Bioinformatics, 31(7), 2015, 1102–1110
doi: 10.1093/bioinformatics/btu782
Advance Access Publication Date: 25 November 2014
Original Paper

OXFORD

Databases and ontologies

A bio-inspired computing model for ovarian carcinoma classification and oncogene detection

Meng-Hsiun Tsai[1,2], Mu-Yen Chen[3,*], Steve G. Huang[4], Yao-Ching Hung[5] and Hsin-Chieh Wang[1]

[1]Department of Management Information System and [2]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung City 402, Taiwan, [3]Department of Information Management, National Taichung University of Science and Technology, Taichung City 404, Taiwan, [4]Institute of Nanotechnology, National Chiao Tung University, Hsinchu City 300, Taiwan and [5]Department of Obstetrics and Gynecology, China Medical University and Hospital, Taichung City 404, Taiwan

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on September 15, 2014; revised on November 18, 2014; accepted on November 19, 2014

Abstract

Motivation: Ovarian cancer is the fifth leading cause of cancer deaths in women in the western world for 2013. In ovarian cancer, benign tumors turn malignant, but the point of transition is difficult to predict and diagnose. The 5-year survival rate of all types of ovarian cancer is 44%, but this can be improved to 92% if the cancer is found and treated before it spreads beyond the ovary. However, only 15% of all ovarian cancers are found at this early stage. Therefore, the ability to automatically identify and diagnose ovarian cancer precisely and efficiently as the tissue changes from benign to invasive is important for clinical treatment and for increasing the cure rate. This study proposes a new ovarian carcinoma classification model using two algorithms: a novel discretization of food sources for an artificial bee colony (DfABC), and a support vector machine (SVM). For the first time in the literature, oncogene detection using this method is also investigated.

Results: A novel bio-inspired computing model and hybrid algorithms combining DfABC and SVM was applied to ovarian carcinoma and oncogene classification. This study used the human ovarian cDNA expression database to collect 41 patient samples and 9600 genes in each pathological stage. Feature selection methods were used to detect and extract 15 notable oncogenes. We then used the DfABC-SVM model to examine these 15 oncogenes, dividing them into eight different classifications according to their gene expressions of various pathological stages. The average accuracyof the eight classification experiments was 94.76%. This research also found some oncogenes that had not been discovered or indicated in previous scientific studies. The main contribution of this research is the proof that these newly discovered oncogenes are highly related to ovarian or other cancers.

Availability and implementation: http://mht.mis.nchu.edu.tw/moodle/course/view.php?id=7

Contact: mychen@nutc.edu.tw

1 Introduction

Ovarian cancer is a common gynecological cancer which is the fifth leading cause of cancer deaths in women in the western world (Jemal *et al*., 2002). The overall 5-year survival rate for ovarian cancer is about 44%, a statistic that has remained unchanged in America for the past 25 years (Meng, 2014). Ovarian cancer patients generally go through three major pathological stages: benign, borderline and invasive. For those whose ovarian cancer is diagnosed in the early pathological stage, the survival rate is higher. However, the transition from a symptomatic benign ovarian tumor (OVT) to a cancerous malignancy is difficult to diagnose and estimate. CA125 is commonly used for ovarian cancer screening, but its specificity or sensitivity is insufficient for early detection (Nossov *et al*., 2008). Many recent scientific studies have noted that gene expressions are highly related to the identification of ovarian cancer stages (Roel *et al*., 2013; Schwede *et al*., 2013). In these studies, the microarray has been used extensively and shown to be a promising tool. Microarray analysis has revealed specific genes, called oncogenes, which biological experiments have apparently indicated are causally related to ovarian cancer. Based on this same analysis, pathologists have further observed the variations of gene expressions in these specific ovarian genes and compared the levels of gene expression in different pathological stages and conditions. Biological and statistical computing can be used to represent gene expressions for such comparisons, particularly during the progressive states as a tumor transitions from benign to invasive ovarian cancer tissue. Therefore, gene expression analysis has become increasingly important because of its excellent performance in both bioinformatics and biomedical studies of the human genome system.

However, gene expression analysis is a complex process and requires a large dataset in order for enough useful information to be extracted. Thus, computer-based analysis and classification of diseases can be helpful for diagnostics. The support vector machine (SVM) is one artificial intelligence method which has recently become popular, mainly because it can process classification and predication via different core parameters and is more capable of making predictions using different types of datasets (Chen, 2012, 2014). In addition, the SVM formalism is also a graphical and mathematical tool for the design, specification, simulation and verification of systems. In recent years, the bio-inspired algorithm has become increasingly useful as an optimization tool (Chen *et al*., 2013). By simulating biologically peculiar social behavior, such as mating or foraging, bio-inspired algorithms quickly search for the optimal solution and achieve the effect of stable convergence. Therefore, this study proposes to apply the parameter optimization method, i.e. a discretized food source for an artificial bee colony (DfABC) algorithm, to an SVM model to develop a hybrid DfABC-SVM classification model which achieves fast convergence and has good prediction capability when handling complex and varying types of ovarian cancer datasets. In addition, the proposed model is compared with those of several pioneering studies regarding ovarian carcinoma prediction.

2 Related work

The microarray is one of the most important tools for studying biological expression. It is a novel biological technique for gathering large numbers of gene expression from cell or tissue samples. A microarray can help cancer researchers quickly identify genes that are differentially regulated during cancerous progression, and discover related oncogenes from diagnosed cancers and other diseases (Slonim and Yanai, 2009). Because of the large number of gene expressions in a microarray, dependence on artificial data not only causes imprecise analysis of the results but also requires that a great deal of time and money be spent on calculation and classification. To solve these problems, recent scholars have used statistical, clustering and classification methods to develop algorithms that apply to microarray analysis (Gopalakrishnan *et al*., 2010).

Recent studies have developed new feature extraction and classification methods to detect various oncogenes by mining the microarray data of ovarian cancer. Shah and Kusiak (2007) integrated the genetic algorithm (GA) and SVM to assist in the prediction of ovarian, prostate and lung cancers. The experimental results showed that the hybrid SVM model had better prediction capabilities than the traditional decision tree classification method. Lee (2008) developed a fuzzy model with SVM to predict ovarian cancer. The experimental results illustrated that the accuracy of the hybrid SVM model surpassed that of both regression analysis and the conventional SVM model. Later, Lai *et al*. (2009) adopted the SVM technique to develop a bio-statistical system for gene expression analysis, and found 21 notable oncogenes. In addition, the average accuracy of the proposed model is 89% with a cross-validation approach. Using a different strategy, Tsai *et al*. (2011) integrated an artificial neural network and *t*-tests to select more important oncogenes and enhance the accuracy rate of ovarian cancer prediction. Tung *et al*. (2012) applied a hybrid neural fuzzy inference system to the diagnosis of ovarian cancer. The experimental results showed that the hybrid system performed better than both the conventional decision tree and the neural network model. Therefore, the adoption of SVM and a hybrid system approach to the classification of ovarian cancer is straightforward in this research.

3 Research methodologies

3.1 Support vector machine

SVM has recently become a popular artificial intelligence method, mainly because SVM can process classification and predication through different kernel functions and suitable parameters. It is also more capable of predicting different types of datasets. SVM was first proposed by Vapnik and Cortes in 1995, and has been applied and improved upon by many scholars in different studies. Chaudhuri and De (2011) proposed integrating fuzzy set theory into the SVM model because better prediction results could be obtained via fuzzy processing of the parameters. Many scholars have also used SVM to work with other artificial intelligence models to improve its accuracy (Sue and Li, 2012).

In the past, the methods for modifying, adjusting and setting SVM core parameters have been time-consuming and have often not been objective. In addition, successful parameter settings cannot be generalized to apply to different types of datasets. Therefore, this research proposes integrating the DfABC parameter optimization method into the SVM model. The goal is to develop a DfABC-SVM classification model which results in fast convergence and accurate predictions, and is capable of handling different types of datasets.

3.2 Artificial bee colony algorithm

The Artificial Bee Colony algorithm (ABC) proposed by Karaboga and Basturk in 2007 is a popular and widely used bio-inspired algorithm derived from observations of the foraging behavior of bees. During the foraging stage, bees are classified into one of three types: Employed, Onlooker or Scout. Each type has a different task and function. The ABC algorithm simulates the pattern and defines half

of the bee group as Employed bees, and the other half as Onlooker bees. The food source represents the relative optimal solution of the desired parameter.

a. Employed bee

Among the different kinds of bees, Employed bees are responsible for going to the food source, returning to the hive, and dancing on a particular area to inform the Onlooker bees regarding the location, distance and nectar content of the food source. After one iteration, the Employed bees will find another new food source near the current food source and compare the fitness values of these two food sources. In the end, they will select the food source with largest fitness value. Equations (1) and (2) are the fitness value and calculation of the new food source, respectively:

$$\mathrm{fit}_i = \frac{1}{f_i + 1} \tag{1}$$

where fit_i is the fitness value of ith food source and $f_i = 1 -$ classification accuracy. The classification accuracy is the value generated by the prediction processed after the parameter combinations of ith food source are substituted into an SVM model. This shows that the food source fitness value of the first iteration is randomly generated.

$$V_{ij} = X_{ij} + \emptyset_{ij}(X_{ij} - X_{kj}) \tag{2}$$

where i represents the location of the current food source and V_{ij} represents the location of the new food source with the jth parameter solution. X_{ij} represents the location of the current food source with the jth parameter solution. $\emptyset_{ij}$ is a randomly generated value between $[-1,1]$. X_{kj} is another food source near to the current food source and is randomly generated with the jth parameter solution.

b. Onlooker bee

Onlooker bees decide which food source to target based on the food source information provided by the Employed bees. The decision is made using a probabilistic selection method; that is, the greater the value of the food source as calculated by probabilistic selection, the more likely it will be chosen by an Onlooker bee. The calculation of probabilistic selection is shown here as Equation (3).

$$\mathrm{prob}_i = 0.9 \times \frac{\mathrm{fit}_i}{\sum_{j=1}^{\mathrm{SN}} \mathrm{fit}_j} \tag{3}$$

where prob_i is the probability that the ith food source will be chosen, fit_j represents the fitness value of each food source, and fit_i represents the fitness value of the ith food source. SN represents the total number of food sources.

c. Scout bee

The purpose of the Scout bee is to avoid the possibility of bees falling into a local optimization. Hence, there is one restricting condition. The meaning of this restriction value is that, if the fitness value of a certain food source has not changed within the restricted number of iterations, it will be considered as falling into local optimization. When that occurs, the Employed bee will give up this food source and shift its role to Scout bee to search for a new food source. Equations (4) and (5) are the calculations for the restriction value and new food source, respectively:

$$\mathrm{Limit} = \mathrm{SD} \times D \tag{4}$$

where SN is the total number of food sources; D is the dimension of the problem to be solved.

$$X_i^j = X_{\min}^j + \mathrm{rand}(0,1)(X_{\max}^j - X_{\min}^j) \tag{5}$$

where j is the number of parameters to be optimized [1,2,……, j]; X_i^j is the value of jth parameter of ith food source; $X_{\min}^j$ and $X_{\max}^j$ are the lower and upper limits of parameter j.

3.3 Discretization of food sources for an ABC

Although ABC is already a stable bio-inspired algorithm, improvements to several aspects of it will significantly increase its parameter optimization effectiveness. Li *et al.* (2012) tried to improve the global optimization of the Employed bee and Onlooker bee in the ABC algorithm and obtained a good result. Kashan *et al.* (2012) also proposed an improved ABC algorithm that is binary-based, overcoming the original ABC algorithm's ability to process only continuous numerical optimizations. In the probabilistic choice method of the original ABC algorithm, the fitness values of food sources are easily distributed diversely. Such a distribution greatly reduced ABC's divergence speed, especially in conditions when several identical excellent food sources exist. If, in the first iteration, the calculated fitness values of five food sources are within 1% of each other, directly applying Equation (1) to calculate the second iteration fitness value will obtain values that are quite close. Under such conditions, the food sources will tend to clump. Thus, the number of iterations increases, affecting the convergence speed of the probabilistic choice stage.

To solve the above problems, this research proposes a new algorithm: a DfABC. The issue of food source fitness values clumping in certain areas can be solved by classifying the fitness values. Moreover, this algorithm provides a new way of calculating a probabilistic choice. Food sources that are classified into different areas will be calculated by different probabilistic choice methods for selection. The classification and probabilistic choice methods are described as follows.

Step 1. Discrete distribution of food source fitness values

This step focuses on making the food source fitness values originally distributed at both ends more discrete, so that fewer food source fitness values are too close, thus resulting in a decrease in iterations. First, we sort the set of food source fitness values so that larger food sources are sorted to the front. Next, in accordance with the size of the food source fitness values, we provide one appropriate weight value w_i, $w_i \in W$ to each food source s_i, $s_i \in S$. If the food source fitness values are the same, the given weight values will also be the same. Finally, we multiply f_i and w_i to calculate the fitness value of the new food source dis_i, $\mathrm{dis}_i \in Dis$. Through this step, food source fitness values that are too close will be separated. Equation (6) is used to make the food source fitness values more discrete, as follows:

$$\mathrm{dis}_i = f_i \times w_i \tag{6}$$

where $\mathrm{dis}_i =$ the fitness value of the new food source while f_i is the fitness value of the original food source; is the w_i weight value that varies in accordance with the size of the fitness value of the original food source.

Step 2. Normalization of the fitness value of new food source dis_i

The ABC algorithm defines the fitness value of a food source as being between [0,1]; hence, in Step 2 the fitness value of a new food source is defined to be between [0,1] through normalization. This attempt uses the common statistical method of normalization

of the maximum/minimum value to adjust the fitness values of all food sources to be between [0,1] and save the adjusted value n_i to the N set, shown here as Equation (7).

$$n_i = \frac{\text{dis}_i - \text{Min}\{Dis\}}{\text{Max}[Dis] - \text{Min}\{Dis\}} \tag{7}$$

Step 3. Grouping the normalized fitness value of food source n_i

To ensure that an excellent food source is easily selected and to avoid Onlooker bees choosing a poor food source, this study processes interval cutting and grouping based on "good" and "bad" food source fitness values. The distribution of food sources is discretized, and the food source fitness values are distinguished as either "good" or "bad." Doing so can avoid the repeated selection of certain food sources, which, in turn avoids the dilemma of local optimization. Thus, Step 3 is further divided into two steps.

Step 3.1 Cutting the intervals to *m* parts

First of all, this research uses the equal-width discretization method to cut the interval [0,1] of the normalized food source fitness value ni, ni ∈ N to the *jm*th interval *jm* ∈ J. The interval distance of each interval is calculated using Equation (8), and the upper limit u_{jm} of each interval *jm* is calculated using Equation (9).

$$\text{Interval\_range} = \frac{\text{Max}\{N\} - \text{Min}\{N\}}{m} \tag{8}$$

herein Interval_range represents the interval distance (boundary range) between each group after cutting; m is the number of intervals of food sources; and max{*N*} and min{*N*} represents the maximum and minimum values of the N set of normalization of food source fitness values, respectively.

$$u_{jm} = m \times \text{Interval\_range} \tag{9}$$

wherein u_{jm} represents the upper boundary value of the interval distance of the *m*th interval; m is the number of intervals of food sources; and Interval_range represents the interval distance (boundary range) between each group after cutting.

Step 3.2 Distributing n_i to appropriate interval j_m

In this step, we distribute the normalized food source fitness value n_i to the appropriate interval in accordance with the upper boundary value of each interval. If n_i is smaller than the upper boundary value of the first interval, n_i belongs to the first interval. If n_i is greater than the upper boundary value of the first interval but smaller than the upper boundary value of the second interval, n_i belongs to the second interval. The rest may be deduced by analogy to distribute all n_i.

Step 4. Probabilistic calculation

This step is to apply the grouped, normalized food source fitness value n_i to different probabilistic choice methods to calculate the probability that a given n_i will be chosen. First, we determine the best interval u_{j1} of food source fitness values and provide it with the best probabilistic choice method. Next, we select a few groups (*g* groups) of good food source fitness values and apply a suitable probabilistic choice method. The remaining food source fitness values will be discarded.

Step 5. Obtain the better probabilistic set

The above steps generate one set of better probabilistic choice values of food sources, which is to be used for the next iteration so that better food sources can be quickly obtained. This increases convergence speed and avoids local optimization. By repeatedly running the above steps until we reach the maximum number of iterations, we will find the optimal parameter combination for this model.

4 Experimental design and results

4.1 Dataset

A cDNA microarray is a kind of biochip, which includes fluorescence spots to represent the expressions of many known genes. Each labeled fluorescence spot is denoted as a gene and is obtained by hybridizing the test and reference sample labeled with Cy3 and Cy5, with the gene expressions represented as the ratio of Cy3 and Cy5 fluorescence intensity. In this article, all gene expressions have been digitalized as numerical data from cDNA microarrays. The ovarian cDNA expression database used in this article was provided by the surgical and pathological units at China Medical University Hospital, Taichung, Taiwan. All microarrary procedures were performed in a dust/climate-controlled laboratory at the China Medical University. The gene expression database contains 9600 gene expressions obtained from the ovarian tissues of 41 patients at different pathology stages and collected over 2001–2003, defined by Federation of Gynecology and Obstetrics (FIGO) classifications as benign OVT (13 samples), borderline tumors (BOT—6 samples), ovarian cancer at stage I (OVCA-I—7 samples) and ovarian cancer at stage III (OVCA-III—15 samples). Despite the small sample population, 9600 genes/chip were obtained for each sample, thus providing sufficient data for analysis. A sequence-verified human cDNA library containing 9600 human cDNA clones was provided by the National Health Research Institute of Taiwan. The clones were originally obtained from the Minimum Information About a Microarray Experiment (MIAME) consortium libraries through its distributor (Research Genetics, Huntsville, AL; Brazma *et al.*, 2001). In addition, given the high cost of the biochips, the results can be considered as representative. Table 1 summarizes the detailed dataset of ovarian carcinoma, and the appendices provide partial datasets. Appendix A represents the gene expressions from the 1st to 27th genes of these 41 patients, with rows representing individual patients and columns representing each gene expression. Appendix B represents the gene expressions from the 9575th to 9600th genes of these 41 patients.

4.2 Feature extraction

This study's experimental gene expression database includes many samples at OVT, BOT, OVCA-I and OVCA-III stages. Based on the distributions of expressions at different cancer stages, this article divides the 15 genes into 8 different classifications, each with a specific goal:

1. Classification I (C_I): to predict each stage throughout the entire dataset.

Table 1. Type of heartbeat

Stage	Encoding	Samples
OVT	ovt1, ovt2, ovt6, ovt9, ovt10, ovt15, ovt2, ovt22, ovt40, ovt38, ovt41, ovt43, ovt44	13
BOT	bot3, bot4, bot5, bot6, bot7, bot8	6
OVCA-I	1-ovca1, 1-ovca7, 1-ovca11, 1-ovca12, 1-ovca20, 1-ovca22, 1-ovca23	7
OVCA-III	3-ovca2, 3-ovca3, 3-ovca4, 3-ovca5, 3-ovca6, 3-ovca9, 3-ovca10, 3-ovca13, 3-ovca14, 3-ovca15, 3-ovca16, 3-ovca17, 3-ovca18, 3-ovca19, 3-ovca21	15

2. Classification II (C_{II}): to predict only the OVT stage throughout the entire dataset.
3. Classification III (C_{III}): to predict only the BOT stage throughout the entire dataset.
4. Classification IV (C_{IV}): to predict only the OVCA-I stage throughout the entire dataset.
5. Classification V (C_{V}): to predict only the OVCA-III stage throughout the entire dataset.
6. Classification VI (C_{VI}): to predict the OVT and BOT stages of ovarian carcinoma, as opposed to the other stages (OVCA-I, OVCA-III), throughout the entire dataset. We believe it is very important to investigate which key gene will influence the ovarian carcinoma to progress from the OVT stage to the BOT stage.
7. Classification VII (C_{VII}): to predict the BOT and OVCA-I stages of ovarian carcinoma, as opposed to the other stages (OVT, OVCA-III), throughout the entire dataset. We believe it is very important to investigate which key gene will influence the ovarian carcinoma to progress from the BOT stage to the OVCA-I stage.
8. Classification VIII (C_{VIII}): to predict the BOT and OVCA-III stages of ovarian carcinoma, as opposed to the other stages (OVT, OVCA-I), throughout the entire dataset.

Selected features are important in enhancing the performance of ovarian carcinoma classification and oncogene detection. After applying the feature selection methods [ID3, J48, Information Gain (IG)], this study obtained the significant features listed in Table 2 for the input variables of the proposed DfABC-SVM model.

Table 2. Features selection and gene description

Method	Classification	Gene ID	Gene Symbols	References
ID3	II	6773	STAT2	Liang *et al.* (2012)
				Uluer *et al.* (2012)
ID3	II	2782	GNB1	Wazir *et al.* (2013)
				Fridley *et al.* (2014)
J48	II	166793	ZNF509	Siggs and Beutler (2012)
J48	III	9397	NMT2	Selvakumar *et al.* (2007)
ID3	IV	23509	POFUT1	Hu *et al.* (2012)
				Li *et al.* (2013)
ID3	IV	2782	GNB1	Wazir *et al.* (2013)
				Fridley *et al.* (2014)
IG	V	5069	PAPPA	Boldt and Conover (2011)
J48				Pan *et al.* (2012)
ID3				
J48	VI	1278	COL1A2	Loss *et al.* (2010)
J48	VI	6387	CXCL12	Obermajer *et al.* (2011)
				Popple *et al.* (2012)
ID3	VI	57569	ARHGAP20	Herold *et al.* (2011)
ID3	VI	5209	PFKFB3	Calvo *et al.* (2006)
				Clem *et al.* (2013)
ID3	VI	6773	STAT2	Liang *et al.* (2012)
				Uluer *et al.* (2012)
J48	VII	596	BCL2	Zhang *et al.* (2011a, b)
				Zhao *et al.* (2012)
				Shirali *et al.* (2013)
ID3	VII	6773	STAT2	Liang *et al.* (2012)
				Uluer *et al.* (2012)
ID3	VII	2782	GNB1	Wazir *et al.* (2013)
				Fridley *et al.* (2014)
J48	VII	1660	DHX9	Zhang *et al.* (2011a, b)
				Jain *et al.* (2013)
IG	VIII	896	CCND3	Beltran *et al.* (2012)
J48				Schmitz *et al.* (2014)
ID3				
IG	VIII	4087	SMAD2	Liu *et al.* (2012)
				Fleming *et al.* (2013)
IG	VIII	8898	MTMR2	Luigetti *et al.* (2013)
ID3	VIII	6773	STAT2	Liang *et al.* (2012)
				Uluer *et al.* (2012)

Note: Signal transducers and activators of transcription 2 (STAT2), Guanine nucleotide binding protein beta polypeptide 1 (GNB1), Zinc finger protein 509 (ZNF509), N-myristoyltransferase 2 (NMT2), Protein O-fucosyltransferase 1 (POFUT1), Pregnancy-associated plasma protein A (PAPPA), Collagen, type I, alpha 2 (COL1A2), Stromal cell-derived factor 1 (CXCL12), Rho GTPase activating protein 20 (ARHGAP20), 6-Phosphofructo-2-kinase/fructose-2,6-biphosphatase 3 (PFKFB3), B-Cell CLL/Lymphoma 2 (BCL2), DEAH (Asp-Glu-Ala-His) box helicase 9 (DHX9), Cyclin D3 (CCND3), SMAD Family Member 2 (SMAD2), Myotubularin-related protein 2 (MTMR2).

4.3 Comparative study

The proposed DfABC-SVM model and other AI approaches were developed in MATLAB R2009 on a PC (Intel I7 six-core CPU, 4 G RAM, Windows 7 OS). Throughout the initial experiment, the parameter values used in the proposed DfABC-SVM were set as follows. The ratio of the training and testing datasets of the model was 7:3. Furthermore, the clusters (m) of the DfABC-SVM was set as 5 of these 8 classifications. The number of iterations was set to 500 rounds, and it can also finish the training if the fitness reach the convergence.

We constructed and tested our proposed DfABC-SVM model. The test results are summarized in Tables 3–5. These three tables summarize the experimental results of a conventional SVM model and three hybrid bio-inspired computing methods for comparison: the GA-SVM (Genetic Algorithm), PSO-SVM (Particle Swarm Optimization) and ABC-SVM (Artificial Bee Colony) models. Our proposed DfABC-SVM model was shown to successfully predict ovarian cancers: the accuracy rates for different classifications were all higher than 83.3%. The prediction performance with training and testing dataset are illustrated in Figures 1 and 2.

In Classification I, the proposed DfABC-SVM model performed better than all other competitors, with a prediction accuracy rate of 91.6% when classifying ovarian cancer from the entire dataset. From Classification II to Classification V, the proposed DfABC-SVM model also outperformed all competitors when classifying ovarian cancer for each stage (from OVT to OVCA-III), with an average prediction rate higher than 83.3%. In addition, the proposed model outperformed its competitors in regards to the ID3 feature selection method. From Classification VI to Classification VIII, the goal was to investigate prediction accuracy within the hybrid stages, and it was also important to know which gene would be the key factor influencing the development of ovarian cancer. As the experimental results show, the prediction accuracy of the proposed model was greater than that of the other models, especially with the IG feature selection method. Therefore, from these results we conclude that the proposed DfABC-SVM model is a good reference model for predicting ovarian cancer.

This research presents several key findings regarding the implications and determinants of ovarian cancer prediction:

1. Our approach only requires 15 genes in eight different classifications fewer than other methods but still obtains highly-accurate ovarian cancer predictions.
2. Experimental results show our proposed bio-inspired computing approach has a high average accuracy rate of 94.76% for

Table 3. Performance measurements with IG

Ovarian Dataset	C_I	C_{II}	C_{III}	C_{IV}	C_V	C_{VI}	C_{VII}	C_{VIII}
SVM	58.3	75	83.3	66.7	75	58.3	66.7	83.3
GA-SVM	66.7	83.3	83.3	75	83.3	66.7	75	83.3
PSO-SVM	66.7	83.3	91.6	75	83.3	66.7	75	91.6
ABC-SVM	75	83.3	91.6	83.3	91.6	66.7	83.3	91.6
DfABC-SVM	91.6	91.6	100	83.3	100	91.6	100	100

Table 4. Performance measurement with C4.5

Ovarian Dataset	C_I	C_{II}	C_{III}	C_{IV}	C_V	C_{VI}	C_{VII}	C_{VIII}
SVM	66.7	75	83.3	58.3	83.3	66.7	75	83.3
GA-SVM	83.3	83.3	91.6	66.7	91.6	75	75	83.3
PSO-SVM	83.3	75	83.3	58.3	83.3	83.3	75	91.6
ABC-SVM	83.3	83.3	91.6	75	91.6	91.6	83.3	91.6
DfABC-SVM	91.6	91.6	100	83.3	100	91.6	91.6	100

Table 5. Performance measurement with ID3

Ovarian Dataset	C_I	C_{II}	C_{III}	C_{IV}	C_V	C_{VI}	C_{VII}	C_{VIII}
SVM	66.7	83.3	83.3	75	83.3	58.3	66.7	83.3
GA-SVM	66.7	91.6	91.6	83.3	83.3	66.7	66.7	83.3
PSO-SVM	75	83.3	91.6	75	91.6	66.7	83.3	91.6
ABC-SVM	83.3	91.6	100	83.3	91.6	58.3	83.3	91.6
DfABC-SVM	91.6	100	100	91.6	100	83.3	91.6	100

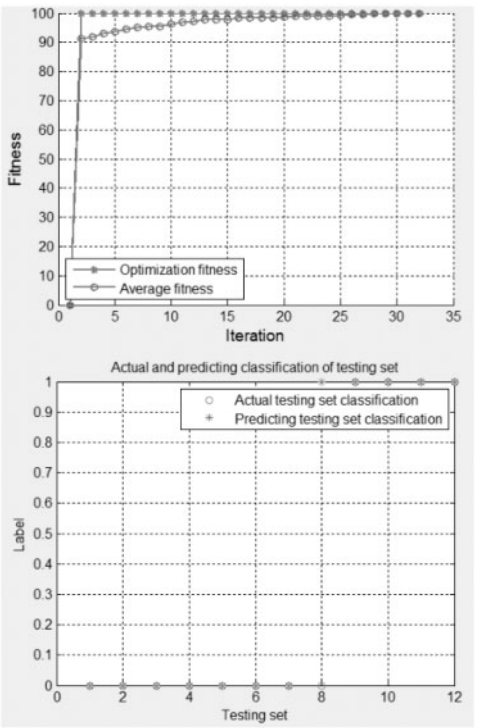

Fig. 1. The training and testing performance with DfABC for C_I classification

predicted target oncogenes in 8 classifications, and achieves a 100% accuracy rate for C_{III}, C_V, and C_{VIII} classifications.

3. This research presents the first indication that two oncogenes (ZNF509 and MTMR2) are associated with ovarian cancer.

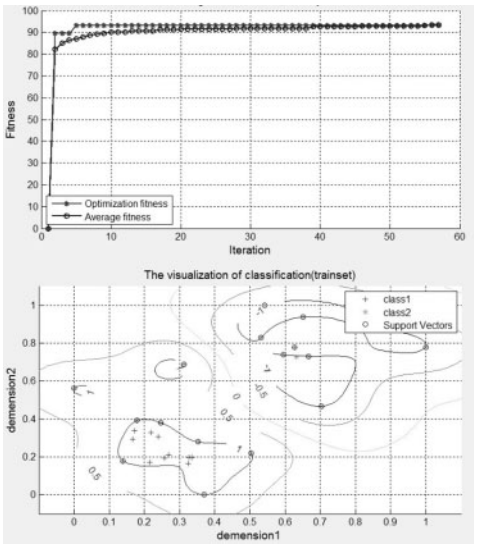

Fig. 2. The training and testing performance with DfABC for C_{III} classification

4. This study empirically demonstrated that a bio-inspired computing approach achieved better prediction accuracy than other evolutionary approaches, such as GA or PSO. Furthermore, bio-inspired computing approaches have rarely been used for prediction, especially in the biomedical domain or for ovarian cancer.

5 Genes discussion and implications

As shown in Table 2, the average expression of gene STAT2 is remarkable at Classifications II, VI, VII and VIII. Classification II is the OVT stage. Classification VII includes the BOT and OVCA-1 stages of ovarian carcinoma, as opposed to other stages (OVT, OVCA-III). STAT2 is a member of the STAT family which is involved in cytoplasmic transcriptional activators and signal transducers. It is activated to regulate gene expression in response to cytokines, interferons and growth factors. After phosphorylation by the receptor associated kinases, activated STAT2 forms a heterodimer with Stat1 and translocates to the nucleus for regulating the transcription of many different genes. According to relevant biological studies, STAT2 is also significantly related to colorectal, skin, cervical and breast cancers (Liang *et al.*, 2012; Uluer *et al.*, 2012).

The average expression of gene GNB1 is especially significant at Classifications II, IV and VII. Classification IV is specifically the OVCA-I stage of ovarian carcinoma. GNB1 is a subunit of the heterotrimeric G-protein and integrates signals between receptors and effector proteins such as modulators or transducers in various transmembrane signaling systems. It is a key player in the membrane-mediated signal transduction cascade and has been recently found to be associated with breast cancer pathology measurements and ovarian cancer outcomes (Fridley *et al.*, 2014; Wazir *et al.*, 2013).

The average expression of gene ZNF509 is especially significant at Classification II. Localized to the nucleus, ZNF509 is a Krüppel C_2H_2-type zinc finger protein that contains a DNA-binding domain and may play a role in transcriptional regulation (Siggs and Beutler, 2012).This study marks the first time that ZNF509 has been shown to have any connection to cancer.

The average expression of gene NMT2 is especially significant at Classification IIII, i.e. the BOT stage of ovarian carcinoma. Cloned

from a human liver library, NMT2 catalyzes the myristoylation of proteins and diversifies their biological functions in signal transduction and oncogenesis (Selvakumar *et al*., 2007). NMT2 activity and protein expressions are higher in human colorectal cancer, gallbladder carcinoma and brain tumors (Selvakumar *et al*., 2007). Attenuation of NMT2 activity may prove to be a novel therapeutic protocol for cancer.

The average expression of gene POFUT1 is especially significant at Classification IV. POFUT1 adds O-fucose through an O-glycosidic linkage to conserved serine or threonine residues in EGF domains, and plays a critical role in the Notch signaling pathway which regulates cell growth, differentiation and cancer metastasis (Hu *et al*., 2012; Li *et al*., 2013). Recent studies have shown that POFUT1 is significantly upregulated in oral cancer (Yokota *et al*., 2013).

The average expression of gene PAPPA is especially significant at Classification VI, i.e. the OVCA-III stage of ovarian carcinoma. PAPPA is a plasma metalloprotease involved in regulating the activity of the insulin-like growth factor (IGF) signal pathway via the cleavage of IGF binding proteins (IGFBPs). Recent reports reveal that overexpression of PAPPA could promote ovarian and lung cancer cell tumorigenicity *in vivo*, thereby establishing a novel tumor growth-promoting role for PAPPA (Boldt and Conover, 2011; Pan *et al*., 2012).

The average expression of the gene COL1A2 is especially significant at Classification VI, i.e. the OVT and BOT stages. COL1A2 is the fibrillar collagen found in most connective tissues. It can strengthen and support many tissues in the body, such as cartilage, bone, tendon and skin. Type I collagen is the most abundant form in the human body. Several mutations in this gene are associated with multiple diseases. It also plays an important role in tumor development. Some studies have related COL1A2 to colorectal, hepatocellular and breast carcinomas (Loss *et al*., 2010).

The average expression for gene CXCL12 is also especially remarkable at Classification VI. The stromal cell-derived factor 1, known as C-X-C motif chemokine 12 (CXCL12), is a chemokine protein and plays a role in many diverse cellular functions such as tumorgenesis, metastasis, inflammation response and angiogenesis. Multiple transcript variants encode different isoforms for this gene. CXCL12 has recently been shown to be related to breast, ovarian, colon, pancreatic and prostate cancers (Obermajer *et al*., 2011; Popple *et al*., 2012).

The average expression of gene ARHGAP20 is also especially significant at Classification VI. Expressed primarily in the brain, ARHGAP20 belongs to the Rho protein family and is involved in the regulation of Rho-type GTPases, an enzyme that catalyzes the hydrolysis of GTP. It is a direct downstream target of Rap1 and induces inactivation of Rho, resulting in neurite outgrowth (Yamada *et al*., 2005). ARHGAP20 has recently been reported as a tumor suppressor gene in human breast cancer and B-cell chronic lymphocytic leukemia (Herold *et al*., 2011).

The average expression of gene PFKFB3 is also especially remarkable at Classification VI. PFKFB3 has the highest kinase/phosphatase activity ratio of all PFK-2 isoforms, consistent with its role as a powerful activator of glycolysis. It is known to be activated by mitogenic, hypoxic and inflammatory stimuli. Importantly, PFKFB3 protein expression is elevated in most tumor types. Silencing of PFKFB3 in tumor cells reduces tumor cell glycolysis and growth (Calvo *et al*., 2006). In particular, high expression of the PFKFB3 protein product was noted in lung, breast, colon, prostatic, pancreatic and ovarian adenocarcinomas (Clem *et al*., 2013).

The average expression of gene Bcl2 is especially significant at Classification VII, which includes the BOT and OVCA-1 stages of ovarian carcinoma. An integral outer mitochondrial membrane protein, Bcl-2 is the founding member of the Bcl-2 family, and regulates programmed cell death or apoptosis. Bcl-2 family members have been widely implicated in cancer. Bcl-2 is often classified as an oncogene because of its specific anti-apoptotic property by preventing the activation of the caspases. The concept that impaired apoptosis is central to tumor development is now widely embraced. Some studies, have related BCL2 to ovarian, prostate, colorectal, liver and lung cancers (Shirali *et al*., 2013; Zhang *et al*., 2011a, b; Zhao *et al*., 2012).

The average expression of gene DHX9 is especially remarkable at Classification VII. DHX9 belongs to the DEXH family of helicase superfamily 2, and performs critical functions for unwinding double-stranded DNA and RNA driven by the hydrolysis of ATP in a 3′ to 5′ direction. It also serves as a transcriptional activator and plays a critical role in RNA metabolism. Recent evidence has suggested that DHX9 plays an important role in the regulation of immune responses (Zhang *et al*., 2011a, b). Furthermore, some studies have reported that dysregulation of DHX9 expression is related to lung and breast cancers (Jain *et al*., 2013).

The average expression of gene CCND3 is especially significant at Classification VIII, i.e. the BOT and OVCA-III stages of ovarian carcinoma. CCND3 is a member of the cyclin protein family involved in regulating cell cycle progression and shows the most broad expression pattern of the D-type cyclins. Its activity is required for the regulation of the G1/S phase transition of the cell cycle. Furthermore, high levels of CCND3 have been related to a poor prognosis for several types of cancer, including melanoma, lymphoma and bladder carcinomas (Beltran *et al*., 2012; Schmitz *et al*., 2014).

The average expression for gene SMAD2 is also especially significant at Classification VIII. SMAD2, known as Smad family member 2, has been found to be activated by TGF-b. It forms heteromeric complexes with SMAD4 and regulates multiple signaling pathways, such as cell proliferation, differentiation and apoptosis.SMAD2 has recently been reported to be associated with gastric, colorectal and lung cancers (Fleming *et al*., 2013; Liu *et al*., 2012).

The average expression of gene MTMR2 is also especially remarkable at Classification VIII. MTMR2 is a protein tyrosine phosphatase which removes phosphate groups from substrates and modifies their activity. Mutations in the gene MTMR2 are responsible for severe autosomal recessive peripheral neuropathies, such as Charcot-Marie-Tooth disease type 4B1 (CMT4B1; Luigetti *et al*., 2013). This study marks the first time that MTMR2 has been shown to have any connection to cancer.

6 Conclusion

In this study, a novel bio-inspired computing model and hybrid algorithms of DfABC and SVM were applied to the classification of ovarian carcinoma and the prediction of oncogenes. The average accuracy of predicted target oncogenes in eight classifications was over 94.76%. According to the literature, 13 oncogenes are reportedly related to various cancers, and five genes (PAPPA, CXCL12, PFKFB3, GNB1 and Bcl-2) are associated with ovarian cancer at different pathologic stages. In addition, two oncogenes (ZNF509 and MTMR2) are first time to associate with the cancer.

This research seeks to develop a new bio-inspired algorithm to quickly search for specific ovarian cancer biomarkers, which is typically a laborious process, and the resulting specificity and sensitivity is typically insufficient for the early detection of ovarian cancer. Improving the survival probability for ovarian cancer patients requires the precise location of efficient tumor markers as tumors transition from benign to invasive. Thus, future work will focus on confirming the expressions of these genes using experimental methods such as real-time PCR, western blotting, and immunohistochemistry. We believe that our discovery of new stage-specific genes can allow for the development of a rapid screening platform for oncogenes and thus improve early diagnosis of various types of cancer.

Acknowledgements

The authors thank the Ministry of Science and Technology of Taiwan [MOST-103-2410-H-025-022-MY2] and National Science Council of Taiwan [NSC-102-2622-E-005-015-CC3 and NSC-102-2622-E-005-011-CC3] which supported part of this research.

References

Beltran,A.J. *et al.* (2012) Fluorescence in situ hybridization analysis of CCND3 gene as marker of progression in bladder carcinoma. *J. Biol. Regul. Homeostat. Agents* **27**, 559–567.

Boldt,H.B. and Conover,C.A. (2011) Overexpression of pregnancy-associated plasma protein-A in ovarian cancer cells promotes tumor growth in vivo. *Endocrinology* **152**, 1470–1478.

Brazma,A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371.

Calvo,M.N. *et al.* (2006) PFKFB3 gene silencing decreases glycolysis, induces cell-cycle delay and inhibits anchorage-independent growth in HeLa cells. *FEBS Lett.* **580**, 3308–3314.

Chaudhuri,A., and De,K. (2011) Fuzzy support vector machine for bankruptcy prediction. *Appl. Soft Comput.* **11**, 2472–2486.

Chen,M.Y. (2012) Visualization and dynamic evaluation model of corporate financial structure with self-organizing map and support vector regression. *Appl. Soft Comput.* **12**, 2274–2288.

Chen,M.Y. (2014) Using a hybrid evolution approach to forecast financial failures for Taiwan listed companies. *Quant. Finance* **14**, 1047–1058.

Chen,M.Y. *et al.* (2013) Credit rating analysis with support vector machines and artificial bee colony algorithm. In: *The 26th International Conference on Industrial, Engineering & other Applications of Applied Intelligent Systems* (IEA/AIE 2013), 17–21 June, Amsterdam, Netherlands. Lecture Notes in Artificial Intelligence (LNAI), 7906, 528–534.

Clem,B.F. *et al.* (2013) Targeting 6-phosphofructo-2-kinase (PFKFB3) as a therapeutic strategy against cancer. *Mol. Cancer Therap.* **12**, 1461–1470.

Fleming,N.I. *et al.* (2013) SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer Res.* **73**, 725–735.

Fridley,B.L. *et al.* (2014) Methylation of leukocyte DNA and ovarian cancer: relationships with disease status and outcome. *BMC Med. Genom.* **7**, 21.

Gopalakrishnan,V. *et al.* (2010) Bayesian rule learning for biomedical data mining. *Bioinformatics* **26**, 668–675.

Herold,T. *et al.* (2011) Expression analysis of genes located in the minimally deleted regions of 13q14 and 11q22-23 in chronic lymphocytic leukemia—unexpected expression pattern of the RHO GTPase activator ARHGAP20. *Genes Chromosomes Cancer* **50**, 546–558.

Hu,Y.Y. *et al.* (2012) Notch signaling pathway and cancer metastasis. Notch Signaling in Embryology and Cancer, Springer, US, pp. 186–198.

Jain,A.A. *et al.* (2013) DHX9 helicase is involved in preventing genomic instability induced by alternatively structured DNA in human cells. *Nucleic Acids Res.* **41**, 10345–10357.

Januchowski,R. *et al.* (2012) MDR gene expression analysis of six drug-resistant ovarian cancer cell lines. *BioMed Res. Int.* 2013. Article ID 241763, http://dx.doi.org/10.1155/2013/241763.

Jemal,A. *et al.* (2002) Cancer statistics, 2002. *CA Cancer J. Clin.* **52**, 23–47.

Karaboga,D. and Basturk,B. (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J. Global Optim.* **39**, 459–471.

Kashan,M.H. *et al.* (2012) DisABC: a new artificial bee colony algorithm for binary optimization. *Appl. Soft Comput.* **12**, 342–352.

Kumar,G. *et al.* (2013) Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework. *BMC Syst. Biol.* **7**(12), doi: 10.1186/1752-0509-7-12.

Lai,C.H. *et al.* (2009) Statistical and SVM-based oncogene detection of human cDNA expressions for ovarian carcinoma. *Int. J. Innovat. Comput. Inf. Control* **5**, 3157–3177.

Lee,Z.J. (2008) An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. *Artif. Intel. Med.* **42**, 81–93.

Li,G. *et al.* (2012) Development and investigation of efficient artificial bee colony algorithm for numerical function optimization. *Appl. Soft Comput.* **12**, 320–332.

Li,M. *et al.* (2013) Mutations in POFUT1, encoding protein O-fucosyltransferase 1, cause generalized Dowling-Degos disease. *Am. J. Hum. Genet.* **92**, 895–903.

Liang,Z. *et al.* (2012) Detection of STAT2 in early stage of cervical premalignancy and in cervical cancer. *Asian Pac. J. Trop. Med.* **5**, 738–742.

Liu,L.C. *et al.* (2012) EGCG inhibits transforming growth factor-β-mediated epithelial-to-mesenchymal transition via the inhibition of Smad2 and Erk1/2 signaling pathways in nonsmall cell lung cancer cells. *J. Agric. Food Chem.* **60**, 9863–9873.

Loss,L.A. *et al.* (2010) Prediction of epigenetically regulated genes in breast cancer cell lines. *BMC Bioinf.* **11**, 305.

Luigetti,M. *et al.* (2013) A novel homozygous mutation in the MTMR2 gene in two siblings with 'hypermyelinating neuropathy'. *J. Peripheral Nervous Syst.* **18**, 192–194.

Meng,Q.H. (2014) Genetic variants in the fibroblast growth factor pathway as potential markers of ovarian cancer risk, therapeutic response, and clinical outcome. *Clin. Chem.* **60**, 222–232.

Nossov,V. *et al.* (2008) The early detection of ovarian cancer: from traditional methods to proteomics. Can we really do better than serum CA-125?, *Am. J. Obstet. Gynecol.* **199**, 215–223.

Obermajer,N. *et al.* (2011) PGE2-induced CXCL12 production and CXCR4 expression controls the accumulation of human MDSCs in ovarian cancer environment. *Cancer Res.* **71**, 7463–7470.

Pan,H.S. *et al.* (2012) Protein secretion is required for pregnancy-associated plasma protein-A to promote lung cancer growth in vivo. *PloS one* **7**, e48799.

Popple,A. *et al.* (2012) The chemokine, CXCL12, is an independent predictor of poor survival in ovarian cancer. *Br. J. Cancer* **106**, 1306–1313.

Roel,G.W. *et al.* (2013) Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.* **123**, 517–525.

Schmitz,R. *et al.* (2014) Oncogenic mechanisms in Burkitt Lymphoma. *Cold Spring Harbor Persp. Med.* **4**, a014282.

Schwede,M. *et al.* (2013) Stem cell-like gene expression in ovarian cancer predicts type II subtype and prognosis. *Plos one* **8**, e57799.

Selvakumar,P. *et al.* (2007). Potential role of *N*-myristoyltransferase in cancer. *Prog. Lipid Res.* **46**, 1–36.

Shirali,S. *et al.* (2013) Adenosine induces cell cycle arrest and apoptosis via cyclinD1/Cdk4 and Bcl-2/Bax pathways in human ovarian cancer cell line OVCAR-3. *Tumor Biol.* **34**, 1085–1095.

Shah,S. and Kusiak,A. (2007) Cancer gene search with data-mining and genetic algorithms. *Comput. Biol. Med.* **37**, 251–261.

Siggs,O.M. and Beutler,B. (2012) The BTB-ZF transcription factors. *Cell Cycle* **11**, 3358–3369.

Slonim,D.K. and Yanai,I. (2009) Getting started in gene expression microarray analysis. *PLoS Comput. Biol.* **5**, e1000543.

Sue,J., and Li,H. (2012) Financial distress prediction using support vector machines: Ensemble vs. individual. *Appl. Soft Comput.* **12**, 2254–2265.

Tsai,M.H. *et al.* (2011) A statistical and learning based oncogene detection and classification scheme using human cDNA expressions for ovarian carcinoma. *Expert Syst. Appl.* **38**, 10066–10074.

Tung,S.W.T. *et al.* (2012) SoHyFIS-Yager: a self-organizing yager based hybrid neural fuzzy inference system. *Expert Syst. Appl.* **39**, 12759–12771.

Uluer,E.T. *et al.* (2012). Effects of 5-fluorouracil and gemcitabine on a breast cancer cell line (MCF-7) via the JAK/STAT pathway. *Acta Histochemica* **114**, 641–646.

Vapnik,V. and Cortes,C. (1995) Support-vector networks. *Mach. Learn.* **20**, 273–297.

Wazir,U. *et al.* (2013) Guanine nucleotide binding protein β 1: a novel transduction protein with a possible role in human breast cancer. *Cancer Genom. Proteom.* **10**, 69–73.

Yamada,T. *et al.* (2005) RA-RhoGAP, Rap-activated Rho GTPase-activating protein implicated in neurite outgrowth through Rho. *J. Biol. Chem.* **280**, 33026–33034.

Yokota,S. *et al.* (2013) Protein O-fucosyltransferase 1: a potential diagnostic marker and therapeutic target for human oral cancer. *Int. J. Oncol.* **43**, 1864–1870.

Zhang,H. *et al.* (2011a). MiR-148a promotes apoptosis by targeting Bcl-2 in colorectal cancer. *Cell Death Diff.* **18**, 1702–1710.

Zhang,Z. *et al.* (2011b) DHX9 pairs with IPS-1 to sense double-stranded RNA in myeloid dendritic cells. *J. Immunol.* **187**, 4501–4508.

Zhao,A. *et al.* (2012) MicroRNA-125b induces cancer cell apoptosis through suppression of Bcl-2 expression. *J. Genet. Genom.* **39**, 29–35.

	1 Phase	signal transducer and activator of transcription 2	nucleotide binding protein (G protein), beta	4 protein O-fucosyltransferase 1	5 ESTs	6 ESTs	7 6	8 Transcribed sequences	9 hypothetical protein 669	cyclin-dependent kinase-like 2 (CDC2-related	11 CDNA FLJ13276 fis, clone OVARC1001040	12 apolipoprotein A-II	translocated promoter region (to activated MET oncogene	14 D-aspartate oxidase	15 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	16 chromosome 21 open reading frame 5	17 chromosome 20 open reading frame 28	almodulin-dependent protein kinase (CaM	19 KIAA0984 protein	20 protease, serine, 22	21 Homo sapiens clone IMAGE 23915	22 G antigen 5	23 CG016	24 integrin, beta-like 1 (with EGF-like repeat domains)	basic, immunoglobulin-like variable motif containin	26 25	27 EST
1	BOT	2.055	1.166	0.995	0.1808	1.979	0.5634	1.878	0.1105	-0.2232	0.0873	0.02618	-0.1278	0.7689	-0.215	0.04572	0.5897	1.571	-0.6917	1.631	-0.5222	2.335	1.439	1.897	1.822	1.787	-0.5488
2	BOT	-1.497	-1.13	-0.8122	-0.8926	-1.477	-1.289	-0.7211	-2.77	-1.676	-0.7092	-2.102	-0.8805	-0.5614	-1.876	-0.5311	-0.5994	1.879	-1.05	0.6915	-2.028	0.2341	0.1761	0.3634	1.077	-0.2923	-2.445
3	BOT	-1.639	-0.2043	-0.04324	0.25	0.6699	-0.7356	1.479	-1.942	-1.573	-0.6558	-1.074	-1.543	-0.3343	-1.368	-0.5986	-0.1159	0.3746	-1.882	1.719	-3.305	1.602	0.9219	0.05254	0.9566	0.4658	-2.19
4	BOT	-4.281	-1.894	-1.866	-1.597	0.6718	-0.5997	0.6664	-1.09	-0.81	-0.07253	-2.442	-0.6805	0.3515	-0.2616	0.1364	0.5763	0.5553	0.2002	0.142	-1.356	-1.334	-0.4824	-0.9036	0.4374	1.083	0.519
5	BOT	1.429	0.914	0.2379	0.977	1.934	1.15	1.964	0.2585	-0.3166	-0.1286	0.7393	-0.8941	0.5533	0.1202	0.4527	0.5694	1.117	-0.5019	0.7059	0.356	1.234	1.025	0.398	1.03	1.853	0.1891
6	BOT	0.02253	-0.8965	0.4426	-0.7454	-1.148	0.1901	0.2473	-0.6661	-0.7522	-0.04148	-0.5864	-0.9894	0.3485	0.2236	-0.4698	1.195	0.6129	0.4325	0.4594	-0.07194	0.7355	1.171	0.4876	0.8686	0.4567	-1.932
7	OVT	-1.32	-1.2	0.65	0.39	0.78	-0.07	0.19	0.72	0.05	-0.14	-0.23	1.35	0.75	-0.21	-0.29	1.39	-0.26	0.21	-1.46	-0.01	0.72	0.21	-1.03	-1	-0.12	1.42
8	OVT	-1.79	-0.6	-0.56	0.12	-1.18	-0.57	-0.48	0.26	0.44	-0.96	-0.07	-0.05	-1.08	-1.17	-0.39	1.44	0.2	0.41	-0.19	-0.02	-1.15	-0.63	-1.43	-0.27	-0.43	2.39
9	OVT	-0.87	-1.06	-0.84	-0.38	-0.43	-0.62	-1.69	-1.43	-1.85	-0.51	-1.05	-0.22	-1.02	-0.51	0.87	0.25	1.04	-0.3	-0.72	0.32	0.63	0.29	-1.78	-1.12	0	-0.83
10	OVT	-0.96	-0.16	-0.56	0.46	-0.94	-0.49	-1.74	-0.35	0.36	-1.97	0.15	0.11	-0.83	-1	1.03	0.29	0.32	1.05	-0.16	-1.04	-0.25	0.57	-0.33	-0.55	-0.44	-0.12
11	OVT	-0.18	-0.57	-0.55	-0.52	1.2	-0.05	0.24	-1.11	-1.22	-1.19	-0.36	-0.1	-1.07	0.2	0.19	0.08	0.48	0.19	-0.19	0.08	-0.62	0.33	-0.96	0.01	-1.71	0.72
12	OVT	-0.39	-0.13	-0.42	-0.85	-0.32	0.18	0.42	0.41	0.15	-0.68	-0.51	0.03	-0.27	-0.52	0.79	2.4	-1.67	-0.43	-0.05	-0.13	0.81	0.79	-1.6	-0.69	0.96	0.33
13	OVT	-2.03	-0.95	-0.98	-1.18	-1.96	-0.22	0.06	-1.47	0.3	-0.53	-0.74	0.08	-1.06	-0.56	0.83	0.58	0.19	1.5	-0.49	0.77	1.04	0.98	-0.79	-1.36	-0.4	0.27
14	OVT	-2	-0.65	-1.32	-1.9	0.81	-0.58	1.09	0	0.01	-0.75	0.45	0	-0.41	-0.04	1.59	1.28	1.74	1.35	0.2	1.03	0.38	-0.07	-1.06	-0.58	-0.65	1.38
15	OVT	-1.73	0.08	-1.96	-1.78	-1.84	-0.24	-0.72	-0.46	0.12	-0.17	-0.14	0.14	-0.57	-0.11	-0.41	0.82	0.21	-0.41	1.29	0.34	-0.81	-0.08	-1.68	-1.6	0.03	0.64
16	OVT	-1.82	-0.73	-0.44	-0.74	-1.95	-0.34	-0.77	-1.09	-0.07	-0.36	1.3	0.83	0.56	-0.39	1.8	0.27	1.13	1.26	0.21	1.85	0.86	-0.39	-0.94	-2.13	-0.43	-0.11
17	OVT	-1.54	-0.43	-0.97	-0.77	-1.41	0.17	0.15	-1.16	0.39	-0.15	-0.78	0.17	0.35	0.17	0.06	-0.59	0.07	-0.41	-0.54	-0.18	-0.51	-0.31	0.61	-1.92	0.56	0.43
18	OVT	-0.7	0.83	-0.51	1.17	-0.53	-0.04	0.65	0.56	0.55	-0.09	0.97	1.02	0.53	0.02	1.59	1.14	1.08	2.11	0.54	1.43	0.84	1.3	0.44	-0.73	0	0.51
19	OVT	-1.09	-0.15	-1.17	-1.35	-1.32	-0.18	-0.24	-0.33	-1.45	-0.25	0.7	-0.02	0.34	-0.02	0.11	1.25	-0.46	-0.49	0.38	0.34	0.08	0.27	-1.23	-1.98	-0.86	-1
20	OVT	0.15	-1.13	-0.9	-0.23	-0.25	-0.12	-0.24	1.1	0.61	-0.63	0.89	0.17	-0.42	-0.65	0.02	1.03	0.57	0.15	-0.96	0.48	0.84	0.55	-1.57	-0.8	0.02	1.18
21	OVT	-1.49	-1.02	-0.88	-0.16	-1	-0.47	-0.8	0.74	-0.9	-0.92	1.7	-0.58	-1.03	-0.65	0.62	0.2	-0.3	-1.06	-1.35	-0.18	-0.17	0	-1.07	-1.4	-0.21	-1.1
22	1-owca	-1.07	0.43	0.36	-1.02	-0.13	-0.16	1.47	0.66	0.45	-0.23	0	0.73	-0.06	-0.36	0	1.17	0.74	0.76	-0.87	0.05	0.97	1.58	-1.21	-0.12	0.15	1.1
23	1-owca	-0.38	-0.66	-0.35	-1	0.59	-0.34	0.93	0.56	-0.64	-0.64	0.27	-0.49	-0.08	-0.03	0.77	0.78	0.37	0.1	-0.37	1.52	0.43	0.68	-1.69	-1.24	-0.6	1.57
24	1-owca	1.9	-0.02	0.04	0.83	0.76	0.66	-0.23	1.07	0.71	0.64	1.98	0.25	0.39	-0.17	0.37	1.56	0.83	0.39	0.95	0.71	0.41	0.39	-1.32	1.63	1.51	1.27
25	1-owca	-0.98	-0.07	-0.02	0.32	-0.63	-0.13	0.7	0.84	-0.81	-0.49	1.72	0.17	0.27	-0.06	0.93	0.09	0.99	1.34	0.11	-0.22	1.01	1.24	-1.93	0.35	1.6	1.87
26	1-owca	-0.96	-0.26	-1.58	0.05	-0.43	-0.65	0.35	-0.7	-2.13	-0.44	-0.67	-0.09	-1.59	0.11	-1.18	-0.56	-0.19	-0.61	0.15	-0.8	1.38	0.15	-1.82	-0.37	0.16	-0.41
27	1-owca	0.16	-2.02	-0.8	-1.87	-0.09	-0.1	0.31	0.07	1.2	-0.69	0.14	0.48	-0.68	-0.31	-0.42	0.07	-0.04	-0.63	0.32	-1.57	-0.41	-0.24	-0.71	-1.66	1.16	-0.84
28	1-owca	-1.11	-1.77	-1.42	-1.24	-0.13	-0.06	-1.91	0.38	-0.46	-1.29	-1.24	-0.05	-0.87	-0.96	-0.34	-0.02	-0.52	-1	-1.37	-1.62	1.52	-0.78	-0.85	-0.5	0.23	0.28
29	3-owca	0.39	-1.04	-2.1	-0.87	0.52	-0.31	-0.97	0.25	0	-0.24	-1.27	-1.41	-0.12	-0.23	0.23	0.12	0.3	-0.77	-0.17	-1.32	1.19	-0.61	-1.48	-1.52	0.49	0.49
30	3-owca	-0.01	-1.17	-0.96	-1.5	-2.24	-0.47	-1.59	-1.83	-1.77	-0.78	-2.32	-2.32	-1.38	-0.43	-1.56	-1.55	-0.83	-1.61	-0.29	-1.31	2.34	1.14	-1.82	-0.35	-0.79	-0.26
31	3-owca	-0.37	-0.72	-0.47	-0.88	1.05	-0.24	-0.36	0.16	0.13	-0.23	-0.57	-0.41	-0.66	-0.2	0.01	1.21	0.29	-0.4	-0.07	0.18	0.94	-1.5	-1.2	-1.15	-0.11	0.04
32	3-owca	-0.39	0.39	0.38	-0.62	-0.52	-0.38	0.71	-0.12	-1.41	-0.48	-0.95	0.38	-1.22	-0.26	-0.34	-1.07	-2.39	-1.45	-0.51	-2.68	2.35	-0.51	-0.46	-0.36	0.27	-0.11
33	3-owca	-0.87	0.32	-0.84	-0.56	0.5	-0.36	-0.12	0.32	-0.76	-0.28	-0.68	-0.3	-0.41	-0.29	0.66	0.08	-0.15	-1.2	0.34	-0.69	-1.65	0.53	-0.74	-0.35	-0.02	-0.77
34	3-owca	0.93	-0.4	1.24	-0.35	0.95	0.93	0.06	-0.33	-0.24	-0.45	-0.77	0.33	-0.22	-0.17	0.66	0.42	0.2	1.07	-0.19	0.81	1.23	0.98	-0.27	0.41	-0.28	-0.74
35	3-owca	-0.33	0.17	-0.68	-2.02	-0.83	-0.83	-2.19	0.43	0.21	-0.68	-0.51	-0.6	-0.7	-0.47	-0.64	0.69	-0.46	-1.41	0.28	-0.16	1.01	-0.49	-1.12	-1.74	-0.18	-0.19
36	3-owca	1.32	0.97	-0.37	0.01	0.67	-0.36	0.22	0.49	-1.11	-0.33	-0.17	-0.41	-0.27	-0.12	-0.29	0.15	-0.3	-1.26	0.34	-0.23	0.41	0.78	-0.53	0.4	0.23	-0.06
37	3-owca	0.2	0.16	-1.37	-0.84	1.29	-0.25	-0.84	0.02	-0.93	-0.35	-0.7	-0.75	-0.46	-0.99	-0.42	0.84	-0.78	-1.06	-0.08	0.13	2.68	-0.27	-0.55	-0.15	0.32	0.56
38	3-owca	0.47	0.07	-0.6	-1.18	1.83	-0.14	-0.17	0.59	-0.38	-0.45	1.44	-1.24	-0.27	-0.33	-1.03	0.61	-0.07	-1.61	-1.01	-1.93	1.49	0.4	0.17	0.34	0.79	1.47
39	3-owca	1.25	-0.53	-2.13	-1.46	1.03	-0.3	0.82	0.59	0.91	-0.24	-0.95	0.1	-0.31	-0.23	-0.03	0.26	0.96	-0.28	-0.26	-0.79	1.14	0.29	-1.86	-0.27	0.3	-0.47
40	3-owca	1.46	-0.04	-0.14	-0.38	0.32	-0.27	0.23	-0.14	-0.63	-0.24	-0.86	0.58	-0.16	-0.25	-1.56	-0.38	0.76	-1.86	-0.4	-0.52	1.83	0.38	-0.41	-0.01	-0.15	-0.03
41	3-owca	0.09	0.28	-1.23	-0.9	0.28	-0.18	0.41	0.89	0.49	-0.23	1.3	-0.23	-0.13	-0.2	0.98	1.51	0.53	0.69	-0.82	-0.61	0.13	-0.03	-0.94	0.02	-0.09	-0.25

Appendix A: The beginning (from 1st to 27th) of the gene expressions dataset of 41 patients

	9575 KARP-1-binding protein	fucosyltransferase 8 (alpha (1,6) fucosyltransferase	calcium channel, voltage-dependent, alpha 2/delta	9578 tubulin, beta polypeptide	9579 elongation factor RNA polymerase II	9580 oxysterol binding protein	9581 cyclin-dependent kinase 2	pleckstrin homology-like domain, family A, member	9583 serine/threonine kinase 38	9584 M-phase phosphoprotein 6	9585 inositol polyphosphate-5-phosphatase, 40kDa	9586 calponin 3, acidic	9587 cadherin 2, type 1, N-cadherin (neuronal)	9588 protein kinase, cGMP-dependent, type I	antigen (activated B lymphocytes, immunoglobulin	9590 homolog of yeast Mis12	9591 chromosome 14 open reading frame 161	9592 hypothetical protein LOC57019	9593 vacuolar protein sorting 24 (yeast)	, TRF1-interacting ankyrin-related ADP-ribose	m voltage-gated channel, subfamily F,	9596 calmodulin 3 (phosphorylase kinase, delta)	9597 glia maturation factor, beta	9598 ring finger protein 44	9599 9598	9600 cyclin-dependent kinase inhibitor 1C (p57, Kip2)	Williams-Beuren syndrome chromosome
1	-1.767	2.077	-0.1613	-0.8767	1.063	-0.7975	0.843	3.683	0.4465	2.648	-0.4165	-1.551	-0.7451	-1.479	0.8938	-0.9414	1.452	-0.9056	-3.351	-0.8405	0.876	0.8356	1.96	-0.455	0.1162	-1.188	0.7858
2	-2.655	1.989	0.8791	0.7621	0.1377	-1.259	-0.06849	1.869	0.5488	2.664	-0.6242	0.493	-1.099	-4.408	1.211	1.017	0.7031	-1.368	-0.728	0.9985	-0.7056	1.392	0.4617	-1.618	-2.283	1.54	-0.6473
3	-2.823	1.897	1.462	0.5368	1.183	0.4017	0.586	0.9687	1.177	2.83	-2.025	0.9876	-0.2693	-4.334	3.449	1.033	-0.07918	-0.6074	-0.7675	0.1451	0.03441	0.4997	1.858	-1.661	0.8315	1.173	0.529
4	0.3269	2.252	0.3832	-0.03918	0.1187	-1.487	0.867	3.572	2.598	3.611	1.32	2.677	1.127	1.6	1.696	1.33	1.544	1.794	-0.2356	1.822	0.884	2.149	1.144	0.7072	0.06035	0.108	0.4122
5	-1.301	1.251	0.5038	1.069	1.438	0.2915	0.4376	3.437	1.558	2.284	-0.09735	1.811	1.879	0.8426	2.84	2.346	3.592	2.2	1.95	2.134	0.2699	0.262	-0.9138	0.69	0.9705	0.2731	0.5564
6	1.37	3.804	1.309	1.65	2.214	0.8295	1.146	2.504	1.646	-0.1278	1.916	1.805	1.076	-0.5459	2.285	0.7458	1.954	1.11	1.248	1.061	1.562	1.142	1.192	1.923	1.163	1.38	1.406
7	0.77	1.92	1.39	0.95	0.63	0.84	0.12	2.21	-1.18	1.98	0.74	0.68	-2.34	1.03	3.16	2.22	0.32	0.45	-0.58	1.14	-0.62	0.41	0.48	2.42	0.72	-2.1	1.84
8	1.31	1.86	1.11	0.06	0.57	-0.52	-1.03	2.07	-1.08	2.71	0.58	-1.89	-0.05	0.79	3.39	1.86	2.73	-0.02	0.78	1.66	2.46	-0.19	0.98	2.5	1.98	-1.32	-1.4
9	3.4	1.95	0.64	0.07	2.23	-0.28	-1.46	3.26	-0.41	2.04	0.67	2.08	-0.18	0.76	2.54	2.82	2.61	-1.25	0.57	1.06	-0.11	-0.68	2.14	-0.06	3.36	-0.19	-0.71
10	1.35	3.09	2.5	1.61	0.87	1.43	-1.65	3.63	0.91	2.77	1.41	1.89	1.63	0.88	5.77	3.08	3.17	1	0.39	0.96	1.78	0.52	2.11	1.36	2.85	-0.09	-0.22
11	1.65	2.55	0.57	0.2	1.24	0.3	-0.54	1.82	0.43	1.78	3.57	-0.08	0.3	-0.07	3.51	1.25	1.95	0.27	-0.07	0.02	1.01	0.98	1.5	0.04	1.05	0.15	-0.02
12	1.56	2.1	1.99	1.21	1.24	0.62	-1.39	2.53	-0.14	1.84	-0.41	1.01	0.02	0.48	2.22	2.49	1.85	-1.61	0.38	1.51	1.79	1.61	-0.67	1.92	3.31	-0.14	0.06
13	1.05	1.73	2.21	0.34	0.13	1.76	-2.11	3.94	-0.38	1.83	0.29	0.01	0.12	0.07	4.36	2.85	2.01	0.57	0.76	0.4	1.87	-0.54	1.05	-0.27	1.86	1.62	-0.1
14	0.71	2.35	0.42	0.74	1.73	0.42	-1.07	3.77	-0.99	1.43	-0.2	-0.45	-0.97	0.5	1.78	2.36	1.59	-1.32	-0.51	-0.01	0.73	-0.41	2.06	0.73	2.85	0.35	-1.52
15	0.58	2	2.15	1.5	-0.64	2.5	0.52	3.47	0.76	3.44	0.5	1.29	0.61	0.9	4.59	2.87	-0.18	-0.65	0.63	1.06	0.59	2.25	2.83	1.66	2.08	0.88	-0.91
16	0.77	0.54	1.85	1.04	1.4	1.31	1.8	4.38	0.25	1.97	0.6	2.04	1.11	1.11	4.39	3.02	0.52	0.13	-0.08	1.06	1.74	2	3.22	0.27	2.38	0.89	1.45
17	0.41	1.9	3.26	0.31	1.12	-1.03	-0.65	3.22	-0.52	3.25	-0.1	-0.17	-0.87	0.93	3.73	0.86	1.42	-1.13	-0.61	0.52	0.75	2.72	1.79	0.75	4.02	1.02	-0.01
18	0.39	1.72	2.11	0.67	0.36	1.87	0.45	4.29	-1.49	2.86	-0.14	0.68	0.34	0.97	2.32	2.67	0.74	-1.08	1.24	0.59	0.64	0.57	2.1	0.72	2.87	2.1	-0.77
19	1.86	1.87	2.17	2.16	0.23	-0.71	-1.87	2.97	0.09	1.7	-0.06	1.28	0.51	-1.78	3.69	2.32	1.46	-0.11	-0.1	1.41	-0.39	2.04	0.39	0.49	2.21	-0.05	-1.96
20	0.46	0.27	1.89	0.28	1.19	-0.58	-1.23	2.49	-0.98	1.98	0.38	0.58	-0.35	0.28	2.85	1.63	0.21	-0.05	-0.12	0.98	0.49	-0.76	0.38	1.08	2.79	-0.02	0.4
21	1.58	1.91	2.3	1.13	0.45	0.08	-0.16	3.91	0.41	2.58	0.17	1.37	0.6	0.44	5.58	3.06	1.64	0.71	0.73	-0.03	0.46	1.91	1.89	1.6	3.26	-1.15	-0.52
22	1.66	0.28	1.01	1.93	-1.04	-0.01	-2.22	3.19	-0.7	2.04	-0.24	-0.49	0.27	0.29	2.75	0.9	2.6	-0.88	0.23	-0.17	0.59	0.17	0.96	0.46	-0.04	-1.51	2.52
23	1.63	2.34	2.04	-0.31	1.04	1.61	-0.53	4.36	-0.07	3.02	-1	0.11	-0.6	-0.5	3.5	2.1	0.48	-1.57	0.38	0.44	1.73	0.41	2.05	0.87	2.76	-0.4	1.61
24	0.7	0.81	0.45	-1.12	0.18	-0.55	-1.42	0.95	-1.16	1.17	0.35	1.1	-0.77	-0.06	2.35	3.4	2.12	0.25	-0.24	0.85	-0.22	0.34	1.47	-0.76	0.49	-1.95	-0.82
25	2.01	2.16	3.02	0.28	0.17	0.39	-0.37	3.65	-0.45	2.48	-0.66	1.36	0.62	1.83	4.83	1.65	1.42	-0.79	0.26	0.37	2.95	1.67	2.18	0.11	2.22	0.79	0.16
26	2.21	2.12	1.12	0.31	0.28	0.87	0.31	1.99	-1.31	3.45	0.42	-0.75	-0.26	0.67	4.76	0.69	1.91	-0.97	-0.55	0.27	1.87	-0.06	1.97	3.06	1.61	0.79	0.04
27	2.69	1.96	1.2	0.42	0.12	1.31	0.4	2.99	-0.82	1.01	-1.02	0.29	-1.29	-1.76	2.16	2.17	1.36	-1.08	-0.15	-0.53	0.36	-0.11	0.71	0.5	0.2	-1.1	1.39
28	1.5	1.98	2.16	0.63	-0.63	2	0.04	0.82	-0.09	-0.27	0.72	0.37	0.5	-0.26	1.35	1.41	1.1	0.28	0.6	-1.06	0.44	0.99	1.35	0.58	1.24	-0.31	0.48
29	1.66	0.71	0.38	-1.28	-0.4	1.56	-1.35	3.08	-0.88	-0.22	0.26	0.58	-0.19	-0.79	0.91	2.7	1.27	-1.33	0.05	0.07	-0.42	-0.83	1.17	0.97	-0.22	-0.65	-0.99
30	2.17	1.49	0.81	0.34	1.47	1.9	-0.77	4.5	-0.08	0.63	-0.88	-0.68	-0.82	0.3	3.33	1.95	2.05	-0.04	1.11	0.62	1.7	-0.92	0.77	0.84	1.09	-2.59	0.4
31	1.84	3.01	1.09	-0.43	1.36	2.78	-1.15	2.32	-1.06	0.94	-1.55	-1.11	-0.3	-0.11	1.17	2.72	2.22	-1.21	-0.66	0.94	-0.86	-0.31	1.79	1.65	-0.72	-0.09	-1.08
32	0.61	1.43	1.26	0.61	0.04	-0.55	0.14	1.58	0.86	1.31	-0.49	1.74	0.13	-0.32	0.72	0.65	0.98	0.48	0.46	0.19	0.9	0.46	1.85	3.36	2.55	0.6	0.93
33	1.5	1.41	1.9	1.28	0.59	1.15	0	1.93	0.74	-0.54	0.59	1.73	0.86	-0.01	1.91	1.75	-0.2	0.43	0.32	2.83	0.51	0.54	2.04	1.68	0.32	-0.68	0.44
34	1.32	1.02	1.91	-0.05	2.57	0.39	-0.39	2.17	-0.57	2.46	0.25	2.14	1.11	0.54	1.15	1.72	2.01	0.59	1.43	-0.39	1.5	0.81	0.85	0.28	1.25	-1.29	0.16
35	1.88	1.38	0.76	1.14	0.97	0.29	-0.49	0.79	0.14	-0.14	-2.09	-1.84	1.04	0.43	-1.18	0.21	1.55	-1.06	-1.95	0.67	1.1	0.96	1.8	-0.07	3.09	-0.61	1.1
36	-0.31	0.12	1.35	0.53	-0.12	-0.56	-0.45	-0.25	1.03	-0.37	-0.13	1.7	2	-0.75	1.31	0.67	-0.64	1.15	-1.39	-0.57	-0.65	-0.02	0.41	2.31	0.28	-1.12	0.54
37	1.47	1.99	2.18	1.85	0.09	-1.54	-0.9	2.74	1.13	0.89	0.15	2.14	0.15	-0.22	1.38	1.09	-0.39	1.5	-0.84	1.57	2.82	1.19	1.45	1.73	1.52	-0.76	0.17
38	0.75	0.89	1.98	-0.19	-1.29	-1.24	-1.39	-0.6	-0.18	-1.09	1.09	-0.09	-1.57	-0.99	1.28	1	-0.69	-0.81	0.66	-1.27	-0.59	-0.12	0.06	1.22	-0.22	-1.19	-0.73
39	-0.17	0.15	4.12	-0.41	-0.7	-0.56	-0.49	1.31	0.79	0.31	1	0.81	-0.42	-1.19	2.43	4.1	1.37	1.31	-1.34	-1.18	-0.59	-0.32	0.97	0.97	-0.48	-0.49	-0.61
40	1.85	0.42	0.55	0.06	-0.06	-0.33	-0.69	0.49	0.1	0.67	0.01	1.64	0.29	-0.21	1.53	-0.47	-0.09	-0.19	-0.69	0.08	0.69	-0.07	2.38	1.08	-0.02	-0.24	-0.56
41	1.35	1.92	2.38	2.85	-0.89	-0.05	0.99	3.92	1.18	0.09	0.77	0.31	0.64	-0.09	1.01	0.18	1.08	0.68	-0.9	0.74	0.37	0.79	2.01	0.23	0.1	-0.47	1.3

Appendix B: The end (from 9575th to 9600th) of the gene expressions dataset of 41 patients