Internal Report CS Bioinformatics Track 10-02 July 2010

Universiteit Leiden

Opleiding Informatica

Immune Response to Prostate Cancer

- Exploration of Normalization, Feature Selection and Classification Procedures -

Ruifang Li

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)

Leiden University

Niels Bohrweg 1

2333 CA Leiden

The Netherlands

Abstract:

BACKGROUD: Prostate cancer is the most common form of cancer in men with an incidence about 670,000 new cases annually world-wide (*CDC, American and Canadian Cancer Societies*; Boyle, 2004, ERSPC). It is the leading non-skin cancer in men above 65, and one out of six men will be affected by prostate cancer during his life. Screen for prostate-specific antigen (PSA) has led to the earlier detection of disease, but increased serum PSA can be present in non-malignant conditions such as benign prostatic hyperplasia as well. So the supplementary biomarkers are strongly needed to improve the diagnosis and prognosis accuracy. There are various biomarker techniques from genomics, proteomics, pharmacogenetics to integrative approaches. Immune response protein microarray as one of the proteomics techniques with the advantage of taking post-translational modification of protein into account is quickly emerging as a follow-up technology.

METHOD: This work is based on a data set with 120 blood samples from five groups from control to advanced metastasis. Considering the difference between DNA microarrays and protein microarrays, the first step is to compare different well-established normalization methods such as global normalization, quantile normalization, VSN as well as robust linear model normalization on this protein microarray data. After the proper normalization, one-side Wilcoxon test is performed on biomarker selection. In order to overcome the problem of simultaneously multiple-test as well the sensitivity of the hit gene list highly depending on the training samples, the re-sampling tests instead of just selecting the most significant genes by one test is used. Besides, gene selection strategy in the process of classification such as random forest, shrunken centroid as well as recursive feature elimination are also used to derive different hit gene lists, which can be used to verify the biomarker selected in statistical test. All the hit genes from different methods are further checked by online annotation database. Except from biomarker discovery, the obtained gene lists from different methods are also used to clarify the underlying prostate cancer progression by enrichment pathway analysis, and the classification performance of these gene signatures will be evaluated by principle component analysis (PCA), and leave-one-out cross-validation error rate of different classifiers, including K-nearest neighbour (KNN) as well as support vector machine (SVM).

CONCLUSION: Here we show that the quantile followed with robust linear model normalization strategy works better than other counterparts. At the same time, we developed a combinatorial strategy on gene selection, although the variation on hit gene list shows significant different, there are always a few overlapping genes with strong diagnosis potential. Meanwhile, enrichment pathway analysis on different hit gene lists also shed light on prostate cancer progression. Altogether, this study provides insight into immune response protein microarray analysis from normalization, gene selection aspects.

Keywords: prostate cancer, prostate biopsy, prostate, benign prostate hyperplasia, machine learning

CONTENT

1. Introduction

Prostate cancer is the most common form of cancer in men with an incidence about 670,000 new cases annually world-wide (*CDC, American and Canadian Cancer Societies*; Boyle, 2004, ERSPC). It is the leading non-skin cancer in men above 65, and one out of six men will be affected by prostate cancer during his life. There are three main types of prostate disorder:

- **Prostatitis:** an inflammation of the prostate gland in men, which may cause some similar symptoms as cancer, but is not cancer.
- **Benign prostatic hyperplasia (BPH)**: also known as benign enlargement of the prostate refers to the increase in size of the prostate in middle-aged and elderly men. It leads to symptoms of urinary hesitancy, frequent urination, dysuria (painful urination), increased risk of urinary tract infections, and urinary retention. Although BPH causes many same symptoms as cancer, it is not considered to be a premalignant lesion as well.
- **Prostate cancer:** a form of cancer that develops in the prostate. The mortality danger of cancer is obvious much serious than the other two types of prostate diseases. High cure rates are generally achieved with early stage prostate cancer.

Two **screening tests** commonly used to detect prostate cancer in the absence of symptoms:

- **digital rectal exam (DRE)**, a doctor feels the prostate through the rectum to find hard or lumpy areas
- a **blood test** that detects a substance made by the prostate called prostate-specific antigen (PSA)

Together, these tests can detect many "silent" prostate cancers that have not caused symptoms. If prostate cancer is found during screening with the PSA test or DRE, the cancer will likely be at an early, more treatable stage than if no screening were done. Due to the widespread use of PSA testing in the United States, approximately 90 percent of all prostate cancers are currently diagnosed at an early stage, and, consequently, men are surviving longer after diagnosis.

There is no question that the PSA test can help spot many prostate cancers early, but neither the PSA test nor the DRE is 100% accurate. These tests can have abnormal results even when cancer is not present (known as false positive results). In addition, normal results can occur even when cancer is present (known as false negative results). False positive results can lead some men to undergo a prostate biopsy cancer is not present, with the cost of unnecessary pains, infection, and bleeding. False negative results more dangerous delay treatment, and higher mortality. So the additional genes are needed to improve the accuracy of prognosis as well as diagnosis.

Till now the accurate diagnosis of prostate cancer can be confirmed only by biopsy, a surgery on removing tissue samples, usually with a needle. Out of question, this kind of invasive diagnosis will take up more medical resources as well as making patients suffer from pains. So identifying additional novel prognosis and diagnosis biomarker is so important.

DNA microarrays enable us inspect thousands of genes simultaneously at the level of mRNA, which

have found particular value in analyzing clustered gene expression, revealing co-regulated gene networks [4] in the past decades. However, as we know, the human genome contains over 20,000 genes, which is further confirmed by "The International Human Genome Sequencing Consortium" researchers in October 2004 that besides the existence of 19,599 protein-coding genes, another 2,188 DNA segments predicted to be protein-coding genes, but these genes code for more than 200,000 proteins, not to mention post-translational modifications [2]. From this point of view, RNA levels don't correlate exactly with proteins'. What's more, it's protein not DNA that exerts all cellular functions in human bodies. So, high-throughput proteomics technologies are urgently demanded for advanced cancer research. Demand driving inventions, there are quite a few proteomics tools used to diagnose biomarkers for the early detection of cancer.

➢ 2D-PAGE

One traditional way called two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) is to detect biomarkers through changes in serum protein concentration. Unfortunately, this method is labour intensive and requires large amounts of samples, which remain a low-throughput proteomic analysis approach [2].

➢ SELDI-ToF-MS

A newly developed MS-based tool named "SELDI-ToF-MS" implemented in the ProteinChip system from Ciphergen Biosystems Inc. is designed to detect the masses of proteins that are differentially expressed in serum on chromatographic array surfaces, but the protein can't be identified, and further experiments are needed to know that protein. Since the identification of the peak masses in the classifier is not necessary for making a diagnosis, in the original paper there is no report about the identification of these biomarkers. However, such information could lead to better therapeutic interventions as well as understanding oncogenesis of prostate cancer. In Table 1-1 two early studies by this technique are listed, from which we can find even with the same technology, different peaks are recognized as biomarkers to distinguish serum from health and diseased individuals, so one drawback is revealed that this technology is very sensitive to experimental details.

In a further study [8] of these two researches [6] [7], an iso-form of ApoA-II giving rise to a common peak that is specifically over-expressed in prostate cancer. ApoA-II is a protein that is encoded by APOA2 gene in human body, which is the second most abundant protein of the high density lipoprotein particles. From this point of view, another drawback is exposed that is it's a labour intensive work to identify biomarkers. There are quite many mass peaks but not all of them can be identified after tedious and complicated experiments.

Table 1-1

Two prostate cancer biomarker detection studies by SELDI-ToF-MS

Data set	Methods	Biomarkers	Reference
167 PCA	Decision tree	Protein fingerprint pattern of **9** masses, and efforts are	[6]

77 BPH 77 HM	classification	under way to purify, identify and characterize these protein/peptide biomarkers	
197 PCA 92 BPH 96 HM	Boosting tree algorithm	**124** peaks identified in the training set were used to construct the classifier, and the protein identification is in progress.	[7]
PCA: prostate cancer; BPH: benign prostate hyperplasia; HM: healthy men			

➢ Protein microarray

Protein arrays are composed of hundreds or even thousands of proteins immobilized on a solid surface, and there are generally two kinds of protein arrays, named abundance based protein microarray and function based protein microarray, which is compared in table 1-2 below.

Table 1-2

Two types of protein microarray [4]

Name	Feature	Application in prostate cancer research	Example
Abundance-based protein microarray	Rely heavily on the availability of well defined and highly specific ASRs, and currently the most available ASRs are antibodies	Miller et al. [9]used capture microarray containing 184 antibodies identified five proteins (*von Willebrand Factor, immunoglobulinM, Alpha1-antichymotrypsin, Villin and immunoglobulinG*) that had significantly different levels between the prostate cancer samples and the controls	1). Capture microarray 2). Reverse-phase protein microarray
Function-based protein microarray	Microarrays with immobilized any type of functional proteins	The data set of my work comes from Invitrogen's ProtoArray Human Protein Microarrays v4.1 containing over 8,000 purified human proteins immobilized on glass slides	1). Autoantigen array 2). Self-assembling protein microarrays

ASRs: analyte-specific reagents, a class of biological molecules which can be used to identify and measure the amount of an individual chemical substance in biological specimens

➢ Difficulties on finding serum biomarkers

1) Protein concentrations span in a huge range so it's hard to observe many valuable low-level expression biomarkers;
2) Protein concentrations are changed markedly with stress, disease and treatment. So ideally, algorithms employed for serum proteomic profiling should filter out temporal fluctuations in the serum proteome which are unrelated to the disease being considered;
3) Proteins can be modified by cleavage or addition of new function group's changes that may affect detection.

1.1. Objective

When facing gene selection problem, there are two main objectives:

1) To identify a small set of genes for clinical prognosis purpose; in this circumstance, the smallest promising set of genes are selected that can still achieve good predictive performance.
2) To identify a relative large set of genes that are related to outcomes, and these genes may be correlated to each other, but they can shed light on new molecular pathways involved in cancer progression.

In this work, the main goal lies in the first aspect above, i.e. target a few novel prostate cancer-associated genes that can be acted as supplementary biomarkers for PSA in the early-stage prognosis and screen, and in order to realize this goal, the comparison study between group2 and group3, group2 and group4 as well as group3 and group4 are conducted separately.

At the same time, the study on cancer progression pathways is also considered, and besides the group comparisons motioned above, group1 is also compared with group5, i.e. control group vs. the most serious metastasis group, which may shed more light on the cancer progression.

1.2. Problem description

Patients with benign prostatic hyperplasia show same symptoms as malignant disease, and increased serum PSA can be present in non-malignant conditions as well. So one of the most challenging problems are target novel early-stage prognosis biomarkers between benign samples from low-grade ones. However, finding this kind of biomarkers is proving problematic, and over the past decade the Food and Drug Administration (FDA) have approved only a few new diagnostic biomarkers, only PSA has been discovered to be useful in testing for early cancer. The difficulties may lie in several aspects:

1) Technique limitations: from genomics to proteomics techniques, there are always based on some kind of assumption and existed some drawbacks to overcome
2) Sample selection limitation
3) Unavoidable experimental mistakes
4) Unthinkable complexity of diseases

Downstream analysis as the last procedure of biomarker discovery takes the responsibility to minimize the bad effects derived from the above problems.

As we know, many cancer-associated genes remain to be indentified to clarify the underlying molecular mechanisms of cancer progression, especially in the context of complex cellular networks. Cancer cells often bear mutations in the genes responsible for various signal-transduction pathways leading to proliferation in response to external signals. Many growth factors, their receptors, cytoplasmic and nuclear downstream effectors of singling and apoptotic pathways have been

identified as oncogenes or tumor-suppressor genes, which are just revealed the tip of the iceberg from current knowledge. According to study different stages of tumor profiling, we can get some idea on the involved pathways for cancer progression, and at the same time, discover the disease-tailored classification methods.

1.3. Solution approach

Normalization

Protein microarrays are measured with either systematic or random variability, potentially controlled by different set of control proteins spread intra- and inter-arrays, which will affect the statistical power in the following analysis. The normalization methods for DNA microarrays have been well established and described, however, "the assumptions for the analysis of DNA microarray do not always translate to protein arrays [1]". So, some related study on comparing different well-known normalization methods from DNA microarray to protein arrays has been done in paper by Andrea Sboner et.al [1], and in that paper, the authors got the conclusion that RLM (robust linear model) normalization performs better than quantile as well as global normalization methods. At the same time, the performance didn't show much difference on distinctive control proteins by RLM.

In paper [1], the comparisons between different normalization methods are based on the assumption that "an effective normalization procedure should reduce the variability in the signal caused by systematic artifacts without losing useful biological information". So, they defined inter- and intra-array variation coefficient (CV) to measure the variation between control proteins, and definitely both of them should be the smaller the better. Then how to describe the biological difference between two groups? The authors adopted the Fisher's signal-to-noise ratio as the measurement, which is inherent in other common methods such as linear discriminate analysis.

- **Intra-array variation coefficient (CV)**

$$C = \frac{\sigma}{\mu}$$

Where σ is the standard deviation of control protein spot intensities at the same position across different sub-arrays on each array

μ is the mean of the protein spot intensities at the same position across different sub-arrays on each array

- **Inter-array variation coefficient (CV)**

$$C = \frac{\sigma}{\mu}$$

Where σ is the standard deviation of control protein spot intensities at the same position across different arrays

μ is the mean of the protein spot intensities at the same position across different arrays

- **Fisher's signal-to-noise ratio**

$$S = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

Where μ_1, μ_2 are the mean intensities of a normal protein spot in each group

σ_1, σ_2 are the corresponding standard deviations for each normal protein spot

The principle of Fisher's signal-to-noise ratio can be explained in the following graph 1-1, for each normal protein spot in the two groups, the signal intensities will form a distribution as D1 and D2. A good separation between two groups should maintain the distance between mean values of groups as large as possible, i.e. maximize (m1-m2), and at the same time, the difference within each group should be as small as possible. In all, this Fisher's ratio will be the larger the better.

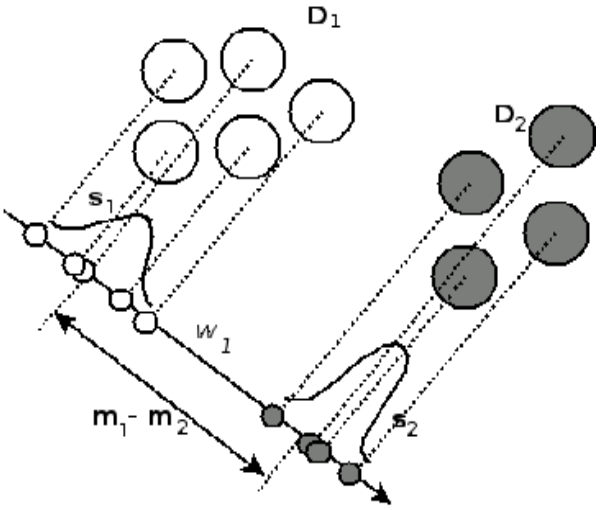


Figure 1-1 principle of Fisher's signal-to-noise ratio

In my work, I also employ the above three measurements as criteria to determine which normalization method is better. Besides the methods used in paper [1], i.e. global, quantile and robust linear model, I also apply VSN (variance stabilization normalization) as well as different combinations of single method such as VSN with RLM and quantile with RLM.

Most of normalization methods are based on the assumption that there are just a few biomarkers and the array distributions are similar or can be considered as identical.

- Global normalization

  This strategy is scaled each array with a factor such that the signal medians of all the considered arrays are the same as the overall median. So in my work I chose this factor as median (all arrays)/median (each array).

- Quantile normalization

  This method assumes that the distribution of actual signals is the same in all samples and adjusts

the observed data accordingly. In one word, the largest signal for each array is replaced by a median valued of the largest signals; the second largest signal is replaced by a median value of the second largest signals, and so forth. I use the implementation provide by Bioconductor package “limma” to do this normalization.

- Variance stabilization normalization (VSN)

  The *vsn* method builds upon the fact that the variance of microarray data depends on the signal intensity and that a transformation can be found after which the variance is approximately constant.

  It is like the logarithm at the upper end of the intensity scale, approximately linear at the lower end, and smoothly interpolates in between. The position of the cross-over point and the slope of the linear part depend on the error distribution of the data. It also incorporates the estimation of "normalization" parameters (shift and scale).

  vsn assumes that less than half of the genes on the arrays is differentially transcribed across the experiment. An advantage of *vsn*-transformation over log-transformation is that *vsn* works also on values that are negative after background subtraction.

- Robust linear model (RLM)

  Taking into account the intra- and inter-array effects simultaneously, I adopt the RLM introduced in paper [1]:

$$\log_2^{(\text{signal intensity})_{ijkr}} = \alpha_i + \beta_j + \tau_k + \varepsilon_{ijkr}$$

  Where α_i is the slide effect of slide i, and for different normalization targets I changes from 1 to the maximum slide number;

  β_j : stands for the sub-array effects, and in this experiments the j varied from 1 to 48;

  τ_k: The effect of protein k, in this work the k is in the range of 1 to different number of control proteins;

  ε_{ijkr}: The random error part is the residue part of parameter estimation of linear model.

Gene selection and group classification

In the gene selection procedure there are two main problems:

1) The hit gene list is sensitivity to the change of training samples, and in other word, the hit gene list is heavily depending on the selected training samples. One way to overcome this problem lies in random re-sampling validation on the selected hit gene list. In my work, instead of using statistical test once, I use random re-sampling on all samples to get a training set to derive

different hit gene lists, and the final reliable hit gene list is the highly overlapping genes of iterations.

2) Using statistical method to obtain hit gene list is the most classical, straightforward as well as well-established one, however, ranking genes by univariate statistic is always criticized as ignoring the relationships between genes, and the selecting an optimal number of genes is another drawbacks of statistical method for gene selection. Recently, more and more research are focus on selecting genes in the process of classification, and the same classification method is repeatedly applied over a training set, and each time some genes will be eliminated from the whole gene set by different criteria which reveals the effect of elimination certain genes. Among these kind of methods, the study on gene selection strategy by random forest is very active, as well as the gene selection by shrunken centroid and recursive feature elimination. In my work, I tried all these three promising methods and using obtained different gene signatures to find the enriched pathway in the process of disease progression.

I work on group pair comparisons between group 2 (benign group), group 3 (low-grade cancer), group4 (locally advanced cancer), intending to explore the potential prognostic biomarkers for early-stage detection. The first and the most important goal are to target a small set of genes performing well on group separation. Then, the comparison between group1 and group5 is conducted by the same workflow. The methods I used in two aspects:

1) The most popular and well-formed methods to derive hit gene list is by statistical method. Either by statistical test or correlation analysis, a list of genes with significantly differential intensity will be selected. In this work, I use Wilcoxon test as well as calculating correlation coefficients to group labels. Considering the reality that on protein microarray the decreasing intensity provides useless information, I adopted one-side Wilcoxon test with 0.05 significant level based on the assumption that the promising prognosis biomarker will show increasing intensity with the disease progression. However, when a large number of tests are made on the same data, suppose the significant level is 0.05, a p-value of 0.05 means a 5% probability that the protein's signal in one group is higher than the other by chance alone. In this experiment there are 8302 proteins are tested, 5% or 415 proteins might be selected as significant ones by chance alone. At the same time, most of the normalization methods are based on the assumption that there are just a few significant changed genes from array to array, so it means that maybe all of the significant proteins derived from the tests are selected by chance only. In order to avoid the problem of multiple-test, I designed a four-step statistical method to acquire a reliable hit gene list.
   - **Step1**: carry out one-side (H1: group2<group3) Wilcoxon test on each normal spot protein with total number of 8302 on all the 44 samples from group2 and group3;
   - **Step2**: plot the histogram of p-values from the tests to evaluate many tests simultaneously. If the null hypothesis is true, the p-values tend to uniformly distributed on the interval [0, 1]. If the null hypothesis is false, i.e. the alternative hypothesis is true, and then the distribution of p-values will tend to have smaller values, which looks like graph in figure1-2. [17]

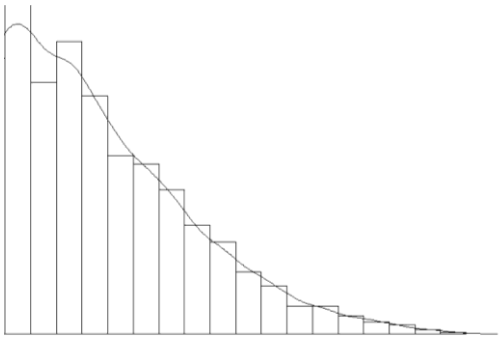


Figure 1-2 a good example of p-values plot

- **Step3**: Considering the hit gene list heavily depending on the training samples, I use random re-sampling on all the samples, and perform the same one-side Wilcoxon test on the randomly selected 90% of samples with 50 times;
- **Step4**: Study the genes emerge at least 40 times over 50 iterations as significant one to distinct the two groups, which is also selected on all the samples.

2) Currently more and more research focus on gene selection in the process of classification, such as random forest, shrunken centroid, as well as recursive feature elimination. All of these methods will return a hit gene list with small or large quantities of genes by minimizing cross-validation error rate in the classification process. The application of the derived gene list from these methods lies in two ways:

- Try to find overlaps with the gene list selected by statistical methods. The ultimate prognosis gene should show significantly increasing intensity with the disease progression, but the gene list selected by classification may contain either intensity-increasing or intensity-decreasing genes. So, on the premises of intensity-increasing, we can further confirm the gene selected by statistical method.
- Try to shed light on the disease progression pathways. The gene list selected by the classification process is also considered as making the largest contribution to discriminating between groups, so it may own huge values for cancer progression pathway analysis. I used a four-step strategy to judge the performance of the obtained gene-signature:
  - Observe the separation of samples on the first two components by PCA

    As a kind of unsupervised methods, i.e. classify samples without information about group labels, the separation of sample just from the data distribution point of view;
  - Minimum leave-one-out cross-validation error rate by k nearest neighbour classifier (KNN)
  - Minimum leave-one-out cross-validation error rate by linear discriminate analysis (LDA)
  - Minimum leave-one-out cross-validation error rate by support vector machine (SVM)

After these procedures, verify the selected genes by online annotation database, including swiss-prot, GO, KEGG etc.

Gene selection by random forest

How to construct a single classification tree?

The workflow in figure 1-3 explain how the classification tree constructed clearly, and the

last step of “tree pruning” is quite important because each subsequent split has a smaller and less representative population with which to work. Towards the end, unique of training records at a particular node display patterns that are peculiar only to those records. These patterns can become meaningless and sometimes harmful for prediction if try to extend rules based on them to larger populations.

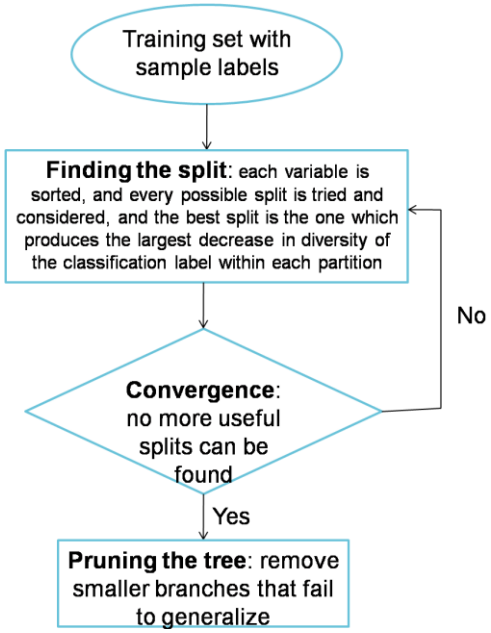


Figure1-3 workflow of classification tree

- *How to construct random forests?*

Assuming we know how to construct a single classification tree, the random forest is just a collection of trees construct on a subset of features, and the size of subset feature is held constant during the forest growing. Each tree is grown to the largest extent possible, and pruning is not taken into account.

At the same time, there is no need for cross-validation or a separate test set because each tree is constructed using a different bootstrap sample from the original sample pool with replacement.

To classify a new object from an input, this object put down each of the trees in the forest, and each tree will gives a classification decision. The final result is given by voting among different classifiers.

So, random forest is a kind of combined classifier getting final decision by majority voting. “Random” here refers to the process to select subset features by random selection, which will affect the correlation between two trees and the forest error rate. Study shows that increasing the correlation between trees will increase the misclassification error rate.

- *How to use random forest selecting genes?*

Random forest returns several measures of variable importance. The most reliable measure

is based on the decrease of classification accuracy when values of variable in a node of a tree are permuted randomly.

Before explaining how to select the gene set, the "**out-of-bag (OOB) error estimation**" should be clear first. Since about 1/3 of the samples are left out of the bootstrap sample and not used in the construction of the tree, so after the K^{th} tree is set up, just put the 1/3 left out samples down the K^{th} tree, and in this way, a test set classification is obtained for each case in about 1/3 of the trees. At the end of the tun, take j to be the class that got most of the votes every time case n was out of bag (oob). The proportion of times that j is not equal to the true class of n averaged over all cases is the oob error estimate.

To select the gene list, we just iteratively fit random forests, and each time built a new forest with reduced variable set discarding 20% of the least important variables. After fitting all forests, the OOB error rates are examined from all the fitted random forests. The solution with the smallest number of genes whose error rate is within u standard errors of the minimum error rate of all forests is selected. In this process the variable importance is no recalculated.

- *Discussion on multiplicity of gene selection by random forest*

Variable selection with microarray data can lead to many solutions that are equally good from the prediction rate point of view, but they share few common genes. This problem has been emphasized by a lot of research [18], although this is not serious when the objective is prediction, it brings doubt on the biological interpretation and clinical practice.

In paper [15] the study about stability of variable selection by random forest is conducted on 10 independent microarray dataset, and the result shows that the stability is quite poor and from time to time a lot of different set of gene combination are selected, and the overlapping is limited. Does this mean the feature selection by random forest is useless?

Also in paper [15], some metrics of this method is introduced:

1) By using random forest, the hit gene list is always very small comparing to other alternative selection methods in the process of classification, while maintaining good prediction performance.
2) The returned set of genes is not highly correlated as other selection methods.

Based on the above advantages of variable selection by random forest, some applications are recommended in paper [15]:

1) When design of diagnostic tools, and just a small set of features is desirable;
2) Surrogate for other gene selection methods returned many correlated gene involving in complex processes.

Taking into account these cons and pros, I will use this gene selection by random forest as the supplementary way for analyzing hit gene list from statistical methods, verifying the one or two promising diagnostic bio-markers derived from statistical analysis above.

Gene selection by shrunken centroid

The main idea of nearest shrunken centroid is to identify a subset of genes that best characterize each class, and a new observation is classified to the nearest centroid.

- *How to get these "de-noised" centroid?*

In one word, these centroids are achieved using soft-thresholding, so that for each gene, class centroids are shrunken towards the overall centroids.

$$\bar{x}'_{ik} = \bar{x}_i + m_k(s_i + s_0)d'_{ik}$$

Where i ---gene index, for each gene of 8302

k --- class index, for each class of group2 or group3

$\bar{x}'_{ik}$ --- the mean intensity in class k for gene i

$\bar{x}_i$ --- the mean intensity for each i across all class

$(s_i + s_0)$ --- within-class standard deviation for gene i plus a positive constant with the same value for genes

$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+$: Δ is the soft-thresholding we are looking for to restrict the gene list.

By changing the value of Δ from small to large, many of the genes are eliminated. What kind of gene will be dropped? For a gene i, d'_{ik} is shrunken to zero for all classes, then $\bar{x}'_{ik} = \bar{x}_i$, i.e. gene i shows the same intensity for all classes, which means it contributes nothing to the nearest-centroid computation, so this gene will be eliminated as useless gene for classification. The Δ is chosen by cross-validation always.

- *How to use shrunken centroid method to get interesting gene list?*

The optimal amount of shrinkage centroid can be determined by cross-validation, and in this work I choose the number of genes with the minimum cross-validation error rate.

The experiment design is give in section 2. In section 3, the results of different gene selection and classification methods are described. Then, some findings derived from results in section 3 are discussed in section 4. After the discussion, in section 5 some conclusions are given.

1.4. Intended audience

This report is addressed to several groups of related audiences:

- Cancer research experimenters devoted to biomarker discovery in laboratory. In this report, some potential prognosis genes are selected by statistical and computational methods, which are urgently waiting for experiment verification.
- Beginner of microarray data analysts. In this report, from the technique background, experiment design to current popular normalization, machine learning and statistical methods on microarray data analysis are explicitly explained step by step, which provides abundance review on microarray downstream analysis.
- Bioinformaticians working on biomarker discovery. The selected potential biomarkers and highlight pathways discovered in this report can supply a result comparison resource for their own findings.
- Protein microarray users. Protein microarray as a kind of fresh techniques is quickly emerging as a follow-up technology but is still lack of methodology. In this report, from upstream experimental design to downstream data analysis main steps are introduced.

The intended audience are supposed to be familiar with basic idea of molecular biology of cancer as well as some knowledge on pattern recognition and statistics. The reference book for the background of molecular biology of cancer is recommended as “Molecular biology of cancer --- mechanism, targets, and therapeutics” written by Lauren Pecorino, published by Oxford University Press. To know more about the background of pattern recognition and statistics, “The elements of statistical learning --- data mining, inference, and prediction” authored by Trevor Hastie et.al and published by Springer is recommend here.

2. Experimental Setting

Immune Response Biomarker Profiling was performed with one hundred and twenty (120) serum samples, all of which were divided into five groups, with 24 each. The reactivity of IgG antibodies in the serum against proteins on ProtoArray Human Protein Microarrays containing 8,302 proteins was investigated. Figure 2-1 explicitly explains how the serum profiling assay conducted and anti-human IgG as one of control protein feature presented on each array was used to identify proper scanning parameters and normalization targets. Since arrays profiled with samples 46, 60, 61, 62, 84, 108, 109 had mean signal intensities that were greater than two standard deviations below the mean for all arrays, these samples were excluded from the analysis. So, there are 113 samples left to find candidate auto-antigens.

- Where are the proteins from?

Human proteins were obtained from **Invitrogen's Ultimate ORF** (open reading frame: a portion of an organism's genome which contains a sequence of bases that could potentially encode a protein) collection or from a **Gateway collection of kinase clones** developed by Protometrix, and the nucleotide sequence of each human protein clone were verified by full length sequencing. By using a proprietary high-throughput insect cell expression system, thousands of recombinant human proteins were produced in parallel. Along with different sets of control proteins, thousands of purified normal proteins are printed on the arrays.

- How to design replicate experiments?

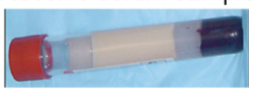
A **technical replicate** involves the multiple labeling or reciprocal labeling of the same sample. The purpose of a technical replicate is to control for technical variability within an experiment. The technical replicates include replicated elements within a single array. And in this experiment, the duplicate spot for each protein on the array belongs to technical replicate.

A **biological replicate** involves isolating samples independently from replicate sources (multiple cell lines, multiple biopsies, multiple patients, etc). The purpose of a biological replicate is to control for biological diversity. In this experiment, 24 biological replicates for each group can be considered as biological replicates.

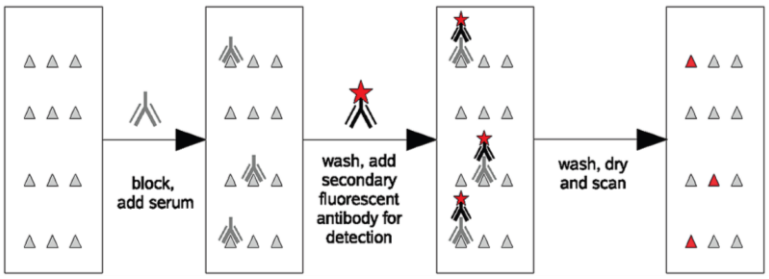
The set of specific replicate experiments will be used to assess intra- and inter-array variability, comparing different normalization strategies as well as eliminate low-quality or questionable array elements under the assumptions that:

1) The reactivity of spotted proteins should ideally remain the same across replicates, so an effective normalization procedure should lower the variability.
2) The separation of groups of samples should be enhanced by normalization.

Receive serum samples

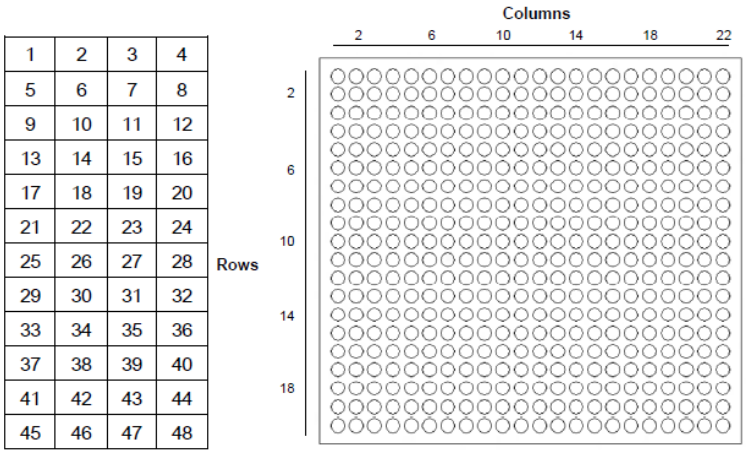


Perform serum profiling on Human ProtoArrays

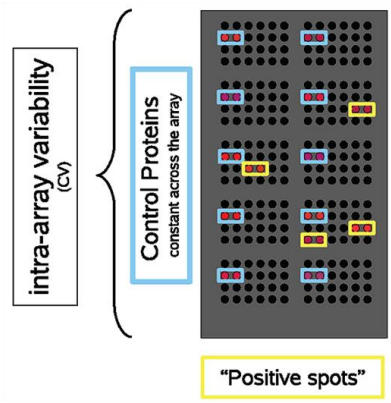


High resolution array images were scanned by
Axon GenePix 4000B fluorescent microarray scanner

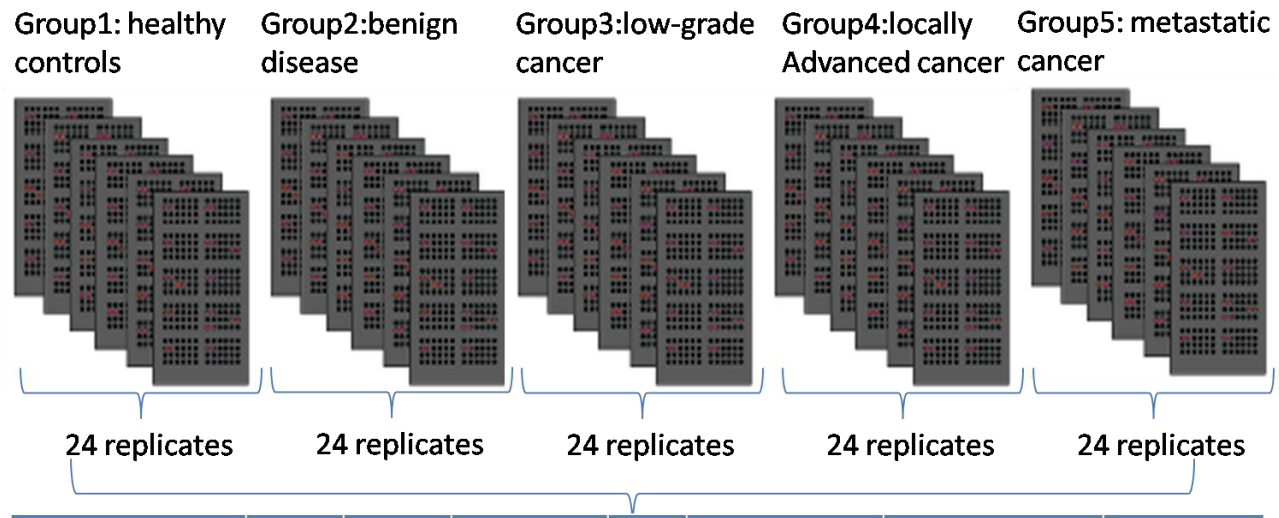
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20
21	22	23	24
25	26	27	28
29	30	31	32
33	34	35	36
37	38	39	40
41	42	43	44
45	46	47	48



One array consist of 48 sub-arrays with 21,120 spots totally
Positive spots: proteins which are reacted with antibodies in the serum or the secondary antibody
Control proteins: constant across all the arrays



Data acquisition: Align array images to protein spot file by GenePix 6.0 software



	BSA	Buffer	ExtraBuffer	GST	Human IgG series	Human IgA series	V5control
Negative control	☺	☺	☺	☺			
Positive control					☺	☺	☺
Gradient control				☺	☺	☺	☺

Figure2-1 Overview of the serum profiling assay

- How to define “groups” among samples?

1) TNM stage:

Stage	Tumor	Nodes	Metastasis
Stage I	T1a	N0	M0
Stage II	T1a	N0	M0
	T1b	N0	M0
	T1c	N0	M0
	T1	N0	M0
	T2	N0	M0
Stage III	T3	N0	M0
Stage IV	T4	N0	M0
	Any T	N1	M0
	Any T	Any N	M1

Evaluation of tumor: (“T”)

T1	T1a	tumor was incidentally found in less than 5% of prostate tissue resected
	T1b	tumor was incidentally found in greater than 5% of prostate tissue resected
	T1c	tumor was found in a needle biopsy performed due to an elevated serum PSA
T2	T2a	the tumor is in half or less than half of one of the prostate glands’ two lobes
	T2b	the tumor is in more than half of one lobe, but not both
	T2c	the tumor is in both lobes
T3	T3a	the tumor has spread through the capsule on one or both sides
	T3b	the tumor has invaded one or both seminal vesicles
T4	T4	the tumor has invaded other nearby structures

Evaluation of regional lymph nodes: (“N”)

N0	there has been no spread to the regional lymph nodes
N1	there has been spread to the regional lymph nodes

Evaluation of distance metastasis: (“M”)

M0	M0	there is no distant metastasis
M1	M1a	the cancer has spread to lymph nodes beyond the regional ones
	M1b	the cancer has spread to bone
	M1c	the cancer has spread to other sites (regardless of bone involvement)

2) Evaluation of **Gleason score**: measure how the tissue is different from normal tissue by its microscopic appearance

G1	Gleason2-4	the tumor closely resembles normal tissue
G2	Gleason5-6	the tumor somewhat resembles normal tissue

G3	Gleason7-10	the tumor resembles normal tissue barely or not at all

3) Evaluation of **PSA**: measure the volume of “prostate specific antigen” by blood test.

PSA is a protein produced by the cells of the prostate gland, which is present in small quantities in the serum of normal men, but is often elevated in the presence of prostate cancer and in other prostate disorders.

4) Evaluation of **PCaV**: measure the “prostate cancer volume”

In this experiment, the five groups are divided by TNM stage, PSA as well as Gleason score.

Group	Name	TNM stage	Gleason score	PSA	PCaV
Group1	healthy			<0.5ng/ml	
Group2	benign		negative prostate biopsy	(3,10) ng/ml	
Group3	local PCa	<=T2	<=6	>3 ng/ml	<0.5ml
Group4	local PCa	>=T2	>6	>3 ng/ml	>0.5ml
Group5	advanced PCa	>T2			

3. Results

Group2 is composed of benign samples and group3 contains the samples with low-grade prostate cancer, and the study on group2 and group3 may help to find novel early-stage prognosis biomarker as well as shedding light on the pathways involved in early-stage prostate cancer progression. So, I use the study strategy introduced in part on group2 vs. group3.

Normalization comparison and array quality assessment

Figure 3-1 is the result of normalization methods comparison between group2 and group3, and we can easily get four conclusions:

1) Comparing to the raw data, the other data sets after different kinds of normalization show improvement in some degree;
2) Just as paper [1] explained, RLM (robust linear model) normalization reveals especially excellent performance on minimizing variation between intra- and inter-arrays control proteins; at the same time, the difference among three sets of control proteins from IgA , IgG to V5 is not very obvious for RLM;
3) VSN, which is ignored in the paper [1], seems to separate two groups the most apparently with the cost of enlarging control spot variation, especially combined with RLM method;
4) Taking into account all conditions, the method integrate quantile with RLM is the most satisfying, i.e. try to minimize the variation between intra- and inter-arrays on all the control proteins without losing group separations.

Based on the assumptions on microarray, there are ways to evaluate the quality of normalization from intra- to inter- aspects. For the intra-array quality assessment, the most common way is using MAplots, M and A are defined as:

$$M = \log_2^{(I_1)} - \log_2^{(I_2)}$$

$$A = \frac{1}{2}(\log_2^{(I_1)} + \log_2^{(I_2)})$$

Where I_1 is the intensity of the study array, I_2 is the median value of all the arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in the mean of M as a function of A.

For inter-array quality assessment, one way is using array intensity distribution in either boxplot or density plot form. Typically, one expects the boxes to have similar size and y position (median). If the distribution of an individual array is very different from the others, this may indicate an experimental problem. After normalisation, the distributions should be similar. At the same time, the distributions of the arrays should have similar shapes and ranges in distribution plots. Arrays whose distributions are very different from the others should be considered for possible problems.

After the normalization by quantile following RLM, the distribution of signal intensities in each array is quite similar, this is verified by boxplot and intensity distribution in figure 3-2. By using the array quality control package ”arrayQualityMetrics” in R, we also get the MA plots for each array, and the bottom of figure 3-2 gives the first several ones, from which we can find the mass of distribution almost concentrated along the M=0 axis. So we can say confidently that the quantile combined with RLM normalizaiton is quite effective on this data set.

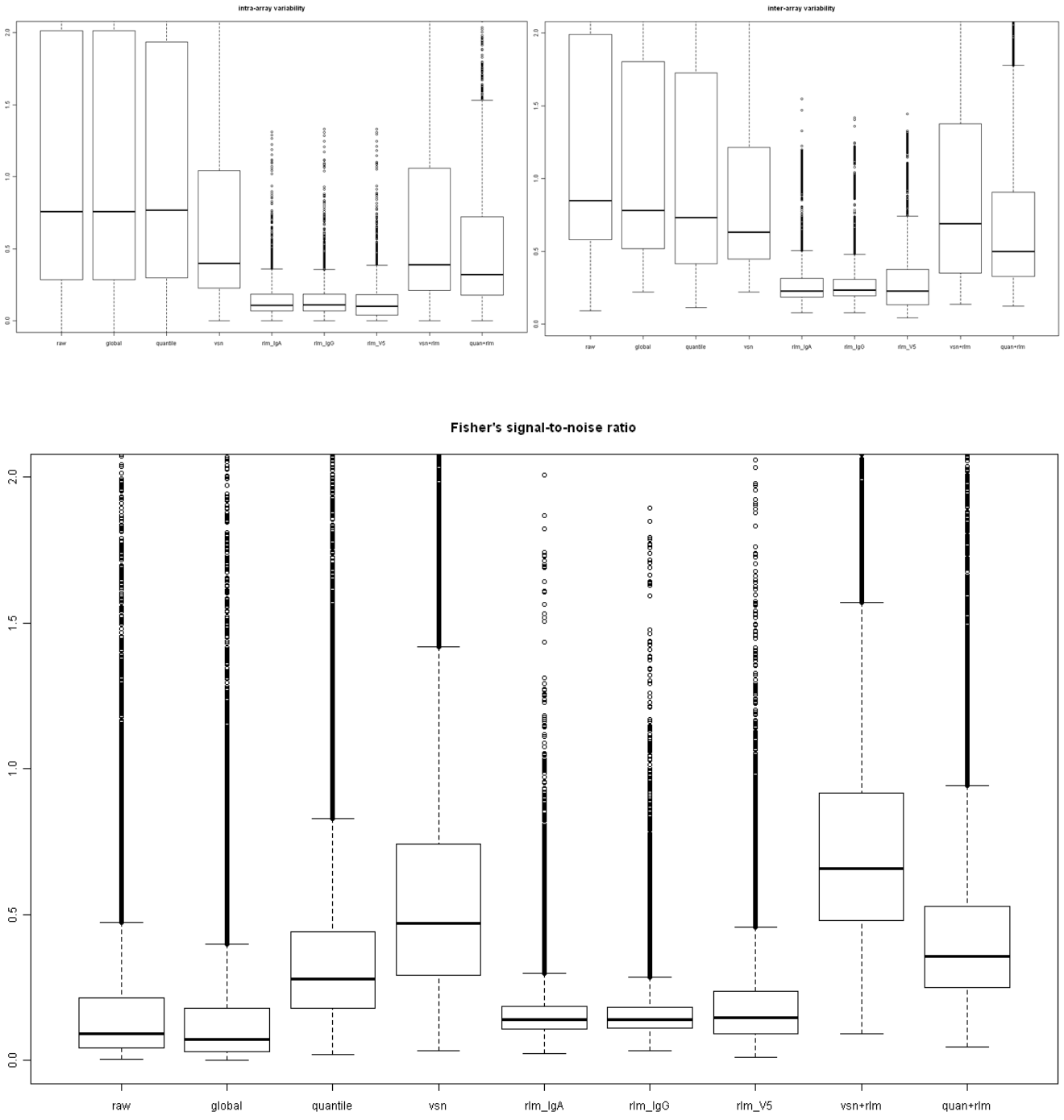


Figure 3-1 Normalization methods comparison on group2 vs group3

(In the left top is the boxplot of intra-array variation coefficient of all the control

proteins, the right top is the boxplot of inter-array variation coefficient of all the control proteins, the bottom large one is the Fisher's signal to noise ratio on all the normal proteins between group2 vs group3. The labels of the methods through three graphs are the same: from raw data to data after [quantile + rlm_IgG] normalizaion)

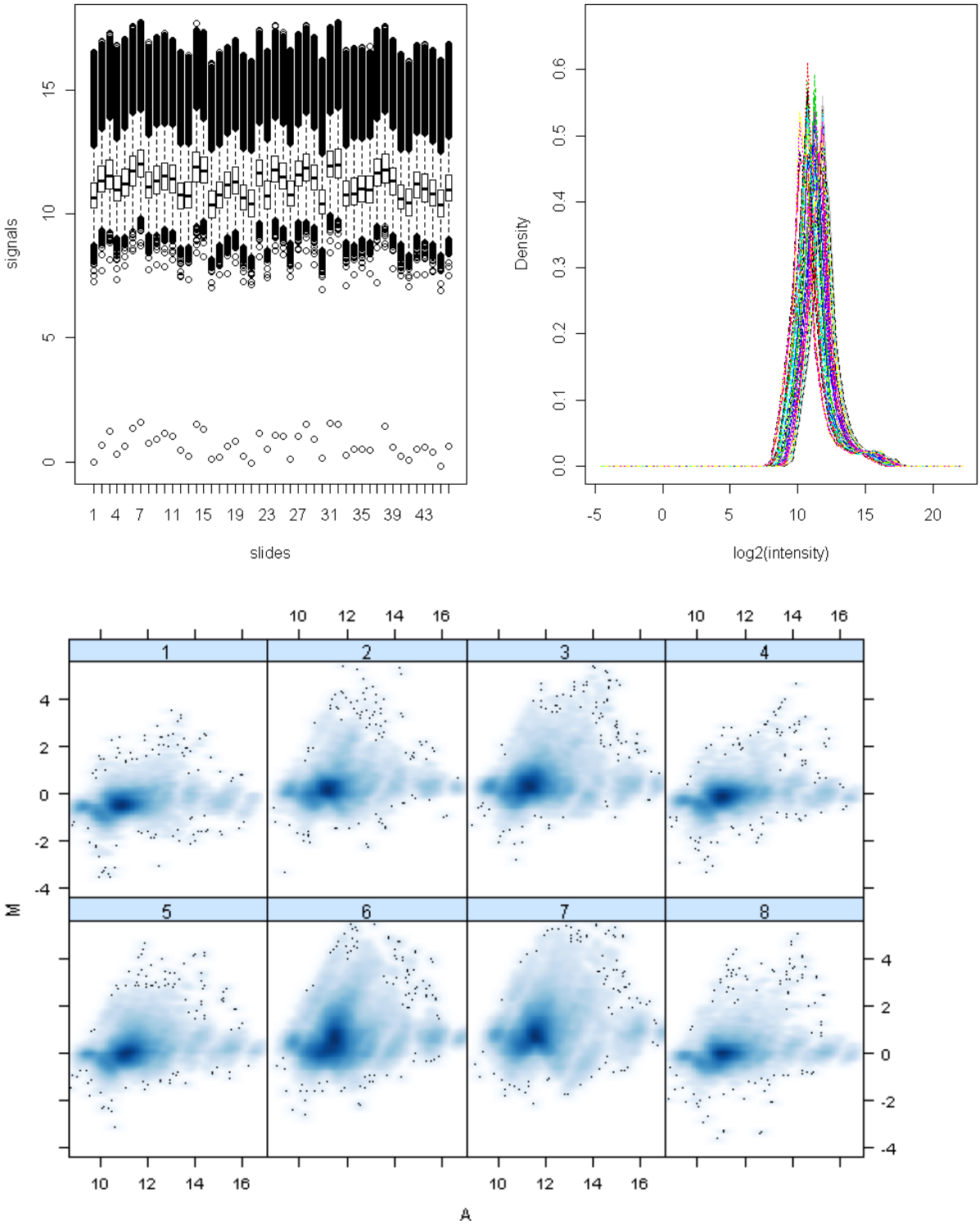


Figure 3-2 inter- and intra-array quality controls

(Top two: intensity distribution of arrays in group2 and group3 after quantile and RLM normalization; bottom: MA plot of the first eight arrays in group2, the remaining one also didn't show abnormal)

- Gene selection by statistical method

  - **One-side Wilcoxon test on all 44 samples from group 2 and group 3**

    By using one-side Wilcoxon test with 0.05 significant level on the 44 samples, 75 genes show significantly different signal intensities between group2 and group3. The gene with the lowest p-value obtains the highest AUC (area under curve) of 0.77 in the ROC curve, which means the single-gene classifier with the classification error rate of 0.23 in this dataset. Instead of using multiple-test, the p-value plot is drawn as figure 3-3. Without uniform distribution, the p-values follow normal-like distribution, so we may say the tests are not very satisfying but acceptable.

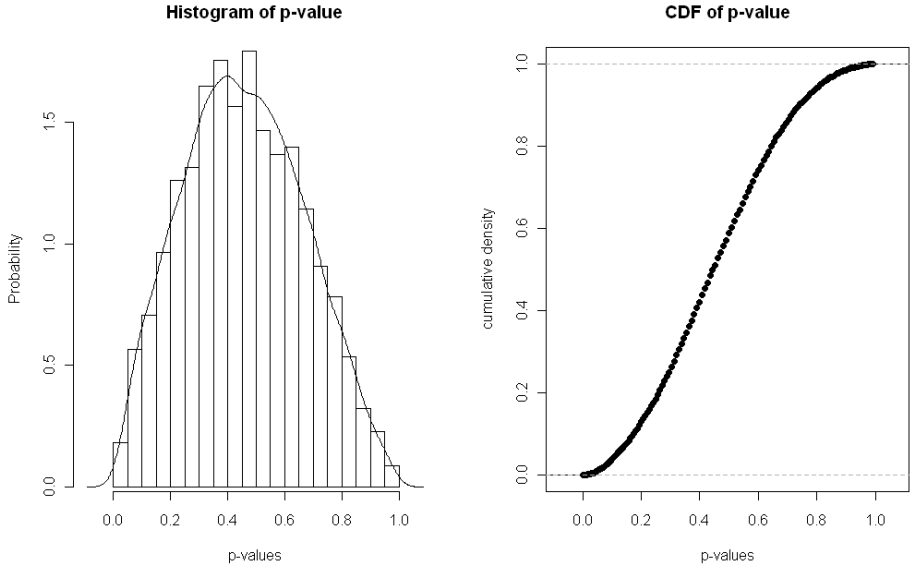


Figure 3-3 plot of p-values from 8302 tests

  - **Random re-sampling tests**

    In order to overcome the problem of strong sensitivity of hit gene list to the training samples, I did one-side Wilcoxon tests on randomly selected 90% of all the samples for 50 times, the result is explained in the following three aspects:

    1) *There are 44 samples in total, and 50 times iteration is not very large, is it possible that the random sampling prefers to certain samples?*

       ---- I check the selection frequency distribution for each sample. All the samples are selected in the range of 39 to 48 out of 50 times, and most of them are chosen at least 40 times. So the sampling is uniform across the samples.

    2) *Does the hit gene list really heavily rely on the training samples selections?*

       ---- I find there are 825 genes with p-value below 0.05 at least once in 50 times iteration, most of which emerge less than 10 times out of 50. The hit gene list size fluctuates from time to time, mostly containing less than 100 genes, but sometimes with more than 400

genes. So, the hit gene list is quite sensitive to the samples selected as training set.

Since the hit gene lists changed so much, is there any relationship between the gene list from re-sampling sets and all the 44 samples? In order to deal with this problem, I draw the plot of emerge frequency over 50 times against the p-value derived from the test on all 44 sample, and I find these two components shows inverse proportion, i.e. the feature with smaller p-value derived from the whole information will also hit most. This is easy to explain, because p-value just describe the probability of getting the hypothesis by chance, and a lower p-value means more confident to say the hypothesis is true. The additional sampling tests just confirm this point, so it also can act as a supplementary method or less stringent t multiple tests when group separation is not so clear.

In order to restrict the set of reliable significant genes, I select the genes emerge at least 40 times over 50 iterations, which is also the gene set with the lowest p-values on all samples. Table 3-1 lists all these 11 genes.

Table3-1 11 hit genes by statistical method

Gene.symbol	Significant times	p-value ranking on 44 samples	p-value on 44 samples
Gene1	50	1	0.001
Gene2	48	2	0.01
Gene3	47	3	0.013
Gene4	45	4	0.01392
Gene5	45	5	0.0158
Gene6	45	6	0.0158
Gene7	45	8	0.0178
Gene8	44	7	0.0158
Gene9	42	9	0.0189
Gene10	42	10	0.0212
Gene11	41	11	0.0238

- **Potential diagnostic signature evaluation**

In order to evaluate the classification performance of the above 11-gene prognostic signature I used a strategy combined unsupervised method principle component analysis (PCA) with different supervised classification methods including linear discriminate analysis (LDA), K-nearest neighbour (KNN) as well as support vector machine (SVM), the most popular and well-performed ones in microarray data classification, to evaluate the effectiveness of these combined genes on classification of samples from gruop2 to group3.

✓ **Step1: PCA on all the 8302 features VS. 11 features from statistical test above**

Figure3-4 shows the result of PCA on all features and 11 the most significant features coming from the above four-step statistical test. In the graph, green dots standing for samples from group2 and red ones for group3, we can obviously find that the samples are hard to be separated based on all features, which is much improved on the selected 11 features. From the bottom graph we can find, samples in group3 are quite well collected but the samples in group2 spread out in the first two components. As we know, the benign disease shares a lot of same symptoms as malignant disease, and at the same time, the benign stage can also be divided into distinctive sub-stages according to the histological and clinical criteria. As a result, the similarity between the group2 and group3 may be quite large, and the selected 11 most significant features can successfully describe the group3 gene profile from group2.

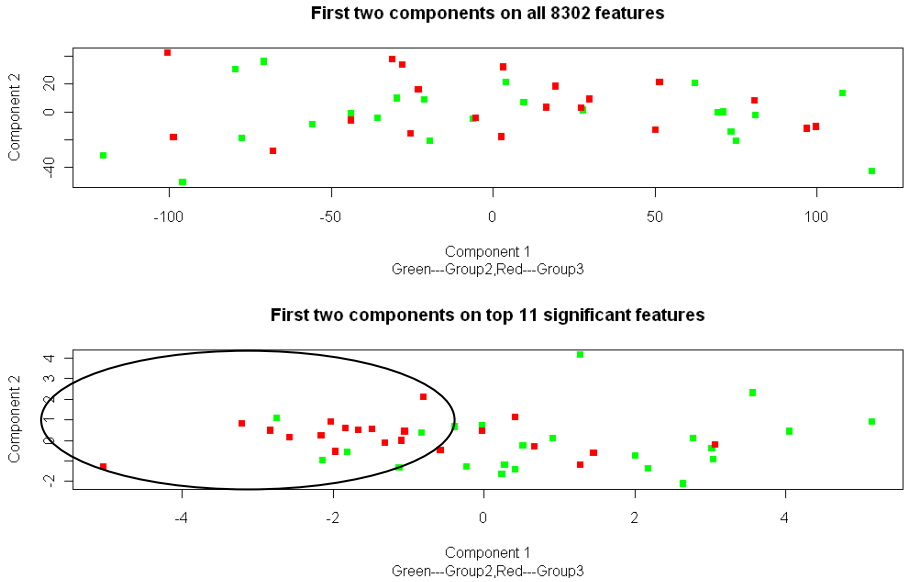


Figure3-4 the first two components of PCA on all features VS. 11 significant features

✓ **Step2: cross-validation error rate evaluation by KNN on all features VS. 11 significant features**

Considering that the sample size in this work is not large enough, and in order to avoid information leak taken by random selection of training set, I decided to use cross-validation error rate as the evaluation criteria for the significant gene set profile working on group separation. The minimum cross-validation error rate by KNN on all features is about 21/44, which is no better than random classification and much higher than the error rate on the significant 11 features selected by statistical test above with 14/44 misclassifications. The error rate of nearly 0.31 for this 11-gene signature however is worse than the single gene classifier with gene1 of 0.23 misclassification error rate.

✓ **Step3: cross-validation error rate evaluation by SVM on all features VS. 11 significant features**

By using the “svm” function in the R e1071 package with linear kernel and C-classification SVM-Type, I get the cross-validation error of 12/44 with 6 misclassification error rate each type.

- Gene selection by random forest

I choose the solution with the smallest number of genes whose error rate is within one standard errors of the minimum error rate of the forests, i.e. 21 in this circumstance. There are 21 variables selected as important nodes to do the classification, gene gene1 and gene7 are overlapped with the above statistical analysis.

Then I check the selected genes by correlation coefficient and the p-value of one-side Wilcoxon test. The correlation is an alternative common way to select the most disease-related genes, and the idea is to calculate each variable vector to the group label. The result shows that:

1) Instead of using only intensity-ascending features, the algorithm also selected features with decreasing intensity;

2) Correlation coefficient is consistent with p-value, i.e. negative correlation means significantly decreasing intensity from group2 to group3, while positive correlation coefficient equals to significantly ascending intensity;

3) All of these 21 genes are high correlated to the group labels, either positive or negative; in other words, all of them show either significantly decreasing or increasing intensities from group2 to group3. So from this point of view, it makes sense to select these genes from the classification procedure.

Now, I use the four step gene-signature evaluation method to verify the performance of this 21-gene profile returned from random forest.

✓ **Step1: PCA on all the 8302 features VS. 21 features from random forest**

From figure3-12 it seems that the samples can be separated on the second components in some degree, but it's hard to find a clear boundary.

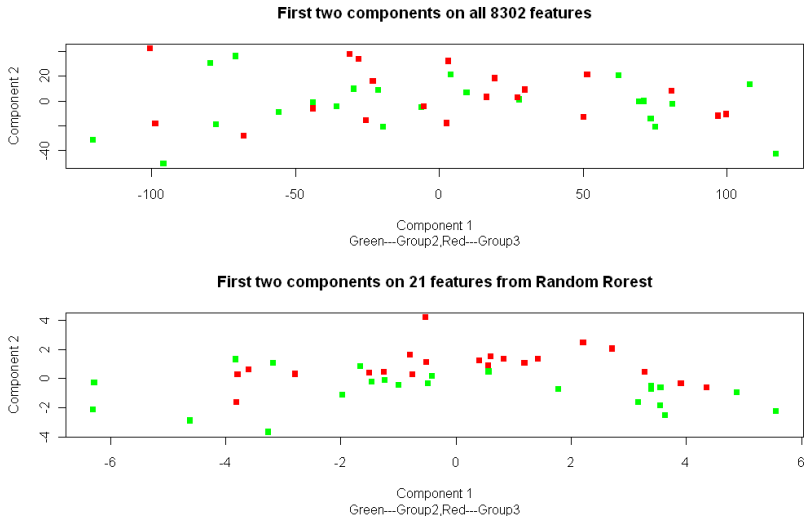


Figure 3-5 the first two components of PCA on all 8302 feature VS 21 features from random forest

✓ **Step2: KNN leave-one-out cross validation error rate evaluation**

The error rate on the 21 genes selected by random forest is 9/44 with 0.2 misclassification error rate; this 9/44 error rate is also lower than the single gene1 classifier.

✓ **Step3: SVM leave-one-out cross validation evaluation**

By using the "svm" function in the R e1071 package with linear kernel and C-classification SVM-Type, I get the cross-validation error of 1/44 with 1 misclassification only, which is again the best one of all the three classification methods.

The 21-gene profile selected by random forest works better than the 11-gene signature derived from the one-side statistical test, which verified the gene selection strategy by random forest is an effective and promising way.

- Gene selection by shrunken centroid

The number of genes around 11 is corresponding to the minimum cross-validation error rate, and the error rate keeps stable as the gene number continuing to decreasing. In order to identify relevant genes for subsequent study, as well as verification of the above discovered genes, I chose the larger gene size with 11 genes, i.e. less stringent threshold. Among the selected 11 genes, gene1 is hit again, and it's the only one overlapped with hit gene list by statistical method above.

Now let's have a look at these selected genes by correlation coefficient and the p-value of one-side Wilcoxon test. From a series of graph we can get the following three conclusions:

1) Instead of using only intensity-ascending features, the algorithm also selected features with decreasing intensity; but the intensity-ascending ones are predominant.

2) Correlation coefficient is consistent with p-value, i.e. negative correlation means significantly decreasing intensity from group2 to group3, while positive correlation coefficient equals to significantly ascending intensity;

3) All of these 11 genes are high correlated to the group labels, either positive or negative; in other words, all of them show either significantly decreasing or increasing intensities from group2 to group3. So from this point of view, it makes sense to select these genes from the classification procedure.

✓ **Step1: PCA on all the 8302 features VS. 11 features from shrunken centroid**

From figure 3-6, it seems that a boundary across the second component can be found to separate these samples, and comparing to well-mixed scenario, the 11-gene profile derived from shrunken centroid seems catch some common characteristics between group2 and group3.

✓ **Step2: KNN leave-one-out cross validation error rate evaluation**

The minimum cross-validation error rate by KNN on all features is about 21/44, which is the same as above parts and much higher than the error rate on the 11 features with 14/44.

✓ **Step3: SVM leave-one-out cross validation error rate evaluation**

By using the "svm" function in the R e1071 package with linear kernel and C-classification SVM-Type, I get the cross-validation error of 5/44 with 4 misclassifications from group2 to group3, and 1 misclassification from group3 to gruop2.

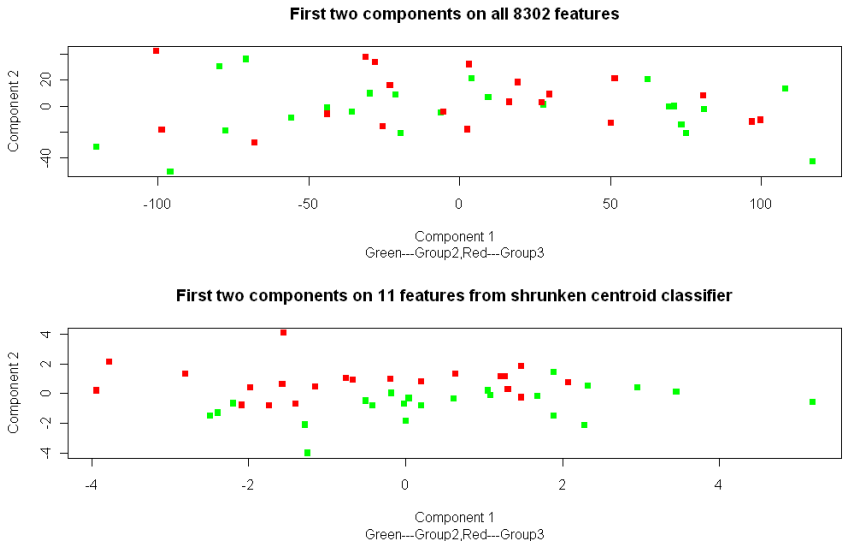


Figure 3-6 the first two components of PCA on all 8302 feature VS 11 features from shrunken centroid

- Enrichment pathway analysis

Now there are three hit gene list derived from different gene selection strategies, i.e. statistical test, random forest as well as shrunken centroid separately. Intuitively just gene1 emerges in all these three lists. In order to shed light on the pathways related to the prostate cancer progression, I have a deeper look at these three hit gene list in the online annotation database swiss-prot, to check the main biological functions and involved pathways.

The result shows that different hit gene lists cover same pathways related the cell cycle regulator, and signal transduction pathway exerting important function to transfer signal from out cellular to nucleus, although there are so few gene overlapped by these gene list.

4. Discussion

Microarray data analysis as a kind of popular and challenging downstream analysis is always demanding and attracting great efforts for bioinformatians. Now the well-accepted workflow for microarray data analysis follows several steps:

1) Normalization;
2) Differential gene selection, and the most common way is by statistical methods, like by using p-value as well as the correlation coefficients. But nowadays criticism points to this classical method because more and more studies found that the hit gene list is always heavily depending on the training samples, especially microarray always limited by not too many training samples. At the same time, the statistical methods always consider the effects of single gene contributed to the result, which obviously disobey the fact of strong relationships among genes. Besides, the randomly selected "optimal" gene set is another critical problem can't be ignored. Recently, the gene selection strategy related to classification is quite striking because it may overcome the problems coming from statistical methods, and the random forest, shrunken centroid as well as the recursive feature elimination by SVM are three activated methods among a lot of counterparts. In my study, I adopted both the statistical methods as well the classification strategy to obtain different hit gene list.
3) After we get the hit gene list, the following step is always related to pathway analysis as well as result verification. In my work, I combined the three different hit gene lists derived from distinctive methods to analyze the enriched pathways, which may shed light on the prostate cancer progression.

In this work, the 120 samples are divided into five groups according to different disease stages, and one of the main purposes of this study is to detect well-performed prognosis biomarkers just as the well-established PSA for prostate cancer detection in the early stage. So, I use the analysis strategy on group2 (benign sample) vs. group3 (low-grade disease) and try to get idea on early-stage progression. Since on protein microarray, the decreasing signal is always considered as noise providing little information, from this point of view, we just focus on the genes with increasing intensity with the development of disease. The other aim of this study may concentrate on clarifying the underlying molecular mechanisms of cancer progression, especially in the context of complex cellular networks, in other words, we want to find disease related pathways in different stages, so I also utilized the same strategy from normalization to enrichment pathway analyzing between group2 vs. group3, group3 vs. group4, as well as gropu1 vs. group5, and from these four group comparison I found that:

i. **Quantile followed by RLM is a stable and good-performed combination for protein microarray normalization**

   From all of these four group comparisons, the quantile + RLM always show low variability on control proteins without losing the sample separation on the normal proteins. RLM is a kind of

normalization taking both intra- and inter-array normalization at the same time, so maybe this is the reason why this combination obtain the best performance.

ii. Gene1 is a very promising biomarker for discrimination of benign from low-grade disease.

From the group comparison between group2 vs. group3 we can find gene1 is hit by three different methods. From the literature, we my easily find that gene1 is also quite active as the biomarker in different studies.

iii. Gene1-related K pathway is important in the prostate cancer development

From the enrichment pathway analysis in four group comparisons I find gene1-related K pathway is important in the prostate cancer development.

As we know, unregulated growth is a quintessential characteristic of cancer, and often an extracellular growth factor stimulates cell growth by transmitting a signal into the cell, and then to the nucleus. Figure 4-1 illustrated one example of epidermal growth factor signalling pathway about epidermal growth factor signalling (EGF), by which a signal from a growth factor outside the cell entering the nucleus where gene expression is regulated. The whole process can be divided into several steps:

1) Binding of the growth factor to the receptor
2) Receptor dimerization
3) Autophosphorylation
4) Activation of intracellular transducers, from RAS (GTP binding proteins) to a cascade of serine/threonine kinases
5) Regulation of transcription factors for gene expression.

In this well-established growth factor signaling pathway we can easily find RAS proteins as well as MAPK signaling pathway together, which can be supported by different hit gene list derived from this study.

From the annotation for gene1 we can find it involved in positive regulation of MAP kinase activity and MAPKKK cascade. From the literature, a previous study showed that gene1 expression is negatively associated with p-MAPK in cholangiocarcinoma, suggesting that K pathways may be involved in gene1-induced tumorigenesis at the first time. [26] Recently, another showed that gene1-dependent ERK1/2 and p38 MAPK activation is required for colorectal cancer cell proliferation in vitro, and in vivo, which further shed light on the linkage of gene1 and K pathways in the regulation of cancer proliferation. [21]

From the K pathway recorded in KEGG database, we can find a geneT-dependent K pathways, considering that geneT has the similar function as gene1, there may be another proof for the verification of this pathway.

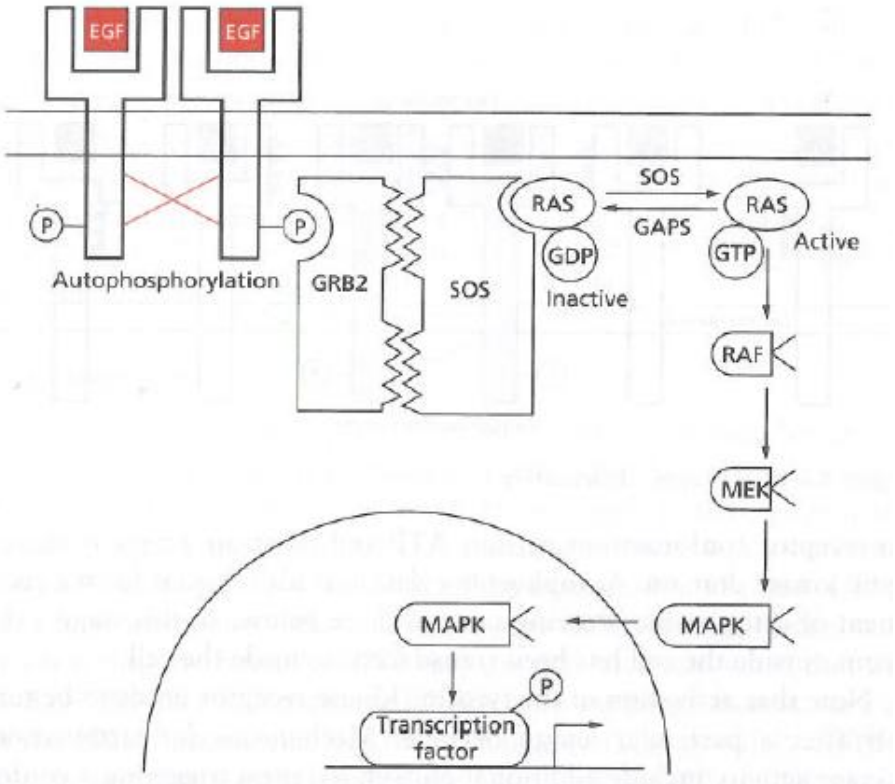


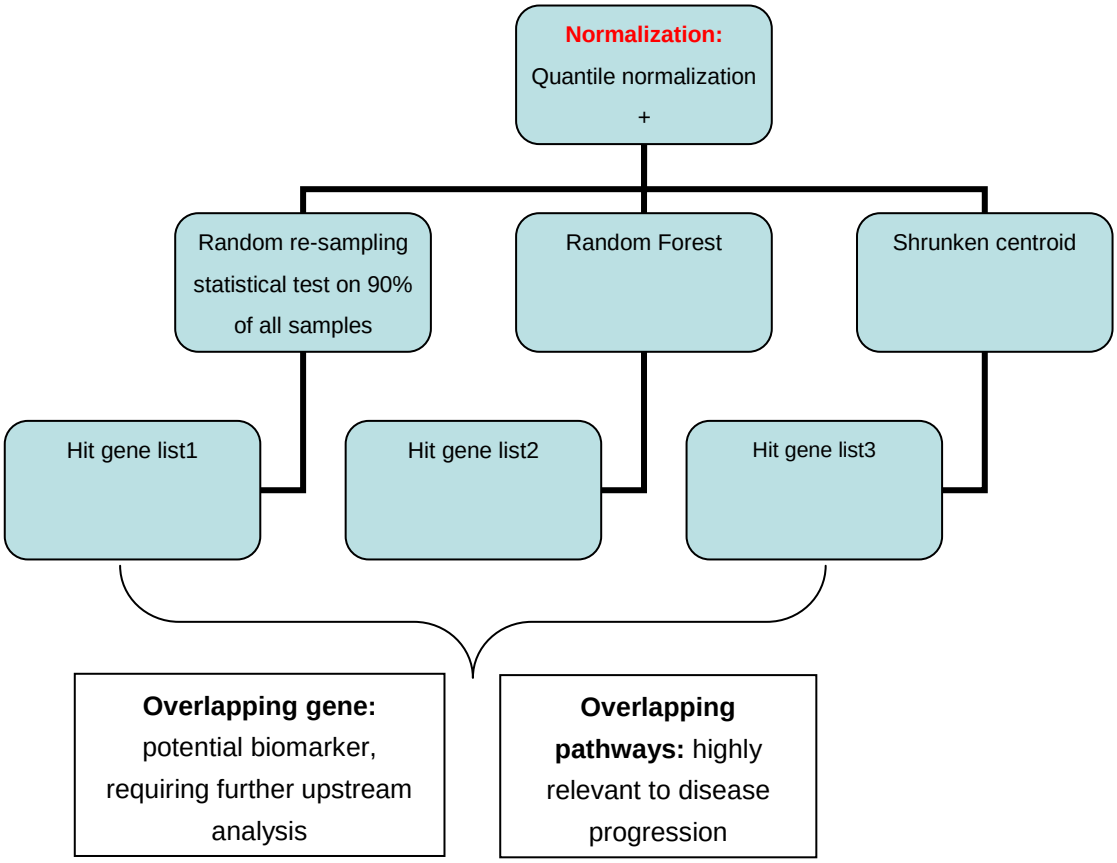
Figure 4-1 the signal transduction pathway of epidermal growth factor (EGF) [27]

iv. Angiogenesis

As one of the hallmarks of cancer [27], cancer cells intend to induce the formation of new blood vessels to supply more oxygen and nutrients for unlimited cell proliferation. Some genes derived from comparison between group2 and group3, group3 and group4 related to angiogenesis are also hit.

5. Conclusions

1. The protein microarray analyzing strategy introduced in this study combining different well-established methods works well on normalization, obtaining hit gene list as well as the enrichment pathway analysis procedures.
2. In this work, different normalization methods are compared, which may provide insight for further protein microarray study.
3. Gene1-related K pathway may play a very important role in the prostate cancer progression, which is valuable for further experiment verification.
4. The hit genes are mostly related to signal transduction, cell cycle, or metabolic process, which is consistent with the known mechanism of cancer disease.
5. We recommend a workflow for immune response protein microarray analysis follows such a strategy:



6. Acknowledgements

This report would not have been possible without the essential and great support of many individuals. Firstly, I want to give many thanks to my supervisor Mr. Matthiew Visser in Philips Research Eindhoven. He gave me this internship chance even I didn't have so much research experience in related works at that time. In the past nine months, he taught me how to conduct a research independently, how to deal with problems during the work and how to present results and findings to others explicitly. At the same time, he spent a lot time and efforts on reading my essay and offering a lot of valuable advice. In all, he takes me to a wonderful research land which I like very much and he strongly encouraged me to pursue a further PhD study as well.

The next thanks go to Wim F.J. Verhaegh and Tim Hulsen in Philips Research Eindhoven. They provided me a lot of instant help and support without hesitate when I needed.

Then I want to express my thankness to my supervisor in Leiden University, Dr. Erwin Bakker, who is responsible for helping me completing the final dissertation and defence. He gave me a lot of space to practice my thought on this work.

Special thanks to Dr. Wolfgang Huber in EMBL. He inspired me a lot from several talks with him, and he recommended me some very useful books and papers which shed light on some important problems in this work.

Let me also say "thank you" to all the students working with me in Philips Research Eindhoven. They brought me a nine-month happy and wonderful time, and relieved me a lot from great pressure of such a challenging work.

Finally never enough thanks to my parents in China who gave me so much confidence when I doubted myself, who provided me so many encouragements when my work was in trouble, out of question, they are the strongest drive to my study and work.

References

1. Andrea Sboner, Alexander Karpikov, Gengxin Chen, Michael Smith, Mattoon Dawn, Lisa Freeman-Cook, Barry Schweitzer, Mark B. Gerstein; Robust-linear-model-normalization to reduce technical variability in functional protein microarrays; Journal of proteome research 2009,8,5451-5464

2. Volker Seibert, Matthias P. A. Ebert and Thomas Buschmann; Advances in clinical cancer proteomics: SELDI-To-mass spectrometry and biomarker discovery; Briefings in functional genomics and proteomics. Vol 4. NO. 1. 16-26, May 2005

3. Paul F Predki; Functional protein microarrays: ripe for discovery; Current Opinion in Chemical Biology; Volume 8, Issue 1, February 2004, Pages 8-13

4. Joshua LaBaer, Niroshan Ramachandran; Protein microarrays as tools for functional proteomics; Current Opinion in Chemical Biology 2005, 9: 14-19

5. Barry Schweitzer, Stephen F Kingsmore; Measuring proteins on microarrays; Current Opinion in Biotechnology 2002, 13: 14-19

6. Adam, B. L., Qu, Y., Davis, J.W. et al. (2002); "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men", Cancer Res., Vol. 1. 62, pp. 3606-3614

7. Qu, Y., Adam, B.L., Yasui, Y. et al. (2002); "Boosted decision tree analysis of surface enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients"; Clin. Chem., Vol.48, pp. 1835-1843

8. Gunjan Malik, Michael D. Ward, Saurabh K. Gupta, Michael W. Trosset, William E. Grizzle, Bao-Ling Adam, Jose I. Diaz, O. John Semmes; "Serum levels of an isoform of Apolipoprotein A-II as a potential marker for prostate cancer"; Clinical Cancer Research, Vol. 11, 1073-1085, February 1, 2005

9. Miller JC, Zhou H, Kwekel J, Cavallo R, Burke J, Bulter EB, The BS, Haa BB; Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers. Proteomics 2003, 3: 56-63

10. Gargi D. Basu, David O. Azorsa, Jeffrey A. Kiefer, Angela M. Rojas, Sukru Tuzmen, Michael T. Barrett, Jeffrey M. Trent, Olli Kallioniemi, Spyro Mousses; Functional evidence implicating S100P in prostate cancer progression; International Journal of Cancer, Vol. 123 Issue 2, pp 330-339, 1st, May, 2008

11. AM Havens, Y Jung, YX Sun, J Wang, RB Shah, HJ Buhring, KJ Pienta, RS Taichman; The role of sialomucin CD 164 (MGC-24v or endolyn) in prostate cancer metastasis; BMC Cancer 2006, 6:195 doi: 10.1186/1471-2407-6-195

12. Robert Calaluce, David J. Bearss, Jean Barrera, Yu Zhao, Haiyong Han, Shaleen K. Bech, Kathy McDaniel, Ray B. Nagle; Laminin-5 h3A expression in LNCaP human prostate carcinoma cells increases cell migration and tumorigenicity; Neoplasia Vol. 6, No. 5, September/October 2004, pp, 468-479

13. JF Knight, CJ Shepherd, S Rizzo, D Brewer, S Jhavara, AR Dodson, CS Cooper, R Eeles, A Falconer, G Kovacs, MD Garrett, AR Norman, J Shipley and DL Hudson; "TEAD Iand c-Cbl are novel prostate basal cell markers that correlate with poor clinical outcome in prostate cancer"; British Journal of Cancer (2008) 99, 1849-1858. Doi: 10.103/sj.bjc.6604774

14. Simona Nanni, Valentian Benvenuti, Annalisa Grasselli, Carmen Priolo, et al; Endothelial NOS, estrogen receptor β, and HIFs cooperate in the activation of a prognostic transcriptional pattern in aggressive human prostate cancer; The journal of clinical investigation,Vol 119, number 5, May 2009

15. Ramon Diaz-Uriarte, Sara Alvarez de Andres; "Gene selection and classification of micro array data using random forest"; BMC Bioinformatics 2006, 7:3 doi:10.1186/1471-2105-7-3

16. Mariusz Lubomirski, et al; "A consolidated approach to analyzing data from high-throughput protein microarrays with an application to immune response profiling in humans"; Journal of computational biology, vol 14, number 3, 2007, page 350-359

17. T. Schweder and E.Spjotvoll; "Plots of P-values to evaluate many tests simultaneously"; Biometrika (1982), 69,3, pp.493-502

18. Benjamin Haibe-Kains, Christine Desmedt, Fanny Piette, Marc Buyse, Fatima Cardoso, Laura van't Veer, Martine Piccart, Gianluca Bontempi and Christos Sotiriou; Comparison of prognostic gene expression signatures for breast cancer; BMC Genomics 2008, 9:394 doi: 10.1186/1471-2164-9-394

19. Shefali Agrawal, Boris W. Kuvshinoff et.al; CD24 expression is an independent prognostic marker in cholangiocarcinoma; The Society for Surgery of the Alimentary Tract (2007) 11: 445-451

20. G Kristiansen, K Schluns, Y Yongwei et.al; CD24 is an independent prognostic marker of survival in nonsmall cell lung cancer patients; British Journal of Cancer (2003) 88, 231 – 236

21. Weifei Wang, Xinying Wang, Liang Peng et.al; CD24-dependent MAPK pathway activation is required for colorectal cancer cell proliferation; Cancer Science Juanuary 2010. Vol. 101, No 1, p112-119

22. Glen Kristiansen , Christian Pilarsky , Janja Pervan et.al; CD24 expression is a significant predictor of PSA relapse and poor prognosis in low grade or organ confined prostate cancer; The prostate, Volume 58 Issue 2, Pages 183 – 192

23. Glen Kristiansen, Carsten Denkert, Karsten Schluns et.al; CD24 is expressed in ovarian cancer and is a new independent prognostic marker of patient survival; American Journal of Pathology, Vol. 161, No. 4, October 2002, p1215-1221

24. Glen Kristiansen, Klaus-Junrgen Winzer et.al; CD24 expression is new prognostic marker in breast cancer; Clinical Cancer Research, Vol. 9, 4906–4913, October 15, 2003, p4906-4913

25. Aigner S, Ramos CL, et.al; CD24 mediates rolling of breast carcinoma cells on P-selectin. EMBO J 1998, 12: 1241:1251

26. Agrawal S, Kuvshinoff BW, Khoury T et al. CD24 expression is an independent prognostic marker in cholangiocarcinoma. J Gastrointest Surg 2007; 11: 445-51

27. Lauren Pecorino; Molecular biology of cancer---Mechanisms, targets, and therapeutics; Oxford university press; 2005