Microarray-based Classification of Histopathologic Responses of Locally Advanced Rectal Carcinomas to Neoadjuvant Radiochemotherapy Treatment

Lokal İleri Rektum Karsinomlarının Neoadjuvan Radyokemoterapi Tedavisine Histopatolojik Yanıtlarının Mikrodizin-Tabanlı Sınıflandırması

Waheed Babatunde YAHYA,[a]
Robert ROSENBERG,[b]
Kurt ULM[a]

[a]Institute for Medical Statistics and Epidemiology,
Technical University of Munich,
[b]Department of Surgery,
Technical University of Munich,
Munich, GERMANY

Geliş Tarihi/*Received:* 26.12.2013
Kabul Tarihi/*Accepted:* 06.02.2014

Yazışma Adresi/*Correspondence:*
Waheed Babatunde YAHYA
Department of Statistics,
University of Ilorin,
P.M.B. 1515, Ilorin,
NİJERYA/NIGERIA
wbyahya@unilorin.edu.ng

ABSTRACT Objective: This paper aims to present preoperative prediction of responses of locally advanced rectal cancer (LARC) patients to neoadjuvant radiochemotherapy (NRC) treatments using their gene expression profiles. **Materials and Methods:** Expression profiles of 24,026 genes were generated on 43 LARC samples using microarray technology. The 43 samples contained two histopathologic response groups of 14 responders and 29 non-responders to NRC treatment. Using a novel k sequential feature selection and classification (k-SS) method, a subset of gene signatures whose expression levels are correlated with the two response groups was selected for pre-therapeutic prediction of the LARC patients. **Results:** Five informative gene chips whose expression profiles are strongly correlated with the two histopathologic response groups of the LARC samples were selected. Some of these are protein encoding genes like SF3A1 that functions in the nucleus and helps to convert pre-messenger RNA to mRNA. The average bootstrap.632+ prediction accuracy based on these genes is about 98%. Results from cluster analysis and PCA showed good discrimination of the two clinical response groups. Although, all the five out-of-bag samples were correctly classified, the classification of the entire 43 LARC samples in a 10-fold cross-validation yielded 86% and 93% correct classifications of responders and non-responders respectively. **Conclusion:** Results from this study equally showed that preoperative prediction of responses of LARC patients to NRC treatment using gene expression profiles is possible. This shall immensely help in clinical diagnosis and treatment of locally advanced rectal carcinomas. Nevertheless, validation studies with larger patient groups might be desirable in future.

Key Words: *k*-SS method; locally advanced rectal carcinomas; neoadjuvant radiochemotherapy; gene expression profiles; misclassification error rate

ÖZET Amaç: Bu çalışma, lokal ileri rektum kanserli (locally advanced rectal cancer-LARC) hastaların neoadjuvan radyokemoterapi (neoadjuvant radiochemotherapy-NRC) tedavisine yanıtlarının, gen ekpresyon profillerini kullanarak, pre-operatif tahminlerinin sunulmasını amaçlamaktadır. **Gereç ve Yöntemler:** Mikrodizin teknolojisi kullanılarak, 24026 adet genin ekpresyon profilleri 43 LARC örneği üzerinde türetilmiştir. Bu 43 örnek, NRC tedavisine yanıt veren 14 ve yanıt vermeyen 29 kişiden oluşan, iki histopatolojik yanıt grubunu içermektedir. LARC hastalarının tedavi öncesi tahmini için, yeni k ardışık özellik seçim ve sınıflama (k sequential feature selection and classification-k-SS) yöntemi kullanılarak, ekspresyon düzeyleri iki yanıt grubu ile ilişkili olan gen imzalarının bir alt kümesi seçilmiştir. **Bulgular:** Ekspresyon düzeyleri, seçilen LARC örneklerinin iki histopatolojik grubu ile güçlü düzeyde ilişkili olan 5 adet açıklayıcı gen çipi seçilmiştir. Bunlardan bazıları nukleusta işlevini yerine getiren ve pre-mesajcı RNA'yı mRNA'ya dönüştürmeye yardım eden, SF3A1 gibi protein kodlayan genlerdir. Bu genlere dayalı ortalama bootstrap 632+ tahmin doğruluğu, yaklaşık olarak %98'dir. Kümeleme analizi ve temel bileşenler analizi sonuçları, iki klinik yanıt grubunun iyi bir şekilde ayrıldığını göstermiştir. Bununla birlikte, beş çanta dışı örneklemlerin tümü doğru sınıflandırılmış olsalar da, 10-katlı çapraz geçerlilikte 43 LARC örneklerinin tümü yanıt verenler ve vermeyenler için sırasıyla %86 ve %93 doğru sınıflandırma yüzdeleri vermiştir. **Sonuç:** Bu çalışmanın sonuçları, eşit olarak LARC hastalarının NRC tedavisine yanıtlarının pre-operatif tahminlerinin, gen ekspresyon profillerini kullanarak yapılmasının mümkün olduğunu göstermektedir. Bu, klinik tanı koymada ve lokal ileri rektum karsinomlarının tedavisinde son derece yardımcı olabilir. Yine de, ileride daha büyük hasta grupları ile doğrulama çalışmalarının yapılması arzu edilebilir.

Anahtar Kelimeler: *k*-SS yöntemi; lokal ileri rektum karsinomları; neoadjuvan radyokemoterapi; gen ekspresyon profilleri; hatalı sınıflandırma oranı

Turkiye Klinikleri J Biostat 2014;6(1):8-23

The recent breakthrough in microarray technology[1] which allows monitoring the expression profiles of several thousands of genes, simultaneously, has been a huge success. This has motivated a number of works in genomic research and many interesting results have emerged through gene expression profiling of different malignancies.[2-4] The management and treatment of cancer tumours depends largely on the specific subtype of cancer.[5] Identification of such cancer subtype might require thorough examination of the tumour cells in a microscope and some other clinical parameters. But studies have shown that cancer might be discovered earlier with microarray analysis than with clinical methods.[6,7] Therefore, another alternative method for identifying tumours with different biology is to monitor the molecular characteristics of cancer through the gene expression profiles measured on the sample specimens.[8-10] Such characteristics could then be used to identify and classify the tumour conditions of the tissue samples.[5,11]

The classification of clinical status or other outcome of interest of biological samples using their gene expression profiles has been given prominent attention in many microarray studies in the recent past. A particular instance of this is the study of Price et al.[12] in which the two tumour types of 68 *Gastrointestinal stromal tumor* (GIST) and *Leiomyosarcomas* (LMSs) tissue samples were identified using their gene expression signatures. Also, in the lung cancer microarray classification work of Gordon et al.,[13] human genome were used to discriminate between the two cancer tumour groups of *malignant pleural mesothelioma* (MPM) and *adenocarcinoma* (ADCA) of the lung.

In the treatment of locally advanced rectal carcinomas however, neoadjuvant radiochemotherapy (NRC) treatment has been clinically identified as a standard therapy apart from primary surgery.[14-16] It has been reported by Samel et al.[17] that NRC induces tumour remission and can prolong survival time of patients with locally advanced adenocarcinoma of the oesophagogastric junction. However, the histopathological and clinical response to NRC treatment has been reported to be relatively lower ranging between 30% to 50% of the patients.[18,19] In other words, only about 30 to 50% of locally advanced rectal cancer (LARC) patients, hereafter tagged the ‘responders’, do respond positively to NRC treatments while about 50 to 70% of them do not normally respond to neoadjuvant treatments hereafter tagged the ‘non-responders’.

If it is possible to identify pre-therapeutically LARC patients that would not benefit from NRC treatments (the non-responders), this would help immensely to place this group of patients on alternative treatment regimens that would be beneficial to them like primary surgery before the cancer tumour could metastasize. This would also protect the patients against possible adverse effects of radiochemotherapy treatments that might not really help to alleviate their health conditions. On the other hand, early identification of LARC patients that would benefit maximally from neoadjuvant treatments (the responders) would save this group of patients the risks of primary surgery in the treatment of their rectal carcinomas. For this group of patients, their treatment regimens might just be limited only to NRC applications.

More generally, the biology of cancer tumours is not identical even within the same type of cell in the same organ. This has led to the development of different therapy measures to address various species of sub-cancer types. In the diagnosis and treatment of locally advanced rectal carcinomas however, a number of clinical methods are often being adopted to predict the responses of LARC patients to NRC treatments.[20,21] Unfortunately, some of these methods have been reported to take a considerable longer period of time before early responses could be detected among the patients that would actually respond to NRC treatment.[19] This allows for the making of early decision on the choice of treatment that would be of benefit to the patients difficult before the cancer tumour could advance.

Due to the above limitations of the clinical methods, a viable alternative as earlier pointed out, is the use of the gene expression profiles to monitor the molecular characteristics of the rectal cancer tumour. This allows for the prediction of responses to NRC treatment. The presence of rectal carcinomas and the responses to NRC treatment might be detected earlier through this microarray base procedure than through the clinical methods.[22-24] All these motivated the present work to employ an efficient microarray based classification method to identify and select the relevant gene expression signatures for proper pretherapeutic prediction of histopathologic responses of LARC samples to NRC treatment. Such identified genetic signatures would provide efficient clinical tools for further chemotherapeutic measures in the treatment of locally advanced rectal carcinomas. The method of analysis employ for this purpose is the *k* sequential feature selection (*k*-SS) method.[25-27]

The *k*-SS method is a fast and efficient algorithm for feature selection and response class prediction in any binary response microarray tumour classification problem. The prediction accuracy of this method is assessed by the estimated average *misclassification error rate* (MER) and by some other performance indices as presented in later sections.

A brief overview of the *k*-SS method as employed here is presented in section two. Further details on this method have been presented in some of the earlier studies.[25-27]

1. MATERIAL AND METHODS

The microarray data discussed in this paper emanated from the clinical study carried out in the Department of Surgery, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany on preoperative endoscopic biopsy specimens of 43 patients with *locally advanced rectal carcinomas*. All the 43 patients underwent neoadjuvant radiochemotherapy treatment followed by surgical resection. Out of these 43 LARC patients, 14 of them (about 33%) demonstrated histopathologic response to NRC treatment and they were therefore tagged the '*responders*' while the remaining 29 patients (about 67%) did not respond to the neoadjuvant therapy treatment and they were tagged the '*non-responders*'.

According to the tumour regression classification of Becker et al.[28] as adapted in Rimkus et al.,[29] the histopathologic responders are defined as rectal cancer sample specimens with less than 10% of viable tumour cells after receiving NRC treatment while the non-responders are those still having at least 10% or more of viable tumour cells after receiving neoadjuvant treatments. Summary of some clinical characteristic of all the 43 LARC patients are presented in Table 1. Further details on these have been presented elsewhere.[29]

Using standard protocols, the expression profiles of 24,026 Affymetrix Gene Chips were meas-

TABLE 1: The description of biological characteristics of the 43 Locally Advanced Rectal Cancer patients employed in this study. Note: s.d. in the parenthesis indicates the standard deviation of the ages.

Clinical Characteristics of All The 43 Locally Advanced Rectal Cancer Patients		
Total Number of Biological Sample		43
Sex	Male	31 (%72)
	Female	12 (%28)
Age	Mean (s.d.)	59 years (6.3 years)
	Minimum	32 years
	Maximum	74 years
Histopathologic Responses to NRC Treatment	Responders	14 (%33)
	Non-responders	29 (67)
Tumour Regression Grades[28]	I (0% - <10% viable tumour cells)	14 (%33)
	II (10% - 50% viable tumour cells)	18 (%42)
	III (>50% viable tumour cells)	11 (%25)

ured on each of the 43 tissue samples of rectal carcinomas. Details of the clinical procedures followed for GeneChip hybridization and isolation as well as amplification of ribonucleic acid (RNA) are discussed in Rimkus et al.[29]

To predict the responses of the 43 sample specimens to NRC treatment efficiently through their gene expression profiles, it is important to first identify the few relevant genes subset among the 24,026 observed gene signatures whose expression levels are jointly correlated with the two histopathologic responses of these samples. This is very important, because not all the 24,026 observed gene chips would possess the required expression profiles that are predictive of the clinical responses of rectal carcinomas to NRC treatment. The identified and selected relevant genetic signatures can then be used to predict the responses of LARC patients to NRC treatment.

1.1. BRIEF OVERVIEW OF THE *k*-SS METHOD

The k sequential feature selection and response class prediction (k-SS) method[25] is a fast and flexible algorithm that sequentially selects relevant subset of gene expression profiles for classifying tumour conditions of biological samples in any binary response microarray classification problem. The k-SS method is a stepwise feature selection procedure that combines the selection of relevant genes subset with the prediction of cancer status of the tissue samples using the selected genes.

Let $X = (X_{i1}, \dots, X_{ip})'$ be $n \times p$ data matrix of p gene expression profiles of n tissue samples ($n < p$) with binary response groups$y_i \in \{0,1\}, i = 1, \dots, n$. As used here, the response variable y_i is coded such that $y_i = 0$ if the i^{th} sample specimen is a responder and $y_i = 1$ if the i^{th} sample is a non-responder to NRC treatments.

The k-SS procedure begins by fitting logistic regression model[30,31] on each gene variable $x_j, j=1, \dots, p,$ through the bootstrap.632+ cross-validation scheme[32-35] in order to examine the discriminatory power of each gene. By bootstrap.632+, a training sample of size n_{tr} is randomly drawn repeatedly in s number of times with replacement and classification rule constructed on each selected training sample. The prediction accuracy of each classification rule is assessed on the validation sample n_{te} through the average *misclassification error rate* (MER)

$$\hat{\hat{\vartheta}}_j = \frac{1}{s}\sum_{r=1}^{s}\left(0.632 * \hat{\vartheta}_{r.test} + 0.368 * \hat{\vartheta}_{r.train}\right) \quad (1.1)$$

averaged over all the s bootstrap samples, where $\hat{\vartheta}_{r.test} = \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} I(y_i \neq \hat{y}_i)$, $r=1, \dots, s,$ is the bootstrap prediction error rate computed over the validation (test) sample n_{te} while $\hat{\vartheta}_{r.train} = \frac{1}{n_{tr}}\sum_{i=1}^{n_{tr}} I(y_i \neq \hat{y}_i)$ is the re-substitution prediction error rate computed over the training sample n_{tr}. The quantity $I(.)$ in $\hat{\vartheta}_{r.test}$ and $\hat{\vartheta}_{r.train}$ is an indicator function whose value is 1 if the predicted class label of the $\hat{y}_i$ i^{th} sample does not equal the true class label y_i of the biological sample and 0 if otherwise. Within the logistic regression set up, the predicted class label is $\hat{y}_i = 1$ if $p(y_i = 1|X = x) \geq 0.5$ and $\hat{y}_i = 0$ if otherwise.

The mixing feature in the estimator of the average bootstrap MER $\hat{\hat{\vartheta}}_J$ in (1.1) is determined by correcting for the fraction of the original data points that are not selected into the training and test samples at each bootstrap sampling. Since the sampling is done n times with replacement, the chance that an element in the original sample data would not be selected into the training set n_{tr} is $\left(1 - \frac{1}{n}\right)^n \approx 1/e = 0.368$ by Taylor's series expansion.[36] This simply shows that at each bootstrap sampling, about $1 - \left(1 - \frac{1}{n}\right)^n = 0.632$ of n would be in the training set n_{tr} while 0.368 of n would be in the test set n_{te}, with $n_{tr} = n_{te} = n$, hence, the term bootrap.632+. Therefore, for each bootstrap sample drawn, the empirical prediction error rate for the bootrap.632+ is computed by the mixture $\hat{\vartheta}_{bootstrap} = 0.632 * \hat{\vartheta}_{test} + 0.368 * \hat{\vartheta}_{train}$,[37,38] thereby correcting for the proportion of original data points that would not be present in the training and test samples used in computation. Here, $\hat{\vartheta}_{train}$ is the estimated re-substitution prediction error rate computed over the training set while $\hat{\vartheta}_{test}$ is the bootstrap prediction error rate computed over the test set. When this bootstrap procedure is repeated s number of times we have the average bootstrap MER estimator given by (1.1).

Now, at the first feature selection stage, the gene variable $X_{(1)}$ that yields the least average bootstrap.632+ MER estimate (according to (1.1)), say $\hat{\hat{\vartheta}}_{(1)}$ among all the estimated MERs $\hat{\hat{\vartheta}}_j, j = 1, \dots, p$, is selected at the first gene selection step since it shows the highest discriminatory power over others. At the second feature selection and classification step, the second best gene predictor $X_{(2)}$ is selected by forming a set of pairs of genes with the first selected gene $X_{(1)}$ and the remaining $p-1$ left out genes. A logistic regression is constructed on each gene pair and the average MER is computed following the above bootrap.632+ procedure. At the end of this exercise, the gene pair $X_{(1)}X_{(2)}$ that yielded the least average bootstrap MER, say $\hat{\hat{\vartheta}}_{(2)}$ is selected.

Before the third best gene $X_{(3)}$, and more generally, the $(k + 1)^{th}$ best gene $X_{(k+1)}$ could be selected after the selection of the k^{th} gene $X_{(k)}$, the marginal gain in the prediction strength $\hat{\Delta}_k = \hat{\hat{\vartheta}}_{(k)} - \hat{\hat{\vartheta}}_{(k+1)}$, due to the inclusion of gene $X_{(k+1)}$ into the classification model is examined by testing the hypothesis H_{0k}: $\Delta_k = 0$ vs. H_{1k}: $\Delta_k > 0, \Delta_k = \bar{\vartheta}_{(k)} - \bar{\vartheta}_{(k+1)}$, , via the test statistic

$$Z_{\hat{\Delta}_k} = \frac{\hat{\Delta}_k - E(\hat{\Delta}_k)}{\sqrt{v(\hat{\Delta}_k)}} \qquad (1.2)$$

where $v(\hat{\Delta}_k)$ is the empirical variance of the estimated MER differences at any steps k and $k + 1$ and $Z_{\hat{\Delta}_k}$ has a skew-normal density with shape parameter $\lambda = 4.0398$ [25,34] and $E(\hat{\Delta}_k) = 0$ under H_{0k}. When the null H_{0k} cannot be rejected, the $(k + 1)^{th}$ gene $X_{(k+1)}$ under consideration is dropped from the classification model and the k-SS algorithm terminates assuming that no other gene variable among the remaining $p - k$ genes is capable of improving the prediction strength of the current classification model that contains the k gene variables. However, if H_{1k}. is accepted, this shows that the gene variable $X_{(k+1)}$ has significantly enhanced the prediction strength of the current classification model and should therefore be retained while the selection of the next best gene $X_{(k+2)}$ begins following the same procedures above.

Finally, the k-SS algorithm performs backward checks on each feature selected beginning from the second selected gene. This enables any previously selected feature to be checked for redundancy anytime a new feature is introduced into the classification model. If a new feature is selected into the model, an average MER, say $\hat{\hat{\vartheta}}_{full}$ is computed for the full model. Each of the previously selected features would be removed from the model and for each removed feature, a new model is fitted using all other features except the removed one and average MER, say $\hat{\hat{\vartheta}}_{remove}$ is computed. If $\hat{\hat{\vartheta}}_{remove} \leq \hat{\hat{\vartheta}}_{full}$, it shows that the removed feature is now redundant in the presence of a newly selected feature in the model, and it is therefore removed from the model at that selection step. If no gene is further rejected this way after terminating the forward sequential selection step, the set of k marker genes selected for classification then becomes the selected k-SS classifiers, as they shall be referred to in later sections.

The k-SS algorithm was developed using R statistical software[39] and the R codes that implement the algorithm can be accessed freely upon request from the corresponding author. This notwithstanding, the R libraries '**sn**' (**library(sn)**) and '**ROCR**' (**library(ROCR)**) are required to run the k-SS algorithm in R.

1.2. DATA ANALYSIS

The rectal cancer microarray data analysed here contained 24,026 genes and 43 tissues samples comprising of 14 responder and 29 non-responder LARC patients to NRC treatments. Each of the 29 non-responders (histopathologic regression grade 2 and 3) was coded 1 while the 14 responders (histopathologic regression grade 1) were coded 0. The 43 rectal cancer samples with their respective sample labels according to their clinical response groups are presented in Table 2.

The sparseness of the biological samples in this study notwithstanding, 38 (about 90%) of all the 43 samples were randomly selected to construct the k-SS classification model for the data while the remaining unused 5 samples (about 10% of the 43

TABLE 2: The sample labels of the 43 locally advanced rectal cancer patients in the clinical study according to their respective histopathologic response groups of 14 responders and 29 non-responders to neoadjuvant radiochemotherapy treatments.

Sample Labels of All The 43 LARC Patients	
14 histopathologic responders	29 histopathologic non-responders
P105, p211, p215, p224, p24, p309, p332, p354, p380, p402, p410, p66, p79, p80	P123, p132, p168, p177, p213, p214, p267, p274, p281, p3, p311, p319, p345, p37, p40, p494, p572, p587, p122, p272, p275, p29, p292, p297, p32, p383, p464, p474, p504

sample) were kept as external test data to assess the generalization error of the final-chosen model in line with the proposal of Hastie et al.[40] This is done to assess the prediction accuracy of the final model developed. The 38 selected samples which comprised of 12 responders and 26 non-responders were further re-sampled into training and validation (test) sets via the bootstrap.632+ cross-validation scheme in line with *k*-SS procedures for model selection and response class prediction as summarized in Section 1.1. The five left-out (external) unused samples comprised of two responders (p332, p80) and three non-responders (p123, p40, p383) LARC samples.

All the 24,026 observed gene expression profiles were pre-processed through normalization according to the scheme adopted by Yang et al.[41] to identify and remove the effects of systematic variations other than the biological differences in the measured fluorescence intensities of genes across the hybridized mRNA samples. The normalized expression profile of each gene was then standardized so that each of them has zero mean and a unit standard deviation. This was aimed at preventing the gene expression measures in a particular array from dominating the overall average expression level.

Due to the curse of dimensionality,[11,42] the larger features and small sample size scenario which often renders the application of some standard regression techniques unsuitable to analyse this kind of microarray data as well as the need to save computation time, we adopted an efficient primary filtering algorithm, the cross-validated AUC (CVAUC) method,[26] to prune the large dimensional space of the 24,026 genes to a manageable size of 263 potential differentially expressed genes. This technique employs the area under the *receiver operating characteristics* (ROC) curve for genes filtering. By this method, any gene variable X_j whose estimated *area under the curve* (AUC), say $\hat{A}_{X_j}$, averaged over a v-fold cross-validation, is greater than 0.5 by $Z_{1-\alpha}$ of its standard error $\sigma_{\hat{A}_{X_j}}$ $\left(i.e.\ \hat{A}_{X_j} > 0.5 + Z_{1-\alpha}\sigma_{\hat{A}_{X_j}}\right)$ was selected into the *k*-SS algorithm for further usage, where $Z_{1-\alpha}$ is the quantile of the standard normal distribution at the threshold value of 10% for Type I error α with v being specified by the user. However, $v = 10$ is often desirable as used here.

The *k*-SS algorithm was implemented on data matrix of 38 samples by 263 preliminarily selected genes according to the procedures highlighted in Section 1.1, over 2000 bootstrap cross-validation runs. This resulted in the selection of the best five relevant gene signatures whose expression levels are correlated with the two clinical response groups of the LARC patients. The value of the Type I error rate, α (the tuning parameter of the *k*-SS classifier) was determined (by cross-validation) to be 0.031 for the *k*-SS test. The five selected marker genes subsets are used to predict the histopathologic responses of LARC patients to NRC treatment and this provided bootstrap prediction accuracy of 98.47% (average MER of 0.0153) by equation (1.1). Details of the results are presented in Section 2. All the data analyses were performed with R statistical software (www.cran.org). Some data preparations were performed within the environment of other software like SPSS version 17,[43] Microsoft Excel (Microsoft Corporation, Redwood, CA) and Stata/SE 10.0.[44]

2. RESULTS

The *k*-SS feature selection and response class prediction method selected, from the rectal microarray data, the best five informative gene chips whose expression levels are strongly correlated with the two histopathologic response groups of the LARC patients. Using equation (1.1), the estimated average (bootstrap.632+) misclassification error rate (MER) provided by the five selected *k*-SS classifiers is 1.53%. This translates to a prediction accuracy of 98.47%, an indication that almost all the 12 responders and 26 non-responders LARC patients in the training set were correctly classified into their respective response groups by the five selected *k*-SS classifiers.

The probe set numbers, the gene symbols and the gene names of the five selected marker genes are presented in Table 3.

Apart from the estimated MER, the prediction performance of the five selected *k*-SS classifiers was also examined through the *receiver operating characteristic* (ROC) curve and the estimated *area under the curve* (AUC). These are simply obtained by considering the number of runs in which each of the patients was correctly or incorrectly classified by *k*-SS classification rule over the 2000 bootstrap.632 cross-validation runs. This enables us to obtain the plot of the cross-validated ROC (CVROC) curve (graph not shown) and compute the respective cross-validated area under the curve (CVAUC) to be 0.9904 ≈ 1.00.

Other performance indices employed to assess the goodness of the five classifiers are the *brier score*, *sensitivity*, *specificity*, *positive predictive value*, *negative predictive value* (the precision) and the *Jaccard* Index the results of which are presented in Table 4. The standard deviations of the estimated performance indices which measure the within-sample spread of the estimated values are equally reported in Table 4. The various results reported by all the performance indices clearly supported that the selected five genes are good predictors of the clinical responses of locally advanced rectal carcinomas to NRC treatments.

TABLE 3: The five selected marker gene signatures from rectal cancer microarray data by k-SS method.

Probe-set Number	Gene Symbol	Gene Name
216457_s_at	SF3A1	Splicing factor 3a, subunit 1, 120kDa
239813_at	FLJ12476	Hypothetical protein FLJ12476
200935_at	CALR	Calreticulin
201591_s_at	NISCH	Nischarin
202269_x_at	GBP1	Guanylate binding protein 1, interferon-inducible, 67kDa

TABLE 4: The estimated performance indices of the five selected k-SS classifiers from rectal cancer microarray data. The standard deviations of various estimated performance indices are also provided.

Performance Indices on k-SS classification rule	Estimated values (in %)	Standard deviations of estimates
Misclassification Error Rate	1.53	0.0577
Brier score	1.77	0.0311
Sensitivity	98.41	0.0241
Specificity	98.08	0.0130
Positive Predictive Value	98.58	0.0215
Negative Predictive Value	98.84	0.0293
Jaccard Index	98.04	0.0361
Cross-validated AUC	99.04	0.0307

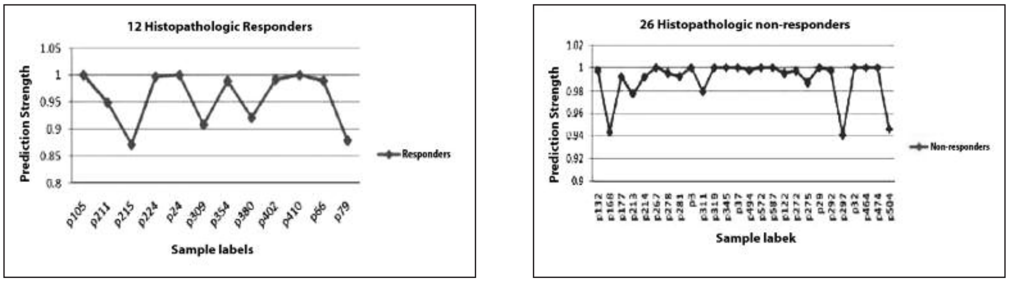

FIGURE 1: The graphs of (joint) prediction strengths (PS) of the five selected k-SS classifiers for the randomly selected 12 histopathologic responders (Fig 1(a)) and the 26 non-responders (Fig 1(b)) to neoadjuvant radiochemotherapy treatments. The graphs show that all the 38 LARC patients are correctly classified into their respective clinical response groups more than 80% of times by the five selected k-SS classifiers.

As a further step, the joint prediction strength (PS) of the five selected *k*-SS classifiers on each of the 43 tissue samples was also examined. The PS, $0 \leq PS \leq 1$, is the proportion of runs, out of the 2000 bootstrap.632+ cross-validation runs, that yielded correct classification of a LARC patient into his/her histopathologic response group by the *k*-SS classifiers. The PS is a measure of the degree of correct classification of any biological sample into its true clinical response group. The graphs of the prediction strength of the five selected *k*-SS classifiers on the randomly selected 12 responders and 26 non-responders from the 43 LARC patients are presented in Fig. 1(a) & (b). From the two graphs, it can be observed that each of the 38 LARC samples was correctly classified more than 80% of time by the five selected *k*-SS classifiers. This again confirms the importance of the selected gene signatures by the instrument of the *k*-SS method.

The discriminatory power of the five selected genes in Table 3 is further established through cluster analysis. Following the procedures adopted by Alon et al.[45] and later by Linn et al.,[4] a two-way *single-linkage hierarchical clustering* (SLHC) was employed using the expression levels of the five selected genes to classify all the 43 rectal carcinomas into their respective response groups of responders and non-responders to NRC treatments. The software 'C*luster 3.0*'[46] which is an enhanced version of the software '*Cluster*' developed by Eisen et al.[47] was employed for clustering. The dissimilarity measure of Euclidian distance was adopted in the SLHC implementation.

In the cluster analysis, the five selected gene chips were organized into groups based on the similarity of their expression levels, thereby facilitating the recognition of functional themes in their gene expression patterns, and the 43 LARC sample specimens were also organized into their distinct histopathologic groups of responders and non-responders to NRC treatment. These results are presented by the dendrogram in Fig. 2. The dendrogram shows the normalized expression profiles of the five selected marker genes with red fluorescent labels representing high expression levels and green fluorescent labels denoting low expression levels of the respective genes as shown by the colour key bar in Fig. 2.

From the dendrogram, except for the responder sample specimen with p215 label (indicated with red arrow), it can be observed that all the 14 responder and 29 non-responder tissue samples are perfectly identified and grouped together by clustering based on the expression levels of the five selected marker genes. More specifically, the following four genes SF3A1, CALR, NISCH and GBP1 are differentially expressed among the two groups of responder and non-responder sample specimens as evident on the dendogram. The expression profiles of the four genes were high (red fluorescent dyes) in all the responder tissue samples while they all have reduced expression measures (green fluorescent dyes) in most of the non-responder samples. These results clearly revealed the significant association between the five selected gene signatures and the two histopathologic response groups of rectal carcinomas to NRC treatments.

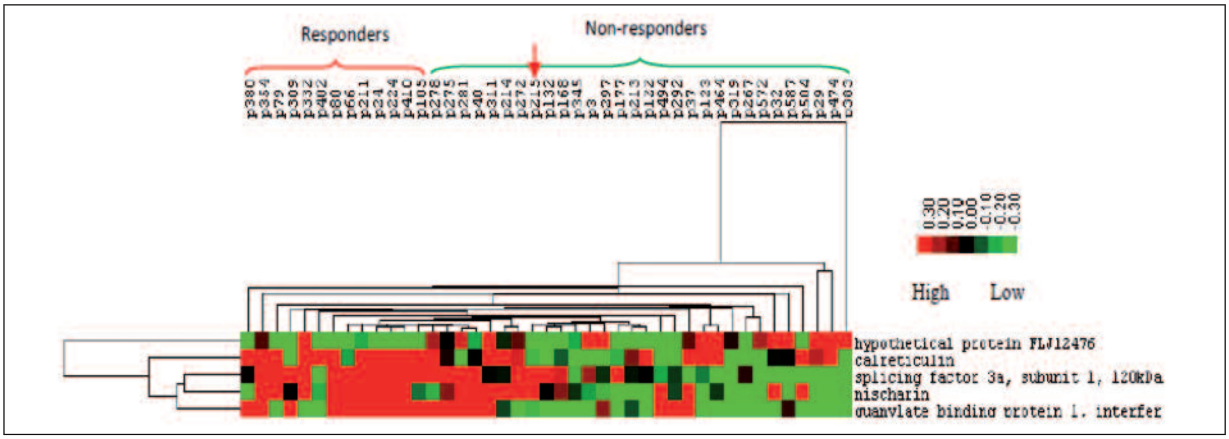

FIGURE 2: The dendrogram of two-way single linkage hierarchical clustering results using the five selected k-SS classifiers from rectal cancer data. The two categories of LARC patients of responders (indicated by red bracket & red arrow) and non-responders (indicated by green bracket) are clearly discriminated.

We present in Fig. 3, the box plots of expression levels of the five selected *k*-SS classifiers in order of selections within the two histopathologic responses (responders and non-responders) of rectal carcinomas to NRC treatments. This is intended to visualize the median intensities of the expression levels of each selected gene subset relative to others within the two clinical response groups.

It can be observed from the box plots in Fig. 3 that, except for the only gene FLJ12476, the median intensities of the expression levels of all other four genes SF3A1, CALR, NISCH and GBP1 are higher in the responder samples than in the non-responder samples. These are the genes with clear evidence of high expression profiles (red fluorescent dyes) among the 14 responders as shown by the clustering results (dendrogram) in Fig. 2. The median expression levels of gene FLJ12476 are slightly higher in some of the non-responder group than in the responder group as shown by the box

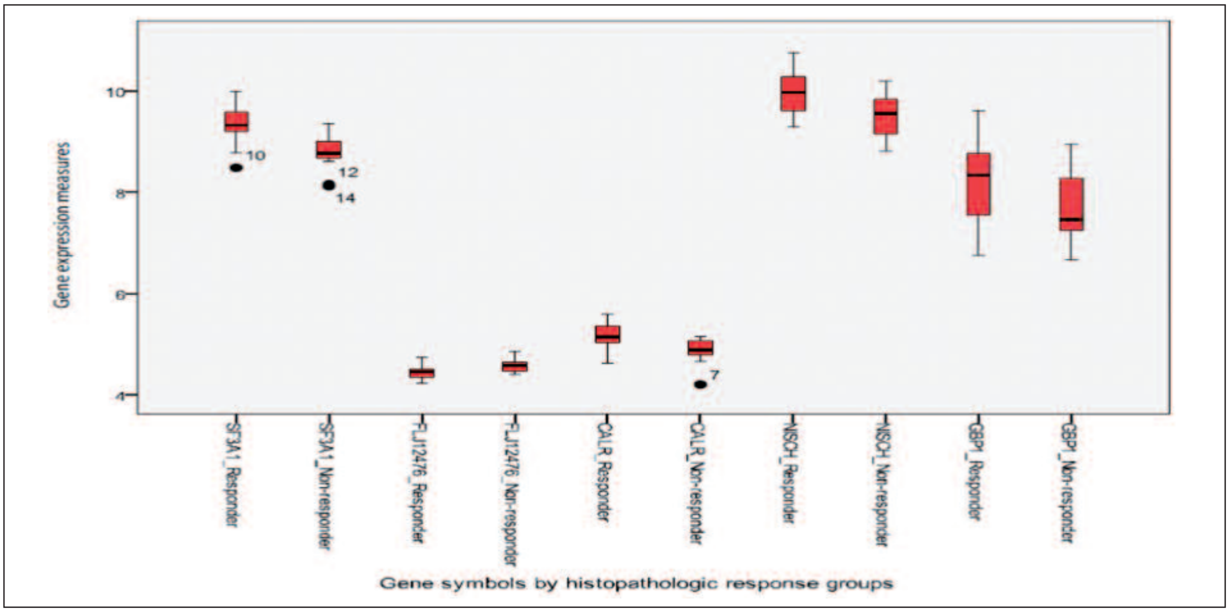

FIGURE 3: The box-plots of expression profiles of five selected genes from rectal cancer microarray data (arranged in order of selection) for the two histopathologic responses (responders and non-responders) of locally advanced rectal carcinomas to NRC treatments. The median expression levels of each gene in the two response groups are shown by the tick black horizontal lines within the boxes.

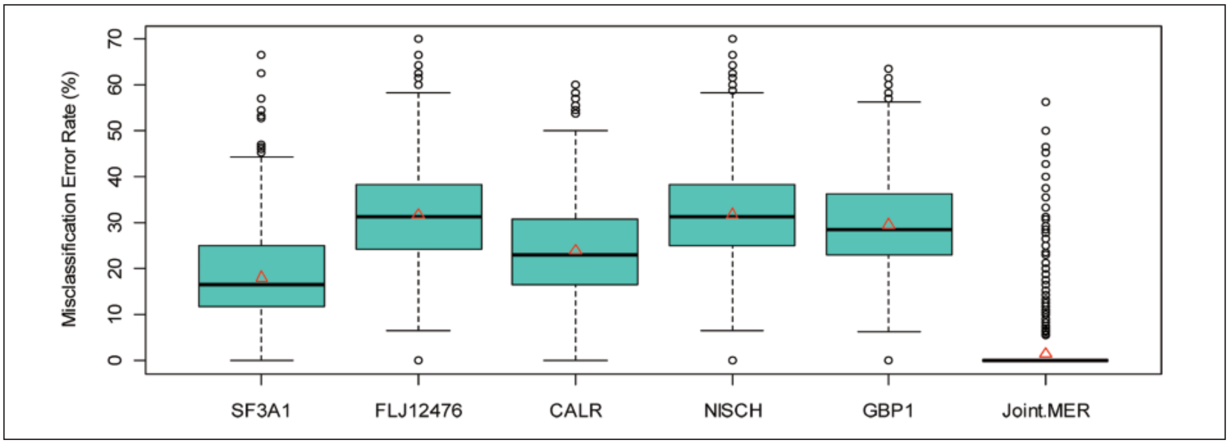

FIGURE 4: Box plots of misclassification error rate (MER) (in %) from univariate analysis of each selected five gene signatures by k-SS classification rule as well as the estimated MER jointly yielded by all the five genes. The red triangular spots represent the average MER of the respective gene variable(s).

plots in Fig. 3. This result agrees perfectly with the results of clustering in Fig. 2 in which apparent up-regulated expression levels of gene FLJ12476 manifested in the nine non-responder samples p311, p3, p123, p464, p32, p587, p29, p474 and p303 than among the responder samples.

Furthermore, we examined the predictive power of each selected genes by reporting their marginal average prediction error rates obtained from their respective univariate classification via the logit model. The estimated average MER values (in %) of these genes based on the results from their respective univariate analyses are SF3A1 (18.09%), FLJ12476 (31.75%), CALR (23.96%), NISCH (31.86%) and GBP1 (29.58%) with their joint estimated average MER being 1.53% as reported in Table 4. The box plot of the MER yielded by each gene from univariate analysis over 2000 bootstrap cross-validation runs is equally presented in Fig. 4. The box plot of the joint MER of all the five selected gene predictors is also presented in Fig. 4. The average MER of each gene variable is indicated by red triangular spot on each of the box plot in Fig. 4.

After the selection of the first gene SF3A1 that yielded average MER of 18.09%, each additional gene selected monotonically reduces this average MER value until the optimal selection level was reached at which the fifth best gene was selected. The reduction in the successive average MERs due to inclusion of additional gene variable in the *k*-SS classification model simply measures improvement in the prediction accuracy of any current *k*-SS$_r$ model that uses *r* gene predictors over the preceding *k*-SS$_{r-1}$ model that uses *r-1* gene predictors for classification, $r \in \{1, \ldots, k\}$.The graph of the average MER at each successive *k*-SS gene selection step is presented in Fig. 5. The graph clearly shows a monotonically decreasing trend of the average MER values as additional gene predictors are selected into the classification model for response class prediction by *k*-SS algorithm.

As typical of any gene expression data set, most of the 24,026 observed gene signatures in the rectal cancer data set are correlated, and by implication, some of the selected five *k*-SS classifiers from this data set are also expected to be correlated. This is simply confirmed by the correlation matrix of the expression levels of the selected five genes as presented in Table 5. As evident in that table, most of the correlation tests are significant ($p < 0.05$). To this end, further analysis carried out on this rectal cancer microarray data set to further depict the significance of the five selected genes subset is the *principal component analysis* (PCA).[48,49]

The PCA procedure was implemented using only the selected five gene variables to form a set of *r* uncorrelated composite variables called the *principal components* (PCs). These principal components are expected to jointly account for most of the variations in the original five gene variables. The

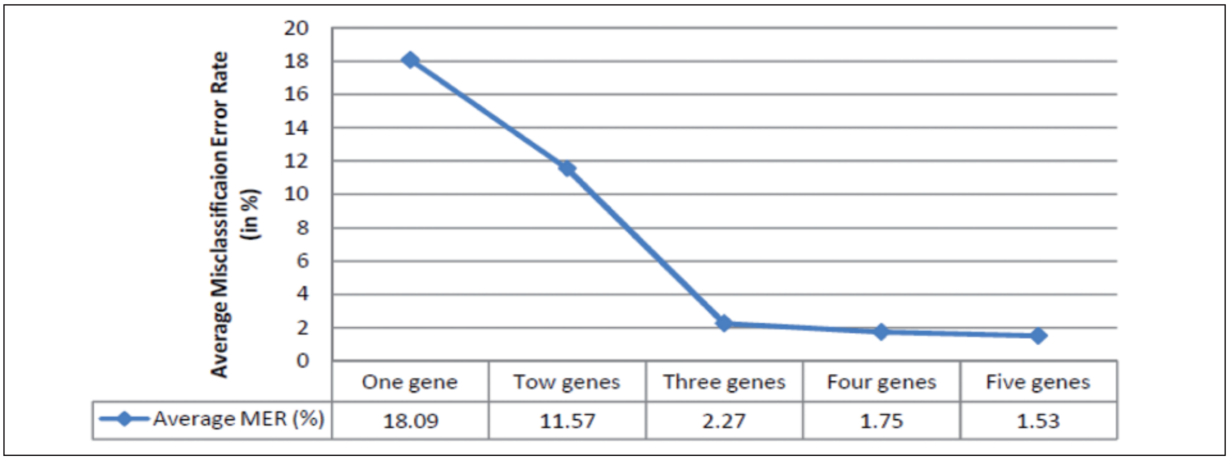

FIGURE 5: The graph of the successive average MER estimates at each k-SS selection step plotted against the numb of genes selected for class prediction. The graph shows a monotonically decreasing trend in average MER values as more genes are selected into the k-SS classification model until the best five genes are selected which provided the optimal average prediction error rate (average MER) of 1.53%.

TABLE 5: Table of correlation matrix of the expression profiles between the five selected gene signatures by k-SS method. The p-values of the Pearson correlation tests are enclosed in brackets. (*) indicates that the correlation coefficient is significantly different from zero at 5% significance level (p < 0.05).

	SF3A1	FLJ12476	CALR	NISCH	GBP1
SF3A1	1.0000				
FLJ12476	-0.5744 (p = 0.0001)*	1.0000			
CALR	0.2455 (p = 0.1126)	-0.2006 (p = 0.1972)	1.0000		
NISCH	0.6055 (p < 0.0001*)	-0.4381 (p = 0.0033)*	0.3145 (p = 0.0400)*	1.0000	
GBP1	0.5533 (p = 0.0001)*	-0.4238 (p = 0.0046)*	0.3628 (p = 0.0168)*	0.3887 (p = 0.0100)*	1.0000

resulting PCs were arranged in descending order of the amount of information in the data they captured as represented by their respective estimated Eigen values. The first few (two or three) PCs in this sequence are expected to possess substantial information that are conveyed by original data.

To this end, we obtained the plot of the first principal components (PC plot) to discriminate between the two clinical response groups of the LARC patients (the responders and non-responders to NRC treatments). This PC plot is presented in Fig. 6 from which perfect separation of the 14 responders from the 29 non-responders could be clearly observed. This is an indication that the five selected marker genes by *k*-SS classification rule from which the PCs are constructed are strongly related to the two clinical response groups of the LARC patients.

Not only that, we also ran principal component analysis using all the 263 genes that were preliminarily selected through the cross-validated AUC (CVAUC) filtering procedure to assess the discriminatory power of all these genes as compared to that of the five selected *k*-SS classifiers.

The plot of the first principal component against the 43 samples in Fig. 7 again shows perfect discrimination of the two histopathologic response groups of the LARC patients to NRC treatment. This result is not quite different from the result

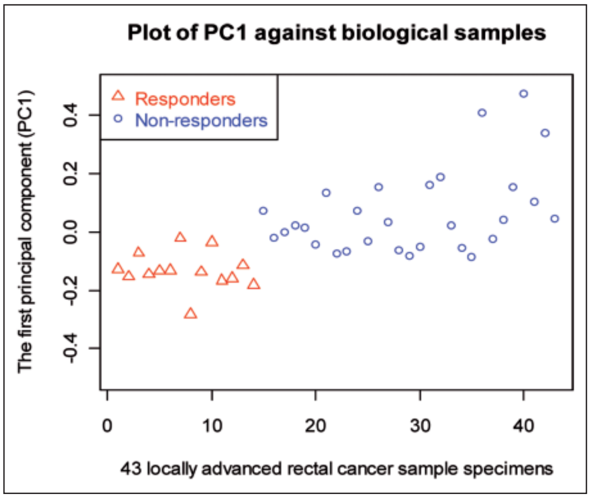

FIGURE 6: The plot of the first principal components (PC1) versus the 43 biological samples to discriminate between the two histopathologic response groups of LARC patients. The PC1 was constructed using the 5 selected marker genes by k-SS classification rule. Clear discrimination between the responders and non-responders is shown by the graph based on the five selected k-SS classifiers.

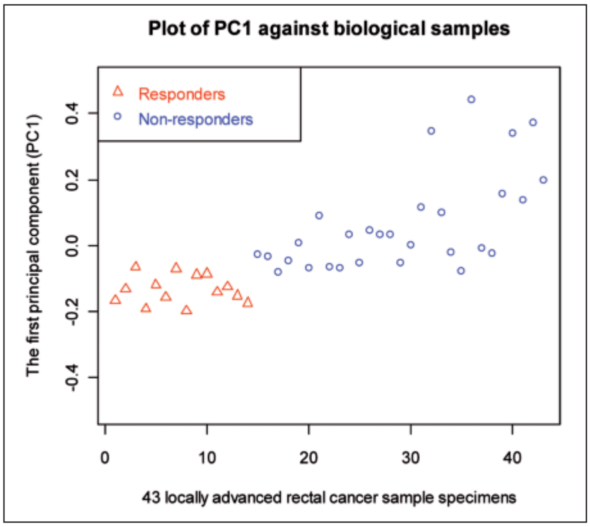

FIGURE 7: The plots of the first principal components (PC1) versus the 43 biological samples to discriminate between the two histopathologic response groups of LARC patients. The PC1 was constructed using all the 263 genes preliminarily selected by filtering procedure of the cross-validated AUC (CVAUC).

achieved using only the five informative genes selected by the *k*-SS method as shown in Fig. 6. This simply shows that the joint expression profiles of the five selected genes are strong enough to provide good prediction of the histopathologic responses of rectal carcinomas to neoadjuvant treatments. In the spirit of parsimony therefore, it would be more economical and a lot easier for molecular biologist to study the biological functions and properties of the five selected genes than to work with all the 263 gene signatures some of which might be biologically uncorrelated to the hispathological responses of rectal carcinomas to NRC treatment.

2.1. ASSESSMENT OF THE SELECTED GENE SIGNATURES

The goodness of the five gene biomarkere-SF3A1, FLJ12476, CALR, NISCH and GBP1 as selected by the *k*-SS method can be simply assessed by employing them to predict the five external test samples that were not used while constructing the *k*-SS classification model. To this end, a logit model was fitted using the 38 by 6 data matrix of randomly selected 38 LARC samples by the five selected *k*-SS classifiers as predictors with the first column of the data matrix representing the true class labels of the 38 LARC samples. These 38 samples specimens excluded the five test samples p332, p80, p123, p40, and p383 as detailed in Section 2. The fitted model was used to predict the histopathologic response groups of these five test samples. The predicted class label $\hat{y}_i$ of a *i* sample in the test set, *i* = 1, ..., 5, is determined via the scheme

$$\hat{y}_i = \begin{cases} 1\ if\ p(y_i = 1|c(x)) \geq 0.5 \\ 0\ if\ p(y_i = 1|c(x)) < 0.5 \end{cases} \qquad (2.1)$$

where $c(x) = \{\text{SF3A1}, \text{FLJ12476}, \text{CALR}, \text{NISCH}, \text{GBP1}\}$ is the vector of expression measures of the five selected gene predictors selected by *k*-SS method. In this context also, y_i is the true class label of the sample with $y_i = 0$ indicating that the *i* sample is a responder while $y_i = 1$ indicates that the *i* sample is a non-responder to NRC treatment as earlier defined in Section 1.1. Out of the five samples in the test set, two of them, p332 and p80 are responders, each carrying the label 0 and the remaining three (p123, p40, p383) are non-responder LARC samples, each carrying the label 1.

Using equation (3.1), the predicted class labels of all the five samples in the test set were computed and the average prediction error rate $\hat{\hat{\vartheta}}_{test}$was estimated by $\hat{\hat{\vartheta}}_{test} = \frac{1}{5}\sum_{i=1}^{5} I(y_i \neq \hat{y}_i)$, where *I*(.) is an indicator function whose value is 1 if the predicted class label $\hat{y}_i$ of the *i* sample in the test set does not equal to the true class label y_i of the i sample and 0 if otherwise. The prediction results obtained based on the above procedures indicated that all the two responder and the three non-responder sample specimens were correctly classified into their respective histopathologic response classes.

To assess the behaviour of the five selected *k*-SS classifiers on all the 43 LARC samples, 10-fold cross-validation technique was adopted in which the entire 43 LARC samples were randomly partitioned into 10 mutually exclusive parts (folds). Nine folds of the data were used in turn to train the five *k*-SS classifiers and the response classes of the unused test samples in the left-out one fold were predicted using the fitted model. This exercise was repeated 10 times such that each of the test sample set was used just once. At each fold v, $v = 1, ..., 10$, the prediction error rate $\hat{\vartheta}_v = \frac{1}{n_v}\sum_{i=1}^{n_v} I(y_i \neq \hat{y}_i)$ was computed and the average of these, over all the 10 folds of the test samples, was determined by $\hat{\hat{\vartheta}}_{test} = \frac{1}{10}\sum_{v=1}^{10}\hat{\vartheta}_v$ as the true prediction error rate of the five selected *k*-SS classifiers on the 43 LARC samples, where $I(.)$ is an indicator function as defined in Section 1.1 and n_v is the number of randomly selected test samples at fold v.

The five selected *k*-SS classifiers based on the above 10-fold cross-validation technique yielded about 90% correct classification of the 43 LARC samples into their respective clinical response groups with just four samples (p79, p309, p122, p272) being misclassified. The estimated prediction accuracy of the responders (sensitivity), based on these results, is about 86% while the prediction accuracy of the non-responders (specificity) is about 93%. The four misclassified samples are two histopathologic responders (p79, p309) and two non-responders (p122, p272) LARC samples.

3. DISCUSSION

It has been fully demonstrated in this work that preoperative prediction of clinical responses to neoadjuvant radiochemotherapy treatments by patients suffering from locally advanced rectal carcinomas using their gene expression signatures is a possible task if proper statistical tools are employed. The low rate of responses by locally advanced rectal cancer patients to NRC treatments has made it a worthy task to device efficient and reliable alternative diagnostic means through which early detection of responses to NRC treatments could be determined among the affected subjects to enable appropriate therapy treatments to be administered at the early stage of cancer tumour formation. One of such alternative diagnostic means, as presented in this work, is the identification of core gene signatures that are correlated with the two clinical response groups (responders and non-responders to NRC treatment) of the LARC patients.

A set of five informative gene signatures has been identified and selected through the instrument of the *k*-SS feature selection and response class prediction method. The selected five genes are found to correlate significantly with the two histopathologic responses of LARC biological samples to NRC treatment as confirmed by various results in Section 2.

From the univariate prediction results of each selected gene as shown by the box plots in Fig. 4, it can be observed that the first selected gene SF3A1 expectedly yielded the least average prediction error rate of about 18% which translates to prediction accuracy of about 82%. The second best selected gene FLJ12476 with relatively lower median expression levels in the two response groups yielded prediction accuracy of about 68% (MER of about 32%). The remaining three genes: CALR, NISCH and GBP1(that are selected in that order) also provided a fairly high marginal prediction error rates of 24%, 32% and 30% respectively relative to average MER of 18% yielded by the first selected gene SF3A1. But interestingly, all the additional four genes (FLJ12476, CALR, NISCH and GBP1) jointly reduce the starting average prediction error rate of 18% yielded by the first gene SF3A1 to about 2%, a reduction of about 89%. These results simply underscored the fact that some genes, probably due to their low expression profiles, might be weaker predictors on their own, but could very well contribute additional useful information when put together with other genes in a classification model[50-52] as evident in this study.

It is quite instructive to remark that the rectal cancer microarray data analysed here have been previously discussed in an excellent work of Rimkus *et al.*[29] where attempt was made to classify the 43 LARC patients into their respective histopathologic response groups using the procedure of *partial least squares* (PLS) with *linear discriminant analysis* (LDA) for classification (Boulesteix and Strimmer[53]). A set of 42 gene subsets were selected using the Welch statistics with a pre-specified threshold *p* value of 0.001. However, no further attempt was made in that work to screen these competing gene candidates for proper identification and selection of the basic ones that are truly correlated with the histopathologic responses of the rectal carcinomas as carried out in the present study.

The prediction accuracy of the responders (sensitivity) and non-responders (specificity) reported by Rimkus et al.[29] were 71% and 86% respectively based on the 42 selected gene subsets. The prediction accuracy of about 90% on both the responders and non-responders LARC samples yielded by the *k*-SS classification model as reported in Section 3.1 using just five informative gene signatures obviously improved on the earlier prediction results reported by Rimkus et al.[29] for this same data. Nonetheless, one of the five vital gene biomarkers (SF3A1) identified, selected used for classification by *k*-SS algorithm was also among the 42 genes selected in the work of Rimkus et al.[29]

In terms of biological functions of the five selected gene signatures, the first selected protein encoding gene SF3A1 together with two other subunits comprises Splicing Factor 3a. As part of the Spliceosom, this Splicing factor helps in converting pre-messenger RNA to mRNA. The second gene FLJ12476 is expressed in fetal and adult testis, in germ cells but not somatic cells. It may play a regulatory role in spermatogenesis, though its function is not fully understood yet. In addition, Calreticulin (CALR), a known chaperone, which together with Calnexin assists with the folding of glycoproteins within the endoplasmatic reticulum. The Nischarin (NISCH) gene initiates signaling cascades for cell survival, growth and migration. It acts as a modulator of Rac-regulated signal transduction pathways while GBP1 (guanylate binding protein 1, interferon-inducible, 67kDa) expression is induced by interferon. The protein converts GTP to GMP, therefore mediating antiviral effects against some viruses.

The novel *k*-SS method adopted here for genes selection and classification of histopathologic responses of LARC samples to NRC treatment has been found to be very efficient in terms of computation costs.[25] The *k*-SS algorithm would run efficiently well on any version of Windows Operating Systems with an average of 1.60GHz and 500 MB RAM or more. On the average, about 25 minutes CPU time may be required by the algorithm to complete a task with a data set of moderate size of between 2,000 to 5,000 gene expression profiles.

4. CONCLUSION

In conclusion, this work has opened up another research opportunity for molecular biologists and medical experts to determine and explore the biological functions of the five selected gene signatures with correlated expression profiles to histopathologic responses of locally advanced rectal carcinomas to NRC treatment. A thorough research work in this direction shall immensely help to exploring maximally the benefits that are inherent in these selected gene biomarkers for proper diagnosis of clinical responses of LARC patients to NRC treatment and many more. Nevertheless, validation studies with larger LARC patient groups might be desirable in future to give more credence to the results obtained in this study.

Acknowledgement

Authors sincerely thank the two anonymous referees and journal editors for their honest criticisms, observations and suggestions all of which have helped to improve the first draft of this paper.

REFERENCES

1. Burnside J, Ouyang M, Anderson A, Bernberg E, Lu C, Meyers BC, et al. Deep sequencing of chicken microRNAs. BMC Genomics 2008;9:185.
2. Cooper CS. Applications of microarray technology in breast cancer research. Breast Cancer Research 2001;3:158-71.
3. Surks HK, Richards CT, Mendelsohn ME. Myosin phosphatase-Rho interacting protein. A new member of the myosin phosphatase complex that directly binds RhoA. J Biol Chem 2003;278(51):51484-93.
4. Linn SC, West RB, Pollack JR, Zhu S, Hernandez-Boussard T, Nielsen TO, et al. Gene expression patterns and gene copy number changes in dermatofibrosarcoma Protuberans. Am J Pathol 2003;163(6):2383-95.
5. Schulz WA. Molecular Biology of Human Cancers: An Advanced Student's Textbook. 1st ed. New York: Springer; 2005. p.1-508.
6. O'Neill G, Catchpoole DR, Golemis EA. From correlation to casualty: microarrays, cancer and cancer treatment. In: Berrer DP, Dubitzky W, Granzow M, eds. 2003;34:S64-S71.
7. Vesterlund J. Feature Selection and Classification of cDNA Microarray Samples in ROSETTA. Uppsala: M.Sc. Thesis, Uppsala Universitet; 2008. p.1-136.
8. Simon RM, Dobbin K. Experimental design of DNA microarray experiments. BioTechniques 2003;Suppl:16-21.
9. Ochs MF, Godwin AK. Microarrays in cancer: research and applications. BioTechniques 2004;Suppl:4-15.
10. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Molecular Biology of the Gene. 5thed. San Francisco: Benjamin Cummings, Pearson Education; 2004.
11. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531-7.
12. Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, Hunt KK, et al. Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas, Proceedings of National Academic of Science (PNAS) 2007;104(9):3414-9.
13. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62(17):4963-7.
14. Theisen J, Kauer WK, Nekarda H, Schmid L, Stein HJ, Siewert JR. Neoadjuvant radiochemotherapy for patients with locally advanced rectal cancer leads to impairment of the anal sphincter. J Gastrointest Surg 2006;10(2):309-14.
15. Rápolti E, Szigeti A, Farkas R, Bellyei S, Boronkai A, Papp A, et al. [Neoadjuvant radiochemotherapy in the treatment of locally advanced rectal tumors]. Magy Onkol 2009;53(4):345-9.
16. Ferrigno R, Novaes PE, Silva ML, Nishimoto IN, Nakagawa WT, Rossi BM, et al. Neoadjuvant radiochemotherapy in the treatment of fixed and semi-fixed rectal tumors. Analysis of results and prognostic factors. Radiat Oncol 2006;1:5.
17. Samel S, Hofheinz R, Hundt A, Sturm J, Knoll MR, Wenz F, et al. [Neoadjuvant radiochemotherapy of adenocarcinoma of the oesophagogastric junction]. Onkologie 2001;24(3):278-82.
18. Rullier E, Goffre B, Bonnel C, Zerbib F, Caudry M, Saric J. Preoperative radiochemotherapy and sphincter-saving resection for T3 carcinomas of the lower third of the rectum. Ann Surg 2001;234(5):633-40.
19. Rödel C, Martus P, Papadoupolos T, Füzesi L, Klimpfinger M, Fietkau R, et al. Prognostic significance of tumor regression after preoperative chemoradiotherapy for rectal cancer. J Clin Oncol 2005;23(34):8688-96.
20. Amthauer H, Denecke T, Rau B, Hildebrandt B, Hünerbein M, Ruf J, et al. Response prediction by FDG-PET after neoadjuvant radiochemotherapy and combined regional hyperthermia of rectal cancer: correlation with endorectal ultrasound and histopathology. Eur J Nucl Med Mol Imaging 2004;31(6):811-9.
21. Gearhart SL, Frassica D, Rosen R, Choti M, Schulick R, Wahl R. Improved staging with pretreatment positron emission tomography/computed tomography in low rectal cancer. Ann Surg Oncol 2006;13(3):397-404.
22. Suit HD, Spiro IJ, Spear M. Benign and low-grade tumors of the soft tissues: role for radiation therapy. Cancer Treatment Research 1997;91:95-105.
23. O'Brien KP, Seroussi E, Dal Cin P, Sciot R, Mandahl N, Fletcher JA, et al. Various regions within the alpha-helical domain of the COL1A1 gene are fused to the second exon of the PDGFB gene in dermatofibrosarcomas and giant-cell fibroblastomas. Genes Chromosomes Cancer 1998;23(2):187-93.
24. Maki RG, Awan RA, Dixon RH, Jhanwar S, Antonescu CR. Differential sensitivity to imatinib of 2 patients with metastatic sarcoma arising from dermatofibrosarcoma protuberans. Int J Cancer 2002;100(6):623-6.
25. Yahya WB, Ulm K, Fahrmeir L, Hapfelmeier A. k-SS: a sequential feature selection and prediction method in Microarray Study. International Journal of Artificial Intelligence (IJAI) 2011;6(S11):19-47.
26. Yahya WB. Genes Selection and Tumour Classifications in Cancer Research: A New Approach. Germany: Lambert Academic Publishing; 2012.
27. Hapfelmeier A, Yahya WB, Robert R, Ulm K. Predictive modeling of gene expression data. Hand Book of Statistics in Clinical Oncology. 3rded. New York: Chapman & Hall/CRC; 2011: p.463-75.
28. Becker K, Mueller JD, Schumacher C, Ott K, Fink U, Busch R, et al. Histomorphology and grading of regression in gastric carcinoma treated with neoadjuvant chemotherapy. Cancer 2003;98(7):1521-30.
29. Rimkus C, Friederichs J, Boulesteix AL, Theisen J, Mages J, Becker K, et al. Microarray-based prediction of tumour response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer. Clinical Gastroenterology and Hepatology (CGH) 2008;6(1):53-61.
30. Yahya WB, Adebayo SB. Modelling the trend and Determinants of breastfeeding initiation in Nigeria. Child Development Research 2013;1-9.
31. Yahya WB, Ulm K. Survival analysis of breast and small-cell lung cancer patients using conditional logistic regression model. International Journal of Ecological Economics and Statistics (IJEES) 2009;14(S09):3-14.
32. Efron B, Gong G. A leisurely look at the bootstrap, the Jackknife and cross-validation. The American Statistician 1983;37(1):36-48.
33. Efron B, Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. Journal of American Statistical Association (JASA) 1997;92(438):548-60.
34. Yahya WB. Sequential Dimension Reduction and Prediction Methods with High Dimensional Microarray Data. München, Germany: 2009. p.1-246.
35. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. 1st ed. New York, NY: Springer; 2013. p.1-440.
36. Dani SG. Ancient Indian Mathematics-A conspectus. Resonance 2012;17(3):236-46.
37. Gerds TA, Schumacher M. Efron-type measures of prediction error for survival analysis. Biometrics 2007;63(4):1283-7.
38. Binder H, Schumacher M. Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. Stat Appl Genet Mol Biol 2008;7(1): Article12.

39. R Development Core Team, R: A language and environment for statistical computing. R foundation for Statistical computing, Vienna, Austria, 2007; ISBN 3-900051-07-0, URL http://www.r-project.org/.

40. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Leaning. 2nd ed. New York: Springer; 2009. p.1-763.

41. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30(4): e15.

42. Nguyen DV, Rocke DM. Tumour classification by partial least squares using gene expression data. Bioinformatics 2002;18(1):39-50.

43. SPSS Inc., Prentice Hall Inc., Chicago, 2008.

44. Stata Corporation, Texas 77845 USA, 2009. http://www.stata.com/

45. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999;96(12):6745-50.

46. de Hoon MJL, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics 2004;20(9):1453-4.

47. Eisen MB, Spellman PT, Brown PO, Bostein D. Cluster analysis and display of genome-wide expression patterns. Proc Nat Acad Sci USA 1998;95(25):14863-8.

48. Lehmann N. Principal components selection given extensively many variables. Statistics & Probability Letters 2005;74(1):51-8.

49. Härdle W, Hlávka Z. Multivariate Statistics: Exercises and Solutions. 1st ed. New York: Springer; 2007. p.1-368.

50. Jaeger J, Sengupta R, Ruzzo WL. Improved gene selection for classification of microarrays. Pac Symp Biocomput 2003:53-64.

51. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Statist Soc B 2011;73(3):273-82.

52. Yahya WB, Oladiipo MO, Jolayemi ET. A fast algorithm to construct neural networks classification models with high-dimensional genomic data. Annals Computer Science Series 2012;10(1):39-58.

53. Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 2007;8(1):32-44.