Use of q-values to improve a genetic algorithm to identify robust gene signatures

D. Urda[(1)], S. Chambers[(2)], I. Jarman[(2)], P. Lisboa[(2)], L. Franco[(1)], J.M. Jerez[(1)]

(1) Department of Computer Science, University of Málaga
Bulevar Louis Pasteur, 35, 29071 Málaga, Spain, {durda,lfranco,jja}@lcc.uma.es

(2) School of Computing and Mathematical Sciences, Liverpool John Moores University
Byrom Street, Liverpool L3 3AF, United Kingdom,
{s.j.chambers,i.h.jarman,p.j.lisboa}@ljmu.ac.uk

Keywords: DNA microarrays, Evolutionary algorithms, t-test, q-values, Feature selection.

Abstract. Several approaches have been proposed for the analysis of DNA microarray datasets, focusing on the performance and robustness of the final feature subsets. The novelty of this paper arises in the use of q-values to pre-filter the features of a DNA microarray dataset identifying the most significant ones and including this information into a genetic algorithm for further feature selection. This method is applied to a lung cancer microarray dataset resulting in similar performance rates and greater robustness in terms of selected features (on average a 36.21% of robustness improvement) when compared to results of the standard algorithm.

1 Scientific Background

DNA microarray technology has been widely used for gene expression profiling and prediction of cancer. Analysis of such data involves facing a problem commonly referred to as the curse of dimensionality [9] where each sample is described by thousands of features (genes) with few samples - often fewer than a hundred - available. Several approaches have been proposed to identify relevant genes with good performance in classifying the disorder under investigation. However, these approaches lack a desirable feature when identifying gene expression profiles - robustness. A common feature of such methods is instability of results with high variability of identified features when repeated executions of the algorithm are made. To tackle this problem, recent works have proposed different methodologies that try to achieve robust feature subset selections with good performance rates in test data [7, 10].

Use of statistical tests with multiple features against some null hypothesis is common practice with the expectation that a proportion of such features would be incorrectly considered significant [8]. In such circumstances it is important to use some form of false discovery rate technique to either adjust the p-values [1] or use a different measure which takes into account false positives such as the q-value [8]. Use of such a measure allows focus to be placed on features which can be considered to satisfy a null hypothesis in further analysis. In the original paper [8] this methodology reduced the number of features identified in the Hedenfalk dataset from 605 to 162 within a total feature set of 3170.

In this paper a modified t-test and q-values [8] are incorporated into a feature selection procedure similar to the genetic algorithm (GA) described in [7] with the purpose of identifying genes that are significant in differentiating lung cancer microarray expressions. In their approach, biological information from KEGG [5, 6] database was included into the GA resulting in more robust feature subsets with good performance

rates. The expectation of introducing a subset of genes, selected using q-values, into the GA would be for better and more robust solutions than the original results from the GA.

The rest of the paper is structured as follows: Section 2 describes the dataset used in this study as well as the methodology applied; Section 3 shows the results obtained in this work and a comparison to previous results of one similar approach; and Section 4 provides some conclusions.

2 Materials and Methods

A freely available[1] high dimensional biomedical dataset has been used throughout this work, comprising 181 tissue samples of two types of lung cancer, malignant pleural Mesothelioma (MPM) and Adenocarcinoma (ADCA) [4]. Samples are unbalanced with 31 corresponding to MPM and 150 ADCA, described in each case by 12533 genes. The Affymetrix ID for the lung cancer DNA microarray dataset is hgu95a and the R package "hgu95a.db" [2] was used to manage and pre-process biological information related to this microarray. For the analysis, the dataset was separated into training and test sets, comprising 80 samples and 101 samples respectively with care taken to keep the same proportion of both MPM and ADCA classes.

The novelty of this approach is the introduction of a more robust statistical method with the expectation of an improvement in the robustness of the final obtained subset of features with direct biological relevance, evidenced by the maintaining of good generalisation in the validation results.

2.1 Significance Testing

A permutation based modified t-test [8] was used to evaluate the null hypothesis that there is no difference in expression between the two different groups (MDM and ADCA) accounting for the different variance within each group. The two sample t-statistic for a given gene is expressed as in (1), and the p-values estimated as per (2). In this case $\bar{x_1}$ and $\bar{x_2}$ represent the means of group 1 and group 2, with s_1^2 and s_2^2 being the respective variances. B is the number of re-samples taken for the modified t-test (a value of B=100 was used), n the number of features and t the value for a given t-statistic (t_1^{0b} to t_n^{0b} are the set of null statistics calculated using the resampling procedure).

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1}$$

$$p_i = \sum_{b=1}^{B} \frac{\#\left\{j : \left|t_j^{0b}\right| \geq |t_i|, j = 1, \ldots, n\right\}}{n \times B} \tag{2}$$

This approach has as the null hypothesis that there is no difference in expression between the two genes and that the t-statistic holds to the same distribution across both [8].

Using the p-values from this study, an FDR-based significance measure, q-values [8], was used to select only those genes significant at the 5% level for inclusion in the model estimation stage. These q-values are an important tool in determining significance of features, and particularly so in genome studies, as they implicitly account for multiple testing, and allow for a more accurate determination of the expected false-positives from the inclusion of a particular feature.

2.2 Model estimation

In this paper, the strategy proposed in [7] is used with some slight changes. The strategy consists of two separate stages:

[1] http://cilab.ujn.edu.cn/datasets.htm

- The 228 pathways identified as related to lung cancer disease are analysed to produce a ranking which allows selection of the best pathways. In contrast to the first stage in [7] only the training dataset of 80 samples is analysed by applying a 10-fold cross-validation strategy. The purpose at this stage is to obtain an accuracy measure and identify the number of keywords for each pathway using a text-mining procedure, following the same process as [7] but refined using genes identified as significant in Section 2.1.

- Using pathways identified from the first stage as being of importance, a genetic algorithm is applied using the fitness function from (3) where $\lambda, \beta \in [0, 1)$ and $\lambda + \beta < 1$, k is the number of selected features, 100 is a normalization factor due to the limited number of active features in a chromosome, and function $score(\mathbf{x})$ which estimates the biological relevance of the selected features. This function has been modified to (4) where M and N are normalization factors representing the number of significant genes on the pathway and total number of significant genes on all pathways respectively. i is the number of selected significant genes included in the pathway being analysed and j the number of selected significant genes not in the pathway such that $i + j = k$. To obtain the accuracy rate for (3), Linear Discriminant Analysis (LDA) [3] is used by applying 10-k-fold cross validation to each chromosome analysed within the GA execution.

- The final step validates the performance of the selected model by performing LDA on the training data and applying the results to the larger test dataset to obtain the prediction accuracy. LDA was chosen in order to make a fair comparison to results previously published in [7] as well as for its simplicity.

$$fitness(\mathbf{x}) = (1 - \lambda - \beta)(1 - ACC(\mathbf{x})) + \lambda \frac{k}{100} + \beta score(\mathbf{x}), \tag{3}$$

$$score(\mathbf{x}) = \frac{\left(1 - \frac{i}{M}\right) + \left(1 - \frac{j}{N-M}\right)}{2}, \tag{4}$$

3 Results

The predictive capability and the number of relevant keywords were calculated for each of the 228 pathways. Table 1 details the best pathways according to Accuracy (Acc) values using the genes identified as being significant at $q < 0.05$ as the first sorting criterion, and the number of keywords found during the text mining of the pathway descriptions in the KEGG database as the second criterion. Bold rows correspond to pathways which ranked in the top 10 found in [7] during the first stage. Those in bold-italic are the six best pathways selected to be analysed on the second stage of the methodology previously. Of note is that the top 10 pathways from the original work are ranked in the overall top 27 pathways ($< 12\%$ of total), and pathway "04610" being the only one exhibiting minimal decline in ranking.

Instead of selecting the best pathways using this ranking as previously done, for comparative purposes the six best pathways from the previous work were selected [7] for analysis using the modified GA presented in Section 2.2 using the test/train datasets for model estimation and validation. This stage of the analysis was repeated 100 times for each of the six pathways to obtain estimates of the model accuracy.

Table 2 shows on average a perfomance comparison using the GA approach published in [7] and our proposal in this paper. In terms of prediction accuracy, both approaches obtain similar performance (approximately 95% depending on the analysed pathway). However, the main advantage of the present approach arises while analyzing the robustness of the subset of features selected. This robustness measure is obtained by

Table 1: Pathways ranked by prediction ability using significant genes and number of keywords found. Bold rows correspond to pathways which ranked in the top 10 according to [7] (See the text for more details).

Rank	*Code*	*Pathway*	*#Genes*	*Acc*	*#Genes 0.05*	*Acc 0.05*	*#Keywords*
1	04020	Calcium signaling pathway	246	0.933	28	0.9975	0/1116
2	***04144***	***Endocytosis***	***244***	***0.99***	***32***	***0.99***	***0/506***
3	04650	Natural killer cell mediated cytotoxicity	172	0.915	13	0.9875	3/871
4	***04010***	***MAPK signaling pathway***	***423***	***0.935***	***37***	***0.9875***	***2/609***
5	**04062**	**Chemokine signaling pathway**	**254**	**0.945**	**29**	**0.9875**	**1/901**
6	04141	Protein processing in endoplasmic reticulum	181	0.908	14	0.9875	1/458
7	**01100**	**Metabolic pathways**	**970**	**0.975**	**146**	**0.9875**	**0/116**
8	00230	Purine metabolism	148	0.93	19	0.9875	0/271
9	04270	Vascular smooth muscle contraction	149	0.973	28	0.9875	0/891
10	**00240**	**Pyrimidine metabolism**	**83**	**0.975**	**15**	**0.9875**	**0/150**
11	04510	Focal adhesion	320	0.955	42	0.985	1/824
12	***05200***	***Pathways in cancer***	***557***	***0.96***	***63***	***0.9825***	***11/4504***
20	***04530***	***Tight junction***	***158***	***0.965***	***27***	***0.9775***	***1/545***
25	**04360**	**Axon guidance**	**166**	**0.975**	**22**	**0.975**	**0/427**
27	***04514***	***Cell adhesion molecules (CAMs)***	***154***	***0.95***	***25***	***0.975***	***0/921***
45	***04610***	***Complement and coagulation cascades***	***73***	***0.978***	***14***	***0.965***	***3/660***

Table 2: Comparison of original approach and proposed approach. Columns 2-4 show the mean of each of No. of Genes,genes in pathway and significant genes in pathway with standard deviations. Additionally, column 5 shows the robustness of results and the last column the accuracy.

	Pathway	*#Genes*	*#Genes in pathway*	*#Genes significant in pathway*	*Robustness*	*Accuracy*
Original GA	04144	4.43±1.00	2.87±1.22	1.75±1.19	0.1225	0.9568±0.0229
	04530	4.22±1.05	2.79±1.11	2.04±1.04	0.14125	0.9630±0.0248
	04514	3.96±1.07	2.32±1.29	1.49±0.95	0.135	0.9463±0.025
	04610	3.85±1.02	2.94±1.24	1.63±1.00	0.24455	0.9398±0.0197
	04010	4.04±1.29	1.71±1.09	0.88±0.83	0.086667	0.9445±0.0275
	05200	3.86±1.52	1.51±1.20	0.68±0.79	0.079	0.9450±0.0249
Our modified GA	04144	4.82±1.31	3.96±1.54	3.66±1.63	0.148	0.9590±0.0201
	04530	4.63±1.54	3.94±1.64	3.86±1.57	0.158	0.9732±0.0192
	04514	5.59±1.55	4.94±1.79	4.89±1.76	0.21591	0.9458±0.0261
	04610	4.29±1.04	3.94±1.23	3.12±1.21	0.29846	0.9439±0.02
	04010	3.54±1.14	2.58±1.33	2.28±1.32	0.15455	0.9431±0.0302
	05200	3.08±1.24	1.63±1.28	1.34±1.20	0.098182	0.932±0.0261

averaging each gene frequency of appearance over the 100 GA executions, discarding those genes that do not appear more than a 5% of the times. In this sense, it should be highlighted that in two out of six pathways analysed, a 78.33% and 59.93% of improvement is reached in terms of robustness (pathways "04010" and "04514" respectively). Pathway "04530" is the one with lowest improvement (just 11.86%), while for the remaining pathways the robustness was approximately increased by a 20%.

The top eleven final selected features for each of the pathways as shown in Table 3 can be directly compared to those obtained in [7]. Consistency is apparent in these as at least four genes are present in the previous work (those highlighted in bold). Furthermore, because of the use of q-values to limit features to those which exhibit significant difference in expression, this approach yields results that contains a larger number of significant selected genes belonging to the top 11 features of a given pathway. These significant genes are also picked by the GA with greater frequency than those shown in the previous work, and thus the robustness of the present method should be higher, as

indeed is as seen in Table 2.

Table 3: Frequency of selection for the most frequently picked features in each pathway previously identified [7] as important. The notes column indicated by† highlights whether the gene is significant in pathway (*), not significant but in pathway (**) or out of pathway(***).

ID	Symbol	Probe Set ID	Freq.(%)	Note†
2520	**CLTB**	**32522_f_at**	13.00	**
1035	KDR	1954_at	14.00	*
633	**ERBB3**	**1585_at**	**15.00**	*
1182	**ERBB3**	**2089_s_at**	**16.00**	*
11052	PARD3	40973_at	16.00	*
2521	**CLTB**	**32523_at**	**22.00**	*
12020	DAB2	479_at	22.00	*
9758	**SH3GLB1**	**39691_at**	**27.00**	*
967	NTRK1	1892_s_at	28.00	*
3893	RAB11FIP5	33882_at	41.00	*
9863	**AP2M1**	**39795_at**	**43.00**	*

(a) Lung Pathway 04144

ID	Symbol	Probe Set ID	Freq.(%)	Note†
6335	PRKCZ	362_at	13.00	*
7453	MYH11	37407_s_at	14.00	*
12312	MYH11	773_at	14.00	*
3916	CLDN3	33904_at	15.00	*
4174	**ACTG1**	**34160_at**	**19.00**	*
8537	**CLDN7**	**38482_at**	**20.00**	*
11052	**PARD3**	**40973_at**	**23.00**	*
8393	**RRAS**	**38338_at**	**26.00**	*
2039	**PRKCD**	**32046_at**	**32.00**	*
3844	**SPTAN1**	**33833_at**	**33.00**	*
5301	**CLDN4**	**35276_at**	**57.00**	*

(b) Lung Pathway 04530

ID	Symbol	Probe Set ID	Freq.(%)	Note†
1248	PECAM1	268_at	24.00	*
1718	HLA-DOA	31728_at	25.00	*
3173	**NEO1**	**33169_at**	**25.00**	*
3916	CLDN3	33904_at	25.00	*
8509	ICAM2	38454_g_at	25.00	*
10372	ICAM3	402_s_at	27.00	*
8537	**CLDN7**	**38482_at**	**29.00**	*
1143	**CDH2**	**2053_at**	**30.00**	*
8508	**ICAM2**	**38453_at**	**34.00**	*
4217	PVRL3	34202_at	38.00	*
5301	**CLDN4**	**35276_at**	**66.00**	*

(c) Lung Pathway 04514

ID	Symbol	Probe Set ID	Freq.(%)	Note†
6581	**F3**	**36543_at**	**11.00**	*
5783	PROS1	35752_s_at	13.00	*
6821	**SERPINA1**	**36781_at**	**14.00**	*
8496	**CD46**	**38441_s_at**	**14.00**	*
12146	VWF	607_s_at	22.00	*
12211	SERPINE1	672_at	23.00	**
9474	**C1R**	**39409_at**	**36.00**	*
5727	**CFI**	**35698_at**	**45.00**	*
8178	**SERPINE1**	**38125_at**	**59.00**	**
5853	**CFB**	**35822_at**	**67.00**	*
9843	**SERPING1**	**39775_at**	**68.00**	*

(d) Lung Pathway 04610

ID	Symbol	Probe Set ID	Freq.(%)	Note†
6700	CD14	36661_s_at	6.00	*
620	PDGFB	1573_at	7.00	*
957	MECOM	1882_g_at	7.00	*
5104	FGF9	35081_at	9.00	*
3250	**MAPK13**	**33245_at**	**11.00**	*
5997	HSPA6	35965_at	14.00	*
909	RRAS2	1838_g_at	16.00	*
8393	**RRAS**	**38338_at**	**16.00**	*
5356	**FLNC**	**35330_at**	**17.00**	*
967	NTRK1	1892_s_at	23.00	*
667	**FGF9**	**1616_at**	**44.00**	*

(e) Lung Pathway 04010

ID	Symbol	Probe Set ID	Freq.(%)	Note†
8813	FADD	38755_at	6.00	*
12200	GAS1	661_at	6.00	***
6616	BIRC2	36578_at	7.00	*
411	RARB	1381_at	8.00	*
7849	SEMA3C	377_g_at	8.00	***
8370	**ALDH1A2**	**38315_at**	**9.00**	***
5104	**FGF9**	**35081_at**	**10.00**	*
967	NTRK1	1892_s_at	11.00	*
1136	**JUP**	**2047_s_at**	**13.00**	*
10532	STAT5A	40458_at	14.00	*
667	**FGF9**	**1616_at**	**16.00**	*

(f) Lung Pathway 05200

4 Conclusion

In this work, a lung cancer disease microarray dataset has been analysed in order to obtain a subset of genes with good predictive performance by using a previously published genetic algorithm modified to include a significance test based on the use of q-values. It has been shown that the inclusion of this information into the GA, to identify those genes having a significant difference in expression, has yielded results that are similar in performance to the original method but exhibiting improved robustness in terms of the selected features with an improvement between 11.86%-78.33% (average 36.21%). This higher robustness observed is achieved as the search in the GA is now guided to genes previously identified as significant without discarding the potential utility of other genes. Moreover, these results are consistent with the original ones since in the top 11 most selected genes, 4 to 6 genes were also included within the results in the original work. Further work should consider a deeper biological analysis of these results and also further investigation of the predictive abilities of pathways either alone or in combination with those genes identified as significant in this study.

Acknowledgments

The authors acknowledge support through Grants TIN2010-16556 from MICINN-SPAIN and P08-TIC-04026 (Junta de Andalucía), all of which include FEDER funds.

References

[1] BENJAMINI, Y., AND HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B 57*, 1 (1995), 289–300.

[2] CARLSON, M. hgu95a.db: Affymetrix Human Genome U95 Set annotation data (chip hgu95a), 2011.

[3] DUDA, R. O., HART, P. E., AND STORK, D. G. *Pattern Classification*, 2nd editio ed. Wiley, 2000.

[4] GORDON, G. J., JENSEN, R. V., HSIAO, L.-L., GULLANS, S. R., BLUMENSTOCK, J. E., RAMASWAMY, S., RICHARDS, W. G., SUGARBAKER, D. J., AND BUENO, R. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res. 62*, 17 (Sept. 2002), 4963–7.

[5] KANEHISA, M., AND GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res. 28*, 1 (Jan. 2000), 27–30.

[6] KANEHISA, M., GOTO, S., SATO, Y., KAWASHIMA, M., FURUMICHI, M., AND TANABE, M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res. 42*, Database issue (Jan. 2014), D199–205.

[7] LUQUE-BAENA, R. M., URDA, D., GONZALO CLAROS, M., FRANCO, L., AND JEREZ, J. M. Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords. *J. Biomed. Inform.* (Jan. 2014).

[8] STOREY, J. D., AND TIBSHIRANI, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A. 100*, 16 (Aug. 2003), 9440–5.

[9] WEST, M. Bayesian factor regression models in the large p, small n paradigm. In *Bayesian Stat. 7 - Proc. Seventh Val. Int. Meet.* (2003), J. M. Bernardo, A. P. Dawid, J. O. Berger, M. West, D. Heckerman, M. Bayarri, and A. F. Smith, Eds., Oxford University Press, pp. 723–732.

[10] XU, J.-Z., AND WONG, C.-W. Hunting for robust gene signature from cancer profiling data: sources of variability, different interpretations, and recent methodological developments. *Cancer Lett. 296*, 1 (Oct. 2010), 9–16.