Integer Programming Models and Algorithms for Molecular Classification of Cancer from Microarray Data

Regina Berretta **Alexandre Mendes** **Pablo Moscato**

Newcastle Bioinformatics Initiative
School of Electrical Engineering and Computer Science
Faculty of Engineering and Built Environment
University of Newcastle
Callaghan, NSW, 2308, Australia
{regina,mendes,moscato}@cs.newcastle.edu.au

Abstract

Novel, high-throughput technologies are challenging the core of algorithmic methods available in Computer Science. Microarray technologies give Life Sciences researchers the opportunity to simultaneously measure thousands of gene expression levels under different conditions or coming from different cell lines. With appropriate data mining models and algorithms, this would lead to a systematic exploration of molecular classification of cancer, just one among many other exciting applications. The aim of this paper is to present a unified mathematical formalization for different feature selection problems and investigate their performance in classification of cancer cell-lines. We also present some results using the NCI60 dataset.

Keywords: Data Mining, Feature Selection, Combinatorial Optimisation, NCI60.

1 Introduction

With the recent introduction of microarray technology, Life Science researchers are now able to simultaneously measure the expression level of thousands of genes in cells from a tissue sample or under different controlled conditions (Quackenbush, 2001). This allows an unprecedented range of possibilities. To analyse this data we can use clustering/ordering algorithms (Cotta *et al*., 2003), classification methods (Dash and Liu, 1997) and/or their combinations as we do in this contribution.

A primary difficulty faced is that the amount of data coming from microarray experiments can be very large. In most cases there are many more genes available (the features of interest) than samples. Typically the ratio between samples and genes is about 1/100. It is possible to find low cardinality explanations for the process of interest but many of these genes may be totally unrelated to the research question. The correlations that lead to these explanations can just be explained due to the low sample to features ratio. This means that clustering and/or classification methods should take this into account.

This paper is concerned with how to reduce the large amount of data coming from microarray experiments, by trying to select relevant genes with the purpose to help understand the reasons behind different outcomes. We thus present new models and algorithms which may have important uses in the wide field of the Feature Selection problem (Dash and Liu, 1997, 2003; Frank, 2002). The Feature Selection problem is an important component in areas such as gene discovery, disease diagnosis, drug discovery, cancer research (Marks, 2000) and predictive genomic medicine.

Feature Selection, as proposed in this paper, is used to reduce the dimensionality of the data but not beyond a point in which we may have missed a subset of genes which actually relate well with the biological process we are trying to uncover. The Feature Selection problem can be defined as trying to find a reduced set of features, which optimizes some goal (consistency, error rate, etc.) (Frank, 2002). While this may be an elusive goal, in this paper we present two mathematical models (see Sections 2.2 and 2.3) which are able to reduce the number of genes selected to groups of between 4.92 and 1.82 percent of the genes, yet maintaining a relatively large number of within class similarities. The solutions presented guarantee an optimal number of within class similarities (under some extra constraints explained in Section 4). Our solutions also guarantee an optimal (maximum) dissimilarity for samples in different classes. What this means is that, if there exists a pair of samples that belong to different classes, differing in *a_max* features then all other pairs of examples from different classes would have at least *a_max* with different types of attribute values.

In this paper, the proposed integer programming models and the mathematical formalization allow a discussion of pre-processing rules used in combinatorial optimization to reduce the size of the instances. The optimality of our solutions has been verified by the utilization of the CPLEX (a mathematical programming software package). We illustrate the usefulness of the proposed approach using a dataset known as NCI60 from Cancer Microarray Project, Stanford, available online (Ross et al., 2000). We present different results and a detailed comparative analysis.

The paper is organized as follows. In Section 2, we present integer programming models for different feature selection problems. In Section 3, we describe the instances we used for our computational tests. Finally, in

Sections 4 and 5, we will present the computational results and conclusions.

2 Mathematical Models

2.1 Min Feature Set

The Min Feature Set problem we consider in this paper can be understood as follows. Consider a matrix $G = g_{ij}$, $1 = i = e$, $1 = j = n$, where e is the number of experiments/samples, n is the number of features (genes) and g_{ij} represents the level of activity of gene j in the experiment i. We are considering that g_{ij} is the result of a measurement and can be represented, without loss of generality, as an integer. Ideally, we should represent g_{ij} values as belonging to a small cardinality domain of different types of values (true/false or high/medium/low, etc.). Consider also a vector $T = t_i$, where $1 = i = e$ and t_i represents the class that corresponds to the experiment i. The objective is to find the minimum cardinality set of features (genes), denoted as S, such that for all pairs of experiments that belong to different classes, there exists at least one feature (gene) that belongs to S, and such that the level of activity is different among experiments for such a feature. In other words,

for all pairs (p,q) with $t_p \neq t_q$

$\exists j \in S$ such that $g_{pj} \neq g_{qj}$.

To illustrate, suppose that one instance of the problem is the following Boolean matrix G and the Boolean vector T below. In this case, the minimum feature set for this instance is $S = \{F4, F5\}$.

G_{5x5}

	F1	F2	F3	F4	F5
E1	1	0	0	0	1
E2	0	1	1	0	1
E3	1	0	0	0	0
E4	1	1	1	1	1
E5	0	1	0	1	1

T_5

Class
0
0
1
1
1

The k-Feature Set problem is NP-complete (Davies and Russel, 1994). Cotta and Moscato (2003) showed that the parameterized version of Min Feature Set problem (when the parameter is the cardinality of the feature set) is W[2]-Complete.

With the purpose to write an integer programming model, first we define a matrix $A = a_{ij}$, $1 = i = m$, $1 = j = n$, where m is the number of pairs of examples that belong to different classes, n is the number of features, and a_{ij} is 1 if g_{pj} ? g_{qj} or 0 if $g_{pj} = g_{qj}$, where t_p ? t_q. In other words, a_{ij} represents whether the features types in the pair of examples that belong to different classes (p,q) are different or not. Using the previous illustrative instance of the problem, the matrix A would be:

	F1	F2	F3	F4	F5
E1,E3	0	0	0	0	1
E1,E4	0	1	1	1	0
E1,E5	1	1	0	1	0
E2,E3	1	1	1	0	1
E2,E4	1	0	0	1	0
E2,E5	0	0	1	1	0

The objective is to choose a minimum subset S of features (columns) which have at least one '1' value in each line. That is, a minimum set of features corresponds to the minimum subset of columns having ones that cover all pair of examples. Notice that the minimum feature set in the example is $S = \{4,5\}$.

An integer programming model for the Min Feature Set can be as shown below, where the variable $x_j = 1$ if the feature j is chosen; and 0, otherwise.

$$Min \sum_{j=1}^{n} x_j \qquad (1)$$

$$\sum_{j=1}^{n} A_{ij} x_j > 0 \qquad i=1,\dots, m \qquad (2)$$

$$x_j = 0 \text{ or } 1.$$

Note that the model (1-2) represents also the Set Covering problem. The Set Covering problem is a classical problem in combinatorial optimization for which many techniques have been developed (Caprara, Toth and Fischetti, 2000).

2.1.1 Reductions for Min Feature Set

Reductions for Min Feature Set are rules that can be applied to an instance to try to eliminate, a priori, some rows and columns from matrix A and, consequently, reduce the instance size. We describe four reduction rules for the Min Feature Set problem below. These reductions rules are the same for the Set Covering Problem and it is possible to find them in references about Integer Programming such as Garfinkel and Nemhauser (1972).

Reduction R0

If $a_{ij} = 0$ for all j, then, the instance is infeasible, since the constraint (2) cannot be satisfied. In other words, if no feature can distinguish a pair of examples that belong to different classes, then the instance is infeasible.

Reduction R1

If $a_{ij} = 0$ for all j ? k and $a_{ik} = 1$, then $x_k = 1$. In other words, if just one feature distinguishes a pair of examples that belong to different classes, then this feature must be in any feasible cardinality solution. In addition, all rows i such that $a_{ik} = 1$, can be deleted, since the feature k will cover these lines. Finally, column k can be deleted.

In the example given, the feature $F5$ should be in the solution, since it is the only one that covers the pair of examples $(E1,E3)$. We can delete row 1 and 4, since the pair of examples $(E1,E3)$ and $(E2,E3)$ are covered by the inclusion of feature $F5$ in our solution.

Reduction R2

A feature j covers a subset W if $a_{ij} = 1$ for all i Î W. If a feature j_1 covers a subset W_1 and j_2 covers a subset W_2 and $W_2 \subseteq W_1$, then feature j_2 is dominated by feature j_1 and consequently, can be deleted.

In the example above, after being updated with the result of reduction R1, the feature $F4$ covers the set $W_4 = \{(E1,E4), (E1,E5), (E2,E4), (E2,E5)\}$. The feature $F3$

covers the set W_3 = *{(E1,E4), (E2,E5)}*. Since, $W_3 \subseteq W_4$, F3 is redundant and can be deleted. Notice that, with the same rule we can delete F1 and F2. Now, using the reduction rule R1, feature F4 is chosen and the instance is solved to optimality (as the reduction rules are safe procedures that do not miss at least one optimal solution of the original instance after they reduce it).

Reduction R3

Let $Q_1 = \{ j / a_{i_1 j} = 1 \}$ and $Q_2 = \{ j / a_{i_2 j} = 1 \}$. If Q_1 Í Q_2 then row i_2 can be deleted. In other words, if a pair of examples i_1 is covered by the set of features Q_1 and a pair of examples i_2 is covered by the set of features Q_2 and Q_1 Í Q_2, we can delete the pair i_2, since it will be covered by any of the features chosen to cover the pair i_1.

In the example, the pair of examples *(E1,E3)* is covered by Q_1 = *{F5}* and the pair of examples *(E2,E3)* is covered by Q_2 = *{F1,F2,F3,F5}*. Since Q_1 Í Q_2, the pair of examples *(E2,E3)* can be deleted from matrix *A*. Notice that when we choose a feature to cover the pair *(E1,E3)* we inevitably will cover the pair *(E2,E3)*.

Although the Min Feature Set is an NP-hard optimization problem, the reduction rules can be very useful in practice to reduce the instance size before we apply a method (either a polynomial-time heuristic or an exact exponential time algorithm) to find one of the optimal solutions.

2.2 Min a-b Feature Set

A generalization of the Min Feature Set is the Min α-β Feature Set introduced by Cotta, Sloper and Moscato (2004). This generalization could be very useful when the dataset is noisy and a larger number of different features needs to be considered.

The problem is defined as follows. We have the same input as for Min Feature Set, i.e., matrix $G = g_{ij}$, $1 = i = e$, $1 = j = n$, where *e* is the number of experiments/samples and *n* is the number of features (genes) and a vector $T = t_i$, where $1 = i = e$ and t_i represents the class (outcome) of the experiment *i*. In addition, the input also includes two integer values a =*1* and $b = 0$. The objective is again to find the minimum set of genes (features) *S*, but the two conditions below also need to be satisfied.

Condition 1 For all pairs of samples that belong to different classes, at least *a* features that belong to *S* have different feature types. In other words,

For all pairs (p,q) with $t_p \neq t_q$,

define $S_1 = \{ j \in S \mid g_{pj} \neq g_{qj} \}$

So, $|S_1| \geq a$.

Condition 2 For all pairs of samples that belong to the same class, at least *b* features that belong to *S* have identical feature types. In other words,

For all pairs (p,q) with $t_p = t_q$

define $S_2 = \{ j \in S / g_{pj} = g_{qj} \}$

So, $|S_2| \geq b$.

To illustrate, consider the same matrix *G* defined previously. Observe that, if we have as input the values *a=1* and *b=1*, the Min *a-b* Feature Set cannot be *S* = *{F4,F5}*, since the examples *E3* and *E4*, which belong to the same class are completely different for the features *F4* and *F5*. For *a=1* and *b=1* the minimum cardinality *(a=1/b=1)* feature set is *S* = *{F1,F3,F5}*.

For an integer programming formulation for this problem, we will define two matrices, *A* and *B*. Matrix *A* will be the same defined before, that is, $A = a_{ij}$, $1 = i = m$, $1 = j = n$, where *n* is the number of features, *m* is the number of pairs of examples that belong to different classes and a_{ij} is *1* if g_{pj} ? g_{qj} or *0* if $g_{pj} = g_{qj}$, where t_p ? t_q. Matrix *B* will be $B = b_{ij}$, $1 = i = m'$, $1 = j = n$, where *n* is the number of features, *m'* is the number of pairs of examples that belong to the same classes and b_{ij} is *1* if $g_{pj} = g_{qj}$ or *0* if g_{pj} ? g_{qj}, where $t_p = t_q$.

Using the previous example, the matrix B would be

	F1	*F2*	*F3*	*F4*	*F5*
E1,E2	*0*	*0*	*0*	*1*	*1*
E3,E4	*1*	*0*	*0*	*0*	*0*
E3,E5	*0*	*0*	*1*	*0*	*0*
E4,E5	*0*	*1*	*0*	*1*	*1*

The mathematical model can be written as:

$$\text{Min} \sum_{j=1}^{n} x_j \qquad (3)$$

$$\sum_{j=1}^{n} A_{ij} x_j \geq a \qquad i=1,\dots,m \qquad (4)$$

$$\sum_{j=1}^{n} B_{ij} x_j \geq b \qquad i=1,\dots,m' \qquad (5)$$

$$x_j = 0 \text{ or } 1$$

2.2.1 Reductions for Min *a-b* Feature Set

We define below reduction rules for Min *a-b* Feature Set as described before for Min Feature Set Problem. Consider the following definitions:

$$Q_a^i = \{ j / a_{ij} = 1 \} \text{ and } Q_b^l = \{ j / b_{lj} = 1 \}.$$

The sets Q_a^i and Q_b^l represent the features that can cover a pair of samples *i* and *l*, respectively, from matrix *A* and *B*. Let r_a^i be an integer that represents the number of features that remain to cover the pair the samples *i* by *a*. Equivalently, r_b^l represents the number of features that remain to cover the pair the samples *l* by *b*. At the beginning of the application of the reduction rules $r_a^i = a$ and $r_b^l = b$.

Reduction R0

If $|Q_a^i| < r_a^i$, for at least one row *i* from matrix *A*, then the instance is infeasible, since the constraint (4) cannot be satisfied. Analogously, if $|Q_b^l| < r_b^l$, for at least one row l from matrix B, the instance is infeasible, since constraint (5) cannot be satisfied. In other words, if at least there is one pair of examples that belong to different classes and does not have at least r_a^i features that have different types for them, then the instance is infeasible (analogously for the within class similarity constraint).

Reduction R1

If $|Q_a^i| = r_a^i$, for any pair of examples i, then $x_j = 1$ for all j Î Q_a^i. In other words, if a pair of examples i is covered by exactly r_a^i features, then all these features should be in any optimal solution. Next, for all j Î Q_a^i, it is necessary to update all $r_a^i/a_{ij}=1$ and $r_b^l/b_{ij}=1$. Finally, we can delete all rows i from A such that $r_a^i = 0$, all rows l from B such that $r_b^l = 0$ and all columns j Î Q_a^i.

Analogously, if $|Q_b^l| = r_b^l$, for any l, then $x_j = 1$ for all j Î Q_b^l. In other words, if a pair of examples is covered by exactly r_b^l features, then all these features should be in the solution. Next, for all j Î Q_b^l, it is necessary to update all $r_a^i/a_{ij}=1$ and $r_b^l/b_{ij}=1$. Finally, again, we can delete all rows i from A such that $r_a^i = 0$, all rows l from B such that $r_b^l = 0$ and all columns j Î Q_a^i.

Consider $a=b=1$ in the example. The feature *F5* should be in the solution, since it is the only one that covers the pair of examples *(E1,E3)* when we examine the matrix A. Also we can delete rows *1* and *4* from matrix A, since the pair of examples *(E1,E3)* and *(E2,E3)* are covered by feature *F5* with $a=1$.

In the matrix B we can delete the rows *1* and *4*, since the feature *F5* will cover the pair of examples *(E1,E2)* and *(E4,E5)*. We conclude that features *F1* and *F3* should be in the solution, since only *F1* covers the pair of examples *(E3,E4)* and only *F3* covers the pair of examples *(E3,E5)*. We also can delete the rows *2*, *3* and *6* from matrix A, since features *F1* and *F3* cover all pair of examples that remain in the matrix A. Notice that we could reduce the entire instance and finish with the solution *{F1,F3,F5}*.

Reduction R2

A feature j covers a subset W_a if $a_{ij} = 1$ for all i Î W_a. Respectively, a feature j covers a subset W_b if $b_{ij} = 1$ for all i Î W_b. If a feature j_1 covers a subset W_a^1 and W_b^1 ; j_2 covers a subset W_a^2 and W_b^2 ; $W_a^2 \subseteq W_a^1$ and $W_b^2 \subseteq W_b^1$; and for all i Î W_a^2 we have $|Q_a^i| > r_a^i$ and for all i Î W_b^2 we have $|Q_b^i| > r_b^l$; then j_2 is redundant and can be deleted.

Reduction R3

If $Q_a^{i_1}$ Í $Q_a^{i_2}$ and $r_a^{i_1} = r_a^{i_2}$ or $Q_b^{l_1}$ Í $Q_a^{i_2}$ and $r_b^{l_1} = r_a^{i_2}$, then row i_2 from matrix A can be deleted. In other words, if a pair of examples i_1 is covered by the set of features $Q_a^{i_1}$; a pair of examples i_2 is covered by the set of features $Q_a^{i_2}$ and $Q_a^{i_1}$ Í $Q_a^{i_2}$; then the pair of samples i_2 can be deleted, if the number of features that remain to cover the pair the samples i_1 is greater than the number of features that remain to cover the pair the samples i_2 ($r_a^{i_1} = r_a^{i_2}$). Equivalent interpretation can be done if $Q_b^{l_1}$ Í $Q_a^{i_2}$ and $r_b^{l_1} = r_a^{i_2}$. Analogously, if $Q_b^{l_1} \subseteq Q_b^{l_2}$ and $r_b^{l_1} = r_b^{l_2}$ or $Q_a^{i_1} \subseteq Q_b^{l_2}$ and $r_a^{i_1} = r_b^{l_2}$ then row l_2 from matrix B can be deleted.

2.3 Max Cover *a-b* Feature Set

Another mathematical model we introduce is obtained by fixing the number of features in a value n_{fix} and the objective is to find a set of n_{fix} features that maximize the coverage. The coverage represents the number of pair of examples that belong to different classes (matrix A) plus the number of pair of examples that belong to the same class (matrix B) that the set of features cover, including repetitions. The coverage of a feature j is:

$$c_j = \sum_{i=1}^{m} A_{ij} + \sum_{i=1}^{m'} B_{ij}$$

The mathematical model is described below.

$$\text{Max} \sum_{j=1}^{n} c_j x_j \qquad (6)$$

$$\sum_{j=1}^{n} A_{ij} x_j \geq a \qquad i=1,\dots,m \qquad (7)$$

$$\sum_{j=1}^{n} B_{ij} x_j \geq b \qquad i=1,\dots,m' \qquad (8)$$

$$\sum_{j=1}^{n} x_j \leq n_{fix} \qquad (9)$$

$$x_i = 0 \text{ or } 1$$

This model can be useful, for example, when an instance has more than one optimal solution when we use the model (3-5).

3 The NCI60 Instance

Ross et al. (2000) introduced an important dataset for the molecular classification of different types of cancer. The data corresponds to gene expression in 64 cell lines using DNA microarrays robotically spotting 9,703 cDNAs. The cDNAs included approximately 8,000 different genes. At the time of presenting this dataset, 3,700 of the genes represented previously characterized human proteins and 2,400 were identified only by ESTs. We are working with a dataset available on the authors' website supplement containing gene expression of 6,831 genes corresponding to Figure 2b of their paper.

There are several good reasons to use this instance for our studies. In their original paper, Ross et al. have identified several groups of genes that correspond to some of the tissue characteristics of the cell lines. Of particular interest for the objectives of our paper are two groupings named “Leukaemia Cluster” and “Melanoma Cluster” corresponding to Figures 3a and 3c of Ross *et al*., respectively. These have been visually identified from a hierarchical clustering as a highly-expressed group of genes in the leukaemia-derived and in most of the melanoma-derived cell lines. It is, however, very difficult to identify, from a hierarchical clustering, an analogous group of genes that is highly under-expressed and that is a robust significant marker of differential expression within the same cell-line and that at the same time discriminates well all other types of lines. The approach we present in this paper has been designed to uncover such groups if they exist. To our knowledge, no other method has been able to identify some of the key genes that allow such an interpretation linking both the highly expressed or under expressed gene expression of groups of genes on this dataset.

In addition, Waddell and Kishino (2000) discussed that such a dataset, even if excellent in technical terms (with a claimed coefficient of variation due to experimental errors of approximately between 20 and 30%), may be of low information content. They argue that Ross et al. did not emphasise on the impact of mutation on cell lines upon their analysis. As a consequence, there are cases of genes that were expected to have a clear relationship (for instance, TP53/Waf1 or p16/Rb) which have a weak pair relationship in this instance. It is then possible that the expression profiles, conditioned to the mutation status of group of “key player” genes, would be part of the explanation. On the other hand, the expression profiles on a large number of genes may help to classify cancers even in the presence of large systematic errors. Our approach is designed to give a relatively larger number of genes, uncovering a more informative set of under expressed genes in the NCI60 dataset, which in turn may help to discover the genetic pathways at play in this case.

4 Computational Results

There are three main reasons motivating the design of our computational experiments: a) the discussion of the previous section, b) the possibility of a direct comparison with Ross et al., Figures 3a and 3c (“Leukaemia Cluster” and “Melanoma Cluster”), and c) the absence of clear highly-expressed analogous clusters for Colon and Renal cell-lines. Towards this end, we have developed the following series of experiments to uncover the key genes that could explain these classes.

We have first completed all missing values for the NCI60 dataset using the LSImpute_EMarray algorithm recently introduced by Bø, Dysvik and Jonassen (2004). Our choice was based on its relatively low running time and good performance on the NCI60 dataset as independently verified by the original authors. For the estimation of the missing values we have used the initial set of 64 cell lines and 6,381 genes. We have calculated the standard deviation of the expression values in the instance (0.7904).

After the missing values have been completed we worked with a reduced set, comprising five different groups of similar number of cell lines. These groups have been chosen based on their tissue or origin as well as the similarity of the overall gene expression profile. The five groups and their associated 41 cell lines are described in Figure 1.

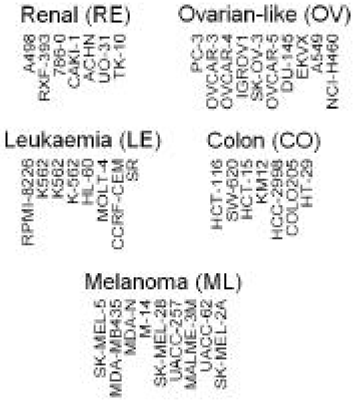

Figure 1. The five cancer groups with the respective cell-lines in the same order they appear in our figures and in Ross et al (2000).

We note that Ovarian-like is a class that contains ovarian cell lines with the addition of NCI-H460, A549, EKVX from Non-Small cell Lung cancer (NSL class in Ross et al) and cell-lines DU-145 and PC-3, from Prostate cell lines, which nevertheless have similar gene expression profiles. Analogously, cell lines NCI-H23 and NCI-H522 (from NSL class in Ross et al.) have not been included due to their dissimilar gene expression patterns with the other cell lines in this group. There are three cell lines for K562 and all are part of the group in our study.

We have then proceeded to establish a series of computational experiences. In each one, only two classes are given. For instance, to identify differentially expressed genes in Melanoma cell-lines, we aim at identifying ‘Melanoma vs. all-others’, where all others in this case correspond to the other four remaining (Renal, Ovarian-like, Leukaemia and Colon) bundled as a single group (not-Melanoma). A threshold of 1.5 times the standard deviation (calculated using the entire dataset) is used to reduce the feature types. This quantizes the gene expression values in only three types (*low, mid, high*), with *‘low’* corresponding to all gene expression values below a threshold of -1.1856, with *‘high’* corresponding to expression values above 1.1856, and with *‘mid’* being assigned to all other expression values.

For each one of these five different instances of the problem Renal (RE), Ovarian-like (OV), Leukaemia (LE), Colon (CO) and Melanoma (ML), we have done the following:

1) We have created an instance following the two modified conditions from the ones presented in Sec. 2.2:

Modified Condition 1

For all pair of cell lines that belong to different classes, there should be at least *a* genes that belong to *S* (which is the *a-ß* feature set we are looking for), such that the level of activity is *markedly different* (low in one and high in the other). In other words:

For all pairs (p,q) with $t_p \neq t_q$,

define $S_1 = \{ j \in S \,/\, g_{pj} = low \wedge g_{qj} = high \}$.

So, $|S_1| \geq a$.

Modified Condition 2

For all pair of cell lines that belong to the same class, there should be at least *ß* genes that belong to *S*, such that the level of activity is either *'high'* or *'low'* in both (but not *'mid'* in both cases). Analogously we can write:

For all pairs (p,q) with $t_p = t_q$, define

$S_2 = \{ j \in S \,/ (g_{pj} = low \vee g_{pj} = high) \wedge g_{pj} = g_{qj} \}$.

So, $|S_2| \geq b$.

2) We have then found, for each of the instances, the maximum number of *a* that could be obtained by any optimal *a-ß* feature set. The obtained values were *24* for RE, *16* for OV, *45* for LE, *16* for CO, and *46* for ML. This means, for instance, that *a priori* we know that there is no pair of cell lines, with one belonging to the seven Renal cell lines and the other belonging to anyone of the other four groups, having more than *25* genes markedly differing in *'high'* vs. *'low'* expression values.

3) We then find, for each of the instances, the size of the minimum cardinality *a-ß* feature set, with *ß=0* and with *a* being fixed to the maximum *a priori* value which is possible for that instance. We have solved each of these problems to optimality using CPLEX (a mathematical programming software package). We found that there exists: a *a=24-ß=0* feature set (with an optimal number of *k=198* genes) for RE, a *a=16-ß=0* feature set with *k=140* genes for OV, *a=45-ß=0* feature set with *k=307* genes for LE, a *a=16-ß=0* feature set with *k=116* for CO, and a *a=46-ß=0* feature set with *k=314* for ML.

4) Finally, we aim to try to find the maximum *ß* achievable by a Max Cover *a-ß* Feature Set (with *a* fixed to the previously obtained maximum *a priori* values), for each of the optimal cardinalities obtained in the previous step. We have solved each of this Max Cover *a-ß* Feature Set problems to optimality so we found that there exist: a Max Cover *24-ß*=0 feature set (with an optimal number of *k=198* genes,) for RE (in this case it was not possible to increase the value of *ß* without increasing the cardinality of the set), a *16-ß=3* feature set with *k=140* genes for OV, *45-ß=8* feature set with *k=307* genes for LE, a *16-ß=4* feature set with *k=116* for CO, and a *46-ß=9* feature set with *k=314* for ML.

These solutions are shown in Figures 2 to 6. An ordering algorithm has been independently applied to each of these subsets of genes to highlight the correlations between genes. It is clear that our method has uncovered a significantly large number of genes that are differentially under-expressed and can contribute to our understanding of the mechanisms that control regulation in these diseases. Our figures illustrate another source of useful information that is obtained by good orderings of the identified genes. For instance, a large number of genes are differentially expressed in Leukaemia and Melanoma (see the lower half of both Figures 2 and 5) yet markedly up-regulated in the other cell-lines. Figure 2 shows Leukaemia cell-lines as highly characterized by a large number of under-expressed genes. This figure contrasts with the solution for the Ovarian-like group (Figure 4) where it seems to be the case that a finer distinction between cell-lines is necessary for proper classification, yet some genes appear to be up-regulated in contrast with down-regulation in the Leukaemia, Colon and Melanoma groups. Finally the results for Melanoma can be seen in the context of a direct comparison with Figure 3c of Ross et al. (2000). We uncover a large number of down-regulated genes, absent in previous articles that also use the same dataset, which may give new insights on the molecular mechanisms of this disease.

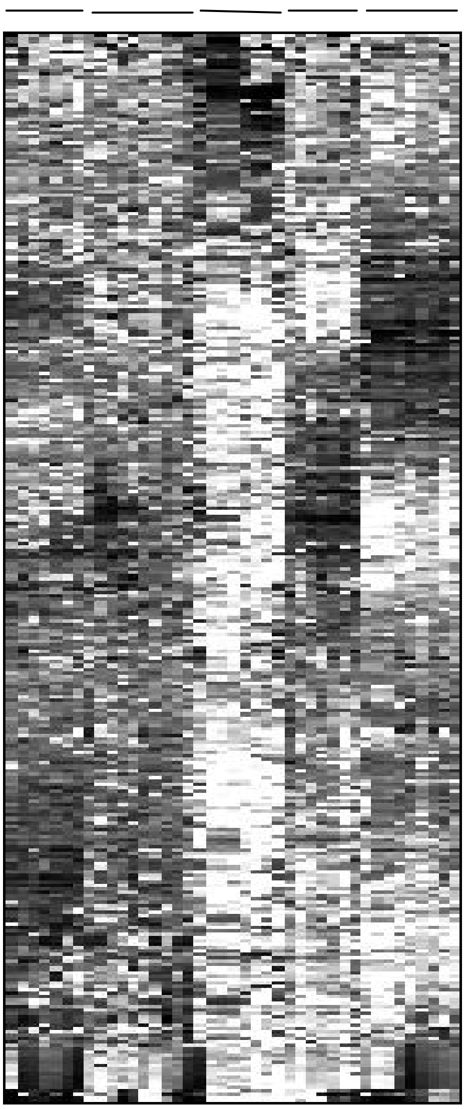

Figure 2. The Leukaemia Max Cover (*a*=45, *ß*=8) gene subset containing 307 up and down regulated genes. The Leukaemia group is located between columns 18

and 25. Although a group of up-regulated genes in these cell lines is clear near the top of the figure, these cell lines markedly differ from other cell lines in being mostly down regulated.

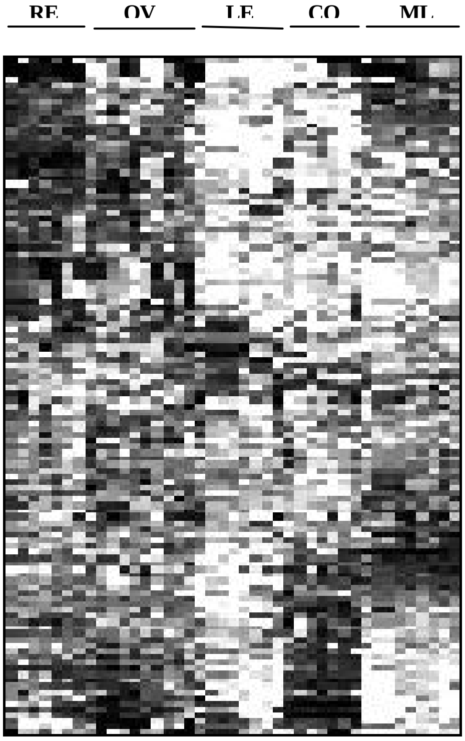

Figure 3. The Colon Max Cover (*a*=16, *ß*=4) gene subset containing 116 up and down regulated genes. The Colon group is located between columns 26 and 32. A group of up regulated genes in these cell-lines and down regulated in the Melanoma class is easy to spot in the lower-right corner of the figure.

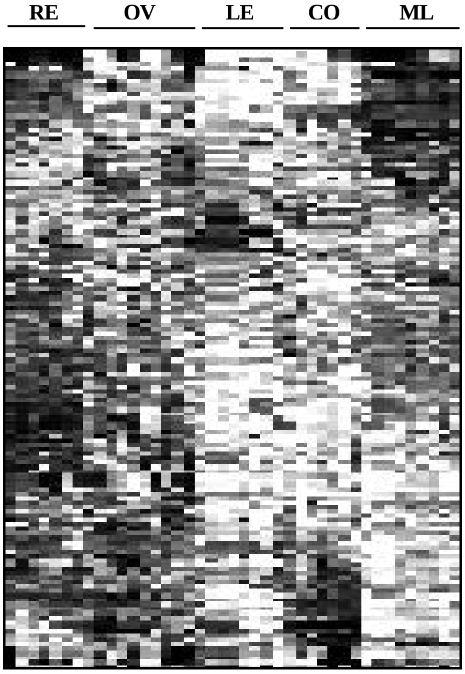

Figure 4. The Ovarian-like Max Cover (*a*=16, *ß*=3) gene subset containing 140 up and down regulated genes. The Ovarian-like group is located between columns 8 and 17. Although several up and down regulated genes help to characterize this group, it is difficult to find a distinguishing subset of genes which differentially are up and down regulated across of all other types of cell lines. This may be a consequence of our decision of grouping different cell lines in this class.

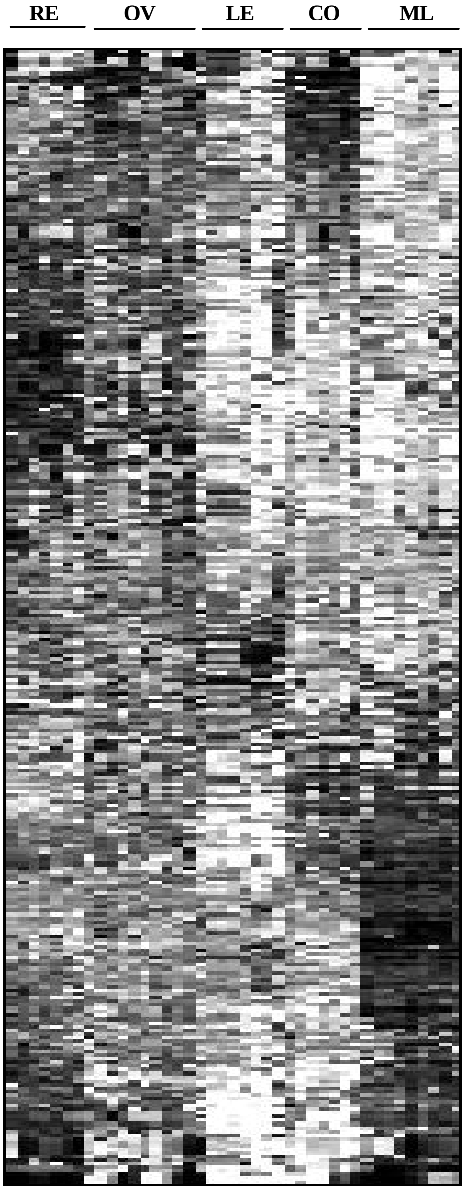

Figure 5. The Melanoma Max Cover (*a*=46, *ß*=9) gene subset containing 314 up and down regulated genes. The Melanoma group is located between columns 33 and 41. Although a group of up regulated genes in these cell lines is clear near the bottom right corner of the figure (and some of these have been previously reported), the solution here presented shows a relatively larger number of down-regulated genes. Near the upper right corner of the figure we can find a subset which is down-regulated, sharing this with the Leukaemia group, yet is markedly different for all other cell lines.

5 Conclusion

We have presented new models and algorithms that have shown to be very useful to address the molecular classification of cancer from microarray data. The methods are general and their applicability is not limited to the field of Bioinformatics. They are mainly based on a generalization of the *k*-feature set problem called *(a-ß) k*-feature set which was recently introduced by Cotta, Sloper and Moscato (2004). The results indicate that the method allows a good balance of discrimination between classes as well as a within-class consistency. This allows Life Science researchers to uncover a larger number of genetic pathways that could lead, in turn, to a broad picture of differential genetic regulation mechanisms.

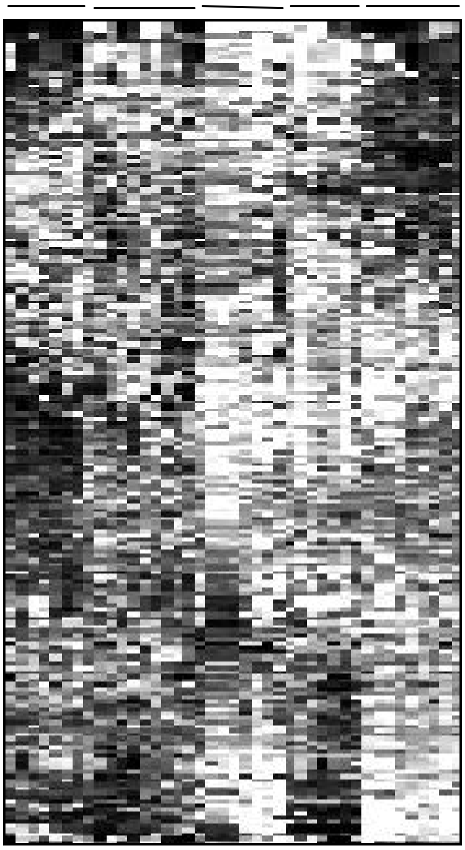

Figure 6. The Renal Max Cover (a=24,ß=0) gene subset containing 198 up and down regulated genes. The Renal group is located between columns 1 and 7. It is possible to identify a subset of genes which are markedly differently expressed between this group and almost all other cell lines. Near the left bottom corner, a group of genes are over expressed and they are in sharp contrast with the Melanoma group.

Our contribution also highlights the importance of safe data reduction methods that keep optimal solutions and maintain the relevant information in the data. It also contrasts with previous research using the same dataset, mainly based on clustering, which has been limited to uncovering highly-expressed genes. This is only one part of the necessary information to understand the genetic network dynamics. Our methods also provide a significant number of down-regulated genes, which have been not previously identified in this dataset. The large number of such genes in the Melanoma group of cell-lines (given in the Appendix) is indicative of the relevance and flexibility of the method, which would help to uncover yet unknown mechanisms that link genes, their products, and disease.

6 References

Bø, T.H., Dysvik B. & Jonassen, I. (2004), LSimpute: accurate estimation of missing values in microarray data with least squares methods, *Nucleic Acids Research* **32**(3), e34.

Caprara, A., Toth, P. & Fischetti, M. (2000), Algorithms for the Set Covering Problem, *Annals of Operations Research* **98**, 353–371.

Cotta, C., Mendes, A., Garcia, V., França, P. & Moscato, P. (2003), Applying Memetic Algorithms to the Analysis of Microarray Data, Proceedings of *EvoBIO2003 - 1st European Workshop on Evolutionary Bioinformatics*, *Lecture Notes in Computer Science* **2611**, 22–32, Springer-Verlag, Colchester, England.

Cotta, C. & Moscato, P. (2003), The *k*-Feature Set problem is W[2]-complete, *Journal of Computer and System Sciences* **67**, 686–690.

Cotta, C., Sloper, C. & Moscato, P. (2004), Evolutionary search of thresholds for robust feature set selection: application to the analysis of microarray data, *Applications of Evolutionary Computing*, G. Raidl *et al.* (eds.), *Lecture Notes in Computer Science* **3005**, 21-30, Springer-Verlag, Berlin.

Dash, M & Liu, H. (1997), Feature Selection for Classification, *Intelligent Data Analysis*, **1**(3), 131–156.

Dash, M. & Liu, H. (2003), Consistency-Based Search in Feature Selection, *Artificial Intelligence Journal* **151** (1-2), 155–176.

Davies, S. & Russell, S. (1994), NP-completeness of searches for smallest possible feature sets, Proceedings of the *AAAI Fall Symposium on Relevance*, New Orleans, USA.

Frank, A. (2002), *A New Branch Bound Feature Selection Algorithm*, Master Thesis, Israel Institute of Technology.

Garfinkel, R.S. & Nemhauser, G.L. (1972), *Integer Programming*, New York: John Wiley & Sons.

Liu, H. & Wong, L. (2003), Data mining Tools for Biological Sequences, *Journal of Bioinformatics and Computational Biology* **1**(2), 139–167.

ILOG CPLEX 9.0, http://www.ilog.com/products/cplex/

Marx, J. (2000), DNA Arrays reveal cancer in its many forms, *Science* **289**, 1670–1672.

Quackenbush, J. (2001), Computational analysis of cDNA microarray data, *Nature Reviews* **2**(6), 418-428.

Ross, D.T., Scherf U., Eisen M.B, et al. (2000), Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**(3), 227–235.

Waddell P.J. & Kishino, H. (2000), Cluster Inference Methods and Graphical Models Evaluated on NCI60

Microarray Gene Expression Data, Genome Informatics **11**, 129–140.

Yu, L. & Liu, H. (2004), Redundancy Based Feature Selection for Microarray Data, SIGKDD, KDD 2004, 22–25, Seattle, Washington, USA.

Appendix

Genes under expressed in the melanoma group (as defined in Section 4). The genes in bold face correspond to those that their average expression in the group is below 1.5 times the standard deviation on the whole dataset.

Gene ID	Protein	Information
R60169	GDA	Guanine deaminase
N74260	**AKR1C1**	**Aldo-keto reductase family 1, member C1**
H68500	**EST**	
T73242	AKR1C4	Aldo-keto reductase family 1, member C4
H14348	**EST**	
AA055808	**TACSTD1**	**Tumor-associated calcium signal transducer 1**
AA046815	**MAL2**	**Mal, T-cell differentiation protein 2**
W40286	**ANXA3**	**Annexin A3**
AA055664	**CDKN2A**	**Cyclin-dependent kinase inhibitor 2A (melanoma)**
AA029948	**LOC255743**	**hypothetical protein LOC255743**
AA036758	S100A4	S100 malignant transformation suppression1
AA056401	**DSP**	**Desmoplakin**
H73761	**LGP1**	**D11lgp1e-like**
W74492	**CLDN4**	**Claudin 4**
AA021558	EST	
N75339	**MAP7**	**Microtubule-associated protein 7**
H28438	SCNN1A	Sodium channel, nonvoltage-gated 1 alpha
H29546	NTSR1	Neurotensin receptor 1 (high affinity)
W90688	MAP7	Microtubule-associated protein 7)
AA053012	**DSP**	**Desmoplakin)**
N98225	**HOOK1**	**Hook homolog 1 (Drosophila)**
AA031287	**SPINT2**	**Serine protease inhibitor, Kunitz type, 2**
R05776	**LLGL2**	**Lethal giant larvae homolog 2 (Drosophila)**
AA035637	**JUP**	**Junction plakoglobin**
W90086	**FLJ22390**	**Hypothetical protein FLJ22390**
AA053218	**GRB7**	**Growth factor receptor-bound protein 7**
AA055668	**MRPL37**	**Mitochondrial ribosomal protein L37**
AA052978	**KRT8**	**Keratin 8**
N39570	EST	
AA037485	p30	Nuclear protein p30
AA054974	ABLIM1	Actin binding LIM protein 1
N30586	NEBL	Nebulette
R14348	MAP3K5	Apoptosis signal regulating kinase
W72586	**MDK**	**Midkine (neurite growth-promoting factor 2)**
N39598	**C11orf9**	**Chromosome 11 open reading frame 9**
N74639	**ACF**	**Apobec-1 complementation factor**
N29319	LIPG	Endothelial lipase precursor
H90431	**ADRB2**	**Adrenergic, beta-2-, receptor, surface**
W81425	CSRP3	Cysteine and glycine-rich protein 3 (cardiac LIM protein)
W92029	**EST**	
N62509	IL20RA	Interleukin 20 receptor, alpha
N64535	**AIG1**	**Androgen-induced 1**
AA029096	PRKCA	Protein kinase C, alpha
AA007361	pp9099	PH domain-containing protein
T77041	MGC45562	Hypothetical protein MGC45562
AA055661	TPD52L1	Tumor protein D52-like 1
AA046274	EST	
H62012	**CCL15**	**Chemokine (C-C motif) ligand 15**
R36703	**EST**	
R34833	**F3**	**Coagulation factor III (thromboplastin, tissue factor)**
H29272	**STYK1**	**Protein kinase STYK1**
AA054706	**EST**	
AA004583	**TFPI**	**Tissue factor pathway inhibitor**
J03037	CA2	Carbonic anhydrase II
AA026089	EGFR	Epidermal growth factor receptor
W40283	IL8	Interleukin 8
T77816	**CCL2**	**Chemokine (C-C motif) ligand 2**
R71338	EST	
H72506	**ANPEP**	**CD13 antigen**
AA002125	**API1**	**Apoptosis inhibitor 1**
N33794	AK3	Adenylate kinase 3
R16561	**API1**	**Apoptosis inhibitor 1**
N35886	**JUB**	**Jub, ajuba homolog (Xenopus laevis)**
T85905	AXL	AXL receptor tyrosine kinase
T84764	FBN1	Fibrillin 1 (Marfan syndrome)
N34799	FOSL2	FOS-like antigen 2
H8719	EST	
T60389	EST	
H24357	NRG1	Glial growth factor 2
AA040872	**CYP1B1**	**Cytochrome P450, family 1, subfamily B, polypeptide 1**
R66239	**PHLDB2**	**Pleckstrin homology-like domain, family B, member 2**
R51025	EML1	Echinoderm microtubule associated protein like 1
N98463	PLOD2	procollagen-lysine (lysine hydroxylase) 2
N71998	**ITGA3**	**Integrin, alpha 3 (antigen CD49C**
AA027942	**MATN2**	**Matrilin 2**
R52480	PAK3	p21 (CDKN1A)-activated kinase 3
W72569	NUDT1	nudix (Nucleoside diphosphate linked moiety X)-type motif 1
AA056022	CSPG2	chondroitin sulfate proteoglycan 2 (versican)
W72468	FAM13A1	Family with sequence similarity 13, member A1
H16591	VCAM1	Vascular cell adhesion molecule 1
H14976	EST	
AA043311	DPYSL3	Dihydropyrimidinase-like 3
AA046572	SERPINE1	Plasminogen activator inhibitor type 1
N50928	SYT6	Synaptotagmin VI
N47888	**DNER**	**Delta-notch-like EGF repeat-containing transmembrane**
AA054564	COL4A1	Collagen, type IV, alpha 1
W48793	CDH2	Cadherin 2, type 1, N-cadherin (neuronal)
T66144	EST	
R21876	EST	
AA017445	TFPI2	Tissue factor pathway inhibitor 2
N20008	PLCB4	Phospholipase C, beta 4
AA046069	FSTL1	Follistatin-like 1
AA004839	**NNMT**	**Nicotinamide N-methyltransferase**
AA040161	**PLK2**	**Polo-like kinase 2 (Drosophila)**
AA018579	**GUCY1B3**	**Guanylate cyclase 1, soluble, beta 3**
H08669	**SPOCK**	**Sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican)**
AA053251	**TMEPAI**	**Transmembrane, prostate androgen induced RNA**
R02280	**CSF1**	**Colony stimulating factor 1 (macrophage)**
H18456	**EST**	
N63138	**PRICKLE1**	**Prickle-like 1 (Drosophila)**
T65562	**CD24**	**CD24 antigen (small cell lung carcinoma cluster 4 antigen)**
AA045437		**Human transglutaminase mRNA**
AA040727	**PLAU**	**Plasminogen activator, urokinase**
W52295	**FGF2**	**Basic fibroblast growth factor**
AA043983	**TNFAIP2**	**Tumor necrosis factor, alpha-induced protein 2**
AA057835	**HIP-55**	**Src homology 3 domain-containing protein HIP-55**
H17799	**EST**	
N99930	**BDG29**	**BDG-29 protein**
AA043311	DPYSL3	Dihydropyrimidinase-like 3
N26801	AVPI1	Arginine vasopressin-induced 1
H11003	**EDN1**	**Endothelin 1**
W93567	**D2S448**	**Melanoma associated gene, p53-Responsive gene 2**
AA029313	**D2S448**	**Melanoma associated gene, p53-Responsive gene 2**
AA029129	**EFEMP1**	**EGF-containing fibulin-like extracellular matrix protein 1**
AA040442	**EFEMP1**	**EGF-containing fibulin-like extracellular matrix protein 1**
H15934	**ITGA6**	**Integrin, alpha 6**
AA031646	NDUFA5	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 5, 13kDa
AA057239	**MAP1B**	**Microtubule-associated protein 1B**
N72559	**RAB31**	**RAB31, member RAS oncogene family**
AA047819	**KIAA1789**	**KIAA1789 protein**
T47150	**MAP1B**	**Microtubule-associated protein 1B**
N20213	**MAP1B**	**Microtubule-associated protein 1B**
R21059	**NFKBIE**	**Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor**
AA045135	RAB31	RAB31, member RAS oncogene family
H65731	CDH13	Cadherin 13, H-cadherin (heart)
N69835	**PBX1**	**Pre-B-cell leukemia transcription factor 1**
AA031267	CNDP2	CNDP dipeptidase 2 (metallopeptidase M20 family)
AA055520	C1S	Complement component 1, s subcomponent
AA026597	EST	
AA046484	SLC16A2	Solute carrier family 16 (monocarboxylic acid transporters)
AA035018	ADAM12	A disintegrin and metalloproteinase domain 12 (meltrin alpha)
AA041382	C1R	Complement component 1, r subcomponent

Gene ID	Protein	Information
AA004204	COL5A2	Collagen, type V, alpha 2
AA029242	EST	
H47744	**PBX1**	**Pre-B-cell leukemia transcription factor 1**
H28104	THY1	Cell surface antigen
W95604	SLC1A3	Solute carrier family 1
AA035639	SET7	SET domain-containing protein 7
W94080	MRPL34	Mitochondrial ribosomal protein L34
R48580	EST	
N94496	ELL2	Elongation factor, RNA polymerase II, 2
N70732	EDG2	Endothelial differentiation gene 2
H04749	FLJ38507	Colon carcinoma related protein
W40153	IF	I factor (complement)
AA037699	LTBP1	Latent transforming growth factor beta binding protein 1
AA045303	**IFITM2**	**Interferon induced transmembrane protein 2 (1-8D)**
AA040523	ANXA1	Annexin A1
H18455	**AGPAT4**	**Lysophosphatidic acid acyltransferase, delta**
W84538	CXCL12	Chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
N63378	**PLAGL1**	**ZAC tumor supressor gene**
R17461	C6orf148	Chromosome 6 open reading frame 148
AA033932	**C20orf112**	**Chromosome 20 open reading frame 112**
N71869	RAFTLIN	Raft-linking protein
T61473	NOD27	Nucleotide-binding oligomerization domains 27
N93476	**EDG1**	**Endothelial differentiation, sphingolipid G-protein-coupled receptor, 1**
AA054556	RAB31	RAB31, member RAS oncogene family
AA046218	PRG1	Proteoglycan 1, secretory granule
R09913	**FADS2**	**Fatty acid desaturase 2**
AA047647	C5orf13	Chromosome 5 open reading frame 13
AA033975	RAC2	Ras-related C3 botulinum toxin substrate 2
R20579	**SOX1**	**SRY (sex determining region Y)-box 1**
H17425	**ITGB2**	**Integrin, beta 2, antigen CD18 (p95)**
AA046482	ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta
W70076	FABP5	Fatty acid binding protein 5 (psoriasis-associated)
AA005018	CGI-49	CGI-49 protein
R78402	FCGR2B	Fc fragment of IgG, low affinity IIb, receptor for (CD32)
W92100	EST	
W78928	GALC	Galactosylceramidase (Krabbe disease)
N41032	CAPG	Capping protein (actin filament), gelsolin-like
W86212	C6orf85	Chromosome 6 open reading frame 85
N52363	ATP11A	ATPase, Class VI, type 11A
W86859	CDH1	Cadherin 1, type 1, E-cadherin (epithelial)
W94793	SOX9	SRY (sex determining region Y)-box 9
AA047106	CAV1	Caveolin 1, caveolae protein, 22kDa