Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases.

Argiris Sakellariou*

National and Kapodistrian University of Athens,
Department of Informatics and Telecommunications
argisake@gmail.com

Abstract. In this dissertation, we address the problem of gene selection from ranked gene lists. We propose a new hybrid feature selection method (m*AP*-KL) that combines successfully multiple hypothesis testing and affinity propagation clustering algorithm along with the Krzanowski & Lai cluster quality index, to select a small yet informative subset of genes. We subject our method across a variety of validation tests on simulated microarray data as well as on real microarray data. The overall evaluation results suggest that m*AP*-KL generates concise yet biologically relevant and informative *n*-gene expression signatures, which can serve as a valuable discrimination tool for diagnostic and prognostic purposes, by identifying potential disease biomarkers in a broad range of diseases. Finally, to provide the research community with the capability to apply m*AP*-KL in any given gene expression dataset, we have implemented this methodology to a Bioconductor/R-package accompanied with extra functionalities.

KEYWORDS: microarrays, gene expression data, significance analysis, hybrid feature selection, biomarkers

1 Introduction

The dawn of DNA microarray technology has improved our potential to comprehend the underlying mechanisms of human diseases and to aid in more accurate classification, diagnosis, and/or prognosis. Because of its high throughput nature, computational tools are essential in data analysis and mining in order to help biomedical researchers to maximize the extracted knowledge from the experimental results. In the area of diagnostics, microarray-derived markers are emerging as a valuable tool. Similar to any other clinical test, the primary goal of molecular tests, including microarray tests, is to provide reliable and timely results for improving patient care. In order to maximize the usefulness of microarrays in the diagnostic/prognostic arena it is important to

* Dissertation Advisor: Sergios Theodoridis, Professor.

minimize the number of biomarkers that need to be tested for an accurate diagnosis to be reached.

The selection of those biomarkers, however, is a challenging process in which feature selection (FS) methods could make a significant contribution. Indeed, from the late 90s a plethora of methods emerged and applied on several microarray studies. Despite differences in their fundamental algorithms, they all share the same objectives: 1) to avoid overfitting and improve prediction performance; 2) to make faster and cost effective models; and 3) to offer a deeper insight into the underlying processes [1]. Nevertheless, selecting those 'significant' genes that perform the same level of classification in relation to a specific disease is far from feasible at the moment and still an open issue.

2 Related work

In reality, every microarray dataset may result to as many significant gene lists to as many FS methods we apply. Even in cases where methods share the same principles the produced gene lists are bound to diverge. Speaking of methods that share common principles, we may define the following broad groups of FS methods. Filtering, wrapper and embedded FS methods are the key categories in the field, each one with the respective advantages and disadvantages. In addition to this classification, a new class of FS methods, hybrid methods, has emerged. Hybrid methods' combine methods of different categories aiming at taking advantage of their pros while alleviating their cons of benefit to the 'significant' gene list selection.

Combining methods is a constructive decision making process based always on scientific assumptions, either biological or statistical, rather than on pot luck. For instance, Jaeger et al. [2] claimed that ranking algorithms produce lists of genes, where the top ranked genes are highly correlated with each other, mainly because they belong to the same pathway. Additionally, Hall in his thesis [3] investigated the hypothesis that "A good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other". Those beliefs were the springboard for several hybrid methods, which combined a ranking (filtering) method and a clustering method to conclude to a list of significant genes.

In particular, Jaeger et al. employed a fuzzy clustering algorithm to prefilter the genes by grouping them according to their similarity. Then, with the aid of a statistical test like t-test or Wilcoxon, selected one or more representative genes from each cluster to form a list of 'significant' yet uncorrelated genes. In this study, the number of clusters to be formed and the number of representative genes remained unaddressed. Similar to Jaeger et al., Hanczar et al. [4] proposed a two step method where an unsupervised clustering method, K-mean, combined with a mathematical notion, prototype gene, that tries to identify the representative genes of each cluster. Analogous issues to Jaeger et al. appeared in this study, and characterized as objectives for future work by the researchers. An alternative algorithmic approach, where ranking of genes precedes any other method is described in the mRMR [5] method. Particularly, the initial ranking through t-test or F-test is then combined with a se-

quential iteration between pairs of the ranked genes, to conclude to a subset of 'significant' genes according to some criteria, maximum relevance and minimum redundancy. One considerable drawback of this approach is that the redundancy criterion may exclude genes that considered important from a biological point of view. Another interesting approach, HykGene [6], proposed a three step gene selection, which incorporates a filtering algorithm, a hierarchical clustering on the top-ranked genes and finally a sweep-line algorithm that first identifies the clusters from the dendrogram and then selects one representative gene per cluster.

Taking into account the promising classification results of those combined methods as well as their intrinsic limitations, we considered a new hybrid method, m*AP*-KL [7]. In the proposed approach, the genes are first ranked according to their differential expression using a multiple hypothesis t-test, which controls successfully the Type I error. Then the top *N* ranked genes are held and grouped to clusters with the Affinity Propagation (AP) clustering algorithm [8]. Prior to AP a clustering index algorithm determines the number of clusters among the top-*N*-genes. The output of this method is a subset of genes, one exemplar per cluster that best describes the phenotypes' characteristics.

3 Proposed Hybrid Feature Selection method (m*AP*-KL)

A FS method, in microarray gene expression data, should be independent of platform, disease and dataset size. Our hypothesis is that among the statistically significant ranked genes in a gene list, there should be clusters of genes that share similar biological functions related to the investigated disease. Thus, instead of keeping N top ranked genes, it would be more appropriate to define and keep a number of gene cluster exemplars. We propose a hybrid FS method (m*AP*-KL), which combines multiple hypothesis testing and AP clustering algorithm along with the Krzanowski & Lai cluster quality index [9], to select a small yet informative subset of genes.

3.1 The Filtering method

The proposed methodology combines ranking/filtering and cluster analysis to select a small set of non-redundant but still highly discriminative genes. In relation to the filtering step, we first employ the maxT [10] function to rank the genes of the training set and then we reserve the top *N* genes ($N = 200$) for further exploitation. Our decision on which FS method to employ follows the findings of an analysis that we carried on FS methods [11]. Specifically, we assessed the classification performance of five different FS methods on data from ten different neuromuscular diseases. Each method yielded a different ranked list of genes, which was then used iteratively from top to bottom, in the range of 2 to 400 genes, to compose a new classification scheme in each iteration. The evaluation of the classification performance of all the produced schemes per FS method is depicted in Figure 1, and shows that the maxT achieved an average discrimination accuracy of 95%, between normal and disease samples.

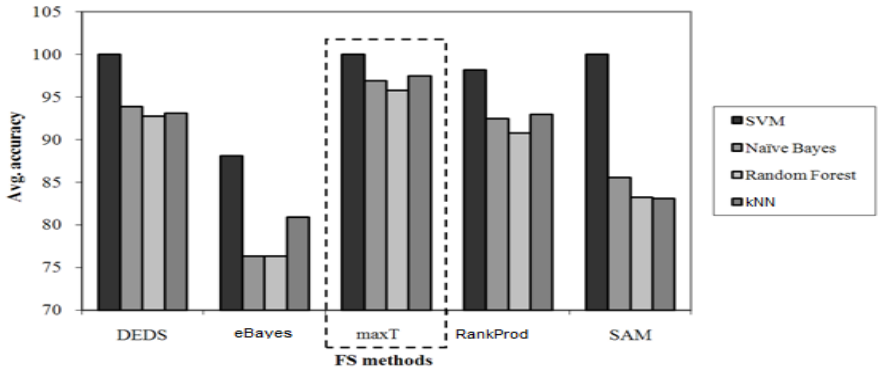

Fig. 1. The overall classification accuracy of five feature selection methods on ten datasets of neuromuscular disease data according to four classification algorithms

3.2 The Clustering Quality index

In the sequel, prior to clustering analysis we define the number of clusters, which in essence will be the number of representative genes that finally will compose our subset. The decision about which quality index to use, was based first on the indices comparison results of the Tibshirani et al. [12] survey as well as on several trials on simulated clustering data that also proved the efficiency of the index. Hence, we employed the index of Krzanowski and Lai to determine the number of clusters solely on the disease samples of the training test set.

This is actually a very fine detail in our methodology, since it has a direct impact on the clusters identification and consequently on the selected genes. However, we came across a dilemma regarding the part of the data that it would be the most proper and advantageous to apply the index. The first option was to search for the clustering structure solely in the samples belonging to the normal/control phenotype, whereas the second alternative was to investigate the samples in the disease phenotype. We finally reckoned that what actually is of interest for the identification of significant genes relevant to a disease, is the disease part of the data because all the information about the 'triggered' molecular processes is definitely present in it.

3.3 The Affinity Propagation Clustering Algorithm

The final step of our methodology involves the cluster analysis through the AP clustering method. The AP algorithm appeared in the late 20s and according to a benchmark analysis [13] across 15 other clustering algorithms, including k-means and k-medians clustering, hierarchical agglomerative clustering e.t.c., excelled at finding the more accurate clustering solution. Besides its intrinsic belief that initially all data points (genes) are considered as potential exemplars and its efficient convergence to the final clustering, urged us to adopt AP as an indispensable part of our methodology. Thus, we pass into AP the number of k clusters according to the Krzanowski and

Lai index and then let AP to detect those n clusters where $(n = k)$ clusters among the top N genes (a pre-defined number). The algorithm converges to the requested number of clusters (most of the times) and provides us with a list of the most representative genes of each cluster, the so called exemplars. These n exemplars are expected to form a classifier that shall discriminate between the normal and disease classes in a test set. Finally, we formulate the updated train and test sets by keeping only those n genes, and proceed with the classification process. The general flowchart of our methodology appears in Figure 2.

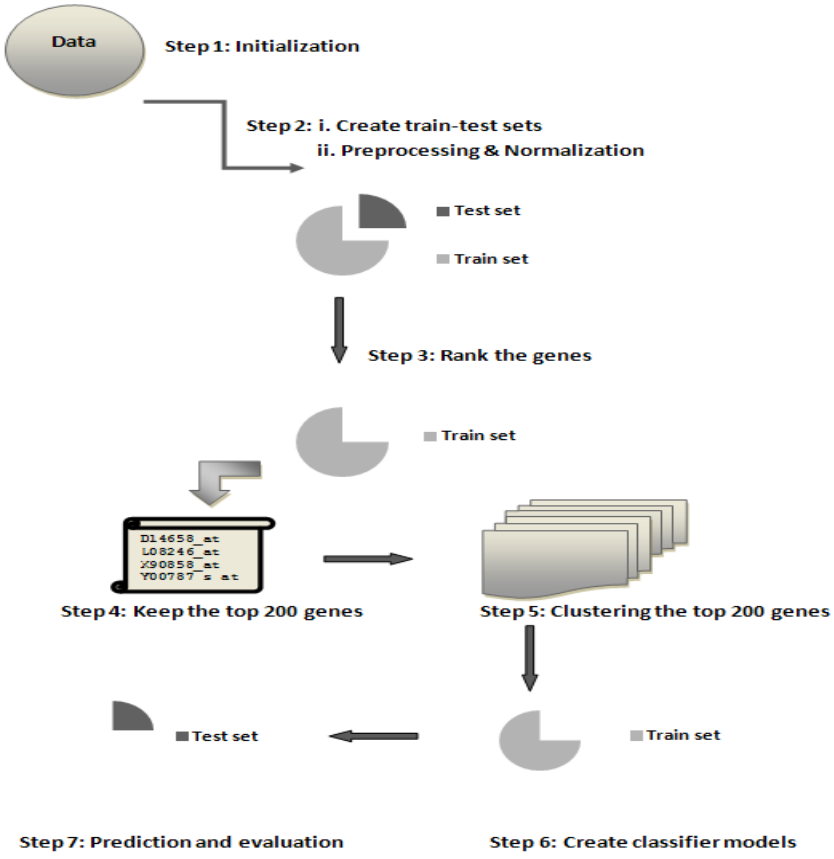

Fig. 2. The mAP-*KL* methodology flowchart

3.4 The Implementation of m*AP*-KL into an R-package

To provide the research community with the capability to apply mAP-KL in any given gene expression dataset, we have implemented this methodology to an open-source Bioconductor/ R package accompanied with extra functionalities such as data

sampling preprocessing, classification, network analysis, gene annotation analysis, pathway analysis and reporting that collaborate through five built-in classes, Figure 3. The centric idea during the package's design was to build functions that either can shape an extensive analysis pipeline or used as standalone modules. For instance, a user may import any dataset of raw gene expression data and apply with a single command eight at maximum different preprocessing methods. Then, may analyze any of the preprocessed data with the mAP-KL method and conclude to lists of significant genes (exemplars). Classification assessment, annotation analysis, pathway analysis and network characteristics are some of the possible analyses that a user may apply on these exemplars. On the other hand, a user may as well employ any of the available functions to exploit a particular functionality for example, to partition a dataset into train and validation sets, to obtain annotation info for a given list of probe ids, and so on.

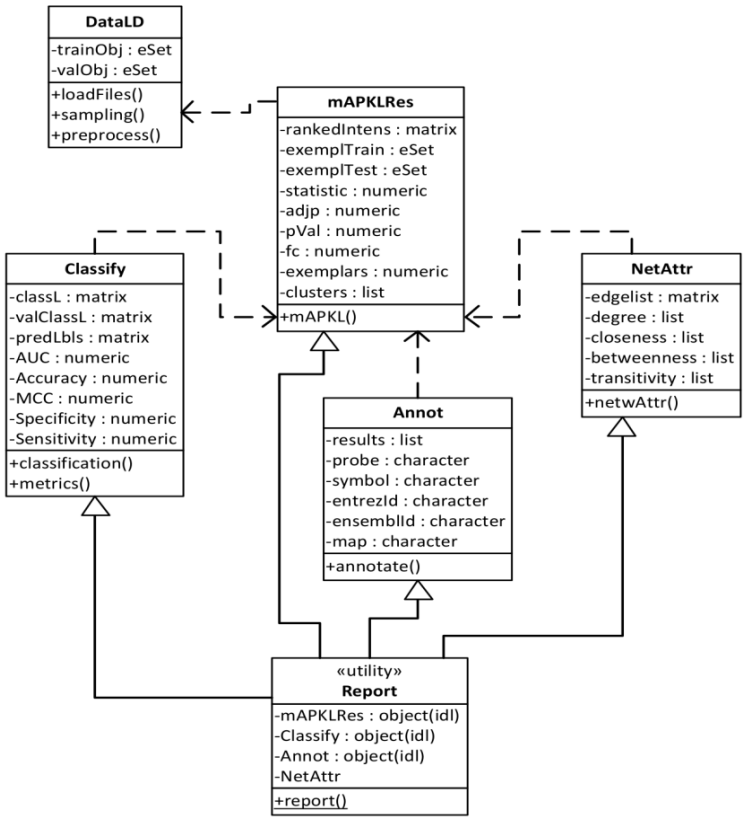

Fig. 3. A UML schematic representation of the classes and functions of the mAPKL.

4 Results and Discussion

We subjected our method to a series of evaluation tests on simulated microarray data in the first part and real microarray data in the second. Regarding the real microarray data we employed datasets of six neuromuscular diseases as representatives of small cohorts and four cancer datasets with numerous samples per phenotype. We designed and executed an elaborate set of analytical experiments with 5-CV on the training set and hold-out validation on a separate set using three different classifiers, RF – SVM – KNN, to assess its performance across whole genome expression datasets from both small and large patient cohorts. Moreover, on those microarray datasets we also applied 12 other feature selection/elimination approaches and compared the classification results using several metrics, for example AUC, TNR, TPR. In particular, we employed six univariate filter methods (eBayes [14], ODP [15], maxT [10], SAM [16], SNR and t-test [17]), one multivariate filter algorithm (cat [18]), three dimension reduction approaches (BGA-COA [19], PCA [20], PLS-CV [21]), one embedded method (Random Forest [22]), and one hybrid method (HykGene [6]).

Apart from the classification analysis, we investigated the produced gene lists from a biological perspective. The power of any FS approach is evident not only from its classification performance, but also from the biological relevance to the respective pathological phenotypes. Therefore we engaged the produced gene lists from mAP-KL and the methods that excelled in the classification process, (eBayes, PLS-CV, SAM, BGA-COA, RF-MDA), as well as the maxT method which is the ranking method of mAP-KL, into a series of validations. During those validations, we tried to unravel the 'semantics' behind those gene lists and its association with the respective diseases.

4.1 Assessing the Classification Performance on Microarray Data

The overall results, based on the RF classifier, as summarized in Figure 4 places mAP-KL at the top shelve among 12 other FS algorithms developed for the mining of gene expression data. In particular, the mAP-KL method achieved the second best mean AUC in neuromuscular diseases i.e. 0.91 and the sixth best in cancer data. Eventually, the classification performance of mAP-KL across all ten diseases reached the AUC score of 0.86, which is the third best AUC score with the minimum standard deviation value compared to the methods with better classification performance e.g. eBayes, PLS-CV. Hence, we may firmly state that the combination of a univariate and a clustering method isolates subsets of genes that may discriminate unknown samples from a variety of diseases and number of samples quite accurately.

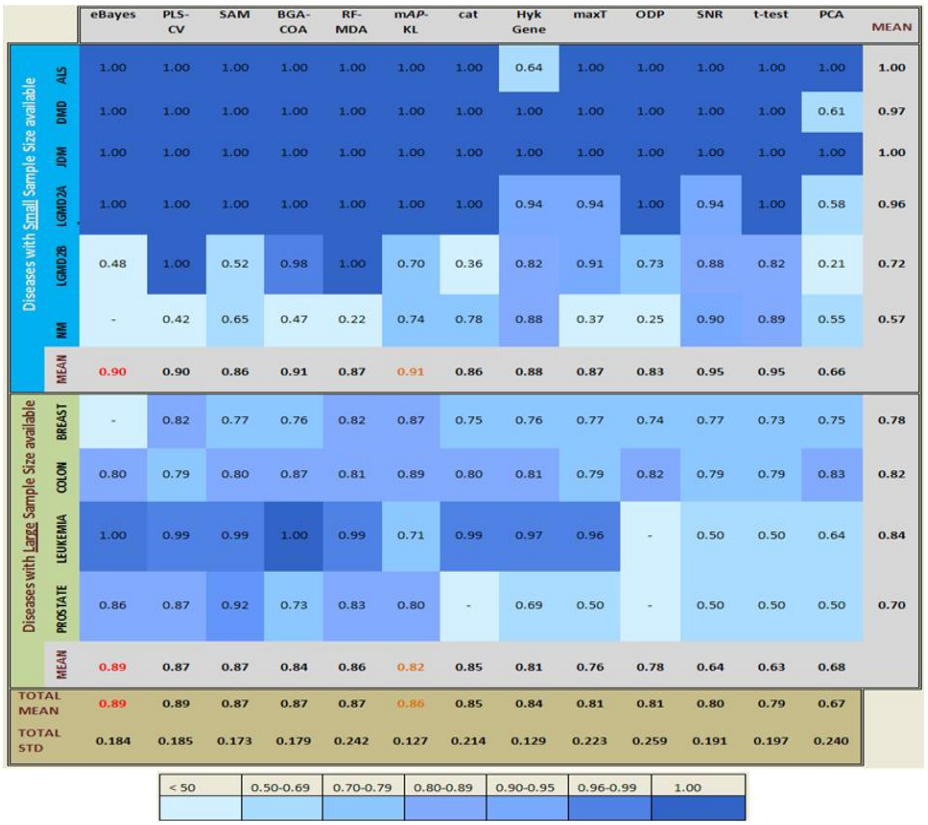

Fig. 4. The overall classification results (AUC metric) with RF classifier

4.2 Biological relevance of discriminatory gene lists

Typically, the initial product of an FS method is a list of ids rather than gene symbols, since the expression data stem from microarray chips technology. Therefore, a necessary action that we typically take is to match those probe ids with the relevant gene symbols. Another interesting thing from chip technology is that one gene symbol is regularly represented by more than one probe ids. Thus, an over or under expressed gene may be present in a top ranked list more than one times according to the chip specifications. As a result, those multiple instances of a gene shall be removed from any top ranked list to conclude to a list of unique top genes. This is an essential step regarding the anticipated gene enrichment since a top list of 20 or 50 probe ids may for example represent 14 or 35 unique gene symbols. Furthermore, gene chips include internal and external spiked in controls responsible for the hybridization quality that should be not included in the top ranking of any differential analysis. For all those reasons, the 'degree of uniqueness' (DoU) of a top ranked list is a first validation measure directly connected to the list's potential from a biological standpoint.

In the following tables, we have cited the number of probe ids and the respective number of gene symbols per method and per dataset. In the last column, we have calculated the DoU value as the average of the division between gene symbols and probe ids. The closest to the unit the more unique is the ranked list. Regarding the neuromuscular data, Table 1, the m*AP*-KL achieved the highest score with the maxT being quite close. In relation to cancer data, Table 2, the eBayes method surpassed the other methods although its average quantity is based on three rather than four datasets. The m*AP*-KL placed second setting a direct inference about the high "uniqueness" of the produced lists.

Table 1. The DoU of seven FS methods across neuromuscular data

FS	ALS		DMD		JDM		LGMD2A		LGMD2B		NM		DoU
	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	
mAP-KL	21	20	14	14	21	20	6	6	15	15	18	18	**0.984**
maxT	20	20	20	20	20	20	20	20	20	20	20	18	0.983
RF-MDA	20	20	20	20	20	20	20	19	20	20	20	18	0.975
SAM	20	14	20	20	20	18	20	16	20	16	20	20	0.867
eBayes[1]	20	17	20	20	20	18	20	16	20	15	-	-	0.860
PLS-CV	20	13	20	20	20	19	20	18	20	16	20	17	0.858
BGA-COA	20	15	20	17	20	18	20	14	20	17	20	17	0.817

[1] The eBayes method evaluated in five datasets

Table 2. The DoU of seven FS methods across cancer data

FS	Breast		Colon		Leukemia		Prostate		DoU
	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	
eBayes[1]	-	-	20	18	20	18	20	19	**0.917**
mAP-KL	6	4	20	16	5	5	12	12	**0.867**
PLS-CV	20	14	20	18	20	19	20	17	**0.850**
BGA-COA	20	12	20	18	20	19	20	18	**0.838**
SAM	20	11	20	18	20	18	20	19	**0.825**
maxT	20	11	20	16	20	17	20	20	**0.800**
RF-MDA	20	9	20	14	20	18	20	19	**0.750**

[1] The eBayes method evaluated in three datasets

A second validation criterion is the enrichment of the unique gene symbols in relation to the associated pathways. At this point is crucial to refer to another parameter before mentioning the results of this validation measure, which are the protein-coding-genes (P-C-Gns) in the ranked list. In essence, not all of the known genes are protein coding and thus involved in molecular functions. Pathway analysis tries to simplify the complexity at the cellular level through the representation of a series of

steps where "each step is an event that transforms input physical entities into output entities" [23]. Such entities are definitely the produced proteins, among other small molecules or particles, and as a consequence only the protein coding genes are requisite for a pathway analysis.

Through a plethora of pathway analysis tools, we utilized the 'Reactome' pathway database [23], which is a curated and peer reviewed database of pathways and reactions in human biology. We uploaded the top lists of the selected FS methods for all diseases and evaluated their pathway enrichment. During the pathway evaluation, we took into consideration the DoU and the number of protein-coding genes parameters as well as the number of pathways according to the 'Reactome' database. The final pathway enrichment (PE) score for each FS (m) is the average of the summation of pathways per protein-coding genes multiplied by the DoU for all diseases (d)

$$PE_m = \sum_{d=1}^{10} \frac{Protein\text{-}coding\text{-}genes_d}{Pathways_d} \times DoU.$$

We summarized the results, Table 3, where the FS methods are in descending order based on their average PE score. In accordance with the pathway analysis, the maxT method appears to achieve the highest PE score across all diseases. Besides is the method with the second highest DoU score marginally behind mAP-KL. However, this significant advantage over mAP-KL and RF-MDA that follow is mainly due to the weird PE score in prostate cancer (4.33), where the maxT identified three (3) pathways with 13 unique genes. Albeit, those three methods appear to constitute a group with PE scores close to unit, which is a satisfactory if not intriguing case for biologists.

Table 3. The overall pathway analysis results

FS	Pathway Analysis											
	ALS	DMD	JDM	LGMD2A	LGMD2B	NM	Breast	Colon	Leukemia	Prostate	Mean	Stdev
maxT	1.00	1.08	1.08	0.43	1.36	1.01	0.47	0.80	0.79	4.33	1.24	1.12
mAP-KL	1.43	0.78	1.38	0.43	0.88	1.40	0.67	0.63	0.80	1.17	0.95	0.36
RF-MDA	0.75	1.10	1.40	0.74	0.63	1.80	0.54	0.63	0.80	1.03	0.94	0.40
eBayes[1]	0.37	1.50	0.90	0.64	0.67	-	-	1.08	1.26	0.86	0.91	0.36
PLS-CV	0.37	0.89	1.21	0.66	0.90	0.85	0.98	0.90	1.07	1.04	0.89	0.23
SAM	0.29	1.13	1.00	0.64	0.80	1.08	0.46	1.15	0.98	1.27	0.88	0.32
BGA-COA	0.68	1.06	0.63	0.70	1.19	0.85	0.60	0.90	1.14	1.00	0.87	0.22

[1] The eBayes method evaluated in eight datasets

5 Conclusions

We proposed a hybrid FS method (m*AP*-KL), which clearly demonstrates how effective the combination of a multiple hypothesis testing approach with a clustering algo-

rithm can be to select small yet informative subsets of genes in binary classification problems. Particularly, across a variety of diseases and datasets, mAP-KL achieved competitive classification results compared to other FS methods and specifically to HykGene method, which follows a similar philosophy i.e. first ranking and then clustering. The advances of m*AP*-KL over HykGene or other similar approaches stem from three key characteristics; the data-driven nature, the affinity propagation clustering, and the classifier independence. Indeed, the engagement of a cluster quality index, the Krzanowski and Lai, diminishes any fuzziness and provides the clustering algorithm with a representative number of potential clusters. Moreover, in m*AP*-KL the data determine the size of the subset i.e. the structure of the data dictate the number of clusters and the clustering algorithm decides on the representatives upon each cluster. Contrary to other methods, for example HykGene, where a classifier is wrapped around its method, in our case no classifier takes part during the subset construction. This methodological characteristic is of great importance since our subsets lack of any overfitting phenomenon pertinent to classifiers.

Relevant to the identification of clusters, the employment of AP clustering algorithm, deals effectively with the issue of representative genes per cluster. Other comparable approaches to m*AP*-KL admitted considerably difficulties on selecting effectively one or more representative genes per cluster. Besides, the AP follows a gene-network mechanism by considering initially all genes as nodes in a network. The resultant exemplars are the central genes within a cluster of genes and probably the key nodes within a network of genes. Therefore mining the exemplars can be considered as the forefront of a network inference process rather than just the outcome of a FS approach. As such, we intent to construct networks based on the top *N* genes of our methodology and then to exploit the network characteristics of the exemplars. An initial attempt towards this direction is already available in the mAPKL package, though more network inference methods for the reconstruction of gene regulatory networks and methods for functional enrichment will be engaged in the near future.

References

1. Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507-17, Oct 1 2007.
2. J. Jaeger, R. Sengupta, and W. L. Ruzzo, "Improved gene selection for classification of microarrays," *Pac Symp Biocomput*, pp. 53-64, 2003.
3. M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. Thesis, Computer Science, The University of Waikato, Hamilton, New Zealand, 1999.
4. B. Hanczar, M. Courtine, A. Benis, C. Hennegar, K. Clement, and J.-D. Zucker, "Improving classification of microarray data using prototype-based feature selection," *ACM SIGKDD Explorations Newsletter*, vol. 5, p. 7, December 2003 2003.
5. C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J Bioinform Comput Biol*, vol. 3, pp. 185-205, Apr 2005.
6. Y. Wang, F. S. Makedon, J. C. Ford, and J. Pearlman, "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," *Bioinformatics*, vol. 21, pp. 1530-7, Apr 15 2005.

7. A. Sakellariou, D. Sanoudou, and G. Spyrou, "Combining multiple hypothesis testing and affinity propagation clustering leads to accurate, robust and sample size independent classification on gene expression data," *BMC Bioinformatics,* vol. 13, p. 270, 2012.
8. B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science,* vol. 315, pp. 972-6, Feb 16 2007.
9. W. J. Krzanowski and Y. T. Lai, "A criterion for determining the number of groups in a data set using sum of squares clustering," *Biometrics,* vol. 44, pp. 23-34, 1988.
10. Y. Ge, S. Dudoit, and T. P. Speed, "Resampling-based multiple testing for microarray data analysis," *Test,* vol. 12, pp. 1-77, 2003.
11. A. Sakellariou, D. Sanoudou, and G. Spyrou, "Investigating the minimum required number of genes for the classification of neuromuscular disease microarray data," *IEEE Trans Inf Technol Biomed,* vol. 15, pp. 349-55, May 2011.
12. R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of data clusters via the gap statistic," *Journal of the Royal Statistical Society: Series B,* vol. 63, pp. 411-423, 2001.
13. D. Delbert, "Affinity Propagation: Clustering Data by Passing Messages," Doctor of Philosophy, Graduate Department of Electrical & Computer Engineering, University of Toronto, 2009.
14. G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Stat Appl Genet Mol Biol,* vol. 3, p. Article3, 2004.
15. J. D. Storey, "The optimal discovery procedure: a new approach to simultaneous significance testing," *Journal of the Royal Statistical Society: Series B* vol. 69, pp. 347-368, 2007.
16. V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A,* vol. 98, pp. 5116-21, Apr 24 2001.
17. J. Gould, G. Getz, S. Monti, M. Reich, and J. P. Mesirov, "Comparative gene marker selection suite," *Bioinformatics,* vol. 22, pp. 1924-5, Aug 1 2006.
18. V. Zuber and K. Strimmer, "Gene ranking and biomarker discovery under correlation," *Bioinformatics,* vol. 25, pp. 2700-7, Oct 15 2009.
19. A. C. Culhane, G. Perriere, E. C. Considine, T. G. Cotter, and D. G. Higgins, "Between-group analysis of microarray data," *Bioinformatics,* vol. 18, pp. 1600-8, Dec 2002.
20. I. T. Jolliffe, *Principal component analysis,* 2nd ed. New York: Springer, 2002.
21. A. L. Boulesteix, "PLS dimension reduction for classification with microarray data," *Stat Appl Genet Mol Biol,* vol. 3, p. Article33, 2004.
22. L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.
23. I. Vastrik, P. D'Eustachio, E. Schmidt, G. Gopinath, D. Croft, B. de Bono, *et al.*, "Reactome: a knowledge base of biologic pathways and processes," *Genome Biol,* vol. 8, p. R39, 2007.