ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

TC-VGC: A Tumor Classification System using Variations in Genes' Correlation

Eunji Shin[a], *Youngmi Yoon*[b], *Jaegyoon Ahn*[a], *Sanghyun Park*[a,*]

[a] *Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea*

[b] *Division of Information Technology, Gachon University of Medicine and Science, 534-2 Yonsu-dong, Yonsu-gu, Inchon 534-2, South Korea*

ARTICLE INFO

Article history:
Received 24 June 2010
Received in revised form
11 January 2011
Accepted 7 March 2011

Keywords:
Tumor classification
Microarray data analysis
Gene–gene correlation
Cancer specific genetic network

ABSTRACT

Classification analysis of microarray data is widely used to reveal biological features and to diagnose various diseases, including cancers. Most existing approaches improve the performance of learning models by removing most irrelevant and redundant genes from the data. They select the marker genes which are expressed differently in normal and tumor tissues. These techniques ignore the importance of the complex functional-dependencies between genes. In this paper, we propose a new method for cancer classification which uses distinguished variations of gene–gene correlation in two sample groups. The cancer specific genetic network composed of these gene pairs contains many literature-curated prostate cancer genes, and we were successful in identifying new candidate prostate cancer genes inferred by them. Furthermore, this method achieved a high accuracy with a small number of genes in cancer classification.

1. Introduction

DNA microarray technologies make it possible to simultaneously monitor the expression levels of thousands of genes [1]. The large amount of data generated by microarray experiments has stimulated the development of many computational methods to study different biological processes at the gene expression level. These microarray techniques are expected to result in precise cancer detection and classification.

The major difficulty of microarray data analysis is the large number of genes compared to the limited number of samples in a typical experiment. Furthermore, many of the genes are 'noise' genes that are not relevant in differentiating between normal and tumor samples. One of the major challenges in designing an accurate classifier using microarray data is identifying the optimal subset of relevant genes. This is known as gene selection and corresponds to feature selection in the field of pattern classification.

Existing gene selection methods just select the marker genes which are differentially expressed in normal and tumor tissues. They do not consider the gene–gene correlations between two groups of samples. However, variations in gene–gene correlations can be a good guide for making a cancer diagnosis. For example, when a transcription factor activates or represses two genes, A and B, simultaneously, the expression levels of A and B reveal that they have high correlation. If gene B is affected by a specific disease, the transcription factor continues to activate or repress gene A, while it can no longer influence gene B.

Most of microarray studies often focus on the identification of differentially expressed genes or the construction of prediction rule. However, gene itself which does not appear to be oncogenic could be highly relevant to be cancer specific when they are considered with others. In this paper,

[*] *Corresponding author*. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail address: sanghyun@cs.yonsei.ac.kr (S. Park).

0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.03.002

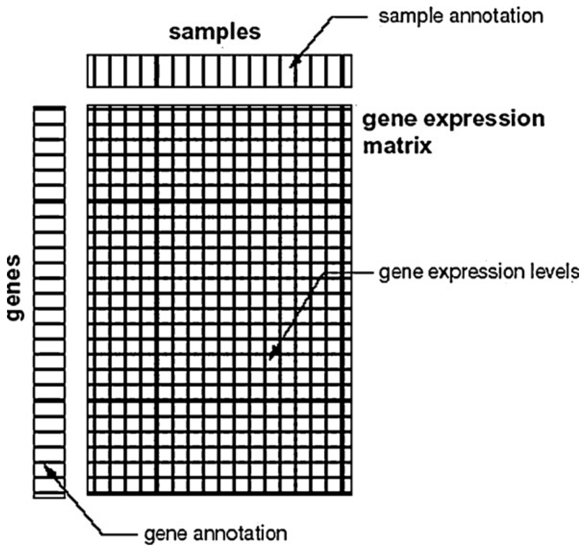


Fig. 1 – Microarray data set.

we are proposing the TC-VGC (Tumor Classification using Variations in Genes' Correlation) that considers complex functional-dependencies between genes using the correlation coefficients of gene pairs. This is a new cancer classification method which uses substantially smaller number of genes and retains a high predictive accuracy. Given a new patient's sample, the method predicts the class of the sample using variations in gene correlations, with high accuracy. Furthermore, the cancer specific genetic network composed of these gene pairs contains many literature-curated prostate cancer genes, and we were successful in identifying new candidate prostate cancer genes inferred by them.

2. Related works

2.1. *Microarray data set*

The gene expression data which is generated from microarrays is organized as matrices. Rows represent genes and columns represent various samples such as normal or tumor tissues. Values in each cell represent the expression level of the particular gene in the sample. Fig. 1 shows an example of a gene expression matrix [2].

2.2. *Existing cancer classification algorithms based on microarray*

Machine-learning methods are used to classify and cluster data. These methods are especially useful in cancer diagnosis and detection [3–5]. The current trend is to apply a computational approach to traditional biological research, which is then used to understand biological processes.

Many researchers have applied a machine-learning approach to microarray data analysis. Examples of these techniques include the SVM (Support Vector Machine), k-NN (K-Nearest Neighbors), Random Forest, and Bayesian networks [6]. Pirooznia et al. recently introduced the commonly used classification methods, and applied these methods to publicly available datasets [7]. Results revealed that SVM classification and RBF Neural Nets had the best accuracy. Wang et al. demonstrated that feature subset selection algorithms, namely wrappers, filters and CFS (Correlation-based Feature Selection), can be very useful in extracting relevant information in microarray data analysis [8]. They exhibited that the filters and CFS are recommended for fast analysis of data. However, for better classification accuracy and fewer genes that could be further used for a cancer diagnosis toolkit, the wrapper approaches are more recommended.

SVM is becoming increasingly popular classifiers for many data, including microarrays. Guyon et al. proposed an SVM-RFE (Support Vector Machine Recursive Feature Elimination) algorithm to recursively classify the samples using SVM and to select genes according to their weights in the SVM classifiers [9]. Duan et al. proposed MSVM-RFE [10]. This technique trains multiple linear SVMs on subsamples of training data and computes the feature ranking score of SVM-RFE. The performance of MSVM-RFE is better than that of SVM-RFE and fewer genes were needed in MSVM-RFE than in SVM-RFE.

Pan et al. proposed a comprehensive KNN/LSVM classification approach [11]. This technique used a combination of a k-NN majority voting approach and a local Support Vector Machine approach which makes optimal decisions at the local level. The goal of this technique is to classify cases based on a local approach without the time burden which is usually necessary to run traditional algorithms. Diaz-Uriarte and Alvarez de Andres investigated the use of a Random Forest for classification of microarray data including multi-class problems [12]. Random Forest is a classification algorithm that is well suited for microarray data even when most predictive variables are noise and the amount of input data is large.

The main challenge in classifying gene expression data is to solve the curse of dimensionality problem. In addition, irrelevant and redundant features increase the search space and make patterns more difficult to detect and make it more difficult to capture rules necessary for prediction or classification. Albrecht has proven that selecting the best feature subset is an NP-complete problem [13]. To overcome these problems, feature selection is an indispensable task in classification to identify a smaller subset of relevant genes for building robust learning models. Gheyas and Smith proposed a hybrid method for feature subset selection consisting of a combination of simulated annealing and a genetic algorithm and compare its performance to a variety of other greedy and stochastic search algorithms [14]. Jaeger et al. reduced the number of relevant genes by eliminating highly correlated ones [15]. Kim et al. attempted to use several methods for extracting informative features and combining classifiers learned from the negatively or complementarily correlated features [68]. Berretta et al. suggests feature set model to extract differentially expressed genes [16]. The genes that must be in a different state in any pair of samples with different labels and in the same state in any pair of samples with the same label are chosen to be a feature set. Dougherty and Brun considered optimal feature sets in the framework of a model in which the features are grouped in such a way that intra-group correlation is substantial whereas inter-group correlation is minimal [17].

Principle components analysis (PCA), a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set, is one of the frequently used methods for feature selection of microarray data [18]. Independent component analysis (ICA) is a method for finding underlying factors or components from multidimensional statistical data. Although ICA is similar to PCA, ICA has some advantage over PCA because it exploits higher order statistics and has no restriction on orthogonal transformations [19].

Heretofore, method for microarray dataset analysis has focused on selecting genes that exhibit extremely large differential expressions between different phenotypes. They ignore the impact of the complex structures found in microarray experiments. Tan et al. proposed the k-TSP method, which is an algorithm that finds k top-scoring gene pairs and addresses the stochastic dependence between different genes [20]. Furthermore, they lack the ability to select genes that change their relationships with other genes in different biological conditions. Wei and Li presented an alternative formulation of cancer classification problem that is based on the biologists' intuition that samples within the same tumor class tend to be more similar in gene expression than samples from different tumor classes [21]. Ravetti et al. also uncovered genes which are highly correlated to the progression of the disease and presented a clear pattern of either up or down regulation with increasing AD severity [22].

In this study, we explicitly used gene–gene correlations for all gene pairs to boost statistical power. The performance of our system was compared with the performance of the previously mentioned algorithms.

3. Cancer classification algorithm

3.1. *System overview*

This section describes the details of the classification algorithm for microarray data (see Fig. 2). We perform the repetitive 10-fold cross validation while varying parameter ranges. The parameter set with best accuracy in the parameter ranges was chosen as the optimal values for parameters. The independent test is used to predict the class of independent samples using optimal parameter set. For both cross validations and independent tests, the following processes are performed.

Firstly, we extract cancer-specific gene pairs by calculating the correlation coefficients of all gene pairs for normal and tumor training sample set separately. Because gene pairs with a greater correlation difference between normal and tumor samples are regarded as more discriminative one for classification, we select the gene pairs with high coefficient differences as the cancer-specific gene pairs.

Secondly, hub genes are used to select gene pairs which consists the best classifier; at least one gene of the gene pair must be a hub gene. Then we calculate the weight of these gene pairs. We assign bigger weight to a pair of genes as the similarity of variations in expression values of the two genes gets increased. Consequently, our classifier is composed of gene pairs with highest weight.

Lastly, we predict the class of test samples by considering the level of correlation change when it is added to the training samples.

3.2. *Definition of notation*

In this section we define the notations which will be used throughout the paper.

g_i	A gene in a microarray
ns_i	A normal sample in a training dataset
ts_i	A tumor sample in a training dataset
us_i	A sample in a test data set
n_{ij}	Expression value of gene g_i in sample ns_j
t_{ij}	Expression value of gene g_i in sample ts_j
u_{ij}	Expression value of gene g_i in sample us_j
$\rho_n(g_i, g_j)$	Correlation coefficient between two genes in normal samples
$\rho_t(g_i, g_j)$	Correlation coefficient between two genes in tumor samples
$\rho'_n(g_i, g_j)$	Recalculated correlation coefficient after including a test sample into the normal training data set
$\rho'_t(g_i, g_j)$	Recalculated correlation coefficient after including a test sample into the tumor training data set
NC	A set of cancer-specific gene pairs which have a strong correlation in the normal samples and shows significantly different correlation coefficient values between two classes
TC	A set of cancer-specific gene pairs which have a strong correlation in the tumor samples and shows significantly different correlation coefficient values between two classes
N_{diff}	Sum of absolute differences between $\rho_n(g_i, g_j)$ and $\rho'_n(g_i, g_j)$ for all gene pairs in *NC*
T_{diff}	Sum of absolute differences between $\rho_t(g_i, g_j)$ and $\rho'_t(g_i, g_j)$ for all gene pairs in *TC*
cor	Parameter for deciding whether the gene pairs have a strong correlation
sig	Parameter for deciding whether the differences in the correlation coefficients for the gene pairs in the two class groups are significant
nk	Parameter for representing the number of classifier component selected in *NC*
tk	Parameter for representing the number of classifier component selected in *TC*

3.3. *Extracting cancer specific gene pairs*

To extract cancer specific gene pairs, we find out whether two genes are related to each other for all gene pairs. We calculated the degree of correlation between two genes using the SCC (Spearman's Correlation Coefficient) [24] as follows:

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2 - 1)}$$

$$d_i = x'_i - y'_i$$

where x'_i and y'_i are ranks of two expression values x_i and y_i at sample i, respectively.

The SCC provides a measure of the similarity of two ranked lists. The Spearman correlation is less sensitive than the Pearson correlation to strong outliers ρ ranges from −1.0 to 1.0. If ρ equals −1.0, the two genes have a perfect negative correlation and if ρ equals 1.0, the two genes have a perfect positive

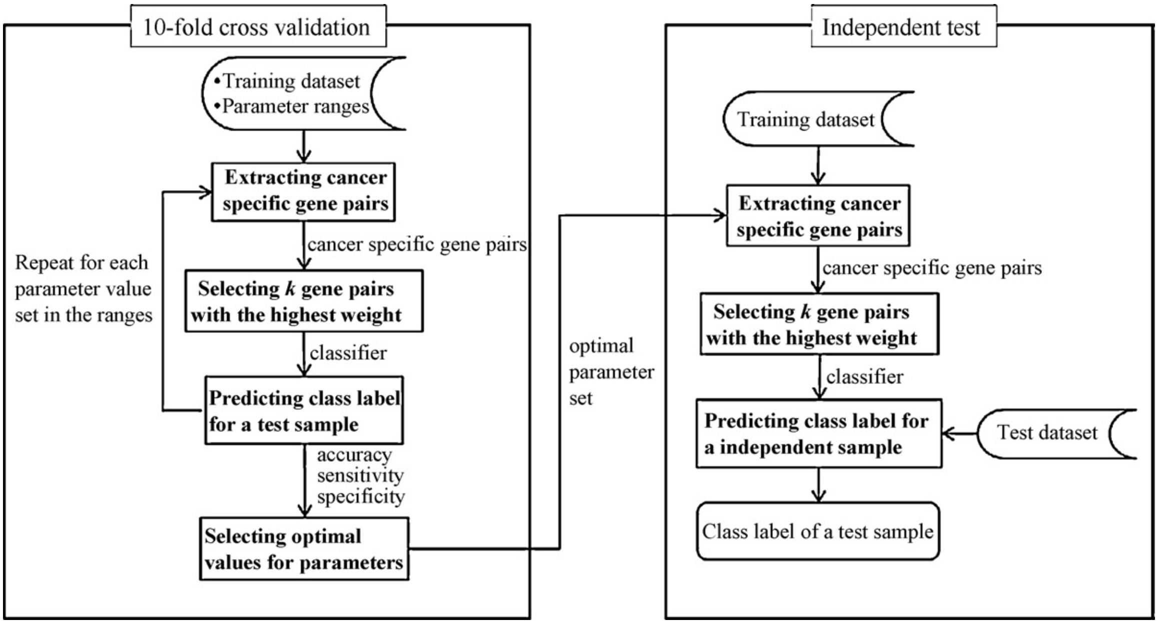

10-fold cross validation
•Training dataset
•Parameter ranges
Extracting cancer specific gene pairs
cancer specific gene pairs
Repeat for each parameter value set in the ranges
Selecting k gene pairs with the highest weight
classifier
Predicting class label for a test sample
accuracy
sensitivity
specificity
Selecting optimal values for parameters
optimal parameter set
Independent test
Training dataset
Extracting cancer specific gene pairs
cancer specific gene pairs
Selecting k gene pairs with the highest weight
classifier
Predicting class label for a independent sample
Test dataset
Class label of a test sample

Fig. 2 – Overview of TC-VGC.

correlation. If ρ equals 0, there is no correlation between two genes. As the absolute value of ρ gets bigger, the relationship between the two genes gets stronger.

In order to find the candidate cancer-specific gene pairs, we computed the correlation coefficients of all gene pairs for the normal and tumor samples separately. The gene pair can be selected as the candidate cancer-specific gene pair when the correlation of the gene pair is strong in only one class of samples and the difference of the correlation coefficient between the two classes is significant. Among the cancer-specific gene pairs, *NC* is a set of gene pairs, which are highly correlated in the normal samples only, and the gene pairs in *TC* are highly correlated in the tumor samples only.

The threshold *cor* is the parameter for deciding whether the correlation between two genes was strong enough. We consider the gene pair to have strong correlation if their absolute value of correlation is greater than *cor*. The threshold *sig* is the parameter for determining if there is a significant difference in the correlation coefficients between the two class groups. If the gene pair has strong correlation in only one class, the gene pair whose difference of coefficient values exceeds *sig* could be selected as the candidate cancer-specific gene pair.

3.4. *Building classifier*

This section describes how to select the best gene pairs among large number of candidate cancer-specific gene pairs. The selection is based on whether a gene is hub-gene, and weight of a gene pair.

Firstly, we extract the hub genes with more than five interactions [23]. A hub is more likely to be essential than a non-hub simply because the hub has more interactions and thus a higher chance to engage in an essential interaction [25]. In a gene co-expression analysis context, Horvath et al. defined a 'hub' gene of an expression module to be a gene with strong intra-module connectivity [26].

Secondly, hub genes are used to select gene pairs which consists the best classifier; at least one gene of the gene pair must be a hub gene. Here, the hub gene must have more than five interactions [14]. We calculate the weight of these gene pairs. Even though a gene pair exhibits strong correlation, expression values of one gene have big variation across samples, while expression values of the other gene have little variation. This gene pair is not informative because the slight variation of the other gene could be just from experimental noise. Accordingly we assign bigger weight to a pair of genes as the similarity of variations in expression values of the two genes gets increased. The weight between the gene pairs is calculated by following formula:

$$\text{weight}(g_i, g_j) = \left\{ \begin{array}{ll} 1/D_N(g_i, g_j) & \text{if} \quad g_i, g_j \in NC \\ 1/D_T(g_i, g_j) & \text{if} \quad g_i, g_j \in TC \end{array} \right\}$$

$$D_N(g_i, g_j) = \frac{1}{n-1} \sum_{k=1}^{n-1} |r_{g_i ns_k} - r_{g_j ns_k}|$$

$$D_T(g_i, g_j) = \frac{1}{n-1} \sum_{k=1}^{n-1} |r_{g_i ts_k} - r_{g_j ts_k}|$$

$$r_{g_i ns_k} = |n_{ik} - n_{ik+1}|$$

$$r_{g_i ts_k} = |t_{ik} - t_{ik+1}|$$

$D_N(g_i, g_j)$ is the average of differences between variations of expression values of two consecutive normal samples in g_i and g_j $D_T(g_i, g_j)$ is likewise calculated with respect to the tumor class. Below is an example that calculates the weight of the

Please cite this article in press as: E. Shin, et al., TC-VGC: A Tumor Classification System using Variations in Genes' Correlation, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.03.002

Table 1 – Classifier construction algorithm.

```
Input:
• $cor(\delta)$
• $sig(\delta)$
• $nk$, the number of classifier component selected in $NC$
• $tk$, the number of classifier component selected in $TC$
• microarray data set
Output: classifier
1. repeat for each gene pair $(g_i, g_j)$
2.   compute the correlation coefficients for the normal and tumor
     classes, $\rho_n(g_i, g_j)$ and $\rho_t(g_i, g_j)$
3.   if $(|\rho_n(g_i, g_j)| > cor$ and $|\rho_n(g_i, g_j)| - |\rho_t(g_i, g_j)| > sig)$ then
4.     insert the gene pair $(g_i, g_j)$ into candidate gene pair set $NC$
5.   else if $(|\rho_t(g_i, g_j)| > cor$ and $|\rho_t(g_i, g_j)| - |\rho_n(g_i, g_j)| > sig)$ then
6.     insert the gene pair $(g_i, g_j)$ into candidate gene pair set $TC$
7. end
8. repeat for the gene pairs that at least one gene among gene
   pairs in $NC$ is hub gene
9.   repeat for all samples of the gene pair
10.      calculate the absolute difference of expression value
         between two consecutive samples on a gene $g_i$,
         $r_{g_i ns_k} = |n_{ik} - n_{ik+1}|$
11.      calculate the absolute difference of expression value
         between two consecutive samples on the other gene $g_j$,
         $r_{g_j ns_k} = |n_{jk} - n_{jk+1}|$
12.      $D_N(g_i, g_j) += |r_{g_i ns_k} - r_{g_j ns_k}|$
13.  end
14.  $weight\ (g_i, g_j) = 1/\text{mean}\ (D_N(g_i, g_j))$
15. end
16. for all gene pairs in $TC$, the process from step 8 to 15 is repeated
17. select the $nk$ and $tk$ gene pairs with highest $weight$ among the
    candidate gene pairs
18. return classifier
```

gene pair g_a and g_b using Table 2.

$$\text{weight}(g_a, g_b) = \frac{1}{avg}(||n_{a1} - n_{a2}| - |n_{b1} - n_{b2}|| + ||n_{a2} - n_{a3}| - |n_{b2} - n_{b3}|| + \ldots + ||n_{ak-1} - n_{ak}| - |n_{bk-1} - n_{bk}||)$$

In this example, $n_{a1} - n_{a2}$ indicates difference of two expression values of gene g_a on the first and the second samples which are two consecutive samples. The order of samples is randomly determined.

Consequently, our classifier is composed of nk and tk gene pairs with highest weight among the gene pairs. Table 1 shows the algorithm for constructing our classifier.

3.5. *Predicting the class label for a test sample*

We obtain the optimal values for parameters based on 10-fold cross validation. Parameter ranges are initially provided by user. Among parameter value sets with best accuracy in the parameter ranges, the one with the highest weight is selected as the optimal values for parameters.

We predict the class label of an independent sample using the optimal values of parameters and evaluate the performance of classifier. We can apply the optimal parameter to test independent microarray dataset. The procedure to predict the unknown class label of the independent microarray dataset is as follows:

(1) Build a classifier using training dataset and the optimal values of parameters. Calculate the correlation coefficient of all gene pairs in the NC for the normal samples and all gene pairs in the TC for the tumor samples.
(2) Add a test sample into normal and tumor sample set and recalculate the correlation coefficients. Then calculate the absolute differences of the recalculated correlation coefficients with those from step 1.
(3) For all classifiers, calculate the average of the absolute differences. The class which had the smaller average is the predicted class.

Below is an example that estimates a class of the test sample us_k for the gene pair g_1 and $g_2$in the NC and the gene pair g_3 and $g_4$in the TC using Table 2.

$$\begin{aligned}
\rho_n(g_1, g_2) &= \text{SCC}[(n_{11}, n_{12}, \ldots, n_{1p}), \quad (n_{21}, n_{22}, \ldots, n_{2p})] \\
\rho'_n(g_1, g_2) &= \text{SCC}[(n_{11}, n_{12}, \ldots, n_{1p}, u_{1k}), \quad (n_{21}, n_{22}, \ldots, n_{2p}, u_{2k})] \\
\rho_t(g_3, g_4) &= \text{SCC}[(t_{31}, t_{32}, \ldots, t_{3q}), \quad (t_{41}, t_{42}, \ldots, t_{4q})] \\
\rho'_t(g_3, g_4) &= \text{SCC}[(t_{31}, t_{32}, \ldots, t_{3q}, u_{3k}), \quad (t_{41}, t_{42}, \ldots, t_{4q}, u_{4k})]
\end{aligned}$$

N_{diff} is computed by summing the absolute differences between $\rho_n(g_i, g_j)$ and $\rho'_n(g_i, g_j)$ for all gene pairs included in NC. T_{diff} is computed by the same process for all gene pairs included in TC.

$$\text{N}_{diff} = \frac{1}{nk} \sum_{(g_i, g_j) \in NC} |\rho_n(g_i, g_j) - \rho'_n(g_i, g_j)|$$

$$\text{T}_{diff} = \frac{1}{tk} \sum_{(g_i, g_j) \in TC} |\rho_t(g_i, g_j) - \rho'_t(g_i, g_j)|$$

Table 2 – Example of a microarray dataset.

	Training dataset										Test data
	Normal class					Tumor class					
	ns_1	ns_2	ns_3	…	ns_p	ts_1	ts_2	ts_3	…	ts_q	us_k
g_1	n_{11}	n_{12}	n_{13}	…	n_{1p}	t_{11}	t_{12}	t_{13}	…	t_{1q}	u_{1k}
g_2	n_{21}	n_{22}	n_{23}	…	n_{2p}	t_{21}	t_{22}	t_{23}	…	t_{2q}	u_{2k}
g_3	n_{31}	n_{32}	n_{33}	…	n_{3p}	t_{31}	t_{32}	t_{33}	…	t_{3q}	u_{3k}
⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮	…	⋮	⋮
g_n	n_{n1}	n_{n2}	n_{n3}	…	n_{np}	t_{n1}	t_{n2}	t_{n3}	…	t_{nq}	u_{nk}

Rows represent genes and columns represent samples.

Please cite this article in press as: E. Shin, et al., TC-VGC: A Tumor Classification System using Variations in Genes' Correlation, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.03.002

Table 3 – Algorithm for predicting the class label of a test sample.

```
Input:
• classifier
• $us_k$, a test(unknown) sample
Output: The label of the class (Normal or Tumor)
1. compute the correlation coefficient for each gene pair in
   classifier, $\rho_n(g_i, g_j)$ and $\rho_t(g_i, g_j)$
2. repeat for gene pair selected in NC among classifier component
3.     recalculate the correlation coefficient after add a test sample
       into normal sample set, $\rho'_n(g_i, g_j)|$
4.     $N_{diff}{}^{+} = |\rho_n(g_i, g_j) - \rho'_n(g_i, g_j)|$
5.end
for all gene pairs selected in TC among classifier component, the
   process from step 2 to 5 is repeated
6.if $N_{diff} < T_{diff}$ then
7.     return Normal
8.else
9.     return Tumor
```

Consequently, if T_{diff} is bigger then N_{diff}, we predict the class of the independent sample to be normal, otherwise to be tumor.

$$\text{Prediction} = \left\{ \begin{matrix} \text{Normal} & \text{if} & N_{diff} < T_{diff} \\ \text{Tumor} & \text{if} & N_{diff} \geq T_{diff} \end{matrix} \right\}$$

We measure the accuracy of our classifier by comparing the predicted class and the actual class of the independent test samples. Table 3 shows the algorithm which predicts the class of the test samples.

4. Experimental results

4.1. Experimental environment

Table 4 shows the information regarding the well-known prostate, colon and lung cancer microarray data sets that were used in this experiment. For convenience, we represented each data set by an abbreviation of the first author's name from the published papers. We used publicly available colon and prostate cancer microarray data. "Rha" is colon cancer microarray data with a cDNA platform. "Singh", "Welsh", "LaTulippe" and "Stuart" are prostate cancer microarray data with an Affymetrix HG_95AV2 platform. "Landi" (GSE10072), and "Hou" (GSE19188) are lung cancer microarray data using HG-U133A and HG-U133_Plus_2 Affymetrix chips, respectively. "Hou" is expression data for early stage non-small cell lung cancer. "Rha" consists of two types of microarray data, "paired" and "unpaired". For the "paired" data, normal and tumor tissue samples were collected from the same person. For the "unpaired" data, only tumor samples were collected.

Table 5 – Type of experiment.

Type of experiment	Training dataset	Test dataset
10-Fold cross validation experiment	Rha_paired	Rha_paired
	Singh	Singh
	Stuart	Stuart
	Landi	Landi
	Hou	Hou
Independent experiment	Rha_paired	Rha_unpaired
	Singh	Welsh + LaTulippe
	Stuart	Welsh + LaTulippe

Experiments consist of two types (Table 5); 10-fold cross validation and independent test. The 10-fold cross validation experiments are used to obtain the optimal values for parameters which will be used in the experiment for predicting the class of independent samples. The independent experiments measure the accuracy of the TC-VGC.

In the independent experiments, "Singh" and "Stuart" were used as the training data and integrated data from "Welsh" and "LaTulippe" were used as the test data. Because the scales of the data are not same, we normalized all of the data by applying a Z-transform [34].

4.2. *Determining the optimal values for parameters*

We measured accuracy, sensitivity, and specificity in order to compare the performance of our system with others using the 10-fold cross validation. These measurements are defined as follows:

$$\text{Accuracy} = \frac{\text{The number of correctly predicted samples}}{\text{The number of total samples}}$$

$$\text{Sensitivity} = \frac{\text{The number of predicted tumor samples}}{\text{The number of tumor samples}}$$

$$\text{Specificity} = \frac{\text{The number of correctly predicted normal samples}}{\text{The number of normal samples}}$$

Table 6 shows the 10-fold cross validation results of our system. We performed the experiment by changing parameters, *cor*, *sig*, *nk* and *tk*. Each row represents *cor* and each column represents *sig*. We set the values of *nk* and *tk* at 2, and increased them by 1 until they reached the highest prediction accuracy. In the following, we only showed experimental result having the smallest *nk* and *tk* values in the parameter ranges, representing the highest accuracies. The value in each cell indicates the accuracy. If there are a number of parameter value sets with the highest accuracies, we selected

Table 4 – Microarray data used in the experiment.

Data	Number of genes	Number of normal samples	Number of tumor samples	Total number of samples
Rha [27]				
Paired	7311	127	131	258
Unpaired	7311	0	297	297
Singh [28]	8828	50	52	102
Welsh [29]	8828	9	24	33
LaTulippe [30]	8828	3	23	26
Stuart [31]	8828	50	38	88
Landi [32]	13260	49	58	107
Hou [33]	20826	65	91	156

Please cite this article in press as: E. Shin, et al., TC-VGC: A Tumor Classification System using Variations in Genes' Correlation, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.03.002

Table 6 – The experimental results of our system. We performed the experiment by changing parameters, *cor*, *sig*, *nk* and *tk*. Each row represents *cor* and each column represents *sig*. We set the values of *nk* and *tk* at 2, and increased them by 1 until they reached the highest prediction accuracy. In the following, we only showed experimental result having the smallest *nk* and *tk* values in the parameter ranges, representing the highest accuracies. The value in each cell represents the accuracy and the shaded cell means the optimal parameter value set with the highest weight among such set with the highest accuracies.

				sig				
Dataset	Number of *nk*	Number of *tk*	*cor*	0.8–0.6	0.7–0.5	0.6–0.4	0.5–0.3	0.4–0.2
Rha_paired	2	3	1–0.9	0	0	0	0	0
			0.95–0.85	99.6	99.6	99.6	98.4	99.2
			0.9–0.8	99.6	99.6	99.6	99.6	99.2
			0.85–0.75	100	**100**	100	100	99.6
			0.8–0.7	100	100	99.2	99.6	100
Singh	6	6	1–0.9	0	89.0	91.2	94.1	94.1
			0.95–0.85	86.3	84.3	90.2	91.2	88.2
			0.9–0.8	94.1	93.1	91.2	89.2	92.2
			0.85–0.75	93.1	91.2	90.2	90.2	**95.1**
			0.8–0.7	93.1	91.2	89.2	89.2	92.2
Stuart	8	8	1–0.9	0	0	0	0	**83.8**
			0.95–0.85	0	0	73.8	72.5	76.3
			0.9–0.8	83.8	77.5	68.8	71.3	72.5
			0.85–0.75	78.8	81.3	70.0	76.3	71.3
			0.8–0.7	80.0	76.3	73.8	72.5	63.8
Landi	6	6	1–0.9	0	0	0	0	92.5
			0.95–0.85	0	93.5	96.3	95.3	91.6
			0.9–0.8	93.5	94.4	97.2	**99.1**	91.6
			0.85–0.75	95.3	98.1	94.4	94.4	93.5
			0.8–0.7	92.5	92.5	94.4	92.5	92.5
Hou	4	4	1–0.9	0	0	96.8	93.6	96.2
			0.95–0.85	96.8	96.8	93.6	92.9	91.7
			0.9–0.8	92.3	94.8	91.7	96.8	**98.1**
			0.85–0.75	96.8	96.2	95.5	94.9	91.7
			0.8–0.7	87.2	91.7	92.3	92.9	91.0

the parameter value set with the highest weight among these. The shaded cell means the optimal parameter value set.

4.3. *Accuracy of our cancer classification method*

In the following, we predicted the class labels of independent test samples using the optimal values of parameters which build the classifier with the least number of gene pairs among the parameter ranges obtained from Section 4.2. Table 7 shows the optimal values of parameters, the number of genes, the accuracies, and the real names of the selected gene pairs as the classifier for each of the experiments.

We successfully classified the “Rha”, “Singh”, and “Stuart” datasets with accuracies of 99.6%, 96.6%, and 98.3% with only 5, 12, and 16 gene pairs and 9, 17, and 21 genes, respectively. All of these datasets indicate that our method can build a classifier with fewer number of gene pairs. This ensures very high prediction accuracy.

4.4. *Comparison to other cancer classification methods*

We compared our method with other cancer classification methods to determine whether TC-VGC is competitive. Four current classifiers, k-NN (k-Nearest Neighbor), Naïve Bayes, Random Forest and SVM (Support Vector Machine), were implemented in Weka (Waikato Environment for Knowledge Analysis) [35], a publicly available open-source software package. We also implemented these classifiers after applying the Relief-F [36] and SymmUncert (SU) [37] methods, which are popular feature selection methods. For k-TSP (k-Top Scoring Pair) we used the executables provided by Tan et al. [20].

Experimental works for getting optimal user parameters for each method were made. Table 8 shows the 10-fold cross validation results of above methods. k-NN runs while changing *k*, the number of the neighbor. Random Forest runs while changing *maxDepth*, the depth of the tree. The experiment to obtain the optimal parameter for Naïve Bayes was not done because it does not require a user parameter. SVM runs while changing the *kernel* type. We also applied two feature selection methods while varying the ratio of selected features among all genes between 1 and 20 percentages. Table 9 clearly shows improvement after applying the feature selection methods.

We compared independent test results of our method to existing classification methods. Because “Rha unpaired” data does not include normal samples, the accuracy rate is the same as the sensitivity, and specificity is not applicable. As displayed in Table 10, the performance of our system is comparable or superior to the results from other systems. Moreover,

Table 7 – Experimental results using independent test dataset.

Training dataset	Test dataset	*cor*	*sig*	*nk*	*tk*	# genes	Accuracy	Classifier	
								g_i	g_j
Rha (paired)	Rha (unpaired)	0.85–0.75	0.7–05	2	3	9	99.6%	AI361330 AA977679 AA936799 AA877166 AA490471	AI681730 AI630998 AA877166 AI986457 AA449742
Singh	Welsh + LaTulippe	0.85–0.75	0.4–0.2	6	6	17	96.6%	HIST1H2BM SPN SPAR HIST1H2BM MAGED1 RPS7 ZNF821 ZNF549 MLN CENPI MAPK10 ROM1	BYSL HIST1H2BM PPM1F ACTB HIST1H2BM PPM1F CFD CFD CFD TFAP2C AKR1A1 CFD
Stuart	Welsh + LaTulippe	1–0.9	0.4–0.2	8	8	21	98.3%	IFIT1 IFIT1 IFIT1 IFIT1 IFIT1 IFIT1 SCN2B SMC3 CD1A SIX6 CD1A CD1A CD1A MOCS1 CD1A CD1A	MRPL9 TBC1D30 SMAP1 CLNS1A MMP20 CALCOCO2 GCNT1 GCNT1 UBD PRRG1 SIGLEC6 MUC3A EVI1 KCNJ4 TOP2B VPS41

the sensitivity and specificity rates for the other methods were not well balanced. For example, for SVM the sensitivity was 97.8% but specificity was 33.3% for "Stuart" dataset. Also, for k-NN the sensitivity was 70.2% and specificity was 100% for "Singh" dataset. Low sensitivity could be fatal for a patient if the system erroneously diagnoses a cancer patient as a normal patient.

Consequently, our method successfully classified all microarray data with very high accuracy. It also produced a well-balanced sensitivity and specificity.

Table 8 – 10-Fold cross validation results of the comparison algorithm.

Comparison algorithm	Parameters	Rha	Singh	Stuart
k-NN	1	92.6	**78.4**	65.9
	5	93.4	77.4	**70.4**
	10	93.8	74.5	69.3
	20	**94.2**	72.5	62.5
	30	93.0	69.6	62.5
Naïve Bayes		**96.6**	**96.6**	**86.4**
Random Forest	1	95.0	73.5	67.0
	5	**96.5**	80.4	**69.3**
	10	95.7	**82.4**	65.9
	20	95.7	**82.4**	65.9
	30	95.7	**82.4**	65.9
SVM	Poly	**98.1**	**89.2**	**78.4**
	NormalizedPoly	97.3	86.3	75.0
	puk	84.1	50.1	56.8
	RBF	96.9	83.4	55.7

Please cite this article in press as: E. Shin, et al., TC-VGC: A Tumor Classification System using Variations in Genes' Correlation, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.03.002

Table 9 – Independent results of the comparison algorithm that apply feature selection methods.

Dataset	Feature selection	The ratio of selected features	k-NN	Naïve Bayes	Random Forest	SVM
Rha	Relief-F	0.01	84.2	84.2	88.9	93.9
		0.05	82.2	94.9	91.6	95.9
		0.1	80.3	**96.6**	94.6	96.0
		0.2	**86.9**	**96.6**	**96.6**	**96.3**
	SU	0.01	**87.9**	83.8	95.3	96.0
		0.05	**87.9**	86.2	91.2	97.3
		0.1	85.5	85.9	94.9	97.6
		0.2	86.5	**93.3**	**96.3**	**98.0**
Singh	Relief-F	0.01	**96.6**	89.8	**96.6**	**96.6**
		0.05	**96.6**	91.5	94.9	**96.6**
		0.1	**96.6**	91.5	88.1	**96.6**
		0.2	93.2	**96.6**	94.9	**96.6**
	SU	0.01	94.9	88.1	88.1	94.9
		0.05	**96.6**	94.9	94.9	94.9
		0.1	93.2	96.6	**96.6**	**98.3**
		0.2	89.8	**98.3**	91.5	96.6
Stuart	Relief-F	0.01	**91.5**	88.1	**84.7**	89.8
		0.05	79.7	86.4	76.3	84.7
		0.1	86.4	**89.8**	81.4	**93.2**
		0.2	88.1	**89.8**	**84.7**	79.7
	SU	0.01	83.1	89.8	**93.2**	89.8
		0.05	86.4	**91.5**	81.4	89.8
		0.1	88.1	**91.5**	71.2	**94.9**
		0.2	**89.8**	**91.5**	79.7	89.8

Table 10 – Comparisons of our method with existing classifications.

Algorithms	Rha		Singh			Stuart		
	Accuracy	Sensitivity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
TC-VGC	99.6	99.6	96.6	97.9	91.7	98.3	97.9	100
k-NN	90.2	90.2	76.3	70.2	100	88.1	91.5	75
k-TSP	97.7	97.7	81.4	93.6	33.3	89.8	87.2	100
Naïve Bayes	96.6	96.6	96.6	97.9	91.7	86.4	95.7	75
Random Forest	97.3	97.3	91.5	93.62	83.3	76.3	80.8	58.3
SVM	96.3	96.3	96.6	95.7	100	84.7	97.8	33.3
Relief-F + k-NN	86.9	86.9	96.6	97.9	91.7	91.5	89.4	100
Relief-F + Naïve Bayes	96.6	96.6	96.6	95.7	100	89.8	87.2	100
Relief-F + Random forest	96.6	96.6	96.6	95.7	100	84.7	82.9	91.7
Relief-F + SVM	96.3	96.3	96.6	95.7	100	93.2	97.9	75
SU + k-NN	87.9	87.9	96.6	97.9	91.7	89.8	93.6	75
SU + Naïve Bayes	93.3	93.3	98.3	97.9	100	91.5	89.4	100
SU + Random forest	96.3	96.3	96.6	100	83.3	93.2	93.62	91.7
SU + SVM	98	98	98.3	97.9	100	94.9	95.7	91.7

4.5. *Biological discussion*

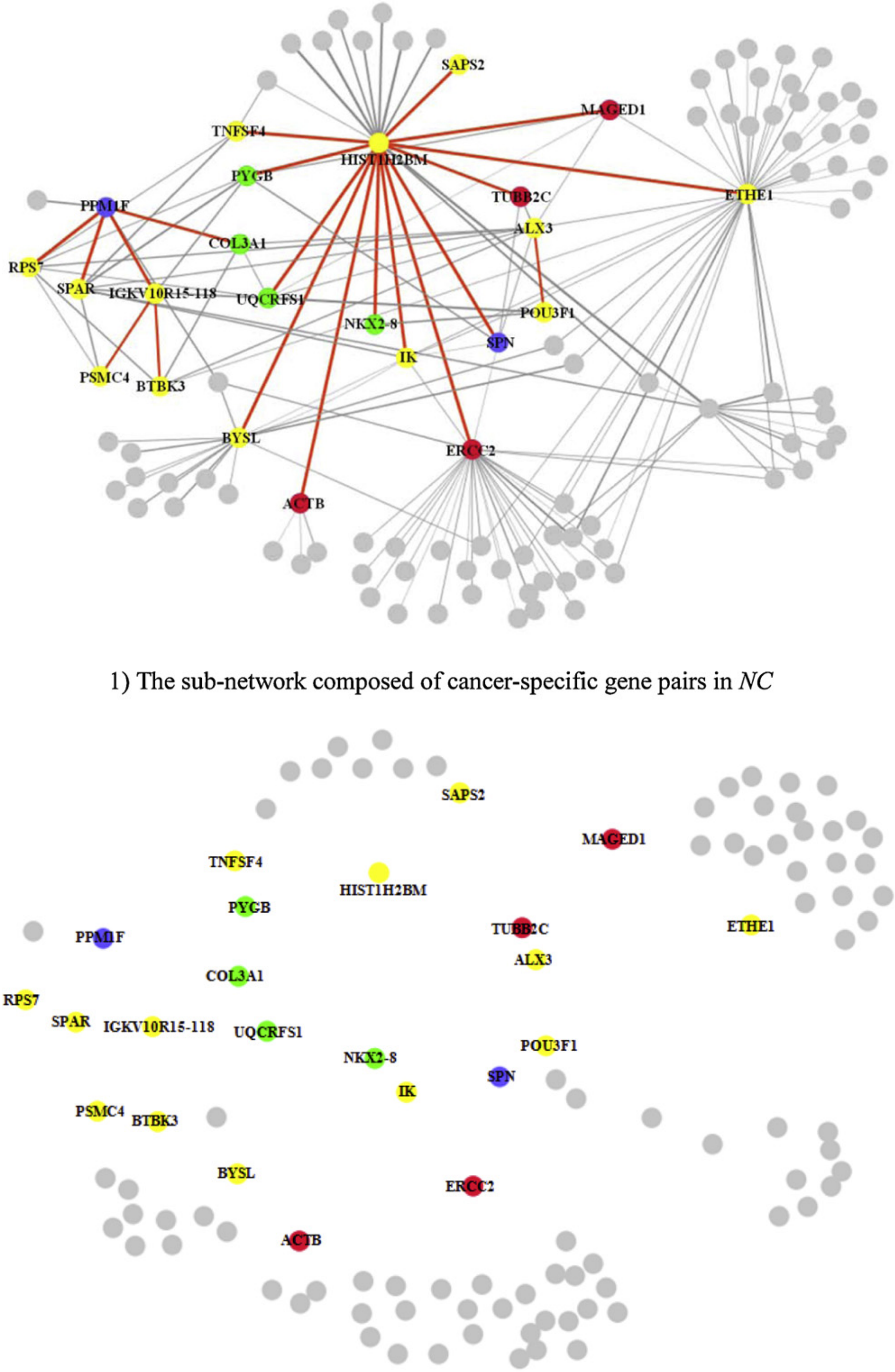
We constructed the network using the Cytoscape software [38]. We also performed literature search for candidate genes using NCBI Entrez [39].

For "Singh" dataset, the network of cancer-specific gene pairs in *NC* contained 58,030 interactions and 4622 genes. Also, the network of cancer-specific gene pairs in *TC* contained 81,196 interactions and 4337 genes. All hub genes and genes linked to those hub genes are present in Supplementary Table 1. Fig. 3 shows the two sub-networks which include top 20 gene pairs of *NC* and *TC*, respectively.

As a result, 48 genes (a total of 40 gene pairs) were listed as candidates. Table 11 shows the genes which have been reported to be associated with cancer in the literature. Among the listed genes in Table 11, the genes that are directly associated with prostate cancer were *SKP2*, *EIF4EBP1*, *TCEB1*, *ERCC2*, *TUBB2C*, *ACTB* and *MAGED1*. Interestingly, several genes in Table 11 have been reported to have a role in cancer, but not previously known to be associated with prostate cancer, such as *TFAP2C*, *SPN*, *PPM1F* and *MB*. Furthermore, many genes among candidate genes have been revealed to be significantly associated with breast cancer (such as *SLC5A5*, *SIX1*, *UQCRFS1*), with lung cancer (such as *NKX2-8*, *PYGB*), with thyroid cancer (such as *SLC5A5*), with cervical cancer (such as *ZNF384*), and with pancreatic cancer (such as *COL3A1*).

As shown in Fig. 3.1, *HIST1H2BM*, a member of the histone H2B family, appeared to be connected to the top 20 gene pairs in highest degree, selected in *NC*. The 8 genes among the 13 genes connected with *HIST1H2BM* are also reported to associate with cancers. *ERCC2* is associated with the develop-

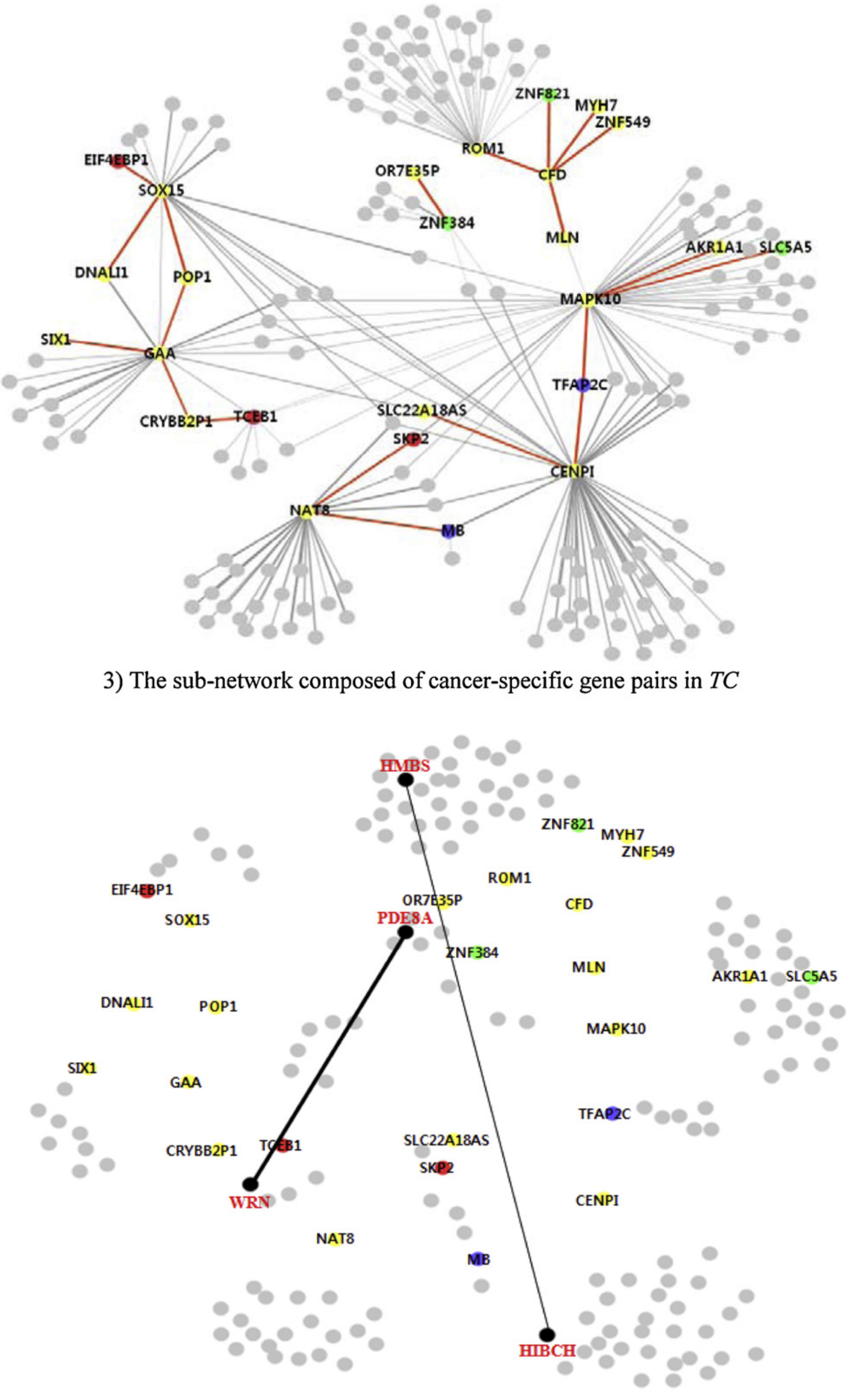
Please cite this article in press as: E. Shin, et al., TC-VGC: A Tumor Classification System using Variations in Genes' Correlation, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.03.002



1) The sub-network composed of cancer-specific gene pairs in *NC*

2) The changes in the interactions of the *NC* genes in cancer state

Fig. 3 – The sub-networks composed of cancer-specific gene pairs in NC and TC. All nodes except for gray nodes are the candidate prostate cancer genes proposed by our method. Red nodes belong to the genes that have been revealed to be directly associated with the prostate cancer. Blue nodes are the genes that have been reported to have a role in cancer. Green nodes are the genes known to be significantly associated with particular type of cancers. Yellow nodes are noble candidate prostate cancer genes which have not been reported to be associated with cancer in the literature. Gray edges indicate the interaction between cancer-specific genes. Red edges indicate the interaction of top 20 gene pairs in NC and TC, respectively. Black edges represent newly created interactions when the interactions of cancer specific genes in one class are applied to the other class. The thicknesses of edges are proportional to their weights. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of the article.)



3) The sub-network composed of cancer-specific gene pairs in *TC*

4) The changes in the interactions of the *TC* genes in the non-cancer state

Fig. 3 – *(Continued).*

Table 11 – Literature summary of the candidate genes.

Genes (NCBI Gene ID)	Literature summary
SKP2 (6502)	Induction of SKP2 may be causally linked with decreased levels of p27 in prostate cancer [40]. SKP2 controls p300-p53 signaling pathways in cancer cells, making SKP2 a potential molecular target for cancer therapy [41].
TCEB1 (6921)	TCEB1 promotes invasion of prostate cancer cells and is involved in development of hormone-refractory prostate cancer [42].
EIF4EBP1 (1978)	Over-expression of EIF4EBP1 is strongly associated with prostate cancer, especially when combined with PTEN and mTOR expression data [43].
ERCC2 (2068)	The ERCC2, a candidate gene for cancer susceptibility regardless of environmental factors [44], may be associated with the development of prostate cancer [45].
MAGED1 (9500)	Phenotype of MAGE-D1were related to confer susceptibility to prostate cancer [46].
TUBB2C (10383)	TUBB2C was found in the nuclei of prostate cancer as well as in adjacent areas of benign prostate hypertrophy [47]. TUBB2C may play a role in assisting rapid cell proliferation [48].
ACTB (60)	At the ACTB promoter, increased H3K4me2 was observed in human prostate cancer cell lines [49].
TFAP2C (7022)	Down-modulation of TFAP2C expression in tumor cells by RNA interference (RNAi) led to enhanced tumor growth and reduced chemotherapy-induced cell death, as well as migration and invasion. Most of these biological modulations were rescued by TFAP2C over-expression [50].
SPN (6693)	The expression of SPN causes the induction of functionally active p53 protein, and Over-expression of SPN causes activation of the tumor suppressor proteins p53 and ARF 1 [51].
PPM1F (9647)	PPM1F is expressed in a wide variety of tumor cell lines, and regulates cancer cell motility and invasiveness [52].
MB (4151)	MB is induced by a variety of signals associated with tumor progression. In other words, MB in tumors including breast, lung, ovary, and colon carcinomas, is expressed at high levels from the earliest stages of cancer development [53].
SLC5A5 (6528)	SLC5A5 expression is also prevalent in breast cancer brain metastases [54].
SIX1 (6495)	Breast cancer patients whose tumors over-expressed SIX had a shortened time to relapse and metastasis and an overall decrease in survival [55]. SIX1 over-expression reinstated an embryonic pathway of proliferation in breast cancer by up-regulating cyclin A1 [56].
UQCRFS1 (7386)	UQCRFS1 gene appears to be involved in development of more aggressive phenotype of breast cancer [57].
NKX2-8 (26257)	Continuous expression of NKX2-8 is also essential to the tumor maintenance of amplified squamous-cell carcinomas cells [58].
PYGB (5834)	PYGB is expressed in NSCLC (non-small-cell lung carcinoma), and is an independent poor prognostic factor [59].
ZNF384 (171017)	ZNF384 expression in pelvic lymph nodes and primary tumors in early stage cervical carcinomas have a correlation with clinical outcome [60].
COL3A1(1281)	The expression of COL3A1 was significantly higher in pancreatic cancer [61].

ment of prostate cancer [45]. *TUBB2C* have been found in the nuclei of prostate cancer [47] and may play a role in assisting rapid cell proliferation [48]. Phenotype of *MAGE-D1* is related to confer susceptibility to prostate cancer [46]. *SPN* reduces the risk of cancer by suppressing the cancer cell because the expression of *SPN* causes the induction of functionally active *p53*, the tumor suppressor proteins [51]. Much to our interest, Shema et al. [62] reported that deregulation of histone H2B monoubiquitination may contribute to cancer development. Intriguing finding is that *HIST1H2BM* which is highly connected to cancer genes is not reported to have any association with cancer, however is thought to be a prostate cancer gene. Note that all interactions in *NC* disappeared when cancer occurs (Fig. 3.2). This suggests that the anomalies of these cancer specific genes are related to the prostate cancer.

As shown Fig. 3.3, *MAPK10* was connected with the 6 genes among the 25 candidate genes in *TC* and 4 genes among the connected genes were already identified in the literature that associated with the cancers. We explained the functions of genes as follows: in prostate cancer, *SKP2* decrease levels of *p27*, which results in reducing cell proliferation [40,63]. *SKP2*, therefore, increased the proliferation and tumorigenic potential of a prostate cancer cell line [64]. *TCEB1* promotes invasion of prostate cancer cells, is involved in development of hormone-refractory prostate cancer [42]. *TFAP2C* regulates tumor growth and chemotherapy-induced cell death, as well as migration and invasion [50]. Moreover, all the genes connected with *MAPK10* were positively correlated with *MAPK10* only in prostate tumor samples. Recently, *JNK* proteins are encoded by three genes (*MAPK8*, *MAPK9* and *MAPK10*), which were revealed to associate with cancers in human [65]. Although *MAPK10* is not reported to have any association with cancer, it is highly connected to cancer genes, and thought to be a prostate cancer gene. Contrary to Fig. 3.2, we can see that all interactions in *TC* disappeared and new interactions between these genes were added in non-cancer state in Fig. 3.4.

As we have stated early, an interaction of which two genes are activated or repressed by a transcription factor (TF) activates or represses two genes, A and B, simultaneously, the expression levels of A and B reveal that they have high correlation. If gene B is affected by a specific disease, the transcription factor continues to activate or repress gene A, while it can no longer influence gene B. In the *NC* sub-network, we found TF that regulates the cancer specific genes. Fig. 4 shows the sub-network of TF and its targets, and cancer specific genes linked to the target genes. In normal state, TF *ARNT* activates *ID2* and *MCL1* simultaneously, and there exists strong interaction between *ID2* and *MCL1* (Fig. 4.1). However, almost all interactions, including *ID2-MCL1* disappeared in cancer state (Fig. 4.2). Coppe et al. discovered that *ID2* were up-regulated during human prostate cancer progression *in vivo* and were

Please cite this article in press as: E. Shin, et al., TC-VGC: A Tumor Classification System using Variations in Genes' Correlation, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.03.002

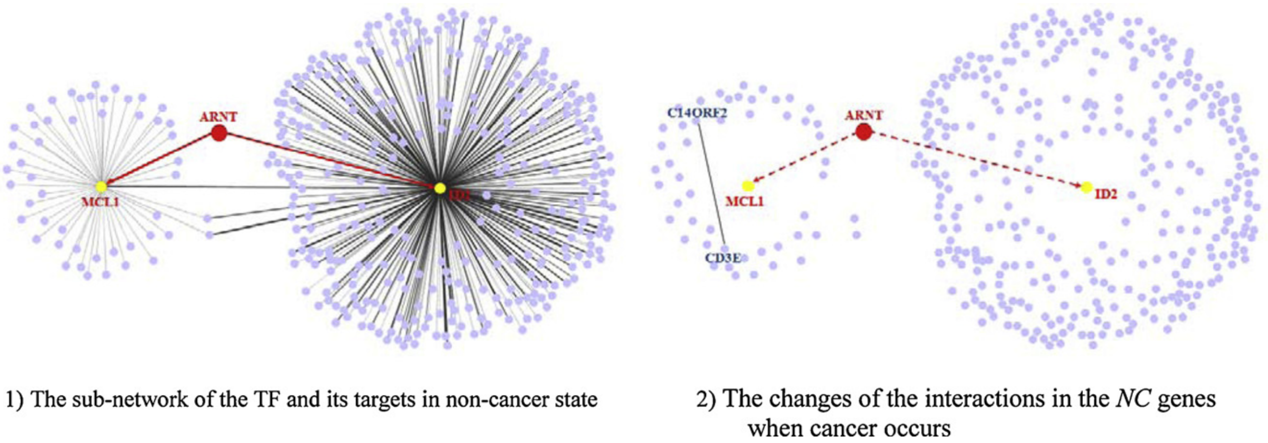

ARNT
MCL1
ID2
C14ORF2
ARNT
MCL1
ID2
CD3E
1) The sub-network of the TF and its targets in non-cancer state
2) The changes of the interactions in the *NC* genes when cancer occurs

Fig. 4 – The sub-network of the transcription factor (TF) and its targets, and cancer specific genes linked to the target genes. Red node is transcription factor (TF). Yellow genes are regulated by TF. The undirected edges represent the interactions between cancer specific genes. The directed edges in red represent the relationship between a TF and its target. The thicknesses of edges are proportional to their weights. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of the article.)

overexpressed in highly aggressive prostate cancer cells [66]. A few studies had explored *MCL1* involvement in prostate carcinogenesis [67]. Therefore we propose that *ID2* and *MCL1* genes could serve as molecular markers of prostate cancer.

There are 30 additional genes which have not been reported to be associated with prostate cancer, but they also have relations with other cancer genes, in this study. Thus, they offer excellent potential to be prostate cancer genes.

5. Conclusion

In this study, we proposed a new cancer classification method using gene pairs with distinguished variations in the gene–gene correlations between two groups of samples. This concept can be explained in the regulatory mechanisms of gene expression. Many genes can be regulated by one transcription factor which controls the expression level of the gene by activating or repressing it. For example, when a transcription factor activates two genes A and B simultaneously, the expression levels of A and B show a positive correlation. If gene B is then affected by a specific disease, the transcription factor can no longer influence gene B, but it continues to activate gene A. The signs of the normal and tumor sample correlations are opposite when two genes are activated by same transcription factor in the normal class, and only one of the two genes in the tumor class is activated by this transcription factor. This situation also happens when one gene is activated and another gene is repressed by same transcription factor in the normal class, and two genes are activated by the same transcription factor in the tumor class. For example, transcription factor *ARNT* activates *MCL1* and *ID2* simultaneously. This leads to strong interaction between *MCL1* and *ID2*, whereas, this interaction disappeared in the tumor state (Fig. 4).

To illustrate and evaluate the efficiency of TC-VGC, we used five real microarray datasets: Rha, Singh, Welsh, LaTulippe, and Stuart. We compared our method with five existing methods: SVM, k-NN, Random Forest, Naïve Bayes, and k-TSP. Compared to current methods, our method builds a classifier that has fewer numbers of gene pairs. It can more accurately classify all datasets. Also, the genes which belong to the set of cancer-specific gene pairs identified by TC-VGC have a higher chance of being clustered within a gene regulatory network, and these clustered gene-regulatory regions can be cancer specific. This annotated network can be expected to clarify the unknown genes which cause various types of tumors.

We expect that the TC-VGC system can be used as an effective classification tool for microarray data with a limited sample size and a large number of genes. We also are confident that our systematic approach will be useful for finding genes of interest from many cancer types. It could be potentially used for revealing new regulatory patterns which are specific to a cancer group.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MEST) (No. 2011-0005154).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2011.03.002.

REFERENCES

[1] D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer, J.M. Trent, Expression profiling using cDNA microarray, Nature Genetics Supplement 21 (1999) p.10–14.
[2] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, et al., Minimum information about a microarray

experiment (MIAME)-toward standards for microarray data, Nature Genetics 29 (2001) 365–371.

[3] A. Bernal, K. Crammer, A. Hatzigeorgiou, F. Pereira, Global discriminative learning for higher-accuracy computational gene prediction, PLoS Computational Biology 3 (2007) 488–497.

[4] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (2000) 906–914.

[5] L. Nanni, A. Lumini, An ensemble of K-local hyperplanes for predicting protein–protein interactions, Bioinformatics 22 (2006) 1207–1210.

[6] M. Kamber, J. Han, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[7] M. Pirooznia, J.Y. Yang, M.Q. Yang, Y. Deng, A comparative study of different machine learning methods on microarray gene expression data, BMC Genomics (2008).

[8] Y. Wang, L.V. Tetko, M.A. Hall, E. Frankb, A. Facius, K.F.X. Mayer, F.W. Mewes, Gene selection from microarray data for cancer classification—a machine learning approach, Computational Biology and Chemistry (2005) 37–45.

[9] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning (2001) 389–422.

[10] K.B. Duan, J.C. Rajapakse, H. Wang, F. Azuaje, Multiple SVM-RFE for gene selection in cancer classification with expression data, IEEE Transactions on NanoBioscience 4 (2005) 228–234.

[11] F. Pan, B. Wang, W. Perrizzo, Comprehensive vertical sample-based k-NN/LSVM classification for gene expression analysis, Journal of Biomedical Informatics 37 (2004) 241–249.

[12] R. Diaz-Uriarte, S. Alvarez de Andres, Gene selection and classification of microarray data using random forest, BMC Bioinformatics 7 (2006).

[13] A.A. Albrecht, Stochastic local search for the feature set problem, with applications to microarray data, Applied Mathematics and Computation 183 (2) (2006) 1148–1164.

[14] I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recognition 43 (1) (2010) 5–13.

[15] J. Jaeger, R. Sengupta, W.L. Ruzzo, Improved gene selection for classification of microarrays, Pacific Symposium on Biocomputing 8 (2003) 53–64.

[16] R. Berretta, W. Costa, P. Moscato, Combinatorial optimization models for finding genetic signatures from gene expression datasets, Bioinformatics 453 (2008) 363–377.

[17] E.R. Dougherty, M. Brun, On the number of close-to-optimal feature sets, Cancer Informatics (2006) 189–196.

[18] I.T. Jolliffe, Principal Component Analysis, Springer Series in Statistics, 2002, p. 28.

[19] G. Hori, M. Inoue, S.I. Nishimura, H. Nakahara, Blind gene classification—an application of a signal separation method, Genome Informatics 12 (2001) 255–256.

[20] A. Tan, D. Naiman, L. Xu, R. Winslow, D. Geman, Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics 21 (2005) 3896–3904.

[21] X. Wei, K.C. Li, Exploring the within- and between-class correlation distributions for tumor classification, PNAS 107 (15) (2010) 6737–6742.

[22] M.G. Ravetti, O.A. Rosso, R. Berretta, P. Moscato, Uncovering molecular biomarkers that correlate cognitive decline with the changes of hippocampus'gene expression profiles in Alzheimer's disease, PLoS 5 (4) (2010) e10153.

[23] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, et al., Dynamic modularity in protein interaction networks predicts breast cancer outcome, Nature Biotechnology 27 (2009) 199–204.

[24] E.L. Lehmann, H.J.M. D'Abrera, Nonparametrics: Statistical Methods Based on Ranks, Prentice-Hall, Englewood Cliffs, NJ, 1998, pp. 292, 300, and 323.

[25] X. He, J. Zhang, Why do hubs tend to be essential in protein networks? PLoS Genetics 2 (6) (2006).

[26] S. Horvath, J. Dong and Y. Yip, Connectivity, Module-Conformity, and Significance: Understanding Gene Co-Expression Network Methods, UCLA Technical Report, 2006.

[27] Yonsei University, Cancer Metastasis Research Center, Yonsei University College of Medicine, South Korea.

[28] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209.

[29] J.B. Welsh, L.M. Sapinoso, A.I. Su, S.G. Kern, J. Wang-Rodriguez, C.A. Moskaluk, Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer, Cancer Research 61 (2001) 5974–5978.

[30] E. LaTulippe, J. Satagopan, A. Smith, H. Scher, P. Scardino, V. Reuter, Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease, Cancer Research 62 (2002) 4499–4506.

[31] R. Stuart, W. Wachsman, C.C. Berry, J. Wang-Rodriguez, L. Wasserman, et al., In silico dissection of cell-type-associated patterns of gene expression in prostate cancer, Proceedings of National Academy of Sciences of the United States of America 101 (2004) 615–620.

[32] M.T. Landi, T. Dracheva, M. Rotunno, J.D. Figueroa, H. Liu, et al., Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival, PLoS One 3 (2) (2008) e1651.

[33] J. Hou, J. Aerts, B.D. Hamer, W.V. Ijcken, M.D. Bakker, et al., Gene expression-based classification of non-small cell lung carcinomas and survival prediction, PLoS One 5 (4) (2010) e10312.

[34] F.N. David, The moments of the z and F distributions, Biometrika 36 (1949) 394–403.

[35] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, San Francisco, 2005.

[36] I. Kononenko, Estimating attributes: analysis and extensions of relief, Proceedings of ECML 784 (1994) 171–182.

[37] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, 1988.

[38] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Research 13 (2003) 2498–2504.

[39] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, et al., Entrez Gene: gene-centered information at NCBI, Nucleic Acids Research 33 (2005) 54–58.

[40] H. Wang, D. Sun, P. Ji, J. Mohler, L. Zhu, An AR-Skp2 pathway for proliferation of androgen-dependent prostate-cancer cells, Journal of Cell Science 121 (2008) 2578–2587.

[41] M. Kitagawa, S.H. Lee, F. McCormick, Skp2 suppresses p53-dependent apoptosis by inhibiting p300, Molecular Cell 29 (2) (2008) 217–231.

[42] S.E. Jalava, K.P. Porkka, H.E. Rauhala, J. Isotalo, T.L. Tammela, et al., TCEB1 promotes invasion of prostate cancer cells, International Journal of Cancer 124 (1) (2009) 95–102.

[43] C.L. Kremer, R.R. Klein, J. Mendelson, W. Browne, L.K. Samadzedeh, et al., Expression of mTOR signaling pathway

Please cite this article in press as: E. Shin, et al., TC-VGC: A Tumor Classification System using Variations in Genes' Correlation, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.03.002

markers in prostate cancer progression, The Prostate 66 (11) (2006) 1203–1212.

[44] D.T. Bau, H.C. Wu, C.F. Chiu, C.C. Lin, C.M. Hsu, et al., Association of XPD polymorphisms with prostate cancer in Taiwanese patients, Anticancer Research 27 (4C) (2007) 2893–2896.

[45] F. Wang, D. Chang, F.L. Hu, H. Sui, B. Han, et al., DNA repair gene XPD polymorphisms and cancer risk: a meta-analysis based on 56 case-control studies, Cancer Epidemiology, Biomarkers & Prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 17 (3) (2008) 507–517.

[46] J. Gudmundsson, P. Sulem, T. Rafnar1, J.T. Bergthorsson, A. Manolescu, et al., Common sequence variants on 2p15 and Xp1122 confer susceptibility to prostate cancer, Nature Genetics 40 (2008) 281–283.

[47] S. Ranganathan, H. Salazar, C.A. Benetatos, G.R. Hudes, Immunohistochemical analysis of β-tubulin isotypes in human prostate carcinoma and benign prostate hypertrophy, Prostate 30 (1997) 263–268.

[48] C. Walss-Bass, K. Xu, S. David, A. Fellous, R.F. Luduena, Occurrence of nuclear beta(II)-tubulin in cultured cells, Cell and Tissue Research 308 (2) (2002) p.215–223.

[49] S. Yamashita, S. Takahashi, N. McDonell, N. Watanabe, T. Niwa, et al., Methylation silencing of transforming growth factor-β receptor type II in rat prostate cancers, Cancer Research 68 (2008) 2112–2121.

[50] F. Orso, E. Penna, D. Cimino, E. Astanina, F. Maione, et al., AP-2alpha and AP-2gamma regulate tumor progression via specific genetic programs, The FASEB Journal 22 (8) (2008) 2702–2714.

[51] L. Kadaja, S. Laos, T. Maimets, Overexpression of leukocyte marker CD43 causes activation of the tumor suppressor proteins p53 and ARF, Oncogene 23 (14) (2004) 2523–2530.

[52] A. Sila, H. Chan, A.X. Loh, H.Q. Phang, E.T. Wong, et al., The POPX2 phosphatase regulates cancer cell motility and invasiveness, Cell Cycle 9 (1) (2010) 179–187.

[53] S.E. Flonta, S. Arena, A. Pisacane, P. Michieli, A. Bardelli, Expression and functional regulation of myoglobin in epithelial cancers, The American Journal of Pathology 175 (1) (2009) 201–206.

[54] C. Renier, H. Vogel, O. Offor, C. Yao, I. Wapnir, Breast cancer brain metastases express the sodium iodide symporter, Journal of Neuro-Oncology 96 (3) (2010) 331–336.

[55] D.S. Micalizzi, K.L. Christensen, P. Jedlicka, R.D. Coletta, A.E. Baron, et al., The Six1 homeoprotein induces human mammary carcinoma cells to undergo epithelial–mesenchymal transition and metastasis in mice through increasing TGF-beta signaling, The Journal of Clinical Investigation 119 (9) (2009) 2678–2690.

[56] E.L. McCoy, R. Iwanaga, P. Jedlicka, N.S. Abbey, L.A. Chodosh, et al., Six1 expands the mouse mammary epithelial stem/progenitor cell pool and induces mammary tumors that undergo epithelial–mesenchymal transition, The Journal of Clinical Investigation 119 (9) (2009) 2663–2677.

[57] Y. Ohashi, S.J. Kaneko, T.E. Cupples, S.R. Young, Ubiquinol cytochrome c reductase (UQCRFS1) gene amplification in primary breast cancer core biopsy samples, Gynecology Oncology 93 (1) (2004) 54–58.

[58] J. Kendall, Q. Liu, A. Bakleh, A. Krasnitz, K.C.Q. Nguyen, et al., Oncogenic cooperation and coamplification of developmental transcription factor genes in lung cancer, Proceedings of the National Academy of Sciences of the United States of America (2007).

[59] M.K. Lee, J.H. Kim, C.H. Lee, J.M. Kim, C.D. Kang, et al., Clinicopathological significance of BGP expression in non-small-cell lung carcinoma: relationship with histological type, microvessel density and patients' survival, Pathology 38 (6) (2006) 555–560.

[60] M. Graflund, B. Sorbe, M. Karlsson, MIB-1, p53, bcl-2, and WAF-1 expression in pelvic lymph nodes and primary tumors in early stage cervical carcinomas: correlation with clinical outcome, International Journal of Oncology 20 (5) (2002) 1041–1047.

[61] E. Ryschich, A. Khamidjanov, V. Kerkadze, Vachtang, M.W. Buchler, et al., Promotion of tumor cell migration by extracellular matrix proteins in human pancreatic cancer, Pancreas 38 (7) (2009) 804–810.

[62] E. Shema, I. Tirosh, Y. Aylon, J. Huang, C. Ye, et al., The histone H2B-specific ubiquitin ligase RNF20/hBRE1 acts as a putative tumor suppressor through selective regulation of gene expression, Genes & Development 22 (2008) 2664–2676.

[63] P. Wang, Q. Ma, J.D. Luo, B. Liu, F.Q. Tan, et al., Nkx31 and p27KIP1 cooperate in proliferation inhibition and apoptosis induction in human androgen-independent prostate cancer cells, Cancer Investigation 27 (2009) 369–375.

[64] E.J. Chenette, Cell cycle: Akt Skps through, Nature Reviews Cancer 9 (2009) 316–317.

[65] E.F. Wagner, A.R. Nebreda, Signal integration by JNK and p38 MAPK pathways in cancer development, Nature Reviews Cancer 9 (2009) 537–549.

[66] J.P. Coppe, Y. Itahana, D.H. Moore, J.L. Bennington, P.Y. Desprez, Id-1 and Id-2 proteins as molecular markers for human prostate cancer progression, Clinical Cancer Research 10 (2004) 2044–2051.

[67] I.T. Cavarretta, H. Neuwirt, M.H. Zaki, H. Steiner, A. Hobisch, et al., Mcl-1 is Regulated by IL-6 and Mediates the Survival Activity of the Cytokine in a Model of Late Stage Prostate Carcinoma, Springer, 2008, pp. 547–555.

[68] K.U. Kim amd, S.B. Cho, Ensemble classifiers based on correlation analysis for DNA microarray classification, Neurocomputing 70 (2006) 187–199.

Please cite this article in press as: E. Shin, et al., TC-VGC: A Tumor Classification System using Variations in Genes' Correlation, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.03.002