annotation vectors are unclear, it can be replaced by a similar algorithm: k-medoids [38]. We believe that the k-medoids algorithm is the best choice of the three for the following reasons: (1) the algorithm shares its main characteristics with the k-means algorithm; both of the algorithms are partitional, crisp (not fuzzy) and minimize the distance between objects that belong to a cluster and the center of that cluster; (2) as with the HFP clustering algorithm, the k-medoids algorithm uses medoids as cluster centers in the place of centroids; it also allows the use of a similarity matrix instead of the data matrix for the input (the object coordinates in the feature space do not need to be available), and it is therefore more suitable for use with the κ similarity that is recommended by the DAVID Functional Annotation Clustering Tool; and (3) although fuzziness is a desirable property because of the biological nature of the gene functions and the resulting enhanced capabilities, e.g., for interpretation of the results, we have experimentally verified that the impact of k-medoids on PA with respect to the HFP clustering algorithm is marginal and appears to be positive.

In the end, our study implements two different clustering algorithms. We used our own Python implementation of the k-medoids algorithm for FC,

whereas GEC employs a Scipy [39] implementation of the k-means algorithm as a benchmark algorithm for GEC. In both FC and GEC, the initial medoids and centroids, respectively, were selected randomly from the considered genes. RC starts with the gene clusters that are found by FC. Then, the genes are randomly shuffled among the clusters. The random shuffling preserves the cluster sizes found in the FC and guarantees that the differences between the RC and FC are not the result of a different number and size of the clusters.

We believe that this strategy results in a less biased analysis than the direct comparison between the most frequently used algorithms for FC and GEC, which are the HFP and k-means algorithms. We have experimentally verified that the answers for the key questions remain the same with regard to FC and RC, which are driven by the HFP clustering algorithm and GEC performed with the k-means clustering algorithm. In this study, we emphasize the hypothesis tests that are reached with similar clustering algorithms in FC, RC and GEC, as they allow for a simpler and more readable formulation of the second technical hypothesis.

2.4 Cluster expressions

After gene clustering, the expression of the gene clusters needs to be computed. The original GE datasets are transformed from the original m-dimensional gene space into q-dimensional cluster space ($m \gg q$). Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ be a sample from the original feature space, where x_{ij}, $j = 1, \ldots, m$ is the expression value of the gene, j, in the sample, i. Next, let $C_1, \ldots, C_q$ be the gene clusters found via a particular clustering algorithm. Then, $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{iq})$ is a sample from the q-dimensional reduced space, where $\tilde{x}_{ij}$, $j = 1, \ldots, q$ is the expression for the value of the gene cluster, j, in the sample, i, which is computed as

$$\tilde{x}_{ij} = \frac{\sum\limits_{g \in C_j} x_{ig}}{|C_j|}. \quad (5)$$

3 Experiments

The goal of the conducted experiments was to compare FC, RC and GEC in terms of the PA of the classifiers learned on datasets that have the dimensions reduced by the given gene clustering approach. In this section, we describe the datasets that were used as well as the experimental framework, and we summarize the results.

3.1 Datasets

For the experiments, we used a set of ten publicly available GE datasets that have two class labels. The key parameters of the datasets are summarized in Table 1. The datasets were normalized by quantile normalization [49] to have the same distribution of GE for each sample in the given dataset. The following criteria were considered during the datasets selection: (1) availability – all of the datasets are publicly available via NCBI GEO [50] and have preferably been used by other researchers as benchmarks; (2) informedness – the GE measured must correlate with the target class somehow; otherwise, no clustering or learning approach will differ from random assortment; (3) difficulty – the relationship between GE and the target class must not be trivial or absolute; if a single gene perfectly splits the samples then there is no room for gene clustering; and (4) platform – we deal with only 2 microarray platforms to accelerate the experiments (RC and FC remain identical for different datasets that use the same platform).

3.2 Design

To compare FC, RC and GEC, we used 10 k values ($k = 2^c, c = 1, 2, \ldots, 10$) that determine the number of clusters, 10 datasets (see Section 3.1) and 5 classification algorithms (see below). For each combination of the gene clustering approach, number of clusters, classification algorithm and dataset, a PA value is computed as follows. At first, 10 partial PA values are computed, each of them is computed via stratified 10-fold cross-validation (as recommended in [51]) with different random seeds for the cluster initialization. Then, the final PA for the given combination is computed as the average of 10 partial PA values. The partial PA values are computed and averaged to avoid bias from random shuffling in RC and from random initialization in FC and GEC. In this way, 500 (10 k values $\times$ 10 datasets $\times$ 5 classification algorithms) final values of the PA for each gene clustering approach are obtained.

The particular classifiers were learned by five different classification algorithms: support vector machines [52] (with linear kernel and hyper-parameters of $C = 1.0$ and $\epsilon = 0.1$), random forests [53] (with 100 random trees from $\sqrt{n}$ random features, where n is the size of original dimension), C4.5 [54], naïve Bayes [55] and nearest neighbor [56]. The support vector machines represent the most frequently used classification algorithm in GE classification [57]. They are known to be able to cope with unfavorable rates of sampling (tissues and other biological situations) and variables (features or genes). The random forests method represents a robust ensemble classification algorithm that is suitable for GE data [58], whereas C4.5 produces decision trees that are instantly readable by a human and are the first option based on interpretability. The naïve Bayes and nearest neighbor algorithms represent classic and computationally efficient classification algorithms that are known to have reasonable accuracy, and in our study, they primarily serve to minimize the learning bias.

TABLE 1
An overview of the key parameters of the benchmark datasets.

Dataset	Reference	Number of samples	Class ratio	Number of features	Platform
ALL/AML	[12]	72	47:25	7,129	GPL80
AML	[40]	64	38:26	22,283	GPL96
Breast cancer	[41]	29	15:14	22,283	GPL96
Gastric cancer	[42]	30	22:8	7,129	GPL80
Glioma	[43]	85	59:26	22,283	GPL96
Hypertension	[44]	20	14:6	7,129	GPL80
MGCT	[45]	27	18:9	22,283	GPL96
Prostate cancer	[46]	20	10:10	22,283	GPL96
Sarcoma/Hypoxia	[47]	54	39:15	22,283	GPL96
Smoking	[48]	44	26:18	7,129	GPL80

We opted for the modifications of the classification algorithms that require as few hyper-parameters as possible to avoid needing another nested cross-validation cycle to optimize them. The nested cross-validation is time consuming, especially for GEC, as it would multiply the number of clustering runs (genome-wide clustering is the most time-consuming step). It also tends to decrease the sample numbers and the variability in the individual stratified folds. The actual applied hyper-parameters are known to be robust at their default setting (support vector machines) or there has been a recommendation for their heuristic prior initialization (random forests). Orange [59] implementation of the classification algorithms was applied.

3.3 Results

By applying the described procedure, 1,500 (3 gene clustering approaches $\times$ 10 k values $\times$ 10 datasets $\times$ 5 classification algorithms) estimations of PA were obtained. The main objective of our study is to compare the individual gene clustering approaches. The hypotheses regarding the equality of the gene clustering approaches in terms of their predictive performance were tested via the Wilcoxon signed-rank test [60], as recommended in [61] in place of the widely used t-test. The hypotheses were tested at a level of significance of $\alpha = 0.05$. If not stated otherwise, the same statistical test and the same α level were used in other experiments too.

First, the medians over the 500 PA values available for the individual clustering approaches can be computed. However, this condensed summary gives only a rough view of the total performance because the PA measured in the different domains is not commensurable and is highly variable; therefore, aggregating it over domains is not meaningful [61]. Instead, mutual direct comparisons should be based on the gene clustering approach rankings, which consider successes and failures rather than the absolute accuracy of the methodology. For example, for the ALL/AML domain, naïve Bayes classifier algorithm and $k = 16$ (16 clusters), the gene clustering approaches had accuracies of FC 90%, RC 83%, and GEC 92%. The ranking is FC – 2nd, RC – 3rd and GEC – 1st; the difference in the PA does not matter. The mean ranks are meaningful even if they are obtained over different datasets. The conclusions with regard to the ranks of the clustering approaches are shown in Table 2 (the final row gives a condensed summary). As outlined in Section 1, our main interest is in paired FC versus RC and in FC versus GEC. The first null hypothesis, that FC and RC have equally predictive performances, was rejected in favor of the alternative hypothesis, FC has a higher predictive performance than RC (one-sided test, p-value $= 0.042$, which is $< \alpha$). The second null hypothesis, FC and GEC are equally predictive, could not be rejected in favor of the alternative hypothesis, FC and GEC have distinct predictive performances (two-sided test, p-value $= 0.85$, which is $> \alpha$).

TABLE 2
Mean ranks of the gene clustering approaches with regard to PA. The table shows the mean domain ranks (averaged over all of the classification algorithms and k values) and the total mean ranks (averaged over all of the domains, last row).

Dataset	FC	GEC	RC
ALL/AML	1.53	1.82	2.64
AML	2.22	1.58	2.20
Breast cancer	2.00	2.10	1.90
Gastric cancer	1.52	2.26	2.21
Glioma	2.32	1.62	2.05
Hypertension	1.50	2.42	2.07
MGCT	2.33	1.46	2.21
Prostate cancer	1.91	1.82	2.27
Sarcoma/Hypoxia	2.00	1.26	2.74
Smoking	1.20	2.98	1.82
All	1.85	1.93	2.21

However, the most relevant conclusions must be drawn from the paired differential analysis that has the largest statistical power. The analysis relates the accuracy values reached by two gene-clustering approaches when the other settings are identical. Our interest is again in paired FC versus RC and in FC versus GEC; therefore, 500 (10 k values $\times$ 10 datasets $\times$ 5 classification algorithms) differential values are obtained for each pair when the differential accuracy for both of the clustering pairs is calculated. The box plots for the particular datasets and the clustering pairs are depicted in Fig. 1.

The following statistical test summarizes the visual differences seen in the results shown in Fig. 1. Prior to the test, the aggregate across the k values and classification algorithms has to be calculated because the runs with different classification algorithms and different k values within a dataset are dependent (that is, a higher accuracy in one predicts a higher accuracy in the others and the same holds true for differences). Then, the final test deals with 10 medians of 50 (10 k values $\times$ 5 classification algorithms) different accuracy values. In other words, it tests a vector of 10 independent median values that are derived for 10 different datasets. The median is used in place of the mean because the differential accuracy for the particular datasets has an asymmetric distribution.

The first null hypothesis, that FC and RC have equally predictive performances, was rejected in favor of the alternative hypothesis, FC has a higher predictive performance than RC (one-sided test, p-value $= 0.019$, which is $< \alpha$). FC performed better than RC on eight out of ten benchmark datasets. Compared with a randomly selected gene set, the functional cluster has increased interpretability and performance.

The second null hypothesis, FC and GEC are equally predictive, could not be rejected in favor of the alternative hypothesis, FC and GEC have distinct predictive performances (two-sided test, p-value $= 0.92$, which is $> \alpha$). FC performed better on five of ten benchmark datasets, and GEC performed better on the other five datasets, which suggests that functional clusters represent an alternative to purely statistical clusters in terms of PA. Note that GEC often identifies gene clusters that share no common annotation pattern and cannot be plainly interpreted. In the case of equally predictive performances, preference is given to the more interpretable option. This option is clearly represented by functional clusters, which are naturally complemented by a shared functional pattern.

4 Discussion

This section provides comments that will aid in the interpretation of the results provided in the previous section, describes the influence from the number of clusters and the classification algorithm and compares two principal approaches for dimensionality reduction. Although the discussed issues can be regarded as technical details with respect to the key questions, they may help place the results into perspective and provide additional details.

4.1 Number of clusters

The clustering algorithms used enabled us to immediately compare the gene clustering approaches based on the functional, gene-expression-based and random gene distances across the considered k values. The differential comparisons can be seen in Fig. 2. The margin between FC and RC is most distinct for lower numbers of clusters and tends to decrease steadily as the number of clusters increase. A few large random clusters have significantly less information than the functional ones, whereas the large number of smaller random clusters can have nearly the same level of informedness as the functional ones. This observation is in agreement with an earlier conclusion that the enrichment of gene expression clusters for biological function is generally the highest at a relatively low number of clusters [62]. FC generates large clusters of genes that tend to share expression profiles, and this relationship decreases as the number of clusters increases. The margin between functional clustering and GEC does not show a strong pattern.

Fig. 3 shows that the PA increases with an increasing number of clusters. The gene clustering approaches are comparable with the full set of features (the dotted line) when the number of clusters reaches approximately 100, which suggests that the original performance can be maintained with a reasonable dimensionality reduction; however, the number of clusters cannot be extremely low without sacrificing PA. Note that the optimal number of clusters differs across domains. As a matter of fact, there are 5 domains with a clear coherent range of the numbers of clusters with the PA of FC higher than the referential one derived from the full gene set. This characteristic is not obvious in Fig. 3 for its aggregation over domains.

4.2 Classification algorithms

We experimented with five diverse classification algorithms (see Section 3.2). None of the methods given below is superior to the others in principle. The main reason for using the pool of learning algorithms is to avoid a dependence of the experimental results on their specific biases. Therefore, the answer given by the pool of methods is more illustrative and robust than the answers provided by any given method. Still, a brief comparison of the classification algorithms can illustrate their differences. Fig. 4 shows the overall performance of the individual algorithms. The only significantly different pairs are random forests versus C4.5 and random forests versus support vector machines (Friedman test [63], p-value $= 0.019$ and p-value $= 0.039$, respectively). The low accuracy of the support vector machines algorithm (with a linear

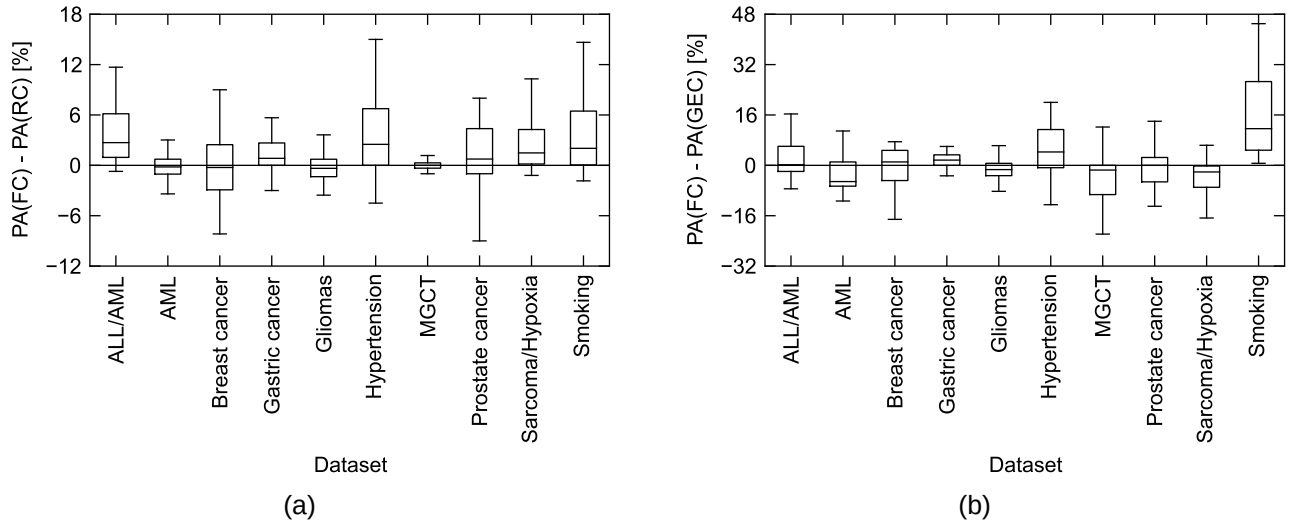

Fig. 1. Box plots for the PA differences for the given datasets and hypotheses: (a) FC versus RC; and (b) FC versus GEC. Each box plot is computed from 50 (10 k values $\times$ 5 classification algorithms) values for the PA difference.

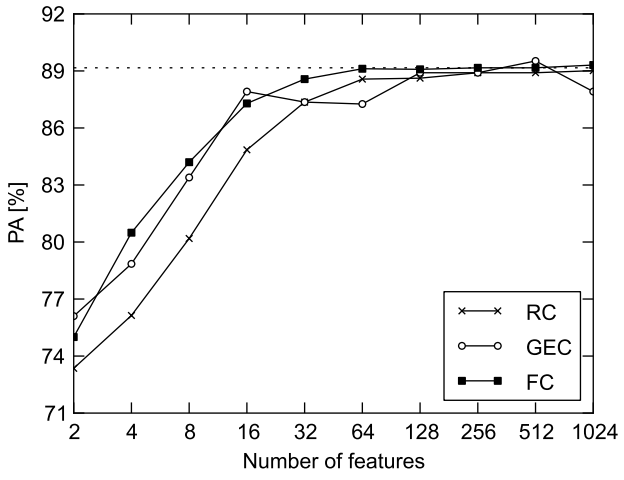

Fig. 3. Medians of the PAs for three ways of gene clustering. The dotted line represents the median of the PAs for the full gene set without dimension reduction. Each median is computed from 50 (10 datasets $\times$ 5 classification algorithms) values for the PA.

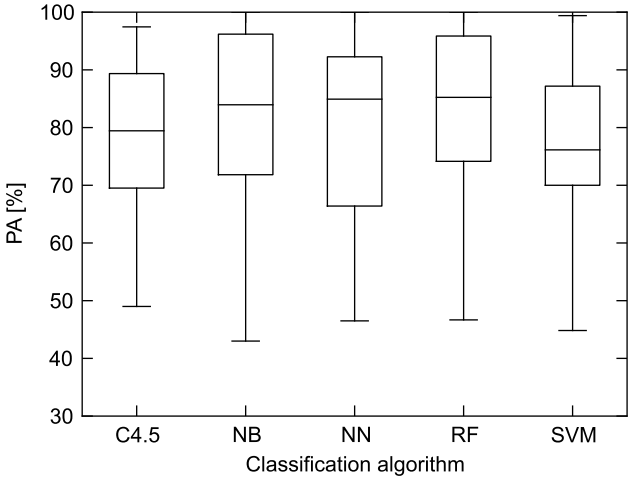

Fig. 4. Box plots for the PAs for the given classification algorithms, namely C4.5, naïve Bayes (NB), nearest neighbor (NN), random forests (RF) and support vector machines (SVM). Each box plot is computed from 300 (3 gene clustering approaches $\times$ 10 k values $\times$ 10 datasets) values for the PA.

kernel) indicates the nonlinearity of the classification problems that are being considered. The improved accuracy of FC with respect to RC is preserved across the classification algorithms; its significance can be proven for the nearest neighbor and random forests algorithms (one-sided test with Bonferroni-Dunn adjustment, p-value $= 0.003$ and p-value $= 0.041$, respectively).

4.3 Feature selection

This paper focuses on clustering as a method that reduces the dimensionality of GE data. The new features that are generated are represented by the cluster centroids, which are extracted from the original features. The parallel approach to dimensionality reduction lies in feature selection (FS); a review of its use in bioinformatics can be found in [64]. FS is frequently implemented with GE data for the selection of differentially expressed genes. Criteria such as the absolute t-test statistic can be used to rank the genes, and permutation tests can help to establish a threshold for genes that are significantly related to the response. To place the algorithms for feature extraction that were discussed and compared in this study into a wider context, we also compared their performance

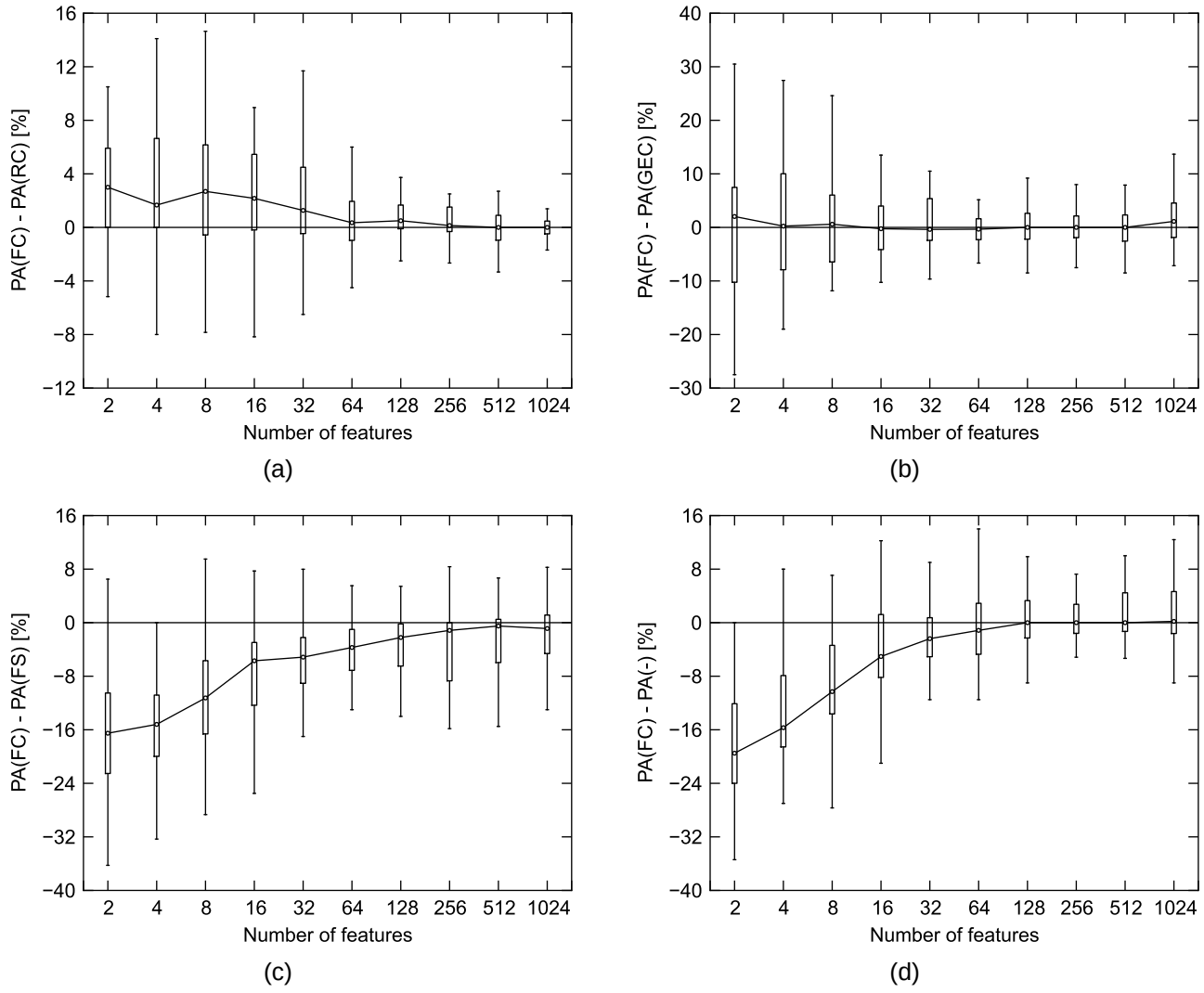

Fig. 2. Box plots for the PA differences for a given number of features and pairs of feature extraction/selection approaches: (a) FC versus RC; (b) FC versus GEC; (c) FC versus FS; and (d) FC versus the full gene set without dimension reduction. Each box plot is computed from 50 (10 datasets $\times$ 5 classification algorithms) values for the PA difference.

against FS. We ranked the genes by t-test, selected the most differentially expressed genes (the thresholds were gradually set to match the number of clusters) and ran the classification algorithms. The process was repeated 10 times for 10-fold cross-validation. As shown in Fig. 2(c), the PA achieved is clearly superior to that achieved by clustering. The null hypothesis that FS and FC have equally predictive performances was rejected (two-sided test, p-value $= 0.002$), which is not surprising because FC ignores the sample class labels, a significant information source for the feature transformation phase. Fig. 5 demonstrates that FS improves a PA in comparison with the full gene set without dimension reduction.

4.4 Functional clustering improvements

Our study did not aim to achieve the maximum PA. To do so, FS would clearly be the first dimension reduction option chosen on the basis of its simplicity and performance. Maximization of the PA by FC would include FS as one of the early steps. We have implemented and tested a simple FC improvement that exploits FS and the sample class labels: (1) in order to reduce noise, the cluster centroids represent only differentially expressed probes (t-test is applied, the probes with p-value $< 0.01 \log_2 k$ are used, the threshold increases with k to minimize empty or trivial centroids); (2) in order to minimize the negative influence of averaging, each cluster is represented by 2 centroids, upregulated and downregulated probes are treated separately; and (3) to keep the number of centroids equal with the number of clusters, the final set of k cluster centroids is made by the most differentially expressed ones. Fig. 6(a) shows that the improvements boost the PA of FC, Fig. 6(b) demonstrates that its performance becomes comparable with

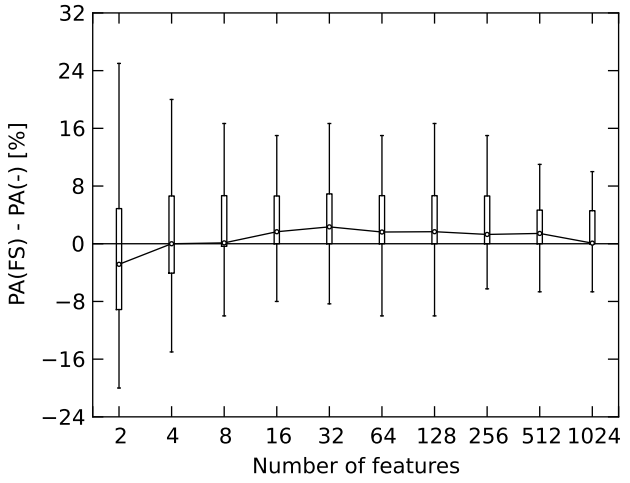

Fig. 5. Differential PA box plots comparing classification based on FS and the full gene set without dimension reduction. Each box plot is computed from 50 (10 datasets $\times$ 5 classification algorithms) PA differential values.

FS. Although it can be argued that FS is still an easier method to reduce dimension, the above described experiment suggests that the approaches that combine FC with FS (and potentially GEC) shall not be ignored.

5 Conclusion

This paper proposes a general methodology to impartially verify the applicability of particular types of gene clustering approaches. The verification is conducted within the predictive classification framework and focuses on prior biological knowledge-based FC. The framework uses three parallel methods of gene clustering. It statistically tests for differences in the PA of machine learning classifiers that are trained on the centroids of particular clusters. We experimentally verified that FC has a higher PA than RC without biological relevance. The effect of prior biological knowledge is remarkable for two main reasons: (1) it can be statistically verified for a limited set of ten GE datasets; and (2) it persists in simplified cluster construction based on GE averaging (see Equation 5), which does not distinguish between gene activation and inhibition. We also showed that FC performs comparably to GEC, which groups genes according to the similarity of their expression profiles.

In addition, we showed that FC can provide a reasonable dimensionality reduction without sacrificing the PA achieved with the full set of features. This observation is promising concerning simplicity of the currently implemented FC, namely the above-mentioned cluster averaging, but also the frequent utilization of genes whose GE profiles have no relation to the phenotype, the imperfections in gene distance calculation and the probes and genes with missing annotations. Another interesting characteristic is that FC is carried out independently of GE data, which makes it an unsupervised and potentially computationally efficient feature extraction technique. Unlike GEC, FC is carried out just once per a particular gene set (platform) and the clusters are immediately applicable across the GE experiments using the particular platform.

At the same time, it holds that FC does not achieve a PA that is comparable to that achieved by FS, and combining the two techniques would maximize performance. It was experimentally demonstrated that FS is a simple method that improves a PA in a vast majority of domains (of course, the conclusion is influenced by the selection of classification algorithms and their noise robustness) and differential expression can hardly be ignored when calculating the cluster aggregates.

There are several directions for future work. First, the current pair of hypotheses can logically be supplemented by a third null hypothesis, there is no synergic action between the knowledge-based FC, GE-based GEC. We showed that both GE data and prior biological knowledge regarding gene roles, functions and interactions can underlie the creation of gene clusters. There are at least three reasons to believe that these algorithms can complement each other: (1) FC corresponds to a universal gene partitioning, whereas GEC provides a local partitioning for specific biological conditions; (2) FC clusters only the genes with an existing annotation, whereas genes without an annotation are left unused or create a cluster without real meaning; GEC uses all of the genes (both with and without annotation), which gives GEC an advantage over FC; and (3) FC deals with human-created annotations, whereas GEC creates ad hoc links based on a limited number of arrays that are known to provide only a noisy image of gene actions. However, the testing of this hypothesis lies beyond the scope of this paper, as there are many ways to aggregate clusters raised from FC and GEC into unified knowledge and statistical groups. Some general ideas regarding clustering aggregation can be found in [65]. In [66], the authors introduce the problem of combining multiple partitions of a set of objects into a single consolidated clustering without accessing the features or algorithms that determined these partitions. A discussion on early, intermediate and late integration of microarray and medical literature data for gene clustering can be found in [67].

Second, the current cluster expression is computed in the most straightforward way by averaging the expression levels of the cluster members. A more complex cluster activity function could also consider the internal structure of the gene set that generates a cluster. The structure could potentially be extracted from the prior biological knowledge, and it could also be (re)invented statistically from GE data. However, preliminary efforts to employ the statistical SVD

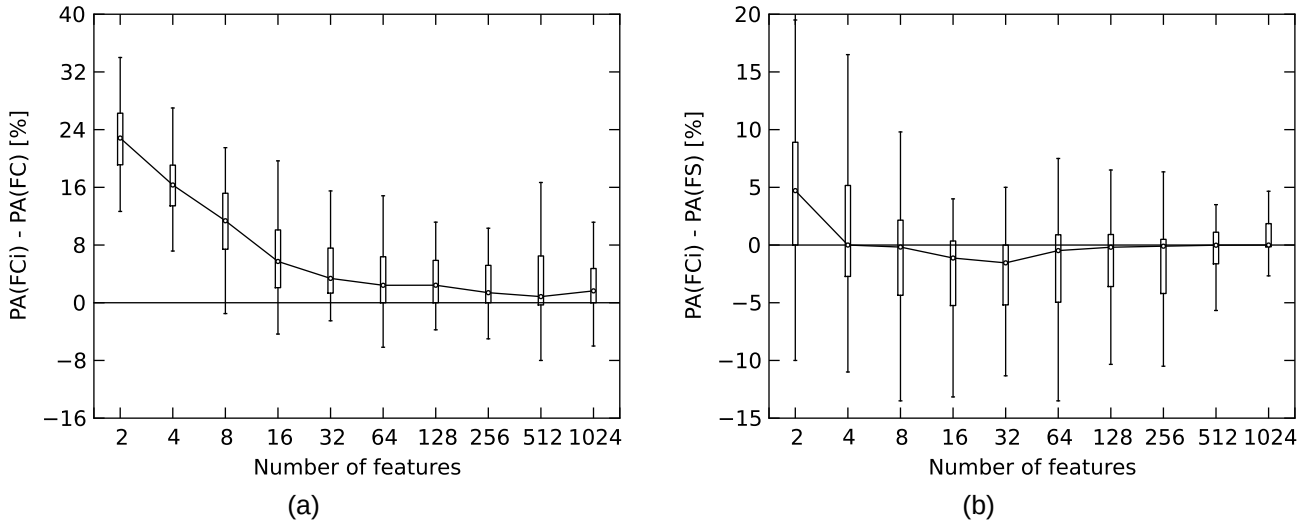

Fig. 6. Illustration of the effects of FC improvements (FCi). Box plots for the PA differences for a given number of features and pairs of feature extraction/selection approaches: (a) FCi versus FC; (b) FCi versus FS. Each box plot is computed from 50 (10 datasets $\times$ 5 classification algorithms) values for the PA difference.

method for constructing metagenes proposed in [68] did not provide a detectable immediate improvement [21].

ACKNOWLEDGMENTS

The work of Miloš Krejník was funded by the Grant Agency of the Czech Technical University in Prague (grant no. SGS10/187/OHK3/2T/13). The work of Jiří Kléma was funded by the Czech Ministry of Education in the framework of the research program, Transdisciplinary Research in the Area of Biomedical Engineering II (MSM 6840770012). We thank Tomáš Sixta for reimplementation of the DAVID clustering algorithm in R. We also thank Robin Healey for English proofreading.

REFERENCES

[1] D. Chaussabel and A. Sher, "Mining microarray expression data by literature profiling," *Genome Biology*, vol. 3, no. 10, pp. research0055.1–research0055.16, 2002.

[2] D. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists," *Genome Biology*, vol. 8(9), no. R183, 2007.

[3] J. Natarajan and J. Ganapathy, "Functional gene clustering via gene annotation sentences, MeSH and GO keywords from biomedical literature," *Bioinformation*, vol. 2, no. 5, pp. 185–193, 2007.

[4] K. Ovaska, M. Laakso, and S. Hautaniemi, "Fast gene ontology based clustering for microarray experiments," *BioData mining*, vol. 1, no. 1, p. 11, 2008.

[5] G. Macintyre, J. Bailey, D. Gustafsson, I. Haviv, and A. Kowalczyk, "Using gene ontology annotations in exploratory microarray clustering to understand cancer etiology," *Biochemistry*, vol. 31, no. 14, pp. 2138–2146, 2010.

[6] P. Khatri, S. Draghici, G. C. Ostermeier, and S. A. Krawetz, "Profiling gene expression using onto-express," *Genomics*, vol. 79, no. 2, pp. 266–270, 2002.

[7] S. Draghici, P. Khatri, R. Martins, G. Ostermeier, and S. Krawetz, "Global functional profiling of gene expression," *Genomics*, vol. 81, no. 2, pp. 98–104, 2003.

[8] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems." *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005.

[9] D. W. W. Huang, B. T. T. Sherman, and R. A. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." *Nucleic acids research*, November 2008.

[10] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," in *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 2000, pp. 54–64.

[11] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[13] J. Lee, J. Lee, M. Park, and S. Song, "An extensive evaluation of recent classification tools applied to microarray data," *Computational Statistics and Data Analysis*, vol. 48, pp. 869–885, 2005.

[14] A. Dupuy and R. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting," *JNCI Journal of the National Cancer Institute*, vol. 99, no. 2, p. 147, 2007.

[15] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[16] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[17] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 23, pp. 15 545–1550, 2005.

[18] I. Dinu, J. Potter, T. Mueller, Q. Liu, A. Adewale, G. Jhangri, G. Einecke, K. Famulski, P. Halloran, and Y. Yasui, "Improving

gene set analysis of microarray data by sam-gs," *BMC bioinformatics*, vol. 8, no. 1, p. 242, 2007.

[19] Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J. Chong, M. Fukayama, T. Kodama, and H. Aburatani, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, no. 8, p. 980, 2007.

[20] M. Holec, F. Železný, J. Kléma, and J. Tolar, "Integrating multiple-platform expression data through gene set features," in *Proceedings of the 5th International Symposium on Bioinformatics Research and Applications*. Springer-Verlag Berlin, Heidelberg, 2009, pp. 5–17.

[21] ——, "A comparative evaluation of gene set analysis techniques in predictive classsification of expression samples," in *International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics (BCBGC-10)*, 2010.

[22] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. P. Vert, "Classification of microarray data using gene networks," *BMC Bioinformatics*, vol. 8, no. 1, pp. 35+, 2007.

[23] E. Lee, H. Chuang, J. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, no. 11, 2008.

[24] S. Efroni, C. F. Schaefer, and K. H. Buetow, "Identification of key processes underlying cancer phenotypes using biologic pathway analysis." *PLoS ONE*, vol. 2, no. 5, 2007.

[25] B. Hanczar, M. Courtine, A. Benis, C. Hennegar, K. Clément, and J.-D. Zucker, "Improving classification of microarray data using prototype-based feature selection," *SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 23–30, 2003.

[26] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[27] J. P. A. Ioannidis, "Genetic associations: false or true?" *Trends Mol Med*, vol. 9, no. 4, pp. 135–8, 2003.

[28] ——, "Why most published research findings are false," *PLoS Med*, vol. 2, no. 8, p. e124, 08 2005.

[29] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations." *Nature Reviews Genetics*, vol. 9, no. 7, pp. 509–515, 2008.

[30] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. R80, 2004.

[31] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, no. 20, pp. 37–46, 1960.

[32] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0.

[33] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. California, USA, 1967, pp. 281–297.

[34] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng, "Evaluation and comparison of gene clustering methods in microarray analysis," *Bioinformatics*, vol. 22, pp. 2405–2412, September 2006.

[35] G. Kerr, H. Ruskin, M. Crane, and P. Doolan, "Techniques for clustering gene expression data," *Computers in biology and medicine*, vol. 38, no. 3, pp. 283–293, 2008.

[36] I. Priness, O. Maimon, and I. Ben-Gal, "Evaluation of gene-expression clustering via mutual information distance measure," *BMC bioinformatics*, vol. 8, no. 1, p. 111, 2007.

[37] F. De Smet, J. Mathys, K. Marchal, G. Thijs, B. De Moor, and Y. Moreau, "Adaptive quality-based clustering of gene expression profiles," *Bioinformatics*, vol. 18, no. 5, p. 735, 2002.

[38] L. Kaufman and P. Rousseeuw, *Finding Groups in Data An Introduction to Cluster Analysis*. New York: Wiley Interscience, 1990.

[39] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online]. Available: http://www.scipy.org/

[40] D. Stirewalt, S. Meshinchi, K. Kopecky, W. Fan, E. Pogosova-Agadjanyan, J. Engel, M. Cronk, K. Dorcy, A. McQuary, D. Hockenbery *et al.*, "Identification of genes with abnormal expression changes in acute myeloid leukemia," *Genes, Chromosomes and Cancer*, vol. 47, no. 1, pp. 8–20, 2008.

[41] A. Tripathi, C. King, A. de la Morenas, V. Perry, B. Burke, G. Antoine, E. Hirsch, M. Kavanah, J. Mendez, M. Stone *et al.*, "Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients," *International Journal of Cancer*, vol. 122, no. 7, pp. 1557–1566, 2008.

[42] Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J. Chong, M. Fukayama, T. Kodama, and H. Aburatani, "Global gene expression analysis of gastric cancer by oligonucleotide microarrays," *Cancer research*, vol. 62, no. 1, p. 233, 2002.

[43] W. Freije, F. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, L. Liau, P. Mischel, and S. Nelson, "Gene expression profiling of gliomas strongly predicts survival," *Cancer Research*, vol. 64, no. 18, p. 6503, 2004.

[44] T. Bull, C. Coldren, M. Moore, S. Sotto-Santiago, D. Pham, S. Nana-Sinkam, N. Voelkel, and M. Geraci, "Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension," *American Journal of Respiratory and Critical Care Medicine*, vol. 170, no. 8, pp. 911–919, 2004.

[45] R. Palmer, N. Barbosa-Morais, E. Gooding, B. Muralidhar, C. Thornton, M. Pett, I. Roberts, D. Schneider, N. Thorne, S. Tavaré *et al.*, "Pediatric malignant germ cell tumors show characteristic transcriptome profiles," *Cancer Research*, vol. 68, no. 11, p. 4239, 2008.

[46] C. Best, J. Gillespie, Y. Yi, G. Chandramouli, M. Perlmutter, Y. Gathright, H. Erickson, L. Georgevich, M. Tangrea, P. Duray *et al.*, "Molecular alterations in primary prostate cancer after androgen ablation therapy," *Clinical cancer research*, vol. 11, no. 19, p. 6823, 2005.

[47] K. Detwiller, N. Fernando, N. Segal, S. Ryeom, P. D'Amore, and S. Yoon, "Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on rna interference of vascular endothelial cell growth factor a," *Cancer research*, vol. 65, no. 13, p. 5881, 2005.

[48] B. J. Carolan, A. Heguy, B.-G. Harvey, P. L. Leopold, B. Ferris, and R. G. Crystal, "Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase l1 gene in human airway epithelium of cigarette smokers," *Cancer Research*, vol. 66, no. 22, pp. 10 729–10 740, 2006.

[49] B. Bolstad, R. Irizarry, M. Åstrand, and T. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, p. 185, 2003.

[50] T. Barrett, D. Troup, S. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "Ncbi geo: mining tens of millions of expression profiles–database and tools update," *Nucleic Acids Research*, vol. 35, no. Database issue, p. D760, 2007.

[51] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International joint Conference on artificial intelligence*, vol. 14. Morgan Kaufmann, 1995, pp. 1137–1143.

[52] V. Vapnik, *The nature of statistical learning theory*. Springer Verlag, 2000.

[53] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[54] J. Quinlan, *C4.5: programs for machine learning*. Morgan Kaufmann, 1993.

[55] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001, pp. 41–46.

[56] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, 1967.

[57] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, p. 262, 2000.

[58] R. Díaz-Uriarte and S. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[59] J. Demšar, B. Zupan, G. Leban, and T. Curk, "Orange: From experimental machine learning to interactive data mining,"

Knowledge Discovery in Databases: PKDD 2004, pp. 537–539, 2004.
[60] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80–83, 1945.
[61] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
[62] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Research*, vol. 12, no. 10, pp. 1574–1581, 2002.
[63] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.
[64] Y. Saeys, I. n. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507–2517, September 2007.
[65] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," in *Proceedings of the 21st International Conference on Data Engineering*, 2005, pp. 341–352.
[66] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.
[67] P. Glenisson, J. Mathys, and B. de Moor, "Meta-clustering of gene expression data and literature-based information," *SIGKDD Explorations*, vol. 5, pp. 101–112, 2003.
[68] J. Tomfohr, J. Lu, and T. B. Kepler, "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, vol. 6, 2005.

Miloš Krejník received the BSc degree in computer technology and the MSc degree with honours in cybernetics and measurement from the Czech Technical University in Prague (CTU), in 2006 and 2008, respectively. In 2007, he was a Researcher in the Gerstner Laboratory at CTU. In 2009–2011, he was a Quantitative Analyst at Analytical Department at RSJ, Prague, Czech Republic. Currently, he is pursuing the PhD degree in artificial intelligence and biocybernetics at CTU. His research interests focus on statistical machine learning in bioinformatics and finance.

Jiří Kléma received the PhD in artificial intelligence and biocybernetics from the Czech Technical University in Prague (CTU) in 2002. In 2005–2006 he carried out post-doctoral training at the University of Caen, France. Currently, he is an Assistant Professor at CTU. His main research interest is data mining and its applications in industry, medicine and bioinformatics. He focuses namely on knowledge discovery and learning in domains with heterogeneous and complex background knowledge. He is a co-author of 15 journal publications and book chapters, a reviewer for several international journals and a member of the Presidium of The Czech Society for Cybernetics and Informatics.