Parsimonious selection of useful genes in microarray gene expression data

Félix F. González-Navarro and Lluís A. Belanche-Muñoz*

Dept. de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya
Ω-Building, North Campus. 08034 Barcelona, Spain.
{fgonzalez,belanche}@lsi.upc.edu

Summary. Machine Learning methods have of late made significant efforts to solving multidisciplinary problems in the field of cancer classification in microarray gene expression data. These tasks are characterized by a large number of features and a few observations, making the modeling a non-trivial undertaking. In this work we apply entropic filter methods for gene selection, in combination with several off-the-shelf classifiers. The introduction of bootstrap resampling techniques permits the achievement of more stable performance estimates. Our findings show that the proposed methodology permits a drastic reduction in dimension, offering attractive solutions both in terms of prediction accuracy and number of explanatory genes; a dimensionality reduction technique preserving discrimination capabilities is used for visualization of the selected genes.

Key words: Biological data mining and knowledge discovery; Gene expression analysis, Tools and methods for computational biology and bioinformatics; Cancer informatics.

1 Introduction

In cancer diagnosis, classification of the different tumor types is of great importance. Traditional methods of tackling the distinction between different types of cancer are primarily based on morphological characteristics of tumorous tissue [8]. Machine Learning methods are now extensively used for this task [3]. Typically, a gene expression data set may consist of dozens of observations with thousands or even tens of thousands of genes. Classifying cancer types using this very high ratio between number of variables and number of observations is a delicate process, because of the high risk of overfitting the data. As a result, dimensionality reduction and in particular *feature selection* (FS) techniques may be very useful. The finding of small subsets of very relevant genes among a huge quantity could derive in much specific and efficient treatments. However, in a FS scenario, gene expression data analysis may entail a heavy computational consumption of resources, due to the extreme sparseness compared to standard data sets in classification tasks [34]. For these reasons, in this work we rely on *filter* measures for feature selection (that are classifier-independent) in order to keep the computational cost within reasonable bounds. We are also concerned with increasing the *reliability* of the FS process, in the sense of reducing the inherent instability caused by the particular choice

* Corresponding author.

of data sample. In addition, in order to further reduce the chance of overfitting the data, we take the decision of using low-complexity classifiers (5 of the 8 used classifiers are linear or quadratic) together with a small subset of highly relevant genes.

Of special importance in a practical medical setting is the *interpretability* of the solutions in terms of the obtained gene subsets. A dimensionality reduction technique that provides a data projection –while preserving the class discrimination achieved by a classifier– is also used in our study. Our experimental findings show that the proposed feature selection methodology offers highly competitive solutions both in terms of prediction accuracy and number of explanatory genes. In particular, we report results that offer better performance in both aspects at least for two of the analyzed microarray tasks. In addition, we provide biological evidence for the three most important genes obtained in each microarray data set.

2 Feature Selection in the Microarray domain

The Bioinformatics community has recognized the FS process as a key issue in gene expression data analysis [25]. In many gene selection methods a list of the *top ranked* genes based on some merit figure is generated, followed by an inductive step where a classifier is incrementally evaluated on the list [31]. Fisher's criterion [15], the *signal-to-noise* ratio [12], the χ^2 statistic [23] or Wilcoxon's rank sum test [9] are popular choices. However, considering individual contributions only can very likely hinder the discovery of interactions between genes.

Mutual Information (MI) has been successfully used in FS for measuring the influence that a feature has over a class or target. Several criteria to evaluate subsets of features employ it, mostly in the *bivariate* case, between a feature and the class. A few use a normalized variant defined by $C_{XY} = \frac{I(X;Y)}{H(Y)}$, where Y is the class, $I(X;Y) = H(Y) - H(Y|X)$ and H denotes the *entropy*. Note that $I(X;X) = H(X)$, since $H(X|X) = 0$ and $I(X;Y) = I(Y;X)$. The expression for C_{XY}, sometimes referred to as the *coefficient of constraint*, can be better understood by analyzing Fig. 1 (left). It can be seen that by increasing $I(X;Y)$, $H(Y|X)$ decreases; in other words, there is a reduction in the uncertainty of Y due to the action of X. The maximum value that $I(X;Y)$ could take is $H(Y)$. This property has been exploited in order to create an index to measure subsets of features with respect to a class or target value [36]. Alternatively, both *relevance* and *redundancy* of genes has been assessed by using MI, configuring a criterion of minimum redundancy-maximum relevance (mRMR) [10]. The reader is referred to [25] for a recent compilation on the use of these measures. The computation of MI can be extended from the bivariate to the multivariate case, of a number $n \geq 2$ of variables against another one, as $I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1}) = H(Y) - H(Y|X_1,\ldots,X_n)$, whereas conditional MI is expressed in the natural way, as $I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$.

Instead of by conditioning, in this work we propose to calculate the MI between a class variable Y and two variables X and Z, as shown in Fig. 1 (right). The shaded area represents $I(Y;X,Z) = H(Y) + H(X,Z) - H(X,Y,Z)$, the information that X,Z explain about Z. Given that $I(Y;X,Z) \leq 1$ and that $H(Y)$ acts as the baseline reference, it is wise to normalize it as $\frac{H(Y)+H(X,Z)-H(X,Y,Z)}{H(Y)}$, obtaining an index that evaluates the influence of two variables with respect to a class. It takes values between zero (no relevance) and one (maximum relevance).

In order for this expression to be of practical use from the FS point of view, it needs to be extended to the multivariate case. The MI between a subset of variables and the class variable is computed by generating first a "super-feature", obtained considering the concatenation of each combination of possible values of its forming features. In symbols, let $X = \{X_1, ..., X_n\}$ be the full feature set and consider a subset $\tau = \{\tau_1, \cdots, \tau_k\} \subseteq X$. A single feature $\mathscr{V}_\tau$ can

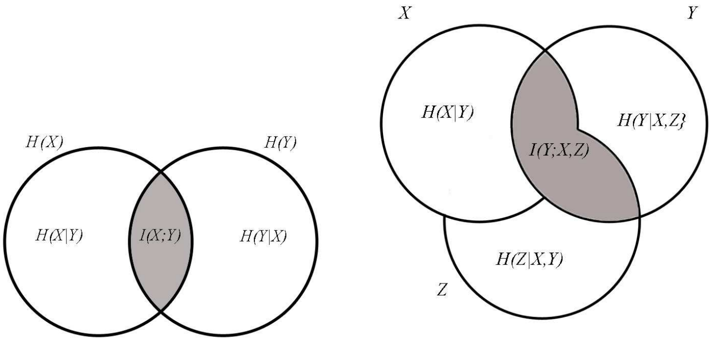

Fig. 1. Entropy concepts. Left: Basic Mutual Information. Right: Mutual information between two variables and class or target variable

be obtained uniquely, whose possible values are the concatenations of all possible values of the features in τ (for completeness, define $\mathscr{V}_{\emptyset} = \emptyset$). An *index of relevance R* of a feature X_i to a class Y with respect to a subset τ is then given by $R(X_i;Y|\tau) = \frac{H(Y)+H(\tau,X_i)-H(\tau,Y,X_i)}{H(Y)}$, for $X_i \in X, \tau \subseteq X \setminus X_i$. The use of the super-feature allows a faster implementation in case of discrete variables, whose development is not essential in this context.

To compute the necessary entropies, a discretization process is needed. This change of representation does not often result in a significant loss of accuracy (sometimes significantly improves it [27], [30]); it also offers reductions in learning time [7]. In this work, the CAIM algorithm was selected for two reasons: it is designed to work with supervised data, and does not require the user to define a specific number of intervals [21]. This way of calculating feature subset relevance is used to evaluate gene subsets, embedding it into a filter *forward-search* strategy, conforming the *BGS*^{3} algorithm –standing for *Best Gene Subset Search Strategy*–, a supervised filter independent of the search strategy and of the *a posteriori* inducer, described in the listing below. This algorithm begins by selecting the feature that maximizes its relevance with respect to the class feature (lines 1-3). Then a forward search is conducted: at every step, the feature providing the maximum value of relevance when added to the current subset is selected (line 6). If, at the end of a step, more that one feature renders the same value for relevance (line 7), the feature that produces the *minimum redundancy* of information is chosen. In case the newly added feature brings a benefit, it is added to the current subset (line 9). The algorithm stops when the index of relevance was not improved, its maximum value has been reached, or it has run out of features, whichever comes first (line 10).

3 Experimental work

The experimental methodology was aimed to achieve results that reflect the true behavior of the system as much as possible; in other words, to obtain *reliably* relevant genes. Bootstrap resampling techniques are used to yield a more stable and thus more reliable measure of predictive ability. The original microarray expression data sets S were used to generate $B = 5,000$

Algorithm 1: *BGS^3* Best Gene Subset Search Strategy.

```
input  : $X = \{X_1, \ldots, X_n\}$: Full gene set;  $C$: Class feature
output: $\Phi$ : Best Gene Subset (BGS)
$\phi \leftarrow \underset{f \in X}{\text{argmax}} \frac{H(C) - H(f,C)}{H(C)}$;
$\Phi \leftarrow \{\phi\}$: Current best subset;
$R \leftarrow \frac{H(C)-H(\Phi,C)}{H(C)}$: Current best relevance;
$exit \leftarrow false$;
repeat
    $\phi \leftarrow \underset{f \in X \setminus \Phi}{\text{argmax}} \left\{ \frac{H(C)+H(\Phi,f)-H(\Phi,C,f)}{H(C)} \right\}$;
    if $|\phi| > 1$ then $\phi^+ \leftarrow \underset{f \in \phi}{\text{argmin}} I(\Phi,f)$ else $\phi^+ \leftarrow \phi$ end;
    $R^+ \leftarrow \frac{H(C)-H(\Phi \cup \phi^+,C)}{H(C)}$;
    if $R^+ > R$ then $R \leftarrow R^+$; $\Phi \leftarrow \Phi \cup \phi^+$ else $exit \leftarrow true$ end
until $R^+ = 1 \lor exit \lor |\Phi| = n$
```

bootstrap samples $S_1, \ldots, S_B$ that play the role of *training sets* in the feature selection process: each relevance value calculated in the algorithm is the *average* across the B bootstrap samples, i.e., the *average behavior* of a feature is used to guide and stabilize the algorithm.

The algorithm is first applied to the discretized bootstrap samples to obtain the Best Gene Subset or BGS (one for each data set). Then the classifier development stage is conducted using those *original* continuous features that are members of their corresponding BGSs. Eight classifiers were evaluated by means of 10 times of 10-Fold Cross Validation (10x10cv), a method suitable to handle small sized data sets: the *k-nearest-neighbors* technique with Euclidean metric (kNN) and parameter $k \in \{1, \ldots, 15\}$, the *Naïve Bayes classifier* (NB), the *Linear* and *Quadratic Discriminant* classifiers (LDC and QDC), *Logistic Regression* (LR), and the *Support Vector Machine* with linear, quadratic and radial kernels (lSVM, qSVM and rSVM) and parameter C (regularization constant) (with $C = 2^k$, k running from -7 to 7). The rSVM has the additional smoothing parameter $\sigma = 2^k$, k running from -7 to 7).

Validation of the described approach uses five public-domain microarray gene expression data sets, shortly described as follows: *Colon Tumor*: 62 observations of colon tissue, of which 40 are tumorous and 22 normal, 2,000 genes [1]. *Leukemia*: 72 bone marrow observations and 7,129 probes: 6,817 human genes and 312 control genes [12]. The goal is to tell acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL). *Lung Cancer*: distinction between malignant pleural mesothelioma and adenocarcinoma of lung [13]; 181 observations with 12,533 genes. *Prostate Cancer*: used in [33] to analyze differences in pathological features of prostate cancer and to identify genes that might anticipate its clinical behavior; 136 observations and 12,600 genes. *Breast Cancer*: 97 patients with primary invasive breast carcinoma; 12,600 genes analyzed.

The results of the FS stage are presented in Table 1. Sizes of final BGSs for each data set are considerable small (as low as 3 genes in the *Leukemia* problem); remarkably, in all cases the maximum relevance is achieved. Computational times are also reported, ranging from few minutes to several hours –8 hours at most[2]. These times should be judged taking into account

[2] These figures were obtained in a standard x86 machine at 2.666 GHz.

Table 1. Gene subsets selected by BGS^3: $|BGS|$ is their number, R_{max} is the final relevance achieved, and 'CPU time' indicates total processing time.

Data set	$\lvert BGS\rvert$	R_{max}	CPU time	Gene Accession Number (GAN) or Gene name
Colon Tumor	10	1	13 min	M26383 M63391 M76378 X12671 J05032 H40095 H43887 R10066 U09564 H40560
Leukemia	3	1	14 min	M23197_at X95735_at U46499_at
Lung Cancer	13	1	4 hrs	37957_at 1500_at 36536_at 35279_at 33330_at 39643_at 32424_at 40939_at 33757_f_at 33907_at 179_at 39798_at 1585_at
Breast Cancer	16	1	4 hrs	AL080059 NM_003258 NM_003239 NM_005192 Contig7258_RC NM_006115 AL137615 Contig38901_RC AL137514 AF052087 U45975 AF112213 AB037828 NM_005744 NM_018391 NM_003882
Prostate Cancer	21	1	8 hrs	37639_at 37720_at 37366_at 31538_at 37068_at 40436_g_at 39755_at 31527_at 1664_at 34840_at 36495_at 33674_at 39608_at 31545_at 914_g_at 41288_at 37044_at 40071_at 34730_g_at 41732_at 41764_at

that 5,000 resamples for each data set are being processed. The composition of each BGS is signaled by the gene identifier. These unique IDs will be used to find biological evidence about the significance of the gene in the disease. Recall that the BGSs are constructed adding at every step the gene most informative to the current subset, every new set having more informative power. It seems therefore sensible, in terms of classification performance, to parsimoniously explore the obtained subsets in an incremental fashion, respecting the order in which the genes were found –which is the order reported in Table 1.

Table 2. Final accuracy results with comparison to other references. (**F**) indicates a Filter algorithm, (**W**) a wrapper and (**FW**) a combination of both. Size of the final gene subset and the used classifier are in brackets.

		Colon		Lung	Breast	Prostate
Work	Validation	Tumor	Leukemia	Cancer	Cancer	Cancer
BGS^3(F)	10x10cv	89.36	97.89	98.84	83.37	93.43
		(9, 3*NN*)	(2, *NB*)	(4, *LR*)	(12, *lSVM*)	(3, 10*NN*)
[5](F)	200x0.632	88.75	98.2	-	-	-
	bootstrap	(14, *lSVM*)	(23, *lSVM*)	-	-	-
[31](W)	10x10cv	85.48	93.40	-	-	-
		(3, *NB*)	(2, *NB*)	-	-	-
[37](W)	100xrandom	87.31	-	72.20	-	-
	subsampling	(94, *SVM*)	-	(23, *SVM*)	-	-
[4](W)	50xholdout	77.00	96.00	99.00	79.00	93.00
		(33, *rSVM*)	(30, *rSVM*)	(38, *rSVM*)	(46, *rSVM*)	(47, *rSVM*)
[17](FW)	10x10cv	-	-	99.40	-	96.30
		-	-	(135, 5*NN*)	-	(79, 5*NN*)
[16](F)	10cv	-	98.6	99.45	68.04	91.18
		-	(2, *SVM*)	(5, *SVM*)	(8, *SVM*)	(6, *SVM*)

The accuracy results are presented in Table 2: shown are the final number of genes obtained in the incremental search from the initial BGSs, giving the best 10x10cv accuracy and the used classifier. To be sure, a non-parametric Wilcoxon signed-rank test is used for the (null) hypothesis that the median of the differences between the errors of the *winner* classifiers per data set and another classifier's error is zero. This hypothesis can be rejected at the 99% level in *all* cases (p-values not shown). It is remarkable that in the *Lung Cancer* data set, as low as 4 genes are required to get almost 99% of accuracy. On the other hand, the *Breast Cancer* data set is one of the most difficult problems, followed by *Colon Tumor* (83% and 89%).

It is a common practice to compare to similar works in the literature. Unfortunately, the methodological steps are in general very different, especially concerning resampling techniques, making an accurate comparison a delicate undertaking. Nonetheless, such a comparison is presented in Table 2. Six references which are illustrative of recent work are included. As stated before, the *Colon Tumor* data set presents difficulties in classification; however, *BGS*^{3} figures are higher than the rest, even with less genes involved and in front of solutions that employ a pure wrapper strategy. For the *Leukemia* data, other references achieve better figures, some of them using a much bigger gene subset –23 or 30 genes–. Two results report two genes in their solutions, [31] and [16]. The former does not match the gene subset obtained by our algorithm, and the latter does not give precise information on the obtained genes. The *Lung Cancer* data set is apparently the easiest to separate. Values as high as 99% are achieved by three of the referenced works, making use of much bigger subset sizes. Incidentally, the solution in [16] agrees in one gene, the 1500_at *WT1-Wilms tumor 1*. The solution offered by *BGS*^{3} in the *Breast Cancer* data set gives the best result among the references, with almost 4% of difference. The *Prostate Cancer* data set is well separated by [17], using 79 genes. *BGS*^{3} separates it using only two genes, with a degradation of 3% of accuracy. No information is provided in this reference about which genes are selected. The good performance achieved with low numbers of selected genes are an indication that these are really good ones in separating the classes. However, even if interpretable by mere inspection of the involved genes, the final selection of genes may still provide few clues about the structure of the classes (cancer types). Visualization in a low-dimensional representation space may become extremely important, helping oncologists to gain insights into what is undoubtedly a complex domain. We use in this work a method based on the decomposition of the scatter matrix -arguably a neglected method for dimensionality reduction- with the remarkable property of maximizing the separation between the projections of compact groups of data. This method leads onto the definition of low-dimensional projective spaces with good separation between classes, even when the data covariance matrix is singular; further details about this method can be found in [22]. Such visualization is illustrated by the plots in Fig. 2. These are scatter plots of 2-D projections of the classes (using the first two eigenvectors of the scatter matrices). In can be seen that separation is in general quite good (although far from perfect); also, the structure itself of the two classes provides clues on the variability of each cancer type.

4 Conclusions

Experimental work in comparison to recent works examining the same data sets reveals that the developed methodology provides very competitive solutions, characterized by small gene subsets and affordable computational demands. The use of resampling methods to stabilize the gene selection process –arguably the most delicate part– has shown to deliver final solutions of very low size and strong relevance. Very noteworthy is the fact that the best classifiers (those that make good use of the gene subsets found in the feature selection phase) are consistently very simple: the *k-nearest-neighbors* technique, *Naïve Bayes classifier*, *Logistic regression* and a linear *Support Vector Machine*. A final concern has been the *practical* use of the results in a medical context, achieved thanks to a dimensionality reduction technique that preserves the class discrimination capabilities. This joint information may become extremely important to help oncologists to gain insights into this highly sensitive domain.

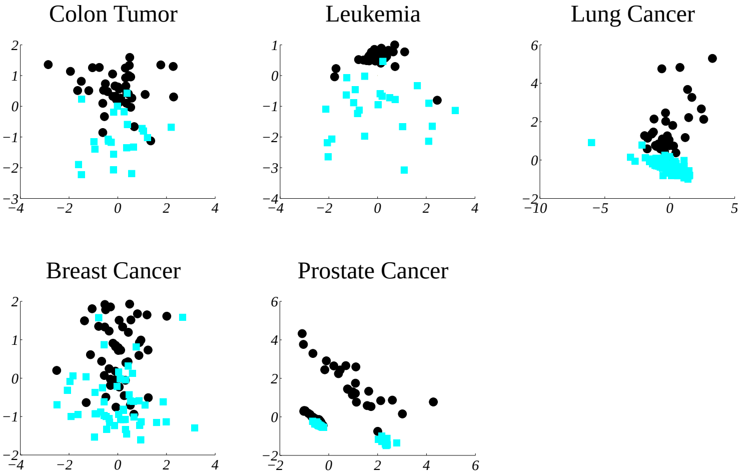

Fig. 2. Visualization of the solutions using the first two eigenvectors of each scatter matrix. In *Colon Tumor* and *Prostate Cancer* circles represent tumorous samples and squares indicates normal tissue; in *Leukemia* circles indicate acute lymphoblastic leukemia cells and squares acute myeloid leukemia cells; in *Lung Cancer* circles are malignant pleural mesothelioma and squares areadenocarcinoma; in *Breast Cancer* circles indicate relapse and squares non-relapse.

References

1. Alon, U., et al.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In: Proceedings of The National Academy of Sciences USA. Volume 96., IEEE (1999) 6745-6750
2. Amin, K., et al.: Wilms' tumor 1 susceptibility (wt1) gene products are selectively expressed in malignant mesothelioma. Amer. J. of Pathology **146**(2) (1995) 344–356
3. Duan, K.B., et al.: Multiple svm-rfe for gene selection in cancer classification with expression data. IEEE/ACM Transactions on Nanobioscience **4**(3) (2005) 228–234
4. Bu, H.L., et al.: Reducing error of tumor classification by using dimension reduction with feature selection. In: Intl. Symp. on Optim. and Sys. Biol. (2007) 232–241
5. Cai, R., et al.: An efficient gene selection algorithm based on mutual information. Neurocomputing **72** (2009) 991–999
6. Chakraborty, S.: Simultaneous cancer classification and gene selection with bayesian nearest neighbor method: An integrated approach. Computational Statistics and Data Analysis **53**(4) (2009) 1462–1474
7. Catlett, J.: On changing continuous attributes into ordered discrete attributes. In: Procs. of the European working session on Machine learning, Springer-Verlag (1991) 164–178
8. Chu, F., Wang, L.: Applications of support vector machines to cancer classification with microarray data. Intl. Journal of Neural Systems **15**(6) (2005) 475–484
9. Chu, W., et al.: Biomarker discovery in microarray gene expression data with gaussian processes. Bioinformatics **21**(16) (June 2005) 3385–3393
10. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. In: Procs. of IEEE Computational Systems Bioinformatics (2003)
11. Dumont, N., Arteaga, C.: Transforming growth factor-β and breast cancer: Tumor promoting effects of transforming growth factor-β. Breast Cancer Res. **2** (2000) 125–132

12. Golub, T., et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286**(5439) (October 1999) 531–537
13. Gordon, G.J., et al.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research **62** (September 2002) 4963–4967
14. Goutebroze, L., et al.: Cloning and characterization of schip-1, a novel protein interacting specifically with spliced isoforms and naturally occurring mutant nf2 proteins. Molecular and Cellular Biology **20**(5) (2000) 1699–1712
15. Hedenfalk, I., et al.: Gene-expression profiles in hereditary breast cancer. The New England Journal of Medicine **344** (2001) 539–548
16. Hewett, R., Kijsanayothin, F.: Tumor classification ranking from microarray data. BMC Genomics **9**(2) (2008)
17. Hong, J.H., Cho, S.B.: Cancer classification with incremental gene selection based on dna microarray data. In: IEEE/ACM Trans. on Comp. Biol. and Bioinf. (2008) 70–74
18. Hong-Qiang, W., et al.: Extracting gene regulation information for cancer classification. Pattern Recognition **40**(12) (2007) 3379–3392
19. Jiang, W., et al.: Constructing disease-specific gene networks using pair-wise relevance metric: Application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. BMC Systems Biology **2** (2008)
20. Johansson, B., et al.: The prostate. Proteomic comparison of prostate cancer cell lines LNCaP-FGC and LNCaP-r reveals heatshock protein 60 as a marker for prostate malignancy **66**(12) (2006) 1235–1244
21. Kurgan, L.A., Cios, K.J.: Caim discretization algorithm. IEEE Trans. Knowl. Data Eng **16**(2) (2004) 145–153
22. Lisboa, P., et al.: Cluster based visualisation with scatter matrices. Pattern Recognition Letters **29**(13) (2008) 1814–1823
23. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric atributes. In: IEEE 7th Intl. Conf. on Tools with Artificial Intelligence, IEEE (1995) 338–395
24. Lurje, G., et al.: Polymorphisms in vegf and il-8 predict tumor recurrence in stage iii colon cancer. Annals of Oncology **19** (2008) 1734–1741
25. Meyer, P.E., Schretter C., Bontempi, G. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2(3), 2008.
26. National center of biothecnology information. `http://www.ncbi.nlm.nih.gov/`
27. Ng, M., Chan, L.: Informative gene discovery for cancer classification from microarray expression data. In: IEEE Machine Learning for Signal Processing (2005) 393–398
28. Öhrvik, A., et al.: Sensitive nonradiometric method for determining thymidine kinase 1 activity. Clinical Chemistry **50**(9) (2004) 1597–1606
29. Plesa, C., et al.: Prognostic value of immunophenotyping in elderly patients with acute myeloid leukemia: a single-institution experience. Cancer **112**(3) (2007) 572580
30. Potamias, G., et al.: Gene selection via discretized gene-expression profiles and greedy feature-elimination. In: SETN. (2004) 256–266
31. Ruiz, R., et al.: Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition **39** (2006) 2383–2392
32. Scherz-Shouval, R., et al.: Reactive oxygen species are essential for autophagy and specifically regulate the activity of atg4. The EMBO Journal **26** (2007) 1749–1760
33. Singh, D., et al.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1** (March 2002) 203–209
34. Tang, Y., et al.: Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. IEEE/ACM Trans. on Comp. Biol. and Bioinf. **4**(3) 365–381

35. Vant'Veer, L., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature (415) (January 2002) 530–536
36. Wang, H.: Towards a Unified Framework of Relevance. PhD thesis, U. of Ulster (1996).
37. Wang, L., et al.: Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics **24**(3) (2008) 412–419

Appendix: Biological evidence

Biological evidence is assembled in the medical literature studying each specific gene. Only the first (i.e. the more relevant) three genes of each subset are presented, for conciseness.

Colon Tumor: **[M26383]** *IL8-Interleukin 8* encodes a protein member of the CXC chemokine family and is one of the major mediators of the inflammatory response [26], associated with a higher likelihood of developing colon tumor recurrence [24]. **[M63391]** *DES-Desmin* encodes a muscle-specific class III intermediate filament. [19] reported that Interleukin 8 and Desmin act as the central elements in colon cancer susceptibility. **[M76378]** *CSRP1-Cysteine and glycine-rich protein 1* may be involved in regulatory processes important for development and cellular differentiation.

Leukemia: **[M23197_at]** *CD33-CD33* is a putative adhesion molecule of myelomonocytic-derived cells that is expressed on the blast cells in patients with acute myeloid leukemia [29]. **[X95735_at]** *ZYX-ZYXIN* is involved in the spatial control of actin assembly and in the communication between the adhesive membrane and the cell nucleus. This is a gene found in many cancer classification studies (*e.g.* [12, 9, 6]) and is highly correlated with acute myelogenous leukemia. **[U46499_at]** *MGST1-Microsomal glutathione S-transferase 1* encodes a protein thought to protect membranes from oxidative stress and toxic foreign chemicals and plays an important role in the metabolism of mutagens and carcinogens [26].

Lung Cancer: **[37957_at]** *ATG4-Autophagy related 4 homolog A*. Autophagy is activated during amino-acid deprivation and has been associated with neurodegenerative diseases, cancer, pathogen infections and myopathies [32]. **[1500_at]** *WT1-Wilms tumor 1* has an essential role in the normal development of the urogenital system; this gene is expressed in several cancer diseases [2]. **[36536_at]** *SCHIP1-Schwannomin interacting protein 1*. The product of the neurofibromatosis type 2 (NF2) tumour suppressor gene, known as schwannomin or merlin, is involved in NF2-associated and sporadic schwannomas and meningiomas [14].

Breast Cancer: **[AL080059]** *TSPYL5-TSPY like 5*. The gene TSPYL5 encodes testis-specific Y-like protein but its role in human cancer has not been fully understood. Gene expression altered by DNA hypermethylation is often associated with cancers. **[NM_003258]** *TK1-Thymidine kinase 1, soluble* is a cytoplasmic enzyme. Studies have shown that in patients with primary breast cancer, high TK values were shown to be an important risk factor in node-negative patients and seemed to be associated with beneficial effect of adjuvant chemotherapy [28]. **[NM_003239]** *TGFB3-transforming growth factor, beta 3* is a potent growth inhibitor of normal epithelial cells. In established tumor cell systems, however, experimental evidence suggests that TGF-bs can foster tumorhost interactions that indirectly support the progression of cancer cells [11].

Prostate Cancer: **[37639_at]** *HPN-Hepsin*. Hepsin is a cell surface serine protease and plays an essential role in cell growth and maintenance of cell morphology and it is highly related with prostate cancer and benign prostatic hyperplasia. **[37720_at]** *HSPD1-Heat shock 60kDa protein 1* encodes a member of the chaperonin family; [20] established HSPD1 as a biomarker for prostate malignancy. **[37366_at]** *PDLIM5-PDZ and LIM domain 5*. Although medical evidence for direct implication of PDLIM5 in prostate cancer was not found, it is a gene recurrently found in several works –e.g. [18].