# Gene Selection for Colorectal Cancer Classification

**[1]Xuxiaoya[a] chenyin[b] zhangwenping[c]**

## Abstract:

In this paper, we address the problem of gene selection for colorectal cancer classification based on the simultaneous expression monitoring of thousands of genes using DNA microarrays. The aim of the present paper is to infer the critical gene for cancer classification from gene data, and to develop its mathematical and computational foundations. In this paper, F-Scores methodology and image processing are applied in Oncogene Selection.

**Keyword: gene mutation identification, SVM, F-Scores, Oncogene Selection, image processing**

## 1. Introduction

At least 50% of the Western population develops a colorectal tumor by the age of 70, and in about 1 in 10 of these individuals, progression to malignancy ensues. As a result, Colorectal cancer is the most common cause of death from cancer after cancer of the lung and breast. It has been realized for many years that cancer has a genetic component and at the level of the cell it can be said to be a genetic disease. Since the early 1960s, molecular techniques are beginning to help in the understanding of the pathogenesis of the disorder and the genetic basis of colorectal cancer is perhaps better understood than for any other cancer.

"For the geneticist, there are accordingly three ways of genetic analysis. Through characters, he can examine function; through their changes he can examine mutation; through their reassortment, he can examine recombination." Francois Jacob, a famous French biologist, said these words in The Logic of Life (p.224). For the mathematician, there is another effective way of genetic analysis. He can find meaning in a genome through expression data.

DNA microarray can provide a broad picture of the state of the cell, by simultaneously monitoring the expression level of thousands of genes. It is of interest to develop techniques for extracting useful information from the experimental data sets. Studies of DNA microarray of normal and colorectal-tumor specimens may shed light on the genetic alterations involved in tumor progression.

There has already been important work in this direction. For example, Vogelstein et al. (1988) inferred from a variety of types of data that the progression of colorectal cancer can be described by a chain of four genetic events. Golub et al. (1999) used DNA chips in the molecular classification of acute leukemias. Alon et al. (1999) applied a two-way clustering method to classify genes into functional groups. Guyon et al. (2002) selected gene for cancer classification using support vector machines.

In Section 2 we develop an effective way to eliminate the irrelevant gene and select several critical gene for colon cancer classification. In Section 3 we create the image of each array and then present an image denoising methodology to process data with inevitably error. We provide a probabilistic framework which is conditional random field (CRF) model. And provide comparing of CRF with F-score method.

## 2. F-Score and LIBSVM

To discriminate the cancer gene sequences from the normal ones is obviously a classification problem. While, to find which genes act a much more important role in cancer discrimination refers more to feature selection. Among the general classification algorithms, SVM is a convenient and effective one. The LIBSVM (Chih-Chung Chang and Chih-Jen Lin) can work fairly well when the training data is limited.

F-score is a simple technique which measures the discrimination of two sets of real numbers (Yi-Wei Chen and Chih-Jen Lin). Given training vectors $x_k \left(k=1,2,3...m\right)$, the F-score of the $i^{th}$ feature is defined as:

$$F(i)=\frac{(\overline{x_i}^{(+)}-\overline{x_i})^2+(\overline{x_i}^{(-)}-\overline{x_i})^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}({x^{(+)}}_{k,i}-\overline{x_i}^{(+)})^2+\frac{1}{n_--1}\sum_{k=1}^{n_-}({x^{(-)}}_{k,i}-\overline{x_i}^{(-)})^2},$$

where n+ and n− number stands for the number of positive and negative instances in the vectors; $\overline{x_i}$, $\overline{x_i}^{(+)}$,

[a] City University of Hong Kong: Computer science Department & Management Science Department;

[b] City University of Hong Kong: Manufacturing Engineering and Engineering Management Department;

[c] City University of Hong Kong: Information System Department;

$\overline{x_i}^{(-)}$ are the average of the $i^{th}$ feature of the whole, positive, and negative data sets, respectively; and $\overline{x_k}^{(+)}$, $i$ is the $i^{th}$ feature of the $k^{th}$ positive instance, and $\overline{x_k}^{(-)}$, $i$ is the $i^{th}$ feature of the $k^{th}$ negative instance.

To judge the performance of the classification when choosing different features, we use Accuracy as a simple measure. Accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

where $TP$ is the number of true positives (number of 'YES' patients predicted correctly), $TN$ is the number of true negatives (number of 'NO' patients predicted correctly), $FP$ is the number of false positives (number of 'YES' patients predicted as 'NO') and $FN$ is the number of false negatives (number of 'NO' patients predicted as 'YES').

In our experiment, we select features mostly with higher F-scores (not the highest), then use SVM to train and predict to find out if these features can bring us good result. Ours procedure is designed as follows:

**Procedure 1:** The procedure of F-scores methodology:

Step 1: Calculate F-score of every feature

Step 2: Pick some possible thresholds by human eye to cut low and high F-scores (noises).

Step 3: For each threshold, do the following

a) Drop features with F-score beyond this threshold.

b) Randomly split the training data into training set $T_k(k = 0,\cdots,5)$ and predicting set $P_k(k = 0,\cdots,5)$. Since the data is limited, the vectors are repeatable for different $T_k$ and $P_k$.

c) Let set $T_k$ be the new training data. Use the SVM procedure in Section 2 to obtain a predictor; use the predictor to predict set $P_k$.

d) Compute the overall Accuracy for each threshold

e) Repeat the above steps until every possible threshold has been tried.

Step 4: Choose the threshold with the highest Accuracy.

Step 5: Drop features with F-score beyond the selected threshold. The left features are the one we need.

**Table1.** Experiments result of Correct Rate of F-scores.

| | Test1 | Test2 | Test3 | Test4 | Test5 |
|---|---|---|---|---|---|
| LOW BOUND | 1.00E-08 | 1.00E-05 | 0.0001 | 0.0001 | 0.001 |
| HIGH BOUND | 0.8 | 0.2 | 0.2 | 0.8 | 0.1 |
| GENE AMOUNT | 1912 | 1783 | 1724 | 1821 | 1324 |
| ACCURACY | 92.771 | 80.645 | 67.857 | 85.507 | 60 |
| | Test6 | Test7 | Test8 | Test9 | Test10 |
| LOW BOUND | 0.001 | 0.001 | 0.01 | 0.01 | 0.04 |
| HIGH BOUND | 0.4 | 0.1 | 0.1 | 0.2 | 0.2 |
| GENE AMOUNT | 1621 | 1324 | 864 | 1018 | 1724 |
| ACCURACY | 81.429 | 60.76 | 65.753 | 86.42 | 87.5 |
| | Test11 | Test12 | Test13 | Test14 | Test15 |
| LOW BOUND | 0.1 | 0.2 | 0.15 | 0.1 | 0.11 |
| HIGH BOUND | 0.2 | 0.3 | 0.2 | 0.15 | 0.19 |
| GENE AMOUNT | 216 | 1137 | 180 | 108 | 192 |
| ACCURACY | 87.5 | 56.522 | 59.756 | 48.649 | 75.676 |

From the table we can easily find that, then the threshold is 0.1 and 0.2, the result is best. So we can choose features between these 216 thresholds.

Though SVM can perform fairly well when the features and training data are limited, the performance decreases when the features and training data are too scarce. So it is different for us to select better ones form these 216 features. Thus, we need to try other methods to boost our performance.

## 3. Oncogene Selection

In this section, we propose a class predictor able to accurately assign samples as cancer or normal. The first issue was to eliminate data noise by an image denoising method. And the second issue was to explore whether there were genes whose expression pattern was strongly correlated with the class distinction to be predicted. In principle, the class discovery techniques above can be used to identify fundamental subtypes of any cancer.

Our initial data set consisted of 62 samples, 40 of them are cancer specimen and 22 of them are normal specimen. Each sample is represented by an expression vector, consisting of its expression level in 1912 genes.

Considering that parallel coordinate plots are used for high-dimensional continuous data, we construct the image of DNA microarrays as parallel coordinate plots. Because parallel coordinate plots are often overrated concerning their ability to depict multivariate features. And they can help to find and understand features such as groups, clusters, outliers and multivariate structures in their multivariate context. The key feature is the ability to select and highlight groups or individual cases in the data, and compare them to other groups or the rest of the data.

A parallel coordinate plot draws an axis for each variable in the plot. As the name suggests, all axes are plotted in parallel. For the DNA dataset $X$ with $m$ people and $n$ gene, each people results in a poly-line. The edges of the polygon are the points $x_{ij}\,(i=1,\cdots,m; j=1,\cdots,n)$, which are plotted at axis $j$ and its coordinate value $x_{ij}$.

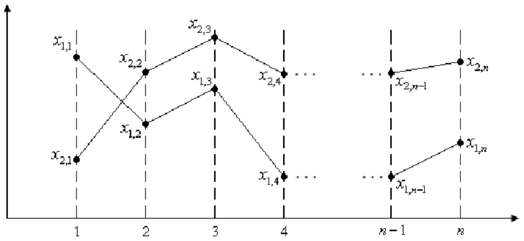

**Figure 1.** It illustrates an exemplary parallel coordinate plot for two $n$-dimensional points $X_1=\left(x_{1,1},x_{1,2},x_{1,3},\cdots,x_{1,n-1},x_{1,n}\right)'$ and $X_2=\left(x_{2,1},x_{2,2},x_{2,3},\cdots,x_{2,n-1},x_{2,n}\right)'$. The coordinate value for each point $x_{1,j}$ (and $x_{2,j}$) is plotted on each axis $j\,(j=1,\cdots,n)$ and then joined by a poly-line.

In this way, we can get an image representation of the DNA microarrays by drawing their parallel coordinate plots. There are 62 1912-dimensional observations plotted in parallel coordinates. 40 of them are cancer specimen and 22 of them are normal specimen. Each poly-line corresponds to a sample, with the columns corresponding to expression levels in different genes. The following is a section of the parallel coordinate plot.

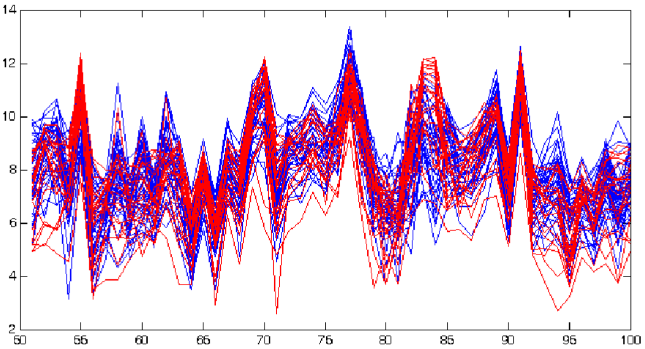

**Figure 2.** It is the parallel coordinate plot of gene 51 to gene 100. All genes are sorted according to their EST name in ascending order. The blue poly-line represent cancer specimen, while the red poly-line represent normal specimen.

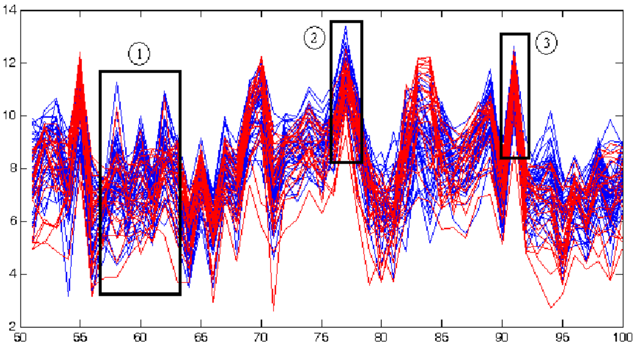

**Figure 3.** It illustrates three kinds of gene. In the first pane, the blue and red poly-lines are messy. It seems that they are disturbed. So we regard the genes in this pane as **noisy gene**. We cannot infer the effect of these genes. In the second pane, trends of all poly-lines are similar. And the coverage of blue poly-lines is higher than that of red poly-lines. So we regard the genes in this pane as **impact gene**. We can infer that these genes may act on colon cancer. In the third pane, trends of all poly-lines are similar. And the coverage of red poly lines overlaps with that of blue poly lines. So we regard the genes in this pane as **irrelevant gene** for colon cancer. We can infer that these genes have no influence with colon cancer.

Since the DNA microarrays are expressed by parallel coordinate plots, we can get the 3-dimension RGB data of the plot by Matlab software and then apply an image denoising methodology to clean the data.

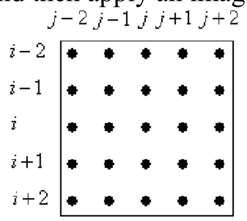

**Figure 4.**It illustrates the image denoising method. The color of points in the plot could be white, red or blue. For point $(i,j)$ in the plot, we zone a $5\times5$ area that taking point $(i,j)$ as the center point. So we have 25 points in this area. Then we calculate the number of white points, the number of red points, and the number of blue points, respectively. And divide 25 to get their percentage. If the percentage of one color is greater than 0.8, we set the color of point $(i,j)$ as that one. Otherwise, we keep its original color. For example, if the percentage of red points in the area is greater than 0.8, we set the color of point $(i,j)$ as red.

We scan all points in the plot. And set the color of each point with the above method. After image denoising, we get a new image of the original plot. The new image is filled by color lump rather than poly-line. For a gene in the plot, we think that the range of red points represents the normal fluctuation range.

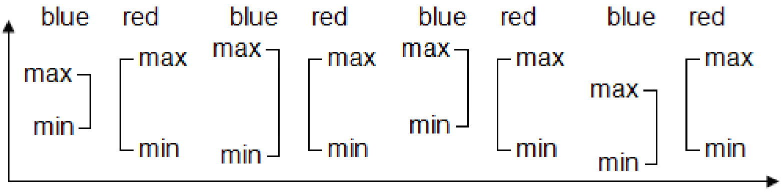

**Figure 5.** It illustrates four kinds of relationship between the range of blue lump and that of red lump. In the first case, the blue range is overlapped or covered by the red range. It means that the gene is irrelevant to cancer. In the other three cases, the range of blue lump is larger, upper and lower than that of red lump, respectively. It means that the gene is impact to cancer.

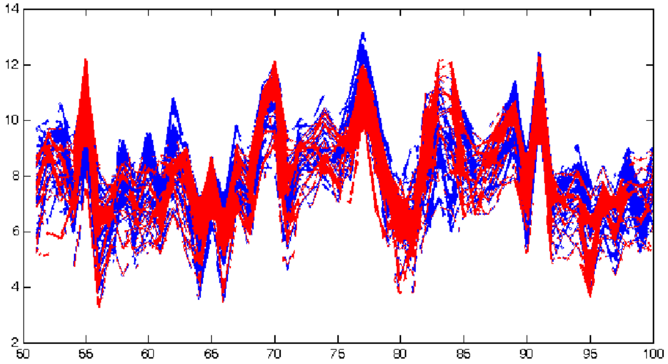

**Figure 6.** It shows the denoised image of **Figure 3.** The new image is filled by color lump rather than poly-line.

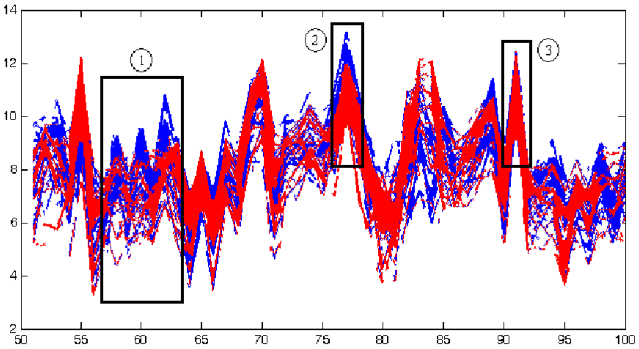

**Figure 7.** It shows the effect of the denoising method. In the first pane, the shape of the color part is clearer than that in **Figure 3.** In the first and second pane, the range of blue points is different with that of red points. So we regard these genes as **impact gene** for colon cancer and infer that these genes may act on colon cancer. In the third pane, the blue part is just overlapped by the red part. So we regard the genes in this pane as **irrelevant gene** for colon cancer and infer that these genes have no influence with colon cancer.

By applying the above method, we regard 187 genes as irrelevant gene and the other 1725 genes as impact gene. For a gene, the difference between the range of blue points and that of red points illustrates its impact extent on colon cancer. So we sort the impact genes according to their range difference in descending order, and select the first 6 genes as critical gene. The result is as follows.

**Table 2.** Critical Gene Selection after Data Denoising

| Internal ID | EST name | GenBank Acc No | Mapped region | Range Difference |
|---|---|---|---|---|
| 1967 | Hsa.11616 | T60778 | 3' UTR | 86 |
| 1635 | Hsa.2097 | M36634 | gene | 60 |
| 1795 | Hsa.3015 | Z18948 | gene | 60 |
| 1978 | Hsa.58 | D00760 | gene | 57 |
| 1668 | Hsa.1454 | M82919 | gene | 54 |
| 625 | Hsa.3306 | X12671 | gene | 54 |

In this case, the error rate on the leave-one-out test is 9/62 errors. The following figure shows the denoised image of these six critical genes.

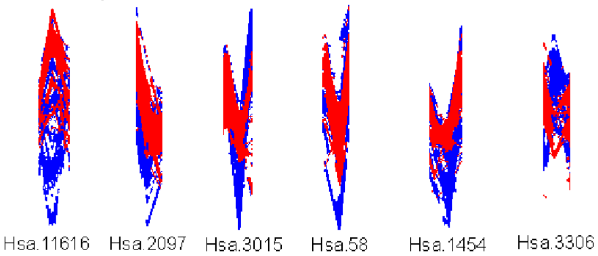

**Figure 8. Six Critical Genes for Colon Cancer Classification.** It shows the image of the six critical genes.

## 4. Estimation by Leave-one-out.

We use leave-one-out test methodology and have result as follow:

**Table 3. Estimation result by leave-one-out.**

| Estimation results | Distant model | Denoise of Gene Data | Denoise of Gene Data | Distant model | Denoise of Gene Data |
|---|---|---|---|---|---|
| Number of Genes | 17 | 12 | 6 | 5 | 5 |
| Accuracy | 50/62 | 53/62 | 53/62 | 48/62 | 52/62 |
| Accuracy Rate | 80.64% | 85.48% | 85.48% | 77.41% | 83.87% |

We can see from the table that Denoise of Gene Data method is adoptable and better that any other distance models.

## 5. Future work:

Given two events A and B, we want to know the probability that B may happen if A happened. It is obviously a conditional probability problem. We can find the probability of B under the condition A, and define it as P(B|A). It is quite similar to the gene prediction. If we know that gene sequence A refers to cancer, and it had mutated, we can take it as a conditional probability problem to find P(label sequence B | observation sequence A) (John Lafferty, Andrew McCallum, Fernando Pereira). Conditional Random field is designed to deal with this conditional probabilistic problem (Hanna M. Wallach).

Taking gene sequence as vectors, then different gene sequences (vectors) can make a matrix (with the same gene in the same row). Then we can employ CRF model to predict the situation of a specific gene sequence according to the sequences both direct related to them and indirect related to them in the position aspects and function aspects.

In fact, we can use CRF model to do more than predict the mutation of genes. We can use it to predict whether a man will have cancer directly. In the training data, we just need take whether a man have cancer as a feature after the possible gene sequence. After training, we can predict what the feature is after other men's gene sequences (containing the corresponding genes in these sequence). For instance, we have 4 genes that may relate to one kind of cancer (say ABCD). We can take whether men with these gene sequences as a feature E (E can be valued as YES, NO, or something else related to the cancer condition). Then we use enough gene sequences (ABCD + E, A,B,C,D and E can be different data in different gene sequences) as train data. When comes a new gene sequence, A'B'C'D'E', we can use CRF to predict what E' is. That's to say we can find the cancer condition of the man with the new gene sequence.

As CRF has so many potential advantages in both gene mutation and cancer prediction, it is an interesting and attractive topic which worth our further discussion and research.

## References

[1] Martin Theus and Simon Urbanek (2008). *Interactive graphics for data analysis: Principles and Examples*, CRC Press.

[2] (John Lafferty, Andrew McCallum, Fernando Pereira) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, ICML 2001

[3] (Andrew McCallum, Dayne Freitag, and Fernando Pereira) Maximum Entropy Markov Models for Information Extraction and Segmentation, 17th International Conf. on Machine Learning, 2000

[4] (Rahul Gupta), Conditional Random Fields

[5] (Hanna M. Wallach) Conditional Random Fields: An Introduction

[6] (Chih-Chung Chang and Chih-Jen Lin) http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[7] (Yi-Wei Chen and Chih-Jen Lin) Combining SVMs with Various Feature Selection Strategies

[8] (Breiman 2001) Breiman, L. (2001). Random forests. Machine Learning 45(5): 5 – 32.

[9] Analysis of Gene and Protein Expression Data, John Wiley & Sons, Inc., Hoboken, New Jersey

[10] (Yun Sing Koh1 Russel Pears2) Rare Association Rule Mining via Transaction Clustering

[11] (Darius M. Dzuida)Data mining for genomics and proteomics : analysis of gene and protein expression data

[12] (Yun Sing Koh, Nathan Rountree.)Rare association rule mining and knowledge discovery : technologies for infrequent and critical event detection

[13] Martin Theus and Simon Urbanek (2008). *Interactive Graphics for Data Analysis: Principles and Examples*, CRC Press.

[14] Richard Desper, Feng Jiang, Olli-P. Kallioniemi, Holger Moch, Christos H. Papadimitriou, and Alejandro A. Schaffer (1998). Inferring tree models for oncogenesis from domparative genome hybridization data, *J Comput Biol.*, 1999 Spring, 6(1): 37-51.

[15] Li Yingxin, Liu Quanjin, Ruan Xiaogang (2005). Analysis of leukemia gene expression profiles and subtype informative genes identification, *Chinese Journal of Biomedical Engineering*, April 2005, 24(2): 240-244.

[16] Debashis Ghosh and Arul M. Chinnaiyan (2002). Mixture modelling of gene expression data from microarray experiments, *Bioinformatics*, 2002, 18(2): 275-286.

[17] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Leine (1999). Broad patterns of gene of expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide arrays, *Cell Biology*, June 1999, 96: 6745-6750.

[18] Isabelle Guyon, Jason Weston, Stephen Barnhill, M. D. and Vladimir Vapnik (2002), Gene selection for cancer classification using support vector machines, *Machine Learing*, 2002, 46(1-3): 389-422.

[19] Desmond Carney and Karol Sikora (1990). *Genes and Cancer*, John Wiley & Sons Ltd., England, 1990.

[20] F. Macdonald, C. H. J. Ford and A. G. Casson (2004). *Molecular Biology of Cancer*, Second Edition, Bios Scientific Publishers, London and New York, 2004.

[21] B. A. J. Ponder and M. J. Waring (1995). The Genetics of Cancer, Kluwer Academic Publishers, USA, 1995.

[22] R. Scott Hawley and Michelle Y. Walker (2003). Advanced Genetic Analysis: Finding Meaning in a Genome, Blackwell Publishing, Germany, 2003.