DEVELOPMENT AND APPLICATION OF BIOINFORMATICS TOOLS FOR DISCOVERING DISEASE MARKERS AND DISEASE TARGETING ANTIBODIES

TANG ZHIQUN

(*B. Eng & M.Med*, HUST)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF PHARMACY
NATIONAL UNIVERSITY OF SINGAPORE

2007

ACKNOWLEDGMENTS

The realization of this thesis was achieved due to the support of a large number of people, all of which contributed in various ways; without them this research would not have been possible.

First and foremost, I would like to express my sincere and deep gratitude to my supervisor, Professor Chen Yuzong, who provides me with the excellent guidance and invaluable advices and suggestions throughout my PhD study in National University of Singapore. I have tremendously benefited from his profound knowledge, expertise in scientific research, as well as his enormous support, which will inspire and motivate me to go further in my future professional career.

I am grateful to our BIDD group members for their insight suggestions and collaborations in my research work: Dr. Yap Chunwei, Dr Han Lianyi, Dr. Lin Honghuang, Dr Zheng Chanjuan, Ms Cui Juan, Mr Ung Choong Yong, Mr Xie Bin, Ms Zhang Hailei, Dr Wang Rong and Ms Jia Jia. I thank them for their valuable support and encouragement in my work.

Finally, I owe my gratitude to my parents, husband and daughter for their love, constant support, understanding and encouragement throughout my life.

TABLE OF CONTENTS

SUMMARY

Thanks to the rapid progress on the research of genomics and genetics, our knowledge on the molecular basis of diseases has been significantly enhanced, which has greatly contributed to the discovery of disease markers for disease differentiation, and to the design of disease-targeting molecules like small-molecule agents or antibodies for disease treatment. The key disease markers determine the characteristics of disease, therefore could be further analyzed the possibility of these markers severing as targets for disease targeting molecule design. The main objective of this dissertation is to develop a disease marker discovery system from microarray data and a bioinformatics tool for disease-targeting molecule prediction.

It is of crucial essence to find the marker genes responsible for disease initiation and progress. The marker genes may benefit early disease diagnosis and correct prediction of prognosis. The expression level of such markers presents potential therapeutic drug targets and may give suggestions to proper treatment regime. Microarray can measure the expression level of thousand of genes at one time, presenting the most important platform for disease diagnosis, disease prognosis and disease marker discovery. Current microarray data analysis tools provided good predictive performance. However, the markers produced by those tools have been found to be highly unstable with the variation of patient sample size and combination. The patient-dependent nature of the markers diminishes their application potential for diagnosis and prognosis. To solve this problem, we developed a novel gene selection method based on Support Vector Machines,

recursive feature elimination, multiple random sampling strategies and multi-step evaluation of gene-ranking consistency. The as-developed program can be utilized to derive disease markers which present both good prediction performance and high levels of consistency with different microarray dataset combinations.

After program implementation, two different cases were tested: colon cancer marker discovery by using a well-studied 62-sample colon-cancer dataset and lung adenocarcinoma survival marker discovery by using an 86-sample lung adenocarcinoma dataset. In the first case, the derived 20 colon cancer marker signatures are found to be fairly stable with 80% of top-50 and 69%~93% of all markers shared by all 20 signatures. The shared 104 markers include 48 cancer-related genes, 16 cancer-implicated genes and 52 previously-derived colon cancer markers. The derived signatures outperform all previously-derived signatures in predicting colon cancer outcomes from an independent dataset. The possibility of the markers as therapeutic target was exploited by a therapeutic target prediction system. Six known targets and 18 potential targets were identified by this system. In the second case, 21 lung adenocarcinoma survival markers were shared by 10 marker signatures. 5 known and 7 novel targets were predicted as therapeutic targets. These results suggested the effectiveness of our system on deriving stable disease markers and discovering therapeutic target.

One major application of marker discovery is the finding of disease targeting molecules for disease prevention and treatment. For this purpose, therapeutic antibodies, a class of effective disease-targeting molecules, were employed to develop a therapeutic antibody prediction system based on antibody-antigen

sequence recognition information. Eventually, an antibody antigen information resource (AAIR) database, which provides information of sequence-specific antibody-antigen recognition and their immunological relevance, was developed. Three classes of information are included in the database. The first class is antigen information consisting of antigen name, sequence, function and source organism. The second class is antibody information containing antibody isotype, source organism, molecular and structural type of antibody. The third one is disease and therapeutic information composed of disease class, targeted disease, diagnosis and therapeutic indication. Currently, AAIR contains 2,777 antibody-antigen pairs covering 159 disease conditions, 2,035 antibody heavy chain sequences, 1,701 antibody light chain sequences, 619 distinct antigen sequences (584 proteins/peptides and 35 other molecules), 254 antigen epitope sequences, and 157 binding affinity constants for antigen-antibody pairs from various viruses, bacteria, tumor types, and autoimmune responses.

The potential application of the data in AAIR for the study of antibody-antigen recognition was demonstrated by applying machine learning models to predict antibody from antigen sequence. It can be concluded from the performance of machine learning models that the information in AAIR is capable of producing comparable and reasonable preliminary results to characterize pair-wise interaction between antibody and antigen, and would be useful for antibody and antigen design.

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Ab-Ag:	antibody-antigen
Ab:	antibody
Ag:	antigen
ALL:	acute lymphoblastic leukemia
AML:	acute myeloid leukemia
ANN:	artificial neural networks
cAMP:	cyclic adenosine monophosphate
cDNA:	complementary DNA
CH:	the constant region of the heavy chain variable sequence
CL:	the constant region of the light chain variable sequence
DNA:	deoxyribonucleic acid
EST:	expressed sequence tag
FDA:	food and drug administration
FN:	false negative
FP:	false positive
HLA:	human leukocyte antigen
IG:	immunoglobulin
KEGG:	Kyoto encyclopedia of genes and genomes database
KNN:	k-nearest neighbors
LS:	least square method
MHC:	major histocompatibility complex
MIAME:	minimum information about a microarray experiment
ML:	machine learning
NCBI:	national center for biotechnology information
NSCLC:	non-small cell lung cancer
NPV:	negative predictive value
NSP:	the number of non-survivable patients
PCA:	principal component analysis
PDB:	protein databank
Pfam:	protein family
PNN:	probabilistic neural networks
PPV:	positive predictive value
Q:	overall accuracy
RFE:	recursive feature elimination
RNA:	ribonucleic acid
SAGE:	serial analysis of gene expression
SCLC:	small cell lung cancer
SE:	sensitivity
SMD:	Stanford Microarray Database
SMO:	sequential minimal optimization
SP:	specificity
SP:	the number of survivable patients
SQL:	structured query language
STDEV:	standard deviation
SV:	support vector
SVM:	support vector machines

TCR:	T-cell receptor
TN:	true negative
TP:	true positive
TTD:	therapeutic target database
VH-VL:	the variable region of the heavy chain sequence and the variable region of the light chain variable sequence
VH:	the variable region of the heavy chain sequence
VL:	the variable region of the light chain variable sequence
WHO:	world health organization

1 Introduction

Functional genomics has been widely applied in determining disease mechanisms and identifying disease markers. The possibility of the marker as a good therapeutic target can be evaluated by how well therapeutic molecules, such as small molecules or antibodies, can target them. However, the disease marker selection, which is critical for disease diagnosis, prognosis, treatment and disease-targeting molecule design, can be a difficult task since human genome contains approximately 25,000 genes (1), which are expressed at different time and are cooperated as an integrated team. The discovery of the disease markers can facilitate disease target identification and disease targeting molecule design. The first section (Section 1.1) of this chapter gives an overview of disease markers and therapeutic molecules. The following two sections of this chapter introduce the current progress in disease marker discovery (Section 1.2) and therapeutic molecules prediction (Section 1.3). The motivation of this work and outline of the structure of this document are presented in Section 1.4.

1.1 Overview of disease markers and therapeutic molecules

Knowing the origin of a disease is the first step in understanding the entire abnormal course of the disease and helping the treatment of the disease. Sometimes it is very easy to determine the cause of certain diseases, such as infectious diseases which are generally caused by virus, bacteria or parasites. However, the sources of some diseases may not be easily identified, especially some genetic diseases resulting from an accumulation of inherited and

environmentally-induced changes or mutations in the genome, such as cancer (2-6), diabetes (7, 8), cardiovascular disorders (9, 10) and obesity (11). For accurate disease diagnosis and proper treatment selection, it is very important to identify the gene markers responsible for disease initiation. Moreover, the discovery of the markers responsible for disease progress is critical because such markers can be used to identify disease stages, subtypes and prognosis effect in an accurate manner. As such, proper treatment regime can be applied and the survivability of the patients can be ultimately extended (12).

The completion of human genome sequencing (1, 13), and the new, cheap, and reliable methods in functional genomics such as gene expression analysis present the potential for disease marker discovery. Most of the markers show significantly different expression profiles between healthy people and patients, or among the patients with different progress stages/subtypes/outcomes, characterizing disease at the molecule level and for diagnosis and prognosis prediction. They can be further analyzed as the potential disease targets which normally play key roles in disease initiation (14) or disease progress (15, 16). The disease targets can be used in developing disease targeting molecules such as antibodies and small molecules based on the antibody-antigen interaction and protein-small molecule interaction (17).

Disease targeting molecule design aims to identify small molecules or antibodies that bind strongly to the disease targets (15, 16). The understanding of the interaction of targets and therapeutic molecules are crucial for disease targeting molecule design. The rapid progress in human genome project and functional

genomics provides an ever-increasing number of potential therapeutic targets, and the computational analysis of protein-protein interaction or ligand-protein interaction should facilitate the therapeutic molecule design.

1.2 Current progress in disease marker discovery

1.2.1 Introduction to disease differentiation

Generally genetic diseases such as cancer are differentiated according to their gross morphological appearance of the cells and the surrounding tissues. However, such a differentiation criterion has some limitations. First, it relies on a subjective review of the tissue, which depends on the knowledge and experience of a pathologist, and may not be consistent or reproducible (18, 19). Second, this method provides discrete, rather than continuous classification of disease into broad groups with limited ability to determine the treatment regime of individual patients. Third, disease with identical pathology may have different origins and respond differently to treatment (20). Last but not the least, current pathology reports offer little information about the potential treatment regime which a disease will respond to. Therefore, new disease differentiation method is needed for accurate diagnosis and treatment.

Fortunately, disease differentiation based on molecular profile of diseases can overcome those limitations (6, 21-24). Microarray technology, which is capable of providing the expression profile information on thousands of genes simultaneously, has become a very important component of disease molecular differentiation. The gene expression profiles can be applied to identify markers

which are closely associated with early detection/differentiation of disease, or disease behavior (disease progression, response to therapy), and could serve as disease targets for drug design (25). This strategy is widely used in cancer research for the identification of cancer markers, and provide new insights into tumorigenesis, tumor progression and invasiveness (5, 6, 26-29).

1.2.2 Approaches of disease marker discovery

1.2.2.1 Traditional gene discovery method

Two approaches, the candidate gene approach and positional cloning approach, have traditionally been used to discover genes underlying human diseases.

Candidate gene method is based on prior biochemical knowledge about the genes, such as putative functional protein domain of genes and tissues in which genes are expressed (30, 31). Genes underlying familial hypertrophic cardiomyopathy (32), Li-Fraumeni syndrome (33), retinitis pigmentosa (34, 35), hereditary prostate cancer risk (31), metastasis of hepatocellular carcinoma (36), and breast cancer risk (37) were discovered in this manner. However very limited well-characterized genes are currently available (30), and most genes can not be analyzed in this manner due to the limitation of biochemical knowledge.

In contrast to candidate gene method, positional cloning identifies genes without any prior knowledge about gene function. This method is performed in patients and their family members using DNA polymorphisms. Alleles of markers that are

in close proximity to the chromosome location of the disease genes can be determined by genetic linkage analysis, and critical region can be defined by haplotype analysis. The candidate genes residing in the critical regions can be identified (9, 30). This method was applied in identifying genes related with asthma (38), cardiovascular disorders (9, 10), and diabetes mellitus (8). However, the nature of positional cloning limits its resolution to relatively large regions of the genome (30). The candidate genes within a certain critical region need to be filtered from the relatively large regions of the genome by identifying mutations in genes that segregate with the disease (30).

1.2.2.2 Proteomics method

Most recent developed proteomics offers the most direct approach to understanding disease and its molecular markers (39-41). Proteomics refers to the systematic analysis of protein, protein complexes, and protein-protein interactions (42). This approach provides complementary information that can be useful in studying disease processes, such as cardiomyopathies (43), autosomal recessive malignant infantile osteopetrosis (44-46), lung cancer (40) and prostate cancer (47). However, this newly-developed and immature method makes limited data available for comparison and analysis.

1.2.2.3 Genomics method

Genomics method is another new gene discovery method. Two kinds of technology, phylogenetic profiles and global profiles of gene expression, are widely used in this approach.

Based on sequencing technology, phylogenetic profiles is a powerful computational strategy that infers gene function from the completed genome sequences (48-51). This technology assumes that function-related genes are evolving in a correlated way, so that they are more likely to share homologs among organisms. Six possible Bardet-Biedel syndrome genes were identified by this technology (52, 53).

Currently the most important method for disease gene discovery is global profiles of gene expression based on genomic knowledge. This method discovers disease genes from the expression level of a set of genes in particular tissues or cell types. Serial analysis of gene expression (SAGE) (54) is a method which produces a snapshot of mRNA population in a sample by a sequence-based sampling technique. Another technology is the newly-developed microarray technology. Probably as the richest source of gene expression data, microarray data is used in this study for gene selection. Microarray measures the expression profiles of thousands of genes at the same time and have been explored for deriving disease genes or disease markers (5, 26, 55-62), elucidating pathogenesis of disease (55, 60, 63-66), deciphering mechanism of drug action (67-69), determining treatment-strategies (70, 71), and characterizing genomic activity during various cellular processes (72-75). The markers in colorectal tumors (76) and non-Hodgkin's lymphoma (77), and prognostic markers of acute myeloid leukemia (78) were identified by using microarray technology.

1.2.3 Brief introduction to microarray technology

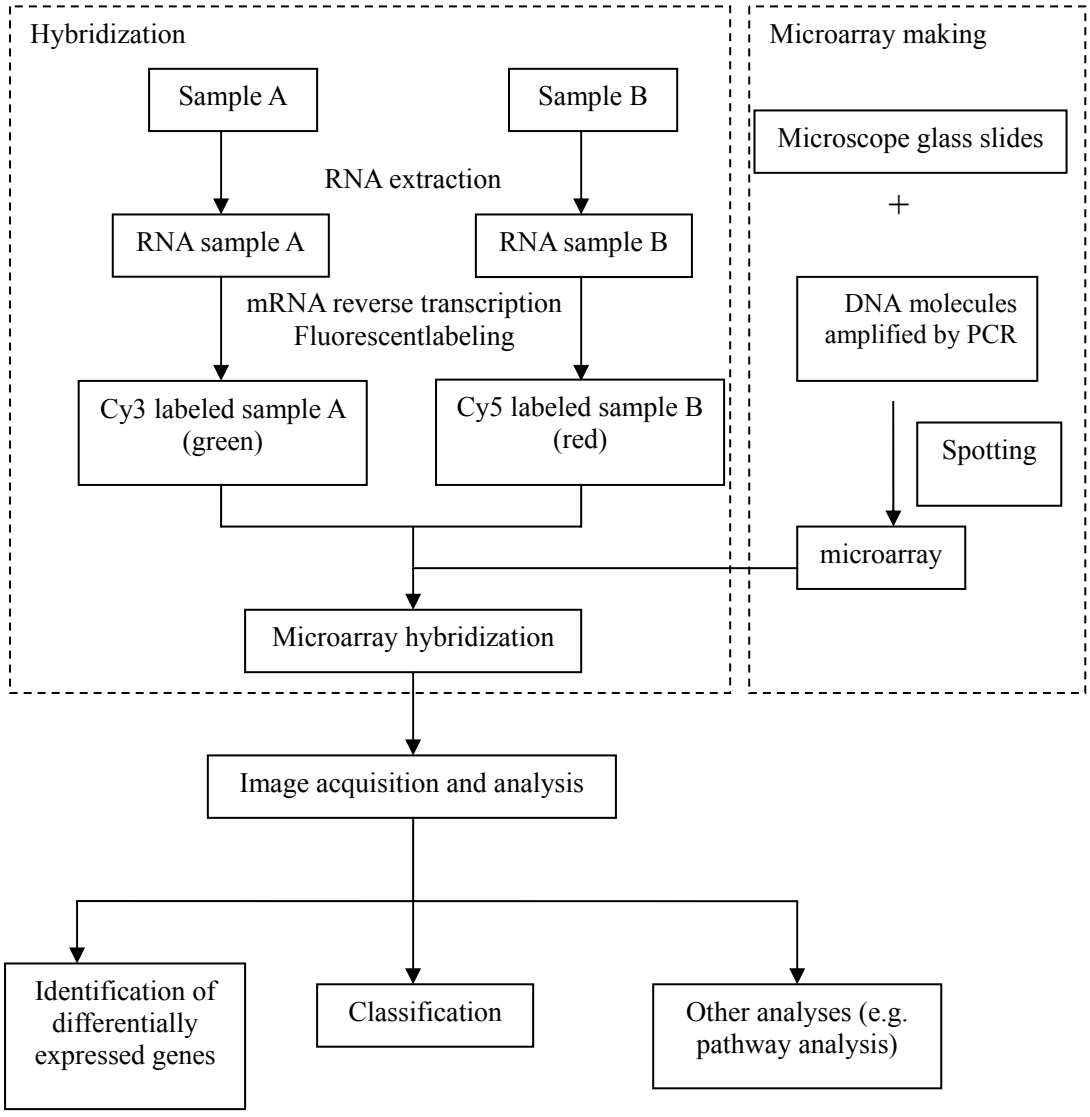
1.2.3.1 Introduction to microarray experiments

Microarray technology, also known as DNA chip, gene ship or biochip, is one of the indispensable tools in monitoring genome wide expression levels of genes in a given organism. Microarrays measure gene expression in many ways, one of which is to compare expression of a set of genes from cells maintained in a particular condition A (such as disease status) with the same set of genes from reference cells maintained under conditions B (such as normal status).

Figure 1-1 shows a typical procedure of microarray experiments (79, 80). A microarray is a glass substrate surface on which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). A microarray may contain thousands of spots, and each spot may contain a few million copies of identical DNA molecules (probes) that uniquely correspond to a gene. The DNA in a spot may either be genomic DNA (81), or synthesized oligo-nucleotide strands that correspond to a gene (82-84). This microarray can be made by the experimenters themselves (such as cDNA array) or purchased from some suppliers (such as Affymetrix GeneChip). The actual microarray experiment starts from the RNA extraction from cells. These RNA molecules are reverse transcribed into cDNA, labeled with fluorescent reporter molecules, and hybridized to the probes formatted on the microarray slides. At this step, any cDNA sequence in the sample will hybridize to specific spots on the glass slide containing its complementary sequence. The amount of cDNA bound to a spot will be directly proportional to the initial number of RNA molecules present for that gene in both samples. Following, an instrument is used to read the reporter molecules and create

microarray image. In this image, each spot, which corresponds to a gene, has an associated fluorescence value, representing the relative expression level of that gene. Then the obtained image is processed, transformed and normalized. And the analysis, such as differentially expressed gene identification, classification of disease/normal status, and pathway analysis, can be conducted.

Figure 1-1 Procedure of microarray experiment



1.2.3.2 Public repository for microarray data

Thanks to the variety of journals and funding agencies which have established

and enforced microarray data submission standards, currently, a wealth of microarray data is now available in different databases such as the Stanford Microarray Database (SMD) (85), Gene Expression Omnibus (GEO) (86), and Array Express (EBI) (87). Table 1-1 gives a list of public available microarray databases. Many of those databases require a minimum information about a microarray experiment (MIAME)-compliant manner in order to interpret the experiment results unambiguously and potentially be able to reproduce the experiment (88). As a public resource, these expression databases are valuable substrates for statistical analysis, which can detect gene properties that are more subtle than simple tissue-specific expression patterns.

1.2.3.3 Statistical analysis of microarray data

Since microarray contains the expression level of several thousands of genes, it requires sophisticated statistical analysis to extract useful information such as gene selection. Theoretically, one would compare a group of samples of different conditions and identify good candidate genes by analysis of the gene expression pattern. However, microarray data contain some noises arising from measurement variability and biological differences (70, 89). The gene-gene interaction also affects the gene-expression level. Furthermore, the high dimensional microarray data can lead to some mathematical problems such as the curse of dimensionality and singularity problems in matrix computations, causing data analysis difficult. Therefore choosing a suitable statistical method for gene selection is very important.

Table 1-1 A list of public microarray databases.

Database	Website*	Description	Organism	References
ArrayExpress	http://www.ebi.ac.uk/arrayexpress/	A public repository for microarray based gene expression data	European Bioinformatics Institute	(87)
ChipDB	http://chipdb.wi.mit.edu/chipdb/public/	A searchable database of gene expression	Massachusetts Institute of Technology	(90)
ExpressDB	http://twod.med.harvard.edu/ExpressDB/	A relational database containing yeast and E. coli RNA expression data	Harvard Medical School	(91)
Gene Expression Atlas	http://symatlas.gnf.org/SymAtlas/	A database for gene expression profile from 91 normal human and mouse samples across a diverse array of tissues, organs, and cell lines	Novartis Research Foundation	(92)
Mouse Gene Expression Database (GXD)	http://www.informatics.jax.org/menus/expression_menu.shtml	An extensive and easily searchable database of gene expression information about the mouse	The Jackson Laboratory, Bar Harbor, Maine	(93)
Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/	Microarray database containing tens of millions of expression profiles	National Center for Biotechnology Information	(86)
GermOnline	http://www.germonline.org/index.html	Information and microarray expression data for genes involved in mitosis and meiosis, gamete formation and germ line development across species	Biozentrum and Swiss Institute of Bioinformatics	(94)
Human Gene Expression (HuGE) Index database	http://www.biotechnologycenter.org/hio/	A comprehensive database to understand the expression of human genes in normal human tissues	Boston University	(95)
MUSC DNA Microarray Database	http://proteogenomics.musc.edu/ma/musc_madb.php?page=home&act=manage	A web-accessible archive of DNA microarray data	Medical University of South Carolina	(96)
RIKEN Expression Array Database (READ)	http://read.gsc.riken.go.jp/	A database of expression profile data from the RIKEN mouse cDNA microarray	RIKEN Yokohama Institute	(97)
Rice Expression Database (RED)	http://red.dna.affrc.go.jp/RED/	Expression profiles obtained by the Rice Microarray Project and other research groups	National Institute of Agrobiological Sciences, Japan	(98)
RNA Abundance Database (RAD)	http://www.cbil.upenn.edu/RAD/php/index.php	A public gene expression database designed to hold data from array-based and nonarray-based (SAGE) experiments	University of Pennsylvania	(99)
Saccharomyces Genome Database (SGD): Expression Connection	http://db.yeastgenome.org/cgi-bin/expression/expressionConnection.pl	A gene expression database of Saccharomyces genome	Stanford University	(100)
Stanford Microarray Database (SMD)	http://genome-www5.stanford.edu/	Raw and normalized data from microarray experiments, as well as their corresponding image files	Stanford University	(85)
Yale Microarray Database (YMD)	http://info.med.yale.edu/microarray/	A microarray database for large-scale gene expression analysis.	Yale University	(101)
yeast Microarray Global Viewer (yMGV)	http://www.transcriptome.ens.fr/ymgv/	A database for yeast gene expression	Ecole Normale Superieure, Paris, France	(102)

*accessible at Apr 06, 2007

The statistical methods in microarray data analysis can be classified into two groups: unsupervised learning methods and supervised learning methods. Unsupervised analysis of microarray data aims to group relative genes without knowledge of the clinical features of each sample (103). A commonly-used unsupervised method is hierarchical clustering method. This method groups genes together on the basis of shared expression similarity across different conditions, under the assumption that genes are likely to share the same function if they exhibit similar expression profiles (104-107). Hierarchical clustering creates phylogenetics trees to reflect higher-order relationship between genes with similar expression patterns by either merging smaller clusters into larger ones, or by splitting larger clusters into smaller ones. A dendogram is constructed, in which the branch lengths among genes also reflect the degree of similarity of expression (108, 109). By cutting the dendogram at a desired level, a clustering of the data items into the disjoint groups can be obtained. Hierarchical clustering of gene expression profiles in rheumatoid synovium identified 121 genes associated with Rheumatoid arthritis I and 39 genes associated with Rheumatoid arthritis II (110). Unsupervised methods have some merits such as good implementations available online and the possibility of obtaining biological meaningful results, but they also possess some limitations. First, unsupervised methods require no prior knowledge and are based on the understanding of the whole data set, making the clusters difficult to be maintained and analyzed. Second, genes are grouped based on the similarity which can be affected by input data with poor similarity measures. Third, some of the unsupervised methods require the predefinition of one or more user-defined parameters that are hard to be estimated (e.g. the number of clusters). Changing these parameters often have a strong impact on the final results (113).

In contrast to the unsupervised methods, supervised methods require a priori knowledge of the samples. Supervised methods generate a signature which contains genes associated with the clinical response variable. The number of significant genes is determined by the choice of significance level. Support vector machines (SVM) (114) and artificial neural networks (ANN) (115) are two important supervised methods. Both methods can be trained to recognize and characterize complex pattern by adjusting the parameters of the models fitting the data by a process of error (for example, mis-classification) minimization through learning from experience (using training samples). SVM separates one class from the other in a set of binary training data with the hyperplane that is maximally distant from the training examples. This method has been used to rank the genes according to their contribution to defining the decision hyperplane, which is according to their importance in classifying the samples. Ramaswamy et al. used this method to identify genes related to multiple common adult malignancies (6). ANN consists of a set of layers of perceptrons to model the structure and behavior of neutrons in the human brain. ANN ranks the genes according to how sensitive the output is with respect to each gene's expression level. Khan et al identified genes expressed in rhabdomyosarcoma from such strategy (27).

In classification of microarray datasets, it has been found that supervised machine learning methods generally yield better results (116), particularly for smaller sample sizes (89). In particular, SVM consistently shows outstanding performance, is less penalized by sample redundancy, and has lower risk for over-fitting (117, 118). Furthermore, some studies demonstrated that SVM-based prediction system was consistently superior to other supervised learning methods in microarray data

analysis (119-121). SVM for microarray data analysis are used in this study.

Feature selection in microarray data analysis

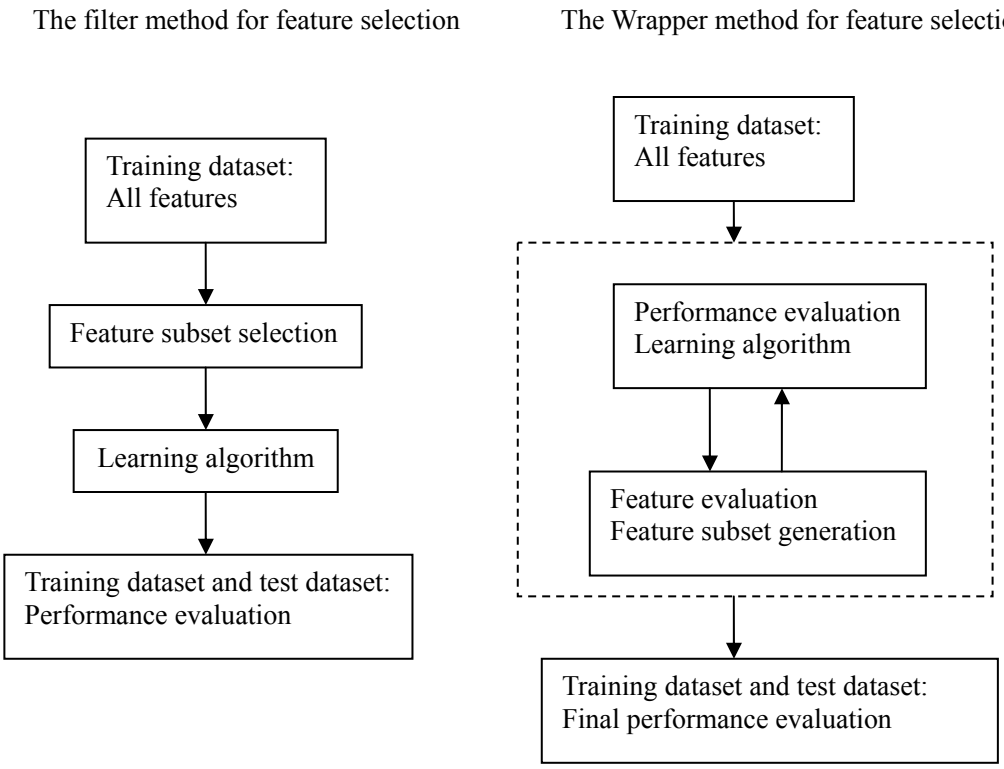
No matter whether the supervised or unsupervised methods are used, one critical problem encountered in both methods is feature selection, which has become a crucial challenge of microarray data analysis. The challenge comes from the presence of thousands of genes and only a few dozens of samples in currently available data. From the mathematical view, thousands of genes are thousands of dimensions. Such a large number of dimensions leads microarray data analysis to problems such as the curse of dimensionality (122, 123) and singularity problems in matrix computations. Therefore, there is a need of robust techniques capable of selecting the subsets of genes relevant to a particular problem from the entire set of microarray data both for the disease classification and for the disease target discovery.

Gene selection from microarray data is to search through the space of gene subsets in order to identify the optimal or near-optimal one with respect to the performance measure of the classifier. Many gene selection methods have been developed, and generally fall into two categories: filter method and wrapper method (124). Figure 1-2 shows how these two methods work.

In brief, the filter method selects genes independent of the learning algorithms (125-127). It evaluates the goodness of the genes from simple statistics computed from the empirical distribution with the class label (128). Filter method has some pre-defined criteria. Mutual information and statistical testing (e.g. T-test and

F-test) are two typical examples of filter method (5, 125, 129-133). Filter method can be easily understood and implemented, and needs little computational time. But the pitfall of this method is that it is based on the assumption that genes are not connected to each other, which is not true in real biological process.

Figure 1-2 Filter method versus wrapper method for feature selection



Wrapper method generates genes from the evaluation of a learning algorithm. It is conducted in the space of genes, evaluating the goodness of each gene or gene subsets by such criteria as cross-validation error rate or accuracy from the validation dataset (134). The wrapper method is very popular among machine learning methods for gene discovery (124, 135, 136). Although the wrapper method needs extensive computational resources and time, it considers the gene-gene interaction and its accuracy is normally higher than the filter method (124, 135, 136). Recursive feature elimination (RFE) is a good example of the

wrapper method for disease gene discovery. The RFE method uses the prediction accuracy from SVM to determine the goodness of a selected subset. This thesis will employ RFE for disease gene discovery from microarray data.

1.2.4 The problems of current marker selection methods

The methodology of SVM and RFE will be discussed in Chapter 2 in details. Here, some problems encountered in current marker discovery from microarray data analysis are discussed. One problem is to specify the number of genes for differentiating disease. The number of derived colon cancer genes and leukemia genes ranges from 1 to 200 (5, 137-142). 50 genes were arbitrarily chosen for differentiating acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) by Golub et al, since they supposed that 50 genes might reflect the difference between AML and ALL (5). In most cases, the gene number was decided by the classification performance of different gene combinations. The gene combination which produced the highest classification accuracy constituted the gene signature. This strategy might produce small sets of genes (one or two genes) that formed accurate classifier (140-142). For example, Slonim et al reported that the classifier consisting of one gene (HOXA9) outperformed all of other classifiers consisting of other gene combinations for recurrence prediction in AML patients (142). Li and Yang showed that one gene (Zyxin) constituted the best classifiers for AML/ALL differentiation (140). Nevertheless these results were only obtained and tested on one dataset. Considering that the number of genes should correlate with the disease situation, the selected genes should be large enough to be robust against noise and small enough to be readily applied in clinical settings. Therefore, it is not appropriate to use the arbitrary gene number.

Similarly, to use just one dataset to decide the optimal gene number may not be satisfactory, because the optimal gene number varies with the different sample sizes and sample combinations (70, 143, 144).

Another problem in gene discovery is the gene signatures were highly unstable and strongly depended on the selection of patients in the training sets (5, 27, 58, 59, 70, 89, 145, 146) (70, 143, 144), despite the use of sophisticated class differentiation and gene selection methods by various groups. The unstable signatures were observed in most microarray datasets including colon cancer, lung adenocarcinoma, non-Hodgkin lymphoma, acute lymphocytic leukemia, acute myeloid leukemia, breast cancer, medulloblastoma, and hepatocellular carcinoma (70, 108, 119, 124, 127, 145, 147-150). While these signatures display high predictive accuracies, the highly unstable and patient-dependent nature of these signatures diminishes their application potential for diagnosis and prognosis (70). Moreover, the complex and heterogenic nature of disease such as cancer may not be adequately described by the few cancer-related genes in some of these signatures. The unstable nature of these signatures and their lack of disease-relevant genes also limit their potential for target discovery. The instability of derived signatures is likely caused by the noises in the microarray data arising from such factors as the precision of measured absolute expression levels, capability for detecting low abundance genes, quality of design and probes, annotation accuracy and coverage, and biological differences of expression profiles (89, 151). Apart from enhancing the quality of measurement and annotation, strategies for improving signature selection have also been proposed. These strategies include the use of multiple random validation (70), large sample

size (152), known mechanisms (153), and robust signature-selection methods which is insensitive to noises (55, 89, 154).

This thesis will explore a new signature selection method aiming at reducing the chances of erroneous elimination of predictor-genes due to the noises contained in microarray dataset. Multiple random sampling and gene-ranking consistency evaluation procedures will be incorporated into RFE signature selection method. The consistent genes obtained from the multiple random sampling method may give us a better understanding to the disease initiation and progress, and may provide potential disease targets.

1.3 Current progress in disease targeting molecule prediction, antibody as a case study

1.3.1 Overview of disease-targeting molecule

As introduced in the previous section, Microarray data can be employed to discover markers closely related to disease initiation and progression and can provide candidate disease targets. The interaction between disease targets and therapeutic molecules is crucial for drug discovery. Therapeutic molecule can attach its specific molecule targets involving in pathogenesis and disease progress without damaging other tissues (155, 156). The rationally design of therapeutic molecules has therefore become a very important area in current drug design.

1.3.1.1 Small molecules

The therapeutic molecules include small molecules and antibodies (15, 16).

Table 1-2 gives an overview of US Food and Drug Administration (FDA) approved anticancer small molecular drugs in recent ten years. A kind of important small molecule drugs for therapeutic application is protein kinase inhibitors, which specifically act on their disease targets - protein kinases (16, 157), which are implicated in a wide range of diseases. Protein kinases can catalyze protein phosphorylation, which is one of the most significant signal transduction mechanisms, and by which crucial intercellular processes are regulated. Currently the protein kinase family is the second largest enzyme family and the fifth largest gene family in the human genome (157). 520 protein kinase genes, corresponding to about 1.7% of all human genes, were identified in humans (157). The key role of protein kinase in regulating signal transduction in the context of multiple cellular processes and environments and the regulatory approval in clinical applications makes kinase as a readily accepted druggable protein (16). Nevertheless, one significant obstacle to the rational design of specific kinase inhibitors is the high level of sequence and structural similarity in the human kinase types (158). Furthermore, kinases tend to conformational changes when drugs bind (158). Currently around 11% successful rate achieved for this kind of drugs (159) from the first use in humans to regulatory approval.

1.3.1.2 Antibodies

Antibodies, another frequently used therapeutic molecules, can specifically act on the disease-causing targets (antigens) (15) on many diseases such as cancer (16), heart disease (160) and rheumatic diseases (161). Antibodies have a unique characteristic that small molecules don’t have, - the ability to exquisitely discriminate diverse disease-related molecules (specificity) and the ability to

tightly bind to their targets (affinity). These two capabilities make antibody fight disease with an efficient, little toxically manner and a good side-effect profile compared to small molecules. Therefore the therapeutic antibodies can achieved 18−29% successful rate (162). This thesis will utilize antibodies as an example for therapeutic molecule design.

Table 1-2 US FDA-approved molecule targeting drugs (small molecules) between 1996 to 2006 (163, 164).

Year	Drugs	Drug Types	Molecular Target	Disease Indication	Therapeutic Application	Company
2006	Sprycel (dasatinib)	Tyrosine kinase inhibitor	BCR-ABL, SRC	Chronic myeloid leukemia (CML)	Treatment of imatinib-resistant chronic myeloid leukemia	Bristol-Myers Squibb
	Sutent (sunitinib)	Tyrosine kinase inhibitor	PDGFR, VEGFR, KIT, FLT3, CSF-1R, RET	Kidney Cancer; Gastrointestinal Stromal Tumors	Treatment of kidney cancer and gastrointestinal stromal tumors	Pfizer
	Gardasil	Quadrivalent recombinant vaccine	Human papillomavirus types 6, 11, 16, and 18	For the prevention of cervical cancer associated with human papillomavirus	For the prevention of cervical cancer associated with human papillomavirus	Merck
2005	Nexavar (sorafenib)	Multikinase inhibitor	VEGFR, PDGFR, c-KIT	Renal Cell Carcinoma	Treatment of Renal Cell Carcinoma	Bayer/ Onyx
	Arranon (nelarabine)[1]	Cytotoxic deoxyguanosine analogue	DNA	Leukemia, lymphoma	For the treatment of lymphoblastic leukemia and T-cell lymphoblastic lymphoma	GlaxoSmithKline
2004	Tarceva (erlotinib, OSI 774)	Tyrosine kinase inhibitor	HER1, EGFR	Non-small cell lung cancer (NSCLC)	Treatment of advanced refractory metastatic non-small cell lung cancer	Genentech, OSI Pharmaceuticals
	Alimta (pemetrexed)	Enzyme Inhibitors	Dihydrofolate reductase, Glycinmaide ribonucleotide formyl transferase, thymidylate synthase	Mesothelioma	For the treatment of malignant pleural mesothelioma	Eli Lilly
	Clolar (clofarabine)	Purine nucleoside analog	DNA	Lymphoblastic leukemia	For the treatment of acute lymphoblastic leukemia in pediatric patients	Genzyme
	Sensipar (cinacalcet)	Allosteric activators	Calcium-sensing receptor	Parathyroid carcinoma	For the treatment of secondary hyperparathyroidism and hypercalcemia in parathyroid carcinoma patients	Amgen
	VESANOID (Tretinoin, ATRA)	Cell Stimulants and Proliferants	Alpha retinoic acid Receptors (RAR)-alpha	Acute promyelocytic leukemia (APL)	For the treatment of acute promyelocytic leukemia (APL)	Roche
2003	Iressa (gefitinib)	Tyrosine kinase inhibitor	EGFR	Non-small cell lung cancer (NSCLC)	The second-line treatment of non-small-cell lung cancer	AstraZeneca

	Velcade (bortezomib)	Proteasome inhibitor	26S proteasome	Multiple Myeloma	Injectable agent for the treatment of multiple myeloma patients who have received at least two prior therapies	Millennium Pharmaceuticals
	Aloxi (palonosetron)	Serotonin 5-HT_3 receptor antagonist (GPCR antagonist)	Serotonin 5-HT_3 receptor (GPCR)	Chemotherapy side effects	For the prevention of nausea and vomiting associated with emetogenic cancer chemotherapy	MGI Pharma, Helsinn Healthcare
	Emend (aprepitant)	P/neurokinin 1 (NK1) receptor antagonists (GPCR antagonists)	Neurokinin receptors (GPCR)	Chemotherapy-induced Nausea and Vomiting	For the treatment of nausea and vomiting associated with chemotherapy	Merck
	Plenaxis (abarelix)	Gonadotropin-releasing hormone (GnRH) antagonist	Gonadotropin-releasing hormone (GnRH)	Prostate Cancer	For treatment of advanced prostate cancer	Praecis Pharmaceuticals
	UroXatral (alfuzosin HCl)	Antagonist of post-synaptic alpha1-adrenoreceptors	Alpha1-adrenoreceptor	Benign Prostatic Hyperplasia	For the treatment of of the signs and symptoms of benign prostatic hyperplasia	Sanofi-Synthelabo
2002	Gleevec (imatinib mesylate)	Protein-tyrosine kinase inhibitor	PDGF, SCF, c-kit,	Positive inoperable and/or metastatic malignant gastrointestinal stromal tumors (GISTs)	Treatment of gastrointestinal stromal tumors (GISTs)	Novartis
	Faslodex (fulvestrant)	Estrogen receptor antagonist	Estrogen receptor	Hormone receptor positive metastatic breast cancer	Treatment of hormone receptor positive metastatic breast cancer	AstraZeneca
	Eligard (leuprolide acetate)	Luteinizing hormone-releasing hormone (LHRH) agonist,	Luteinizing hormone-releasing hormone (LHRH)	Prostate cancer	For the palliative treatment of advanced prostate cancer	Atrix Laboratories
	Eloxatin (oxaliplatin/5-fluorouracil/leucovorin)	Synthases inhibitor	Thymidylate synthetase	Metastatic colon or rectum carcinomas	For the treatment of colon or rectum carcinomas	Sanofi-Synthelabo
	SecreFlo (secretin)	Diagnostic Agents	Secretin receptor	gastrinoma	To aid in the diagnosis of pancreatic dysfunction and gastrinoma	Repligen
	Zometa (zoledronic acid)	Bisphosphonate, Antihypocalcemic Agents	Farnesyl pyrophosphate synthetase	Multiple myeloma; bone metastases from solid tumors	For the treatment of multiple myeloma and bone metastases from solid tumors	Novartis
2001	Gleevec (imatinib mesylate)	Protein-tyrosine kinase inhibitor	c-kit, PDGFR	Chronic myeloid leukemia (CML)	Oral therapy for the treatment of chronic myeloid leukemia	Novartis
	Femara (letrozole)	Enzyme inhibitor	Aromatase enzyme	Breast cancer	First-line treatment of postmenopausal women with locally advanced or metastatic breast cancer	Femara (letrozole) Tablets
	Kytril (granisetron)	serotonin 5-HT_3 receptor antagonist (GPCR antagonist)	serotonin 5-HT_3 receptor (GPCR)	Side effect of cancer therapy	For the prevention of nausea and vomiting associated with cancer therapy	Kytril (granisetron) Solution
	Trelstar LA	Repressor	gonadotropin	Prostate cancer	Intramuscular injection for the treatment of advanced stage prostate cancer	Trelstar LA
	Xeloda[2]	Synthases inhibitor	Thymidylate synthetase	Colorectal cancer	Chemotherapy for the treatment of metastatic colorectal cancer	Xeloda
	Zometa (zoledronic acid)	bisphosphonate	osteoclasts	Hypercalcemia of malignancy	For the treatment of hypercalcemia of malignancy	Zometa (zoledronic acid)

2000	Trelstar (Triptorelin Pamoate)	Repressor	gonadotropin	Prostate cancer	For the palliative treatment of advanced prostate cancer	Debio Rechereche Pharmaceutique, Target Research Associates
	Trisenox (arsenic trioxide)	Homeopathic Agents	PML-RAR Alpha Protein	Acute Promyelocytic Leukemia	For the induction of remission and consolidation in patients with acute promyelocytic leukemia (APL)	Cell Therapeutics
	Viadur (leuprolide acetate)	Testosterone suppressor, luteinizing hormone-releasing hormone (LH-RH) agonist	Gonadotropin	Prostate Cancer	For pain relief in men with advanced prostate cancer	Alza
1999	Aromasin (Exemestane)	Oxidoreductase inhibitor	Aromatase	Breast cancer	Treatment of breast cancer	Pharmacia & Upjohn
	Busulflex	Alkylating agent	DNA	Leukemia	For use for the treatment of leukemia	Orphan Medical
	Doxil (doxorubicin HCl liposome injection)	Nucleic acids intercalator	Topoisomerase II	Breast cancer, ovarian cancer	Treatment for ovarian cancer that is refractory to other first-line therapies	Alza
	Ellence (epirubicin hydrochloride)	Anthracycline cytotoxic agent	DNA Helicase	Breast cancer	For treatment of axillary node tumor involvement for primary breast cancer	Pharmacia & Upjohn
	Ethyol (amifostine)	Radiation-Protective Agents	Alkaline phosphatase	Side effect of cancer therapy	Treatment for xerostomia (dry mouth) due to radiation	U.S. Bioscience, Alza
	Temodar (temozolomide)	Cytotoxic alkylating agent,	DNA	Anaplastic astrocytoma	Treatment for refractory anaplastic astrocytoma	Schering-Plough
	UVADEX (methoxsalen)	Inhibitor	DNA	Cutaneous T-cell lymphoma	Treatment of the skin manifestations of cutaneous T-cell lymphoma (CTCL)	Therakos
	Zofran ODT (ondansetron)	Serotonin 5-HT_3 receptor antagonist	Serotonin 5-HT_3 receptor	Chemotherapy side effect	Treatment for the prevention of chemotherapy and radiation-induced nausea	GlaxoWellcome
1998	Actiq (Fentanyl)	Opiate Agonists	Opioid mu Receptor (OP3)	Cancer Pain	Treatment for Cancer Pain	Anesta Corporation
	Anzemet (Dolasetron)	Serotonin 5-HT_3 receptor Antagonists	Serotonin 5-HT_3 receptor	Side effect of cancer therapy	Treatment for the prevention of nausea and vomiting associated with chemotherapy and surgery	Hoechst Marion Roussel
	Camptosar (Irinotecan)	Enzyme Inhibitors	DNA Topoisomerase I	Colorectal	Treatment for Colon or Rectal Cancer	Pharmacia & Upjohn
	Gemzar (Gemcitabine)	Immunosuppressive Agents	Ribonucleoside-diphosphate reductase large subunit	Lung cancer	Treatment for Lung Cancer	Eli Lilly
	Neupogen (Filgrastim)	Immunomodulatory Agents	Granulocyte colony stimulating factor receptor (CD114 antigen)	Low white blood cell recovery following chemotherapy	Treatment for slow white blood cell recovery following chemotherapy	Amgen
	Nolvadex (tamoxifen citrate)	Nuclear receptor modulator	Oestrogen receptor	Breast cancer	Treatment for Breast Cancer	Zeneca Pharmaceuticals
	Photofrin (Porfimer)	Photosensitizing agent	Low density lipoproteins (LDL)	Lung cancer	Treatment for early-stage, microinvasive endobronchial non-small cell lung cancer	QLT
	Proleukin (Aldesleukin)	Human interleukin 2	Interleukin-2 receptor beta chain (IL-2-RB)	Metastatic melanoma	Treatment for metastatic melanoma	Chiron Corporation

	Valstar (Valrubicin)	Antibiotic	DNA Topoisomerase II	Bladder Cancer	Treatment for Bladder Cancer	Anthra Pharmaceuticals
	Xeloda (Capecitabine)	Antimetabolites	Thymidylate synthase	Breast cancer	Treatment for advanced breast cancer tumors	Roche
	Zofran (Ondansetron)	Serotonin 5-HT_3 receptor antagonist	Serotonin 5-HT_3 receptor	Chemotherapy side effect	Treatment for postoperative vomiting and nausea in adults	GlaxoWellcome
1997	Xibrom (Bromfenac)	Anti-Inflammatory Agents,	COX-1	Side effect of cancer therapy	Management of acute pain	Duract, Wyeth-Ayerst Laboratories
	Femara (Letrozole)	Aromatase Inhibitors	Aromatase	Breast cancer	Treatment for breast cancer	Novartis
	Gliadel (polifeprosan 20 with carmustine implant)	bifunctional alkylating agent	Glutathione reductase (mitochondrial)	recurrent glioblastoma multiforme	Treatment for brain cancer	Rhone-Poulenc Rorer, Guilford Pharmaceuticals
	Intron A (interferon alfa-2b, recombinant) for	Immunomodulatory Agents	Interferon receptor IFNAR2c	Non-Hodgkin's lymphoma	Treatment for non-Hodgkin's lymphoma	Schering-Plough
	Kytril (Granisetron)	Verotonin 5-HT_3 receptor antagonist	Serotonin 5-HT_3 receptor	Side effect of cancer therapy	Prevention of nausea and vomiting associated with chemotherapy	SmithKline Beecham
	Lupron Depot	Gonadotropin releasing hormone (GnRH) analogs	Leutinizing-hormone-releasing hormone	Prostate cancer	Treatment for prostate cancer	TAP Pharmaceuticals
	Neumega (Oprelvekin)	Thrombotics	Interleukin-11 receptor alpha chain (IL-11R-alpha)	Platelet deficiency in cancer patients	Treatment for thrombocytopenia	Genetics Institute
	Taxol (Paclitaxel)	Taxoid antineoplastic agent	Apoptosis regulator Bcl-2 (Tubulin beta-1 chain)	Kaposi's Sarcoma	Treatment for AIDS-related Kaposi's Sarcoma	Bristol-Myers Squibb
1996	Anexsia (Acetaminophen)	Antipyretics	Prostaglandin G/H synthase 1 precursor	Chronic pain	Treatment for chronic pain	Mallinckrodt Group
	Arimidex (anastrozole)	Aromatase Inhibitors	Aromatase	Breast cancer	Treatment for advanced breast cancer in postmenopausal women	Zeneca Pharmaceuticals
	Elliotts B Solution (buffered intrathecal electrolyte/dextrose injection)	Calcium Channel Blockers	Voltage-dependent calcium channel gamma-1 subunit	Leukemia, lymphoma	Treatment of meningeal leukemia or lymphocytic lymphoma	Orphan Medical
	Eulexin (flutamide)	Androgen Antagonists	Androgen receptor	Prostate cancer	Treatment for prostate cancer	Schering-Plough
	Gemzar (Gemcitabine HCl)	nucleoside analogue	DNA	Pancreatic cancer	Treatment for pancreatic cancer	Eli Lilly
	Hycamtin (topotecan hydrochloride)	Topoisomerase I inhibitor	Topoisomerase I	Ovarian cancer	Treatment for metastatic ovarian cancer	SmithKline Beecham
	Kadian (Morphine)	Opiate Agonists	Mu-type opioid receptor	Chronic pain of cancer patients	Treatment for chronic moderate to severe pain	Purepac Pharmaceutical
	Leukine (sargramostim)	Immunomodulatory Agents	Granulocyte-macrophage colony stimulating factor receptor (GM-CSF-R-alpha or CSF2R)	Replenishment of white blood cells	Treatment for the replenishment of white blood cells	Immunex
	Taxotere (Docetaxel)	Radiation-Sensitizing Agents	Apoptosis regulator Bcl-2	Breast cancer	Treatment for locally advanced or metastatic breast cancer	Rhone Poulenc Rorer

	Zoladex (goserelin acetate)	Decapeptide analogue	Luteinizing Hormone-Releasing Hormone (LH-RH) Receptor	Prostate cancer	Treatment for advanced prostate cancer	Zeneca Pharmaceuticals

[1] Nelarabine is demethoxylated by adenosine deaminase to ara-G, and converted by cellular kinases to the active 5'-triphosphate, ara-GTP. Incorporation of ara-GTP into DNA leads to inhibition of DNA synthesis and cell death (165)
[2]Once in the body, Xeloda is converted into fluorouracil (5-FU) by the naturally produced enzyme thymidine phosphorylase (TP).

1.3.2 Introduction to therapeutic antibody

Antibody is a kind of highly specific, naturally evolved molecules that recognize and eliminate pathogenic and disease antigens (166). The past 40 years of antibody research have hinted at the promising of new versatile therapeutic agents to fight cancer, autoimmune disease and infections. Currently antibody is one of the largest classes of drugs (167).

Antibodies are large glycoprotein molecules produced by B lymphocytes of the human immune system, with the capability to recognize a specific molecular structure on a target known as an antigen. The specificity of antibodies is that they are capable of distinguishing the sublet of molecular differences. The basic unit of all antibodies is a four-chain structure, which is composed of two identical light chains (lambda or kappa) and two identical heavy chains (IgA, IgD, IgG, IgE or IgM). Both the heavy and light chains can be divided into two regions based on the variability in the amino acids sequence. The regions include variable region of light chain (VL, approximately 110 amino acids), constant region of light chain (CL, approximately 110 amino acids), variable region of heavy chain (VH, approximately 110 amino acids), and constant region of heavy chain (CH, approximately 330 to 440 amino acids). The antibodies bind to antigens via variable regions. Constant regions interact with other components of the immune system and initiate the appropriate biological response, such as phagocytosis,

cytolysis or initiation of complement cascade followed by cell lysis, to eliminate the target pathogen or neutralize toxins.

Antibody is an essential component of the human immune system and a part of human body's principle defense mechanism against disease, and using antibody to fight disease is just a logical extension of their natural role. Even in one century ago, Paul Enrich proposed that antibody could be used as "magic bullets" to target and treat human diseases. However, only when the hybridoma technology was utilized in monoclonal antibodies production in 1975 and revolutionized the potential application of antibodies both for research, clinical diagnosis and treatment of disease (168), it makes antibody an important drug class (162, 167). The first successful use of a monoclonal antibody for cancer treatment was reported in 1982 (169) and the first US FDA-approved antibody for therapeutic usage was OKT3 in 1986 (170-174). Several years later, another antibody Reopro was approved (175). Currently 18 antibodies have been approved by FDA (Table 1-3) and at least 400 additional antibodies are in clinical development (176). The annual sales of antibody drugs was predicted to reach $16.7 billion in 2008 (177-179).

The successful application of antibody in the therapeutics makes antibody design an impressive research area. The popular wet-lab technologies such as phage-display technology (180) and transgenic technology (181) are available for antibody design. However, much effort is needed to identify the specificities of the antibody for these methods. A key challenge of current antibody rational design is to make an antibody for a specific antigen but not a vast number of other

molecules. Therefore it is very important to dissect the antibody-antigen recognition and interaction.

Table 1-3 US FDA-approved therapeutic antibody drugs.

Year	Drugs	Target Antigen	Type of Antibody	Isotype	Kd (nM)	FDA-Approved Indication(s)	Company	Reference
1986	Orthoclone OKT3 (muromonab-CD3)	CD3	Murine antibody	IgG2a	0.83	For treatment of acute allograft rejection	Johnson & Johnson	(162, 163, 167)
1994	ReoPro (abciximab)	GP IIb/IIIa receptor	Fab fragment of a chimeric antibody	IgG1	5	Used for prevention of cardiac ischemia complications	Johnson & Johnson	(162, 163, 167)
1997	Rituxan/ MabThera (rituximab)	CD20	Chimeric antibody	IgG1, kappa	8	For treatment of CD20-positive, B-cell non-Hodgkin's lymphoma (NHL)	Genentech, Roche, and Biogen Idec	(162, 163, 167)
	Zenapax (daclizumab)	CD25	Humanized antibody	IgG1, kappa	0.3	For prophylaxis of acute organ rejection in renal transplant patients.	Hoffmann-La Roche	(162, 163, 167)
1998	Simulect (basiliximab)	CD25	Chimeric antibody	IgG1, kappa	0.1	For prophylaxis of acute organ rejection	Novartis	(162, 163, 167)
	Synagis (palivizumab)	RSV gpF	Humanized antibody	IgG1, kappa	0.96	For prevention of serious lower respiratory tract disease caused by respiratory syncytial virus (RSV)	MedImmune	(162, 163, 167)
	Remicade (infliximab)	TNF-alpha	Chimeric antibody	IgG1, kappa	0.1	For treatment of rheumatoid arthritis, Crohn's disease, ankylosing spondylitis, psoriatic arthritis, and ulcerative colitis.	Johnson & Johnson	(162, 163, 167)
	Herceptin (trastuzumab)	HER2 protein	Humanized antibody	IgG1, kappa	5	For treatment of metastatic breast cancer	Genentech and Roche	(162, 163, 167)
2000	Mylotarg (gemtuzumab ozogamicin)	CD33	Humanized antibody-drug (cytotoxic antitumor antibiotic calicheamicin) conjugate	IgG4, Kappa	0.08	Treatment of CD33 positive acute myeloid leukemia (AML)	Wyeth Pharmaceuticals	(162, 163, 167)
2001	Campath (alemtuzumab)	CD52	Humanized antibody	IgG1, kappa	10 ~ 32	Injectable treatment of B-cell chronic lymphocytic leukemia	Berlex Laboratories	(162, 163, 167)

2002	Zevalin (ibritumomab tiuxetan)	CD20	Radiolabeled (Yttrium 90) murine antibody	IgG1, kappa	14 ~ 18	Treatment of non-Hodgkin's lymphoma	IDEC Pharmaceuticals	(162, 163, 167)
	Humira (adalimumab)	TNF-alpha	Human antibody	IgG1, kappa	0.1	For treatment of adults with rheumatoid arthritis and psoriatic arthritis.	Abbott Laboratories	(162, 163, 167)
2003	Xolair (omalizumab)	IgE	Humanized antibody	IgG1, kappa	0.17	For treatment of adults and adolescents with moderate to severe persistent asthma.	Genentech, Novartis, Tanox and Roche	(162, 163, 167)
	Raptiva (efalizumab)	CD11a	Humanized antibody	IgG1, kappa	3	For treatment of adults with chronic moderate to severe plaque psoriasis	Genentech and Roche	(162, 163, 167)
	Bexxar (Tositumomab and Iodine I 131Tositumomab)	CD20	Murine antibody	IgG2a, lambda	1.4	Treatment of patients with CD20 positive, follicular, non-Hodgkin's lymphoma following chemotherapy relapse	Corixa	(162, 163, 167)
2004	Avastin (bevacizumab)	VEGF	Humanized antibody	IgG1	1.1	Treatment of metastatic carcinoma of the colon or rectum	Genentech	(162, 163, 167)
	Erbitux (cetuximab)	EGFR	Chimeric antibody	IgG1, kappa	0.2	Treatment of EGFR-expressing metastatic colorectal cancer	Imclone, Bristol -Myers Squibb	(162, 163, 167)
2006	Vectibix (panitumumab)	EGFR	Human antibody	IgG2, kappa	0.05	Treatment of colorectal cancer	Amgen	(163, 182, 183)
	Herceptin* (trastuzumab)	ERBB2	Humanized antibody	IgG1	0.1	A second- or third-line therapy for patients with metastatic breast cancer	Genentech	(163, 184, 185)
	Lucentis (ranibizumab)	VEGF	Humanized antibody fragment	IgG1, kappa		treat the "wet" type of age-related macular degeneration (ARMD), a common form of age-related vision loss	Genentech	(163, 186)

*First approved October 1998, used extended 2006

Much effort has been spent on the recognition of antibody-antigen interaction in structure level (187-193), whereas little research has been conducted on the sequence level to study the interaction between antibody and antigen. However, the availability of structure information of antigen and antibody is much less than

that of sequence information. 42,627 protein structures information exists in Protein Data Bank (PDB) (accessed at 03-Apr-2007) (194). This number is less than 1% of the proteins with sequence information from SwissProt (4,495,647 protein sequences, Release 35.2, 03-Apr-2007) (195). Therefore the antibody rational design may benefit from the huge number of sequence information and the major advances in informatics technology (196). Publicly accessible resources, includes the rapidly increasing number of bioinformatics databases especially immunoinformatics database and their strategies, should be useful for antibody design.

The rapid development of computational tools has also offered a new solution to speed up the antibody design. Since both antibody and antigen are special classes of proteins, the strategies for studying protein-protein interaction may be applied in antibody-antigen interaction to find the mechanism of antibody-antigen interaction and facilitate antibody design.

1.3.3 The need for development of antibody-antigen interaction databases

A number of antibody and/or antigen databases had been developed for providing information about various aspects of antibodies and antigens (Table 1-4). Kabat database (197) is the oldest antibody database started in 1970 (198) and now a comprehensive immunoinformatics database, comprising of nucleotide sequences, sequences of antibody, T cell receptors for antigens (TCR) and major histocompatibility complex (MHC) molecules. VIR II provides an interface of Kabat database with the antibody sequences (107). The ImMunoGeneTics (IMGT) database provides annotated listings and alignments of immunoglobulins, TCR,

MHC of all vertebrate species (199-201). FIMM database contains protein antigens, MHC, T- and B-cell epitopes and relevant disease associations (202). Molecular Modeling Database (MMDB) (203) contains the crystal structure of antibody and HLA obtained from the PDB (194). JenPep is a database of quantitative binding data for immunological protein-peptide interactions (204). IEDB (205) contains data related to antibody, T cell epitopes, MHC binding data for human and some animal species. HaptenDB (206) provides comprehensive information about the Hapten molecules and ways to raise corresponding antibodies. Although these databases provide valuable information about the antibodies and antigens, such as sequences (IMGT, KABAT, FIMM, BCIPEP), structures (IMGT, IEDB, MMDB, SACS), epitope information (IEDB, FIMM, Epitome, CED), binding information (IEDB, JenPep, AntiJen, Epitome) and disease implication (IMGT, FIMM). It tends to be difficult to extract the information of targeted diseases, the therapeutic indications and sequence-level recognition data (i.e. which antibody sequence recognizes which antigen sequence or sequences) from these databases. Although other database such as the epitome database (207) contains sequence-specific information about antibody and antigen interactions, it only covers a limited number of Ab-Ag pairs obtained from protein Databank (194). As a result, there is a need to develop a database capable of providing both easily accessible information and more comprehensive coverage of sequence-specific Ab-Ag recognition to complement existing databases.

Table 1-4 Public antibody and antigen databases.

Database	Description	URL	Reference
KABAT	The oldest antibody database, and now a comprehensive immunoinformatics database including nucleotide sequences, sequences of antibody, TCR and MHC molecules		(197, 198)
SYFPEITHI	Sequences of T-cell epitopes and MHC ligands	http://www.syfpeithi.de/	(208)
VIR	Public 3D-structure of known antibodies and an interface with the antibody sequences in the Kabat database	http://www.ibt.unam.mx/vir/structure/structures.html	(107)
the international ImMunoGeneTics information system (IMGT)	A high-quality integrated knowledge resource specialized in the immunoglobulins (IG), T cell receptors (TCR), major histocompatibility complex (MHC), immunoglobulin superfamily (IgSF) and MhcSF. It consists of sequence databases (IMGT/LIGM-DB, IMGT/MHC-DB, IMGT/PRIMER-DB), genome database (IMGT/GENE-DB) and structure database (IMGT/3Dstructure-DB)	http://imgt.cines.fr/	(199-201)
Functional molecular immunology database (FIMM)	MHC, antigens, T- and B-cell epitopes	http://research.i2r.a-star.edu.sg/fimm/	(202)
Summary of Antibody Crystal Structures (SACS)	All antibody structures in the Protein Databank	http://www.bioinf.org.uk/abs/sacs/	(209)
Molecular Modeling Database (MMDB)	3D macromolecular structures, including antibody, HLA, TCRs	http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml	(203)
JenPep	Quantitative binding data for immunological protein-peptide interactions	http://www.jenner.ac.uk/JenPep	(204)
HLA Sequence Database	HLA sequence	http://www.anthonynolan.org.uk/HIG/index.html	(210)
MHCBN	MHC binding, non-binding peptides and T-cell epitopes	http://www.imtech.res.in/raghava/mhcbn/	(211)
dbMHC	DNA and clinical data related to HLA	http://www.ncbi.nlm.nih.gov/mhc/MHC.cgi?cmd=init	(212)
Blood Group Antigen Gene Mutation Database (BGMUT)	Variations in genes that directly or indirectly affect our blood groups. Allelic genes of blood group antigens	http://www.ncbi.nlm.nih.gov/gv/mhc/xslcgi.fcgi?cmd=bgmut/home	(213)
BCIPEP	B cell epitope database	http://www.imtech.res.in/raghava/bcipep	(214)
VBASE2	Germ-line sequences of human and mouse immunoglobulin variable (V) genes	http://vbase.mrc-cpe.cam.ac.uk	(215)

The Immune Epitope Database and Analysis Resource (IEDB)	Data related to antibody and T cell epitopes, MHC binding data for human and some animal species	www.immuneepitope.org	(205)
TumorAntigen database	A listing of human tumor antigens recognized by T cells	http://sdmc.i2r.a-star.edu.sg/Templar/DB/cancer_antigen/	(216)
HIV molecular immunology database	A collection of HIV-1 cytotoxic and helper T-cell epitopes and antibody binding sites	http://www.hiv.lanl.gov/content/immunology/index.html	(217)
HCV Immunology database	A collection of HCV cytotoxic and helper T-cell epitopes and antibody binding sites	http://hcv.lanl.gov/content/immuno/immuno-main.html	(218)
EPIMHC	A relational database of MHC-binding peptides and T cell epitopes that are observed in real proteins	http://bio.dfci.harvard.edu/epimhc/	(219)
AntiJen	Quantitative binding data for B cell epitopes	http://www.jenner.ac.uk/AntiJen/	(220)
HaptenDB	A listing of haptens, structural similarity searches, and antibody and biological information	http://bioinformatics.uams.edu/raghava/haptendb/	(206)
epitome database	All known antigenic residues and corresponding antibodies from PDB entries	http://www.rostlab.org/services/epitome/	(207)
CED	A conformational epitope database	http://web.kuicr.kyoto-u.ac.jp/~ced/	(221)

1.3.4 Current progress in antibody-antigen interaction prediction

A number of studies have demonstrated the capability of combining experimental results with computational methods in understanding antibody–antigen interactions. Padlan et al (1993) and Garman et al (2000) (187, 188) employed structure model to study IgE/receptor interactions. Friedman et al (1994) (190) utilized Metropolis Monte Carlo algorithm to dock the antigen to antibody, showing distinct roles for the "lock-and-key" (recognition) and the "handshake" (binding) paradigms in antibody-antigen interaction. Irnaten et al (1998) (191) developed a molecule modeling method for predicting epitopes. Tenette-Souaille et al (2000) (222) modeled antibody-antigen complex structure in the absence of X-ray crystallographic information of antibody, which was agreeable with the experimental data. Choulier et al (2002) (192) applied QSAR

model to the prediction of Ab-Ag interaction kinetics as measured by BIACORE. Chen et al (2007) (223) applied SVM to identify linear B-cell epitopes using amino acid pair antigenicity scale.

Since antibody-antigen interactions are representative of a broad class of protein-protein interactions involving both specificity and affinity, the method used in protein-protein interaction prediction could be extended to antibody-antigen interaction. SVM has been recently been used for prediction protein-protein interaction from sequence-based properties (224-227). The sequence information of Ab-Ag interaction can also be used to develop SVM models for prediction antibody sequence from antigen sequence in a manner similar to the prediction of protein-protein interaction from individual sequences. This thesis will use SVM to develop a prototype model for predicting antibody from the sequence-recognition information of Ab-Ag.

1.4 Scope and research objective

The purpose of this thesis is to design bioinformatics tools for disease marker discovery from microarray data and to perform therapeutic molecule prediction for facilitating drug discovery.

A disease marker discovery system is developed by using gene selection strategies from microarray data. This system aims to provide gene signatures which should produce good prediction performance for disease differentiation, and show a certain level of stability with the variation of sampling method. The strategies

include the incorporation of multiple random sampling method and the evaluation of gene-consistency into RFE gene selection procedure. The stable gene signatures may help us understand the mechanism of disease initiation and process, and may provide an insight for diagnosing disease, predicting disease types, prognosis of the outcome of a specific therapeutic strategy, and drug resistance before drug treatment. In this thesis colon cancer gene selection and lung adenocarcinoma survival marker selection are used as case studies to evaluate the performance of the system. The stable gene signatures provide the biologists an opportunity to further investigate the role of derived genes in the initiation and progress of a disease, and give suggestions about potential disease targets for therapeutic molecule design.

This thesis also develops a bioinformatics tool for predicting therapeutic molecules which act on disease targets. Antibody, as a well-established therapeutic molecule, is selected in this study. All therapeutic applications of antibodies are based on their ability to recognize specific target molecules – antigens. Currently much effort has been put to understand the antibody-antigen interaction in order to generation and optimization of antibodies to improve their potential in the prevention and treatment of disease. The rapid advances in informatics technology and computational technology may be helpful for us to understand the antibody-antigen interaction. Therefore, an antibody-antigen sequence-recognition resource will be described and a prediction system of disease gene targeting antibody from antibody-antigen sequence-recognition information will be developed. The prediction system could be helpful for therapeutic antibody design.

This thesis is divided into six chapters. Chapter one provides the background and current progress for disease marker, disease targeting molecules and their discovery method. Chapter two describes the methodology of this study. Support vector machine is described at the beginning of Chapter two since this method is used in both disease marker discovery and disease targeting molecule design. The strategies for disease marker discovery and disease targeting molecule design are described in the following sections of this Chapter. The application of the disease marker discovery is described in Chapter three and Chapter four. Chapter five gives a case study of therapeutic molecule discovery. Chapter six presents conclusion and future work.

2 Methodology

This chapter introduces the methodologies for (1) gene selection from microarray data and (2) bioinformatics tool development for therapeutic molecular prediction. Firstly a classification method –support vector machines, which is used in both gene selection and therapeutic molecule prediction, is described (Section 2.1). The following two sections present other strategies used for marker selection from microarray data (Section 2.2) and bioinformatics tool development for therapeutic antibodies (Section 2.3).

2.1 Support Vector Machines

2.1.1 Theory and algorithm

Support vector machines (SVM) is a relatively new machine learning method proposed by Vapnik (114, 228, 229). It defines a mapping, or a decision function, from feature vector space to the class label space. Over the past decades, SVM has become a popular supervised learning method in variety applications including image classification and object detection (230, 231), text categorization (232), prediction of protein solvent accessibility (233), microarray data analysis (120, 121, 138, 149), protein fold recognition (234), protein secondary structure prediction (235), prediction of protein-protein interaction (224) and protein functional class classification (236).

SVM can be divided into linear and non-linear SVM. Linear SVM directly constructs a hyperplane in the feature space to separare positive examples from

negative examples. On the other hand, non-linear SVM projects both positive and negative examples into a higher-dimensional feature space and then separates them in that space.

Linear SVM is the simpleast form of SVM, in which the data represented as a *p*-dimensional vector (a list of *p* numbers) can be separated by a *p-1* dimensional hyperplane. On each side of this *p-1* hyperplane, two parallel hyperplanes can be constructed (Figure 2-1). The separating hyperplane is the one that maximizes the distance between these two parallel hyperplanes. Many linear hyperplanes (also called classifiers) can separate the data. However, only one achieves maximum separation. Under the assumption that the larger the margin or distance between these two parallel hyperplanes the better the generalization error of the classifier will be (237), the maximum separating hyperplane (also known as maximum-margin hyperplane) is clearly of interest (Figure 2-1).

Mathematically, supposed the training set is composed of *n* examples with two classes, it could represent as

$$\chi = \{(x_1, y_1), ((x_2, y_2)..., (x_n, y_n)\}, \quad i=1, 2, \ldots, n, \qquad (1)$$

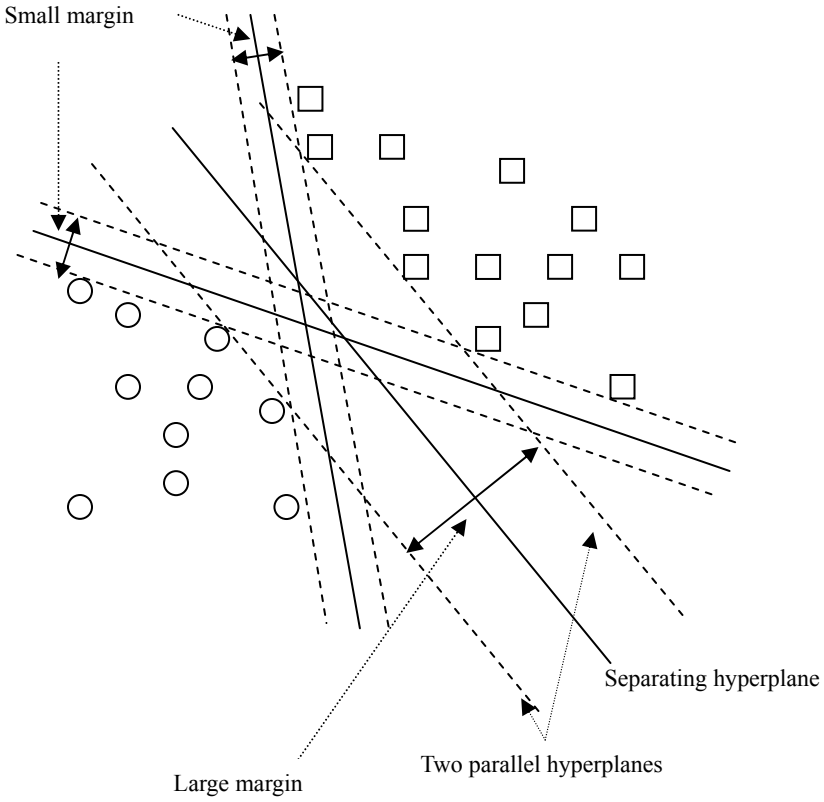
where $x_i \in R^N$ is an N-dimensional real vector and $y_i \in \{-1,+1\}$ indicates class label.

The separating hyperplane can be described by equation:

$$w \bullet x + b = 0 \qquad (2)$$

where $w=(w_1,w_2,\ldots, w_n)^T$ is a unit vector of *n* elements and *b* is a constant, and the relative two parallel hyperplanes can be described by equations

Figure 2-1 Margins and hyperplanes



$$w \bullet x + b = +1 \quad \text{for} \quad y_i = +1 \tag{3}$$

$$w \bullet x + b = -1 \quad \text{for} \quad y_i = -1 \tag{4}$$

If the training data are linearly separable, we can select those two parallel hyperplanes with no data point between them and try to maximize their distance. By using geometry, we find the distance between the two parallel hyperplanes is $2/|w|$. Therefore, to obtain the solution of SVM, $|w|$ should be minimized.

To exclude data points between the two parallel hyperplanes, we need to ensure that for all i either

$$w \bullet x + b \geq +1 \quad \text{for} \quad y_i = +1 \quad \text{or} \tag{5}$$

$$w \bullet x + b \leq -1 \quad \text{for} \quad y_i = -1 \tag{6}$$

It can be rewritten as

$$y_i(w \bullet x_i + b) \geq 1, \quad 1 \leq i \leq n \tag{7}$$

The problem now is to minimize |*w*| subject to the above constraint. More clearly,

$$\text{Minimize } \frac{1}{2}\|w\|^2 \tag{8}$$

Subject to $y_i(w \bullet x_i + b) \geq 1, \quad 1 \leq i \leq n$

This is a quadratic programming (QP) optimization problem.

Such optimization problem could be efficiently solved with the introduction of lagrangian multiplier a_i,

$$L_p(w,b,a) = \frac{1}{2}\|w\|^2 - \sum \alpha_i (y_i \bullet ((x_i \bullet w) + b) - 1) \tag{9}$$

where $\alpha_i \geq 0$.

The solution to this QP optimization problem requires that the gradient of *L(w,b,* α *)* with respect to *w* and *b* vanishes,

$$\frac{\partial}{\partial w} L_p(w,b,a) = 0 \text{ and} \tag{10}$$

$$\frac{\partial}{\partial b} L_p(w,b,a) = 0 \tag{11}$$

resulting in the following conditions:

$$w = \sum_{i=1}^{n} a_i y_i x_i \text{ and} \tag{12}$$

$$\sum_{i=1}^{n} a_i y_i = 0 \tag{13}$$

By substituting Equations (12) and (13) into Equation (9), the QP problem becomes the maximization of the following expression:

$$L_p(w,b,a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i a_j y_i y_j (x_i \bullet x_j) \qquad (14)$$

under the constraints $\sum_{i=1}^{n} a_i y_i = 0$, $0 \leqslant a_i \leqslant C$, i=1, 2, …, n. C is a penalty for training errors for soft-margin SVM and is eual to infinity for hard-margin SVM.

The points located on the two optimal margins will have nonzero coefficients a_i among the solutions to Equation (14), and are called *Support Vectors* (SV). The bias b_0 can be calculated as follows:

$$b_0 = -\frac{1}{2}\left\{ \min_{\{x_i | y_i = +1\}} (w_0 . x_i) + \max_{\{x_i | y_i = -1\}} (w_0 . x_i) \right\} \qquad (15)$$

After determination of support vectors and bias, the decision function that separates two classes can be written as:

$$f(x) = sign\left[\sum_{i=1}^{n} a_i y_i \mathbf{x_i} . \mathbf{x} + b_0 \right] = sign\left[\sum_{SV} a_i y_i \mathbf{x_i} . \mathbf{x} + b_0 \right] \qquad (16)$$

When the examples are inseparable by linear SVM, nonlinear SVM can be applied, which projects the input data to a higher dimensional feature space by using a kernel function *K(x,y)*. The linear SVM procedure is then applied to the featrue vectors in this feature space. After the determination of *w* and *b*, a given vector *x* can be classified by using

$$f(x) = sign\left[\sum_{SV} a_i y_i K(x . x_i) + b_0 \right] \qquad (17)$$

A positive or negative value indicates that the vector x belongs to the posive or negative class respectively.

In equation (17), kernel function *K(x,y)* represents a legitimate inner product in the

input space:

$$K(x, y) = \phi(x) \cdot \phi(y) \quad (18)$$

A number of kernel functions have be used in SVM. Examples of the most popular ones are:

Polynomial kernel $K(x_i, x_j) = (x_i \cdot x_j + 1)^p$ (19)

Sigmoid kernel $K(x_i, x_j) = \tanh(\kappa x_i x_j + c)$ (20)

Radial basis function (RBF) $K(x_i, x_j) = e^{-\|x_j - x_i\|^2 / 2\sigma^2}$ (21)

In practice, RBF kernel is the most widely used kernel function due to three reasons. First, linear kernel and sigmoid kernel can be treated as special cases of RBF kernel since RBF kernel in certain parameters has the same performance as the linear kernel (239) or sigmoid kernel (240). Second, comparing with polynomial kernel, RBF kernel has few parameters which influence the complexity of model selection. Third, RBF function has less computational cost compared with polynomial kernels in which kernel values may go to infinity or zero while the degree is large. Based on these reasons, this thesis mainly used RBF kernel.

Several specialized algorithms can be used to solve the QP problem of SVM by heuristically breaking the problem down into smaller, more-manageable chunks. Table 2-1 listed some popular SVM software tools. In our case, we modified the source code of libSVM to fit our program requirements. libSVM is a sequential minimal optimization (SMO) algorithm(238), which breaks the problem down into 2-dimensional sub-problems that may be solved analytically, eliminating the

need for a numerical optimization algorithm such as conjugate gradient methods.

Figure 2-2 Architecture of support vector machines

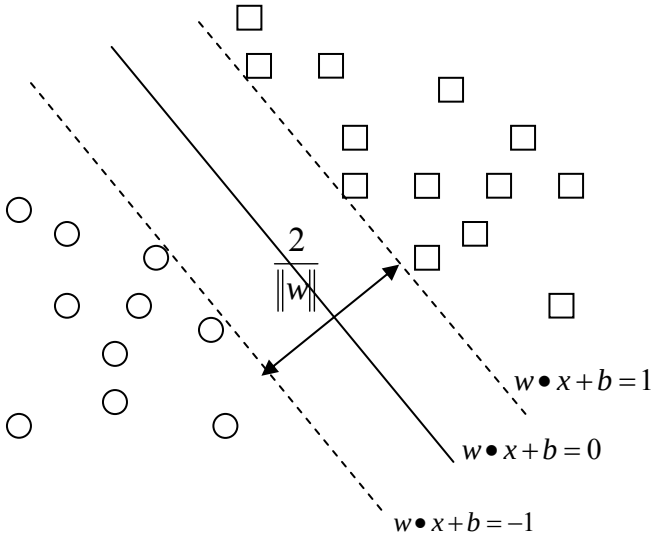


Table 2-1 List of some popular used support vector machines softwares

Software	URL
SVM-Light	http://svmlight.joachims.org/
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/
WinSVM	http://www.cs.ucl.ac.uk/staff/M.Sewell/winsvm/
LS-SVMlab	http://www.esat.kuleuven.ac.be/sista/lssvmlab/
GIST SVM Server	http://svm.sdsc.edu/svm-intro.html

2.1.2 Performance evaluation

The performance of SVM can be measured as true positive TP (the number of positive examples which are correctly predicted as positive), false negative FN (the number of positive examples which are incorrectly predicted as negative), true negative TN (the number of negative examples which are correctly predicted as negative) and false positive FP (the number of negative examples which are incorrectly predicted as positive) (Table 2-2).

The simplest way to evaluate the performance of a classification is overall accuracy (Q), which measures the proportion of the total number of the correctly predicted examples.

$$Q = \frac{TP + TN}{TP + FN + TN + FP} \tag{22}$$

Another two concepts, sensitivity (SE) and specificity (SP), which measure the positive and negative prediction performance respectively, are also frequently used in classification.

$$SE = \frac{TP}{TP + FN} \tag{23}$$

$$SP = \frac{TN}{TN + FP} \tag{24}$$

In some cases such as epidemiology and the evaluation of diagnostic tests (241), positive predictive value (PPV, also called precision rate) and negative predictive value (NPV) are very important.

$$PPV = \frac{TP}{TP + FP} \tag{25}$$

$$NPV = \frac{TN}{TN + FN} \tag{26}$$

Table 2-2 Relationships among terms of performance evaluation.

		Condition		
		True	false	
Test outcome	Positive	True positive (TP)	False positive (FP)	→Positive predictive value (PPV)
	Negative	False negative (FN)	True negative (TN)	→Negative predictive value (NPV)
		↓ Sensitivity (SE)	↓ Specificity (SP)	

2.2 Methodology for gene selection from microarray data

2.2.1 Preprocessing of microarray data

2.2.1.1 Missing data estimation

Missing values is a common issue existing in microarray data. The missing values arise from experimental errors due to spotting problems (cDNA), dust, poor hybridization, inadequate resolution, fabrication errors (e.g. scratch) and image corruption (242, 243). They could also come from the suspicious data with low expression (e.g. background is stronger than signal) or censored data (26). Repeating experiments could be a solution but often not be a realistic option because of economic reasons or limitations in biological material (121, 244). However, many microarray data analysis methods, such as classification, clustering and gene selection methods, require complete data matrix. Therefore in many microarray projects, one needs to determine how to estimate missing values. Proper missing value estimation can significantly improve performance of the analysis methods (245-247). The simplest way is to remove all genes and arrays with missing values, or to replace missing values with an arbitrary constant (usually zero), row (gene) average or column (array) average. The better approaches had also been proposed such as k-nearest neighbors method (KNN) (247), least square methods (LS) (244, 248), and principal component analysis (PCA) (249, 250). Among these estimation methods, KNN is the most widely used and is also a standard method for missing value estimation currently (85, 245, 247).

The KNN-based method for missing value estimation involves selecting *k*

neighbor genes with similar expression profiles to the target gene (the gene with missing values in one or more arrays), and estimating the missing value of the target gene in specific array as the weighted mean of the expression levels of the *k* neighbor genes in this array. A popular KNN-based method is KNNimpute (247), which is the only imputation method available in many microarray data analysis tools for missing value estimation (251-253). KNNimpute can be downloaded from Stanford Microarray Database (85, 254). In this thesis, KNNimpute is employed if the microarray data contains missing values.

2.2.1.2 Normalization

The purpose of normalization is to remove systematic variations from the expression values, so that biological difference can be easily distinguished and the comparison of expression levels across samples can be performed. In microarray experiments, all the values are fluorescent intensities, which are directly comparable. Therefore the normalization among genes and arrays (120) are both possible.

The popular normalization methods for microarray experiments include global normalization using all genes on the array, and housekeeping genes normalization using constantly expressed housekeeping/invariant genes (80). Since Housekeeping genes are not as constantly expressed as assumed previously (255), using housekeeping genes normalization might introduce extra potential sources of error. It was further approved that normalization using a reduced subset of genes was less statistically robust than the normalization using the entire gene set (256). Currently, a typical normalization procedure is (1) normalizing the

expression levels of each sample to zero-mean and unit variance, and then (2) normalizing the expression levels of each gene to zero-mean and unit variance over all the samples. This normalization method have been shown to perform well (257, 258) and is applied in this thesis.

2.2.2 Gene selection procedure

2.2.2.1 Overview of the gene selection procedure

A novel gene selection procedure method based on Support Vector Machines classifier, recursive feature elimination, multiple random sampling strategies and multi-step evaluation of gene-ranking consisitency was established (Figure 2-3):

(1) After preprocessing the original data, by using random sampling method, a large number of training-test sample combinations are generated from the original data set.

(2) The large number of sample combinations is divided into *n* groups, and each group contains 500 sample combinations.

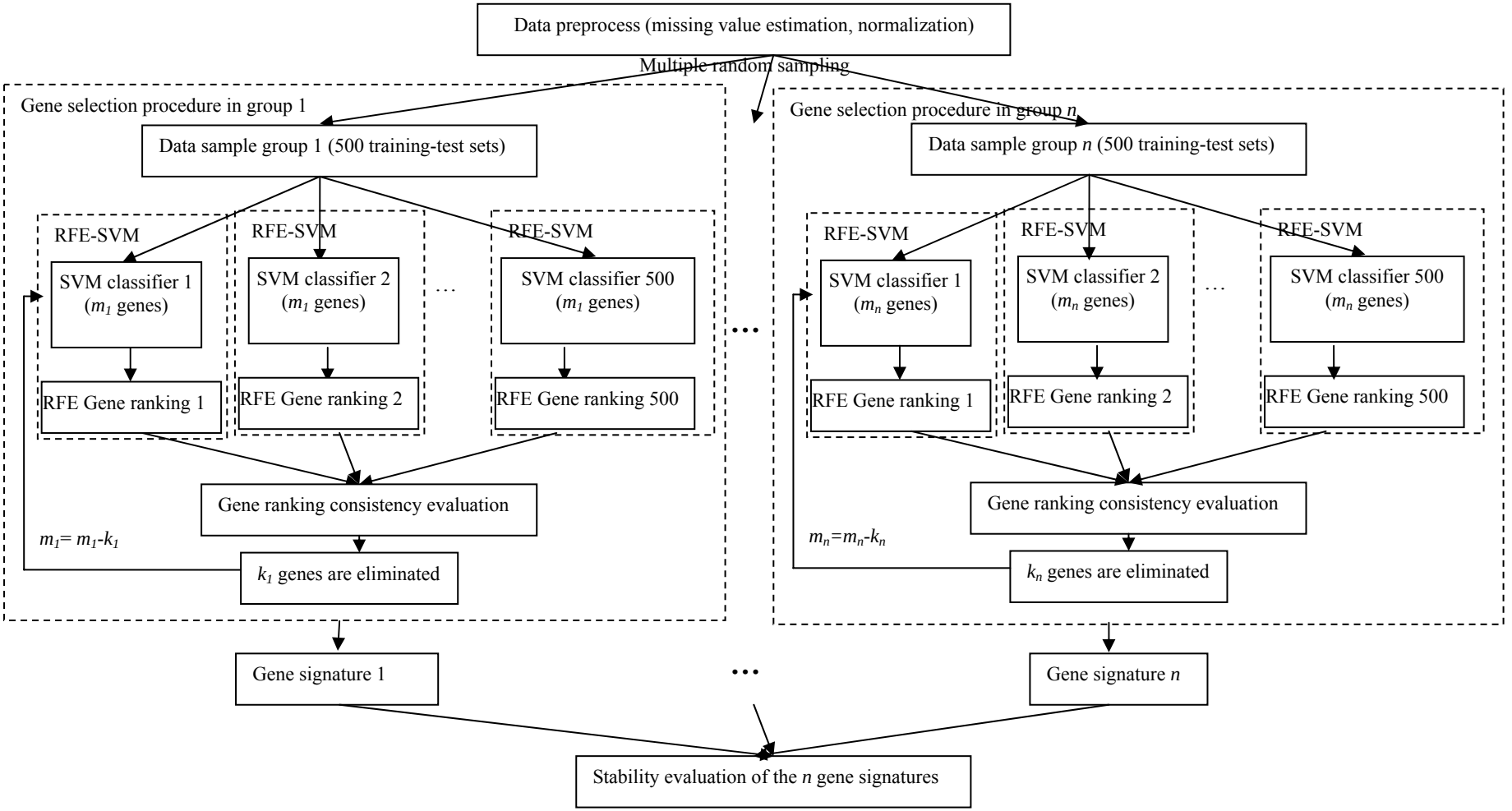
(3) In each training-test sample combination of each group, SVM and RFE are used to classify the samples (SVM classifiers) and rank the genes (RFE gene rank criteria). Therefore 500 gene ranking sequences are formed.

(4) The consistency evaluation can be performed based on the 500 sequences and a certain number of genes (for example, *k* genes) can be eliminated.

(5) Step (3) and (4) can be iteratively done until no gene can be eliminated.

(6) The gene subset which gives us the highest overall accuracies of the 500 test sample sets can be selected as gene signatures of this group. By this way, we can obtain *n* gene signatures.

Figure 2-3 Overview of the gene selection procedure



(7) The stability evaluation of the gene signatures can be performed by looking into the overlap gene rate of the *n* gene signatures.

Below Recursive feature elimination is introduced first and followed by a detailed introduction of the whole feature selection procedure.

2.2.2.2 Recursive feature elimination

During gene selection procedure, the genes ranked according to their contribution to the SVM classifiers. The contributions of genes are calculated by Recursive feature elimination (RFE) procedure, which sort genes according to a gene-ranking function generated from SVM classifier. As described in Section 2.1, SVM training process needs to find the solution for the optimum problem (also known as objective function or cost function) shown in equation (14), which can be rewritten as

$$J = \frac{1}{2}\alpha^T H\alpha - \alpha^T 1 \tag{27}$$

Under the constraints $\sum_{i=1}^{n} a_i y_i = 0$ and $\alpha_i \geq 0$, i=1,2,…n.

Where $H(i, j) = y_i y_j K(x_i, x_j)$, K is the kernel function.

The gene-ranking function of RFE can be defined as the change in the objective function J upon removing a certain gene. When a given feature is removed or its weight w_k is reduced to zero, the change in the cost function $J(k)$ is

$$DJ(k) = \frac{1}{2}\frac{\partial^2 J}{\partial w_k^2}(Dw_k)^2 \tag{28}$$

where the change in weight $Dw_k = w_k - 0$ corresponds to the removal of feature k.

Under the assumption that the removal of one feature will not significantly influence the values of αs, the change of cost function can be estimated as

$$DJ(k) = \frac{1}{2}\alpha^T H\alpha - \frac{1}{2}\alpha^T H(-k)\alpha \quad (29)$$

Where H is the matrix with elements $y_i y_j K(x_i, x_j)$, and H *(-k)* is the matrix computed by using the same method as that of matrix H but with its kth component removed.

The change in the cost function indicates the contribution of the feature to the decision function, and serves as an indicator of gene ranking position (259).

2.2.2.3 Sampling, feature elimination and consistency evaluation

In order to present statistical meaning, gene selection is conducted based on multiple random sampling. Each random sampling divide all microarray samples into a training set which contains half number of samples and an associates test set which contains another half number of samples. This sampling method can be treated as a special case of the bootstrap technique. Many researchers showed that bootstrap-related techniques present more accurate estimation than cross-validation on small sample sets (260, 261). By using this random sampling, thousands of training-test sets, each containing a unique combination of samples, are generated. These thousands of randomly generated training-test sets are randomly divided into several sampling groups, with equal number of training-test sets (such as 500 traing-test sets) in each group. Every sampling group is then

used to derive a signature by RFE-SVM.

In every training-test sampling group generated by multiple random sampling, each training-set (totally 500 training-test sets) is used to train a SVM class-differentiation system and the genes are ranked by using Recursive feature elimination (RFE), according to the contribution of genes to the SVM classifier. In order to derive a gene-ranking criterion consistent for all iterations and all the 500 training-test sets in this group, a SVM class-differentiation system with a universal set of globally optimized parameters, which give the best average class-differentiation accuracy over all of the 500 test sets in this group, is applied by RFE gene-ranking function at every iteration step and for every training-test set.

To further reduce the chance of erroneous elimination of predictor-genes, additional gene-ranking consistency evaluation steps are implemented on top of the normal RFE procedures in each group:

(1) For every training-set, subsets of genes ranked in the bottom (which give least contribution to the SVM classification procedure) with combined score lower than the first few top-ranked genes (which give highest contribution to the SVM classification procedure) are selected such that collective contribution of these genes less likely outweigh top-ranked ones.

(2) For every training-set, the step (1) selected genes are further evaluated to choose those not ranked in the upper 50% in previous iteration so as to ensure that these genes are consistently ranked lower.

(3) A consensus scoring scheme is applied to step (2) selected genes such that only

those appearing in most of the 500 testing-sets were eliminated.

For each sampling group, different SVM parameters are scanned, various RFE iteration steps are evaluated to identify the globally optimal SVM parameters and RFE iteration steps that give the highest average class-differentiation accuracy for the 500 testing-sets.

The several signatures derived from these sampling-groups are then applied to evaluate the stability and performance.

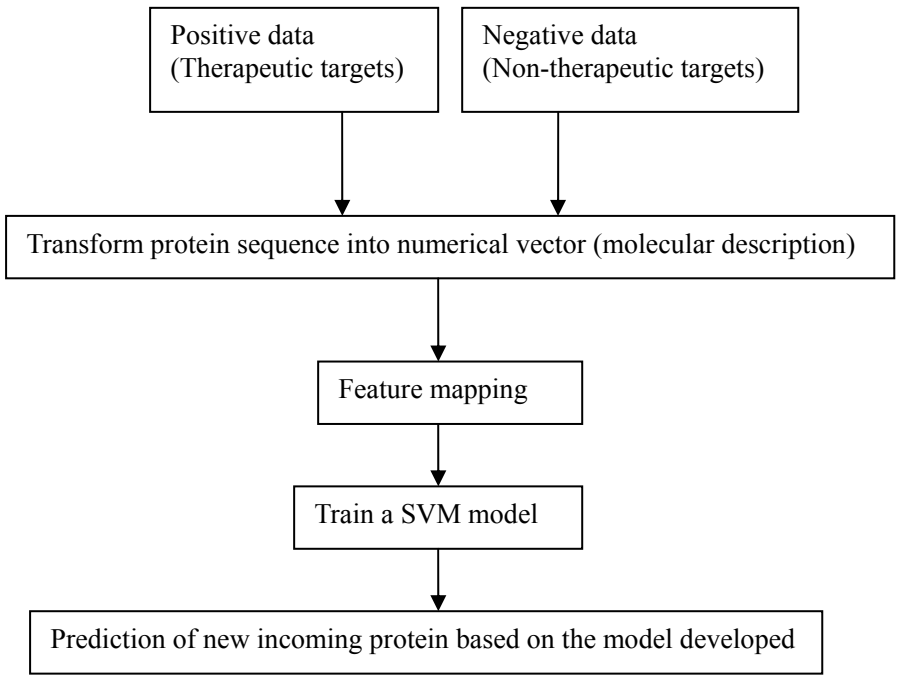
2.2.3 The development of therapeutic target prediction system

To evaluate the possibility of the identified markers as therapeutic targets, a therapeutic target prediction system is applied. Target identification is an important first step in target-based drug discovery processes (262, 263). Computational methods have been used for facilitating target identification by predicting therapeutic targets, whose activity can be regulated by drug-like molecules (264) and play key roles in a disease, from genomic, structural and functional information (264-267). Proteins can be divided into therapeutic target and non-therapeutic target classes. Thus machine learning methods (SVM, PNN, kNN, decision tree) can be used for predicting therapeutic target (268, 269). Among these methods, SVM showed the best performance (268). Therefore, this thesis uses SVM classifier to predict the possible therapeutic targets from identified disease markers.

2.2.3.1 Outline of prediction strategy

One strategy for predicting therapeutic targets from their sequences without sequence similarity is to use a sequence-independent classifier generated from the analysis of known druggable targets that share some characteristics but may be significantly different in sequence, structure and function (264, 270). Each therapeutic target or non- therapeutic target is represented by a feature vector, which is composed of sequence-derived descriptors representing its structural and physicochemical properties. SVM classifies these proteins by projecting their feature vectors into a multi-dimensional space in which therapeutic targets and non-therapeutic targets are separated by a hyperplane. A protein can be predicted as therapeutic target or non-therapeutic target depending on whether its feature vector is projected onto the therapeutic target or non- therapeutic target projected at the side of the hyper-plane (Figure 2-4).

Figure 2-4 Architecture of therapeutic target prediction system



2.2.3.2 The selection of therapeutic target and non- therapeutic target

Sufficiently diverse sets of therapeutic targets and non-therapeutic targets are needed for training a SVM model for predicting possible therapeutic targets. 1,484 therapeutic targets in the TTD database (271) with available sequence information forms the therapeutic target class. There are 6,637 protein families in the protein domain family Pfam database (272) found to contain unknown target at present. Therefore, without substantially reducing SVM prediction performance, putative non-therapeutic targets can be tentatively derived from these non-target representing families (270). Their representative proteins form a non-therapeutic target class.

2.2.3.3 Molecular descriptors

In using SVM for predicting therapeutic targets, each protein is represented by a multi-dimensional feature vector whose components are protein descriptors encoding various constitutional and physicochemical properties of that protein (273). Web servers such as PROFEAT (274) and ProtParam (275) have been developed for facilitating the computation of these descriptors from amino acid sequence of proteins. The descriptors used for predicting therapeutic targets (270) include a constitutional descriptor, amino acid composition, and several physicochemical descriptors that describe the composition, transition and distribution of hydrophibicity (h), polarity (p), polarizibility (z), charge (c), secondary structures (s), solvent accessibility (a), surface tension (t), and normalized Van der Waals volumes (v) (236).

Amino acid composition is the fraction of each kind of amino acid in a sequence $f_k = N_k / N$, where k=1, 2, 3, …, 20 is the index of amino acids, N_k is the number of a particular kind of amino acid and N is sequence length. For computing the descriptors of each of the physicochemical properties, amino acids are divided into three types. For instance, for hydrophobicity descriptors, amino acids can be divided into hydrophobic (CVLIMFW), neutral (GASTPHY), and polar (RKEDQN) types. Three descriptors, composition (C_q), transition (T_q), and distribution (D_q), are introduced to describe global composition of each of the physicochemical properties, where $q = h, p, z, c, s, a, t$ and v. $C_q = (\frac{N_{q_1}}{N}, \frac{N_{q_2}}{N}, \frac{N_{q_3}}{N})$ represent the percentage of each type of residues in a sequence, where *Nqi* is the number of type *i* residue. $T_q = (\frac{T_{q_{12}}}{N-1}, \frac{T_{q_{13}}}{N-1}, \frac{T_{q_{23}}}{N-1})$ characterizes the percent frequency of transition between different types of residues, where T_{qij} is the number of type i to j transitions and N-1 is the total number of transitions. These transitions are considered to be undirected such that $T_{q_{ij}} = T_{q_{ji}}$. $D_q = (D_{q_1}, D_{q_2}, D_{q_3})$ with $D_{q_i} = (\frac{P_{q_{i0}}}{N}, \frac{P_{q_{i25}}}{N}, \frac{P_{q_{i50}}}{N}, \frac{P_{q_{i75}}}{N}, \frac{P_{q_{i100}}}{N})$ measures the chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular group is located respectively, where $P_{q_{ik}}$ is the length within which k% of type i residues are located. Overall, each physicochemical property is represented by 21 elements: 3 for C_q, 3 for T_q and 15 for D_q. The complete feature vector is composed of

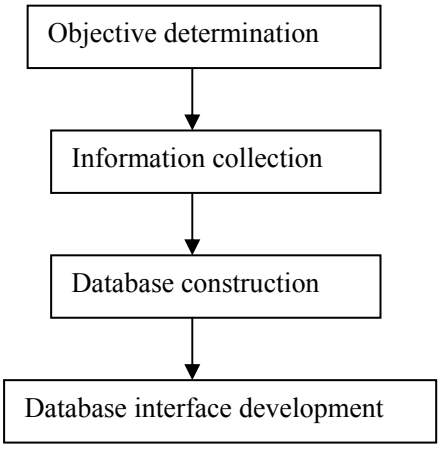
188 elements that include 20 elements for amino acid composition and 8×21 elements of the physicochemical properties. All generated vectors have equal length.

2.3 Methodology for therapeutic molecule prediction

2.3.1 Database development

In order to develop bioinformatics tools for therapeutic molecule design, a bioinformatics database is developed. Database is an organized collection of data and relationships among the data items. Generally database development is a complicated and time-consuming process, including determination of database objective, collection of related information, design of database scheme, development of database, design of database interface (Figure 2-5). A stage-by-stage discussion of general database development is discussed below.

Figure 2-5 Flowchart of database design



2.3.1.1 The objective of database development

Clear objective of the database will help us focus on the relevant information and discard the unnecessary parts. Generally, a successful database should meet the expectations of their corresponding researchers, afford them what they want, and help them further information they want to know more about.

As described in Chapter one, the antibody-antigen interaction plays a very important role in disease treatment, vaccine design and therapeutic antibody design. It is essential for biomedical researcher to know more about the disease-related antibody, and their counterpart – antigen, especially their interaction in molecular level.

2.3.1.2 The collection of related information

Normally, a knowledge-based database is supposed to provide enough domain knowledge around a specific subject. For instance, therapeutic antibody database will let users know about some biological information such as antibody information, the corresponding antigen target information, targeted disease and potential diagnosis, and therapeutic indications of the antibody. Thus, for every database entry, there are several different knowledge domains. Some of them provide basic introduction to entries themselves, and some others give information derived from entries or relevant to entries.

The information mentioned above can be selected from a comprehensive search of available literatures including journal articles and a large number of other publications. With respect to different type of information, we use different

collecting methods. The subject of databases, such as disease target antibody is the primary focus. Thus, in the first step, we collect reliable subject information. At present, no single index or library is available and almost all the relevant information is scattered in various biological and medical literatures. Therefore, literature information extraction is the only feasible way to collect the essential biological and medical information. It is generally agreed that literatures are typically unstructured data source. In addition, the names of the subject, which may be in some synonymous terms, various abbreviations, or totally different expression, are difficult to be recognized by automatic language processing. A fully automated literature information extraction system, thus, cannot be invented to gather useful information from literature efficiently.

In this study, automatic text mining methods with manual reading process was combined. Simple automated text retrieval programs developed in PERL were used to screen the literature that contained the key word related to searching the subject via Medline (276). Since the purpose of the therapeutic antibody database is to provide sequence-specific recognition of antibodies and their recognitions, at the first step, only those literatures that contain both the antibody sequence and the corresponding antigen or antigens were selected and constructed as the basis of the database. The sequence of the corresponding antigen or antigens were obtained either from the respective literature (if it is described in that literature) or retrieved from the protein sequence database Swiss-Prot based on the name and host species of the antigen or antigens. Other useful information was picked up manually from Medline by using keyword searching. If necessary, the full literature was referred to facilitate information searching.

After information collection, a consideration how to store, organize and manage the data by using database techniques should be considered. In the next section, the database construction is described.

2.3.1.3 The construction of database

There are a number of different ways to construct database to store and present data and data relationships. Some of the more common database types include hierarchical database, object database and relational database.

The hierarchical database arranges data in a tree format. This database organizes data into different groups, which in turn may be divided into different subgroups. These subgroups also can have sub-subgroups. This forms a hierarchy of parent and child data segments. This strategy may make it difficult to identify complex multiple relationships between individual data items because there is no obvious link between two data if they are not in the same parent data segment.

Object database stores data in discreet, self-contained units – objects, which have specific data, attributes and behaviors. Object-oriented programming languages are normally used to access the data. This tightly entwined property of database and its application increases the complexity of accessing the data outside of the application, and limited the application of object database.

While hierarchical database arranges data in a tree format, relational database arranges them in a tabular format. A relational database creates formal definitions

of all the included items in a database, setting them out in tables, and defines the relationship among them. Using IDs or keys, the tables can be related between each other. Such database is called 'relational' because they explicitly define these connections. Currently relational database is the most common form of database.

The relational model has been used in our therapeutic antibody databases. It represents relevant data in the form of two-dimension tables. Each table represents relevant information collected. The two-dimensional tables for the relational database include entry ID list table (Table 2-3), main information table (Table 2-4), which contains a record for the basic information of each entry, data type table (Table 2-5), which demonstrates the meaning represented by different number, and reference information table (Table 2-6), which gives the general reference information following by different PubMed ID in Medline (277).

Table 2-3 Entry ID list table

Entry ID	Entry Name
...	…

Table 2-4 Main information table

Entry ID	Data type ID	Data content	Reference ID
…	…	…	…

Table 2-5 Data type table

Data type ID	Data type
...	…

Table 2-6 Reference information table

Reference ID	Reference
...	…

Table 2-7 Logical view of the database

Entry ID	Data type ID	Data content	Reference ID
…	…	…	…

Data type ID	Data type
...	…

Reference ID	Reference
...	…

Figure 2-7 is the general logical view of database we developed. It shows the organization of relevant data into relational tables. In these tables, certain fields may be designated as keys, by which the separated tables can be linked together for facilitating to search specific values of that field. Commonly, in relational table, the key can be divided into two types. One is primary key, which uniquely identifies each record in the table. Here it is a normal attribute that is guaranteed to be unique, such as entry ID in Table 2-3 with no more than one record per entry. The other is foreign key, which is a field in a relational table that matches the primary key column of another table. The foreign key can be used to cross-reference tables. For example, in tables of our databases, there are two foreign keys: Data type ID and Reference ID. According to Figure 2-7, a connection between a pair of tables is established by using a foreign key. The two foreign keys make three tables relevant. Generally, there are three basic types of relationships of related table: one-to-one, one-to-many, and many-to-many. In our case, these databases belong to one-to-many case.

Most relational databases now make use of Structured Query Language (SQL) to handle queries. SQL is widely used by relational databases to define queries and help to generate reports. SQL has become a dominant standard in the world of

database development, since it allows developers to use the same basic constructions to query data from a wide variety of systems.

By using relational database software (e.g. Oracle, Microsoft SQL Server) or even personal database systems (e.g. Access), the relational database can be organized and managed effectively. This kind of data storage and retrieval system is called Database Management System (DBMS). An Oracle 9i DBMS is used to define, create, maintain and provide controlled access to the antibody-antigen interaction databases and the repository. All entry data from the related tables described in previous section are brought together for user display and output using SQL queries.

2.3.1.4 The design of the database interface

The database provides information for users to access. So it is very important to design a friendly interface which mediates the interaction between the database and its users.

Normally two or three layers are used in most bioinformatics database. The main user interface provides querying tools to find specific entries. The entry layer provides detail information of the entries. An optional searching lists the searching results with some specific matching rules.

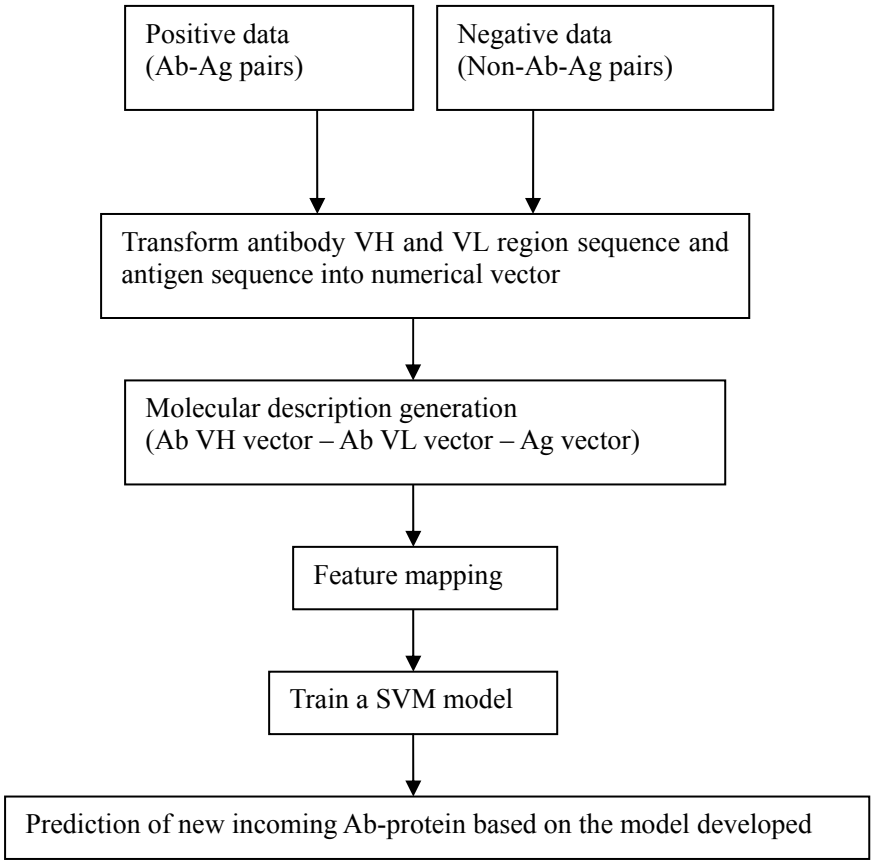
2.3.2 Predictive system development

2.3.2.1 Outline of prediction strategy

Knowledge of sequence-level Ab-Ag recognitions can be used to develop machine learning (ML) models for predicting the sequence of an antigen-targeting antibody based on the sequence of a specific antigen in a manner similar to the prediction of protein-protein interactions (224-226, 278, 279). In this study, each protein-pair is represented by a feature vector composed of sequence-derived descriptors which indicate the structural and physicochemical properties. Popular ML models, such as SVM which was introduced in Section 2.1 in this chapter, are used to predict interacting protein-pairs by projecting their feature vectors on a multi-dimensional space of protein features in which interacting and non-interacting protein-pairs are separated by a hyperplane (280). Known interacting protein-pairs and non-interacting ones are applied to train a SVM model in order to determine the hyper-plane. A protein-pair is predicted to be interacting or non-interacting depending on whether its feature vector is projected on the interacting side of or on the non-interacting side of the hyper-plane. For instance, given the sequence of a protein A, the sequence of its interacting protein B can be determined by searching protein databases or generated sequences in such a way that the pair of protein A and the target sequence (protein B, in this occasion) can be classified into the interacting class.

This approach can be potentially extended to the prediction of antibody sequence from antigen sequence or vice versa. The architecture of antibody prediction system was shown in Figure 2-8.

Figure 2-8 Architecture of disease targeting antibody prediction system



2.3.2.2 Selection of antibody-antigen pairs and non-antibody-antigen pairs

The sequences for both antibody-antigen pairs (positive data) and non-antibody-antigen pairs (negative data) are required for building such model. The diversity of such data is very important for an unbiased model development. Known Ab-Ag pairs in the antibody-antigen information resources can be employed as the training sets to develop machine learning prediction models. The putative non-Ab-Ag pairs can be tentatively generated from random combination of antibodies with antigens of other antibody. Most of these putative non-Ab-Ag pairs are expected to be valid because multi-antigen antibodies or multi-antibody

antigens constitute a small percentage of Ab-Ag pairs. Therefore, the prediction error due to the potential inclusion of "wrong" non-Ab-Ag pairs in SVM training process is expected to be relatively low.

2.3.2.3 Feature vector construction

In using SVM for predicting Ab-Ag pairs, each Ab-Ag or non-Ab-Ag pair is represented by a multi-dimensional feature vector composed of sequence-derived descriptors that encode constitutional and physicochemical properties of antibody and antigen (273). An antibody consists of two chains: heavy chain and light chain, and each chain has variable region and constant region. Consequently, the descriptors of each antibody are derived from the combination of the variable region of the heavy chain sequence and the variable region of the light chain (VH-VL). The descriptors of each Ab-Ag pair are generated by the descriptors of the antibody followed by those of the corresponding antigen. The descriptors of antibody VH and VL region and antigen sequence are generated as the similar manner as the molecular descriptors generated for therapeutic target prediction (Section 2.2.3.3 in this chapter). These descriptors include a constitutional descriptor, amino acid composition, physicochemical descriptors, transition and distribution of hydrophobicity, polarity, polarizability, charge, secondary structures, solvent accessibility, surface tension, and normalized Van der Waals volumes (236).

3 Colon cancer marker selection from microarray data

This chapter describes a disease–related gene selection method from microarray data. Colon cancer-related gene selection are used here as a case study. By using randomly sampling method, we generate 20 sets of colon cancer gene signatures. The predictive ability of the cancer genes shared by all of the 20 sets is evaluated by SVM models on an independent dataset collected from Stanford Microarray Database. Unsupervised hierarchical clustering analysis provides additional indication of the predictive ability of selected signatures. A therapeutic target prediction system is further applied to identify the possible cancer genes and possible therapeutic targets from the selected markers.

3.1 Introduction

Knowing what causes a disease is the first step in understanding the abnormal course of disease and helps the treatment of the disease. It has been found that some diseases such as cancer (2-4) and diabetes (7) are caused by some specific gene defects or mutations. Therefore, disease gene discovery is very important for disease diagnosis and treatment.

The simple and direct way to identify disease genes is through analyzing the change of expression level across a series of samples. There are around 25,000 genes in human genome (1). Therefore, Microarray becomes a very important tool for disease gene discovery because microarray can measure the gene expression profiles of tens of thousands of genes at one time. By discovering the differences

in gene expression between normal and disease tissues, we can focus on the genes with different expressions and those genes that might be activated or inactivated in association with a particular disease.

Since disease is a kind of broad class, to be specific, colon cancer was chosen in this study. Our rational is that the study on colon cancer gene discovery will provide a platform to study other diseases in similar way.

Cancer is the leading cause of death around the world. In United States, each year more than 1 million individuals were diagnosed as some kind of tumor (281, 282). Around 38% of cancers become metastasis and cause the death of the individuals (282). Colon cancer is the third leading cause of death from cancer in the US. It is estimated that colon cancer affects or will affect more than 145,000 individuals in 2007 (282). Up to 30% of these cases exhibit familial clustering, which means that tens of thousands of individuals have a disease with a potentially definable genetic component (283). Identification of the genes that cause the colon cancer is thus very important for colon cancer diagnosis and colon cancer treatment. Currently microarray technique has becoming a powerful tool for colon cancer diagnosis and colon cancer genes identification (76, 108, 124, 127, 137, 138, 147-150, 259). However, comparing colon cancer gene signatures from these groups (119, 124, 127, 147-150), there are little overlap between one and another. From the same 62-sample dataset (108), 10 different colon cancer gene signatures were derived by using different sampling methods (Table 3-1) (119, 124, 127, 147-150). Only 1~5 of the 4~60 selected colon cancer genes in each signatures were present in more than half of the other 9 signatures (Table 3-1), and 2~20 of the colon cancer genes in each signatures were cancer-related (Table 3-2) (119, 124, 127, 147-150).

Table 3-1 Statistics of the colon cancer gene signatures for differentiating colon cancer patients from normal people by 10 different studies that used the same microarray dataset. The dataset is from (108).

Study (Reference)	No. of selected genes in signature	Class differentiation method	Signature selection method	Validation method	Prediction accuracy	No. of genes selected by other N studies									
						9	8	7	6	5	4	3	2	1	0
Zhou and Mao 2005 (147)	15	LS-SVM	LS Bound measure	bootstrap	<85%	0	0	0	1	0	1		1	1	11
Ding and Peng 2005 (127)	60	NB, SVM, LDA, LR	Filter method (MRMR)	LOOCV	93.55%	0	0	0	3	2	1	1	4	4	45
Isabelle Guyon 2002 (119)	7	SVM (linear kernel)	Wrapper method (RFE)	LOOCV	98%	0	0	0	0	0	1	1	2	0	3
Inza, Larranaga et al. 2004 (124)	5	decision tree1	wrapper method	LOOCV	87.1%	0	0	0	3	2	0	0	0	0	0
Inza, Larranaga et al. 2004 (124)	4	decision tree2	wrapper method	LOOCV	88.81%	0	0	0	0	0	1	1	2	0	0
Bo and Jonassen 2002 (148)	50	linear discriminant	gene pair ranking	LOOCV L-31-OCV	87.8% 85.9%	0	0	0	3	2	1	2	7	8	27
Huang and Kecman 2005 (149)	10	SVM1	Wrapper method (RFE)	LOOCV	Not indicated	0	0	0	3	2	0	0	5	0	0
Huang and Kecman 2005 (149)	10	SVM2	Wrapper method (RFE)	LOOCV	88.84%	0	0	0	3	0	0	0	2	2	3
Huang and Kecman 2005 (149)	10	SVM3	Wrapper method (RFE)	LOOCV	88.1%	0	0	0	3	2	0	0	4	0	1
Liu, Krishnan et al. 2005 (150)	6	clustering method	Filter method (mutual information)	LOOCV	91.9%	0	0	0	2	2	0	0	0	2	0
Total Number of uniquely selected genes = 107				Number of unique genes selected by only one study =83											

1. LS Bound measure: a hybrid of filter and wrapper methods
2. SVM: Support Vector Machines
3. LS-SVM: Least Square SVM
4. MRMR: Minimum Redundancy-Maximum Relevance feature selection framework
5. NB: Naïve Bayes classifier
6. LDA: Linear Discriminant analysis
7. LR: logistic Regression
8. LOOCV: Leave-One-Out Cross Validation
9. RFE: recursive feature selection
10. SVM1, SVM2, and SVM3: Support Vector machines classifiers with different parameters
11. Decision tree 1, decision tree 2: decision tree classifiers with different parameters
12. NB, SVM, LDA, LR: the authors tried these four methods, and choose those one which showed the highest accuracy

Table 3-2 Distribution of the selected colon cancer genes of the 10 studies in Table 3-1 with respect to different cancer-related classes

Study (Reference)	Cancer genes					Tumor marker	Interacting partner of cancer gene	Cancer pathway affiliated gene	Gene having possible implication in any cancer
	Anticancer target	Oncogene	Tumor-suppressor genes	Angiogenesis genes	other types				
Zhou and Mao 2005 (LS-SVM) (147)	2	0	0	0	5	0	0	0	0
Ding and Peng 2005 (NB, SVM, LDA, LR) (127)	1	0	1	0	5	1	4	2	1
Isabelle Guyon 2002 (SVM) (259)	1	0	0	0	0	0	0	1	0
Inza, Larranaga et al. 2004(decision tree1) (124)	0	0	0	0	1	0	0	1	0
Inza, Larranaga et al. 2004 (decision tree2) (124)	0	0	0	0	0	0	0	0	0
Bo and Jonassen 2002 (linear discriminant) (148)	2	2	2	0	4	0	7	3	0
Huang and Kecman 2005 (SVM1) (149)	0	0	0	0	2	0	2	1	0
Huang and Kecman 2005 (SVM2) (149)	0	0	2	0	1	0	1	2	0
Huang and Kecman 2005 (SVM3) (149)	0	0	0	0	1	0	2	1	0
Liu, Krishnan et al. 2005 (clustering method) (150)	1	0	0	0	0	0	1	1	0

This discrepancy increases the difficulty of applying those cancer genes in clinics.

In this chapter, we explored a new gene selection method aiming at reducing the chances of erroneous elimination of predictor-genes. We employed the recursive feature selection method based on a model built from support vector machines to identify novel molecular signatures with respect to the interactions among genes. Derived from the consensus scoring of multiple random sampling and the evaluation of gene-ranking consistency embedded in the recursive feature selection system, totally 104 genes were selected after 20 times of experiments. The gene signatures are fairly stable with 80% of top-50 and 69%~93% of all predictor-genes shared by all 20 signatures. These shared predictor-genes include 48 cancer-related and 16 cancer-implicated genes, as well as 50% of the previously-derived predictor-genes. The derived signatures outperform all previously-derived signatures in predicting colon cancer outcomes from an independent dataset collected from the Stanford Microarray Database. The differential expression and function analysis of the identified marker genes implies that the selected genes should play important roles in colon cancer initiation and progress.

3.2 Materials and methods

3.2.1 Colon cancer microarray datasets

Two independent data sets of colon cancer were used for colon cancer gene discovery and for validating the effect of our selected genes. The first dataset was reported in previous publication (108). Another dataset was collected from

Stanford Microarray Database (SMD) (284).

The dataset for colon cancer gene discovery was Alon's dataset (108, 285), which contained the expression levels of 40 colon cancer patients and 22 normal patients. This dataset was obtained by using the Affymetrix Hum6000 array (Affymetrix Inc) (84). This array contained about 65,000 features, each containing around 107 strands of a DNA 25-mer oligonucleotide, and these features represented the sequences from about 3,200 full-length human cDNAs and 3,400 ESTs that have some similarity to other eukaryotic genes (therefore some genes are non-human). 2000 genes with the highest minimal intensity across the samples were pre-selected from the 6500 genes by Alon et al (108, 285).This colon microarray data have been analyzed in several previous studies using a number of statistical approaches (124, 127, 147-150, 259).

In order to evaluate the performance of selected genes, the expression profiles of 34 colon cancer cell lines and 8 normal colon tissues were collected from Stanford Microarray Database (SMD) (284). Appendix S1 shows the detailed information of collected samples.

3.2.2 Colon cancer gene selection procedure

By using repeated random sampling (70), 10,000 training-testing sets were generated, each constituted a training set which contains 31 samples and an associates test set which contains the other 31 samples from Alon's colon cancer dataset (108). These 10,000 randomly generated training-testing sets were randomly placed into 20 sampling groups, and each group contains 500

training-testing sets.

Each of the 20 sampling groups was used to derive a signature. In the 500 training-testing sets in every sampling group, each training-set was used to select genes by RFE based on SVM system. For all iterations and testing-sets, SVM system employed a set of globally modified parameters which gave the best average class-differentiation accuracy over the 500 testing-sets.

On every sampling group, three gene-ranking consistency evaluation steps were implemented on top of the normal RFE procedures in all sampling groups:

(1) For every training-set, subsets of genes ranked in the bottom 10% (if no gene was selected in current iteration, this percentage was gradually increased to the bottom 40%) with combined score lower than the first top-ranked gene were selected such that collective contribution of these genes less likely outweighed higher-ranked ones.

(2) For every training-set, the step (1) selected genes was further evaluated to choose those not ranked in the upper 50% in previous iteration so as to ensure that these genes were consistently ranked lower.

(3) A consensus scoring scheme was applied to step (2) selected genes such that only those appearing in >90% (if no gene was selected in current iteration, this percentage was gradually reduced to 60%) of the 500 training-sets were eliminated.

3.2.3 Performance evaluation of signatures

The predictive capability and robustness of gene signatures was evaluated by

using several microarray data analysis methods on independent SMD microarray datasets (Appendices S1) (284) and the Alon's microarray dataset (108). The microarray data analysis methods included hierarchical clustering and SVM.

By using hierarchical clustering analysis, the performance of gene signatures was analyzed. As a popular unsupervised method, hierarchical clustering analysis groups genes and samples which have similar expression in the microarray data. Typically, the analysis begins with each gene/sample considered as a separate cluster. They are successively merged until one large cluster comprising the whole dataset is achieved. Later, these clusters are displayed in the form of a branching tree diagram, which can be broken into distinct clusters by cutting across the tree at a particular height. Hierarchical cluster analysis was carried out using the selected signatures by the software from Eisen et al (109, 286). The results from hierarchical clustering were displayed by TreeView, which was also provided by Eisen et al (109, 286).

In SVM evaluation system, 500 random-generated training-test set were generated. The performance of the gene signatures was evaluated by overall accuracies Q (Equation 2-22) from the associated 500 test sets of the 500 SVM classification systems.

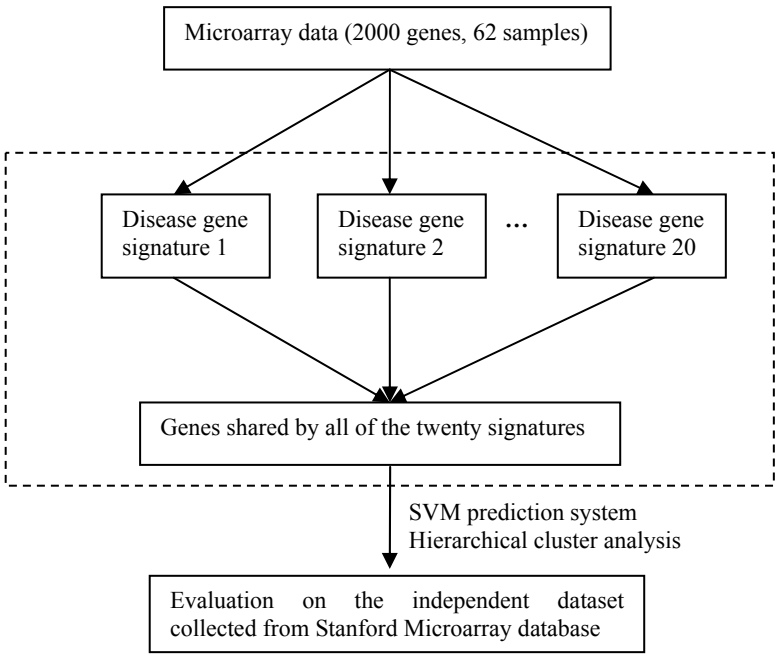
3.3 Results and discussion

3.3.1 System of the disease marker selection

The aim of this study was to identify the important gene signatures with

regard to the intrinsic complex interactions of genes in disease initiation. Moreover, considering the noise in the microarray data arising from measurement variability and biological differences, the selected important gene signatures should be stable with regarding to such kind of variations. Based on the above concerns, recursive feature elimination method based on SVM was used to identify the different signatures from the multiple random combinations of samples. 20 sets of survival marker signatures were obtained by using RFE-SVM from 500 training-testing datasets with random sampling methods. SVM classifiers and hierarchical cluster analysis were used to evaluate the prediction system constructed from selected signatures (Figure 3-1).

Figure 3-1 The system of colon cancer genes derivation and colon cancer differentiation



3.3.2 Consistency analysis of the identified disease markers

The consistency level of the 20 derived signatures was estimated from the percentage of predictor-genes shared by them. 104 genes were shared by all of 20

signatures (Table 3-3) in which the number of disease genes ranged from 112 to 157 (Table 3-4, Appendices Table S2), indicating that 69%~93% of all genes in each signature were shared by 20 signatures. Moreover 80% of the top-50 ranked genes in each signature were shared by 20 signatures. Comparing to 10 sets of signatures derived from the same dataset of 62 samples (108) by other groups, our selected signatures are stabler. Whereas, the results from other groups showed that only 1~5 of the 4~60 selected predictor-genes in each of these sets are present in more than half of the other 9 sets (Table 3-1) (119, 124, 127, 147-150).

There are two aspects explaining why our selected gene signatures possess better stability. First, a SVM class-differentiation system with a universal set of globally optimized parameters, which gave the best average class-differentiation accuracy over the 500 testing-sets, was used to derive RFE gene-ranking function at every iteration step and for every testing-set. As such, the effect from the parameter-dependence of conventional gene selection can be reduced dramatically. In earlier studies using RFE or other wrapper methods for selecting signatures, non-predictor-genes have been eliminated in multiple iterations, and at every iteration step a different class-differentiation system, characterized by a different set of optimized parameters, has been constructed (117, 259). As gene-elimination is parameter-dependent, these selected predictor-genes are likely path-dependent and heavily influenced by sampling method, composition, order of gene evaluation, computational algorithm and parameters. These characteristics partly explain the highly unstable and patient-dependent characteristics of the previously-derived signatures (108). Second, an additional gene-ranking consistency evaluation was performed on top of the normal RFE procedure to reduce the variations of erroneous eliminations of predictor-genes.

The optimal SVM parameters for the 20 sample-sets were in a narrow range of 17~18 and the highest average accuracies were 92.2%~92.8% for colon cancer patients and 90.4%~91.1% for normal people respectively (Table 3-5). At these parameters, the accuracies for the individual testing-sets ranged from 82.4~100% for colon cancer patients and 77.0%~100% for normal people. Further deviation from these optimal parameters had relatively small effect on prediction accuracy and composition of predictor-genes. The relatively small variations of optimal SVM parameters and prediction accuracies across the 20 sampling-sets suggests that the performance of the SVM class-differentiation systems constructed by using globally optimized parameters and RFE iteration steps are fairly stable across different sampling combinations.

Our signatures include 52 of the 104 previously-derived predictor-genes, and those selected by a higher number of other studies tend to be ranked higher by our gene-ranking function (Table 3-3, Appendix Table S2). Regardless of their possible roles in cancer, these genes have shown proven capability for colon cancer outcome prediction. It is not surprising that they are included in our signatures.

Table 3-3 Gene information for colon cancer genes shared by all of the 20 signatures

EST accession number (Included in this number of previously derived signatures by other groups)[1]	gene Name	Gene description	Gene aliases	relationship with cancer[2]	Interacting partner (bold: partner related to cancer) [3]	Other Information[4]
U14631 (1)	HSD11B2	Hydroxysteroid (11-beta) dehydrogenase 2	AME, AME1, HSD11K, HSD2	successful tumor target		Decrease of HSD11B2 mRNA abundance and enzyme activity is associated with colorectal cancer (287)
R46753 (2)	CDKN1A	Cyclin-dependent kinase inhibitor 1A	CAP20, CDKN1, CIP1, MDA-6, P21, SDI1, WAF1, p21CIP1	research tumor target, Cancer pathway affiliated gene (p53 signaling, PI3k signaling, G1-S phase transition (Rb), Cyclins and Cell Cycle Regulation)	**TP53,SMAD4, CDKN1B, CDK4, STAT3, CDK2, CCNB1, CCND1, MYC, CCNE1, CCNA2, CCNA1,CREBBP, HIF1A, HDAC1, RB1, TP73,** SET, PCNA, CSNK2A1, CSNK2B, PSMA3, POLD2, SMARCA4, MUC1, HIST4H4, GADD45A, MITF, CDC2, APEX1, SP1, FOXO1A, IBRDC2, PIAS2, PIM1, PPARBP	This protein is a regulator of the cell division cycle (288, 289). Its expression of level is tightly controlled by the tumor suppressor protein p53 and can mediate the p53-dependent cell cycle G1 phase arrest in response to a variety of stress stimuli (290).
T41204 (0)	MMP9	Matrix metallopeptidase 9	CLG4B, GELB	research tumor target	LCN2, COL4A6, CD44, PZP, COL1A2, BTC, TIMP3, COL4A1, COL4A2, COL1A1, COL4A3, MMP7, COL4A5, COL4A4, AREG, RECK, FN1, THBS1, CXCL5, THBS2	The balance between MMPs/TIMPs regulates the extracellular matrix (ECM) turnover and remodeling during normal development and pathogenesis (291)
J03040 (0)	SPARC	Secreted protein, acidic, cysteine-rich (osteonectin)	ON	research tumor target	**VEGF, TGFB1,** Collagen, COL13A1, Fibrinogen, HSPG2, TGM2, PLG, Procollagen type I, PLAT, Laminin, PDGF alpha, THBS1, SDC2,	This protein inhibits cell-cycle progression, and influences the synthesis of extracellular matrix (ECM). It was suggested that this protein plays a dual role in the VEGF functions, tumor angiogenesis, and extravasation of tumors mediated by the increased permeability of endothelial barrier function (292)
X14958 (1)	HMGA1	High mobility group AT-hook 1	HMG-R, HMGIY, MGC12816, MGC4242, MGC4854	Oncogene	**JUN,** HIPK2, POU3F1, SUFU, INSR locus, ATF2, INSR, Casein kinase 2, NFYA, CEBPB, SP1, IRF1,RNF4	HMGA1 is a novel MYCN (myc myelocytomatosis viral related oncogene) target gene relevant for neuroblastoma tumorigenesis (293)
J04102 (1)	ETS2	V-ets erythroblastosis virus E26 oncogene		oncogene, Cancer pathway affiliated gene(PI3k signaling)	**BRCA1, ETS1, SMARCA4, ERG, JUN, SRC, EP300,** NCOA3, TTRAP, EAPII, ZMYND11, CDK10, GATA3, NCOR1, SMARCA4, NR3C1	

X77548 (0)	NCOA4	Nuclear receptor coactivator 4	ARA70, DKFZp762E1112, ELE1, PTC3, RFG	oncogene, research tumor target	RXRA, PPARG, TFIIB, RNF14, PCAF, AR	This protein induces the secretion of myeloid growth and chemotactic factors and potent mitogenicity of this protein correlates with its prevalence in tall-cell variant of papillary thyroid carcinoma (294)
R67343 (0)	CNNM4	Cyclin M4	CNNM4, ACDP4, KIAA1592	tumor-suppressor gene		Cyclins activate crucial protein kinases and thereby help control progression from one stage of the cell cycle to another (295)
X12369 (2) Z24727 (0)	TPM1	Tropomyosin1	HTM-alpha, TMSA, TPM1-alpha	tumor-suppressor genes	EPB41, TPM2	Silencing of the TPM1 gene by DNA methylation alters tumor suppressor function of TGF-beta (296). Loss of expression of TPM1, a novel class II tumor suppressor that induces anoikis, was observed in primary breast tumors (297).
U15212 (0)	CDX1	Caudal type homeobox transcription factor 1	MGC116915	tumor-suppressor gene		DNA methylation down-regulates CDX1 gene expression in colorectal cancer cell lines. CDX1 inhibits proliferation of intestinal epithelial cells and regulates intestine-specific genes involved in differentiation. The expression of CDX1 is aberrantly down-regulated in colorectal cancers and colon cancer-derived cell lines (298)
X63629 (2)	CDH3	Cadherin 3, type 1, P-cadherin	CDHP, HJMD, PCAD	tumor-suppressor gene, Cancer pathway affiliated gene (WNT signaling, Cell adhesion molecules (CAMs))	**Caspase, CTNNB1,CTNND1,** Alpha catenin	Cadherin cell–cell adhesion proteins play an important role in modulating the behavior of tumor cells. Smad4 induces the tumor suppressor E-cadherin and P-cadherin in colon carcinoma cells (299). It was reported that CDH3 may present at an increased level in colon cancer cells (300, 301).
R33481 (0)	CREB5	CAMP responsive element binding protein 5	CRE-BPA	tumor-suppressor gene	**JUN, CREB5**, ATF2	This protein contributes to ovarian tumorigenesis in humans (302)
H49870 (1) L07648 (1)	MXI1	MAX interactor 1	MAD2, MGC43220, MXD2, MXI	tumor-suppressor gene (transcriptional repressor)	**Mmip1, MAX**	Expression of the gene, which produces an oncogenic transcription factor, is tightly regulated in normal cells but is frequently deregulated in human cancers. Defects in this gene are frequently found in patients with prostate tumors
T51493 (0)	PPP2R5C	Protein phosphatase 2, regulatory subunit B (B56), gamma isoform	B56G, MGC23064	tumor-suppressor gene, Cancer pathway affiliated gene(MAPK	PPP2CA, PPP2R1B	This protein is an important regulator of Wnt/beta-catenin pathway activity in colorectal cancer cell (Oncogene. 2006 Jul 31) and is functionally inactivated in blast crisis CML through the inhibitory activity of the BCR/ABL-regulated

				signaling; Cyclins and Cell Cycle Regulation, Wnt / beta-catenin pathway)		SET protein (303)
X07767 (0)	PRKACA	Protein kinase, cAMP-dependent, catalytic, alpha	MGC102831, MGC48865, PKACA	tumor-suppressor gene, Cancer pathway affiliated gene (MAPK signaling pathway, Apoptosis, Wnt signaling pathway, Hedgehog signaling pathway)	**EGFR,** MGMT, PKIA, PKA, AKAP8L, AKIP1a, NMT1, NPR1, IFNAR1, cPKA-RI, SLC9A3R2, C11orf17, AKAP14, NIN	This protein is a signaling molecule important for a variety of cellular functions
Z49269 (0)	CCL14	chemokine (C-C motif) ligand 14	CC-1, CC-3, CKb1, HCC-1, HCC-3, MCIF, NCC-2, NCC2, SCYA14, SCYL2, SY14	angiogenesis gene, Cancer pathway affiliated gene (PI3k signaling, JAK/STAT Pathway, Cytokine-cytokine receptor interaction)	CCBP2, CCR5, CCR3, DDX39, CCR1	Chemokines play an important role in leukocyte mobilization, hematopoiesis, and angiogenesis. Tissue-specific expression of particular chemokines also influences tumor growth and metastasis (304)
L34657 (0)	PECAM1	platelet/endothelial cell adhesion molecule	CD31, PECAM-1	angiogenesis gene, , Cancer pathway affiliated gene(Cell adhesion molecules (CAMs))	LYN, YES1, HCK, ITGA5, CSK, LCK, PLCG1, PTPN6, CTNNB1, SRC, CD38, FYN, DSP	This protein participates in adhesive and/or signaling phenomena required for the motility and organization of endothelial cells (305)
T54303 (0)	KRT8	Keratin 8, cytokeratin 8; keratin, type II cytoskeletal 8	CARD2, CK8, CYK8, K2C8, K8, KO	tumor marker, Cancer pathway affiliated gene(Cell Communication)	**EGFR,** HSPA5, BYSL, KRT18, ANXA1, PLAT, PKP1, PNN	This protein alters the epidermal cell differentiation, favors the neoplastic transformation of cells, and is ultimately responsible of the invasive behavior of transformed epidermal cells leading of conversion of benign to malignant tumors (306)
T51571 (0)	S100A11	S100 calcium binding protein A11 (calgizzarin)	MLN70, S100C	Tumor marker	NCL, ANXA1, Actin, S100B	
M94132 (1)	MUC2	Mucin 2	MLP	Tumor marker (colon)	GALNT14, PLEKHM1, GALNT12	MUC2 expression was regulated in human colon cancer cells at the level of transcription via AP-1 activation (307). Reduction of MUC2 expression may be associated with the occurrence and progression of colorectal carcinomas (308)
H78386 (0)	IL1R2	Interleukin 1 receptor, type II	CD121b, IL1RB, MGC47725	immune tolerance, research therapeutic target, Cancer pathway affiliated	IL1RAP, IL1A, IL1RN	.

				gene (MAPK signaling pathway, Cytokine-cytokine receptor interaction)		
M23115 (0)	ATP2A2	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2	ATP2B, DAR, DD, MGC45367, SERCA2	Interacting partner of cancer gene	**BCL2,** TNFRSF1A, GSH, PLN, TRADD, S100A1, TRAF6, TNFRSF1B	This enzyme is involved in regulation of the contraction/relaxation cycle
T62947 (3)	C15orf15	Chromosome 15 open reading frame 15	HRP-L30-iso, L30, RLP24, RPL24, RPL24L	Cancer pathway affiliated gene (Ribosome)		
D45887 (0)	CALM2	Calmodulin 2 (phosphorylase kinase, delta)	CALM2, CAMII, PHKD, PHKD2	Interacting partner of cancer gene, Cancer pathway affiliated gene (Calcium signaling pathway)	**MYOD1,** NEUROD1, ESR2, ASCL2, IQCB1, PPEF1, GRM7, KCNQ2, KCNQ3, KCNQ5, GRM5, TCF4, MYF6, MYF5, PCNT, MARCKS, INVS, CALD1, RAB3B, PPEF2, EDF1, ESR1, MYOG	Upregulated of CALM2 might be associated with the oncogenesis of ALCL (309)
U26312 (0)	CBX3	Chromobox homolog 3	HECH, HP1-GAMMA, HP1Hs-gamma	Interacting partner of cancer gene	**PIM1,** CBX5, SP100, MKI67, CBX3, CBX1, DNMT3B, C20orf172, TRIM24,MIS12, LBR	
R33367 (0)	CD46	CD46 molecule, complement regulatory protein	MCP, MGC26544, MIC10, TLX, TRA2.10	Interacting partner of cancer gene	**SRC**, YES1, TSPAN4, C4B, U48, DLG4, ITGB1, MSN, LLGL1, C3	The level of CD46 protein expression is associated with tumor epithelial cell population, suggesting that CD46 should be evaluated as a novel prognostic indicator (310)
X54942 (2)	CKS2	CDC28 protein kinase regulatory subunit 2	CKSHS2	Interacting partner of cancer gene	**CCNB1**, ESPL1, CCNB1, CDC2	CKS2 protein binds to the catalytic subunit of the cyclin dependent kinases and is essential for their biological function.
T64885 (0)	CNOT1	CCR4-NOT transcription complex, subunit 1	AD-005, CDC39, DKFZp686E0722, FLJ90644, KIAA1007, NOT1, NOT1H	implication in cancer	CNOT8	The Ccr4-Not complex is a global regulator of gene expression that is conserved from yeast to human, as a regulatory platform that senses nutrient levels and stress (311)
U37012 (0)	CPSF1	Cleavage and polyadenylation specific factor 1	CPSF160, HSU37012	Interacting partner of cancer gene	**RNA polymerase II,** TAFII100, TAFII28, TAFII18, MCM2, hCIP1, Poly A polymerase, HEAB, BAT2, TAFII20, TAF7	
M76378 (5)	CSRP1	cysteine and glycine-rich protein 1	CRP, CRP1, CSRP, CYRP, D1S181E, DKFZp686M148	implication in cancer	Alpha-actinin	This protein may be involved in regulatory processes important for development and cellular differentiation with critical functions in gene regulation, cell growth, and somatic differentiation
M63391 (7)	DES	desmin	CMD1I, CSM1, CSM2, FLJ12025, FLJ39719, FLJ41013, FLJ41793	Cancer pathway affiliated gene (Cell Communication)	S100A1, DMN, S100B, CAPN1, DSP, SPTAN1, SYNC1, Syncoilin	

X60489 (0)	EEF1B2	Eukaryotic translation elongation factor 1 beta 2	EEF1B, EEF1B1, EF1B	implication in cancer	VARS, AARS, EF-1-beta, HARS	This protein is involved in regulation of the cell cycle, normal and pathological (312)
R54097 (2)	EIF2S2	Eukaryotic translation initiation factor 2	DKFZp686L18198, EIF2, EIF2B, EIF2beta, MGC8508	Interacting partner of cancer gene	**NCK1,** EIF2B5, CSNK2A1, CK2beta, CSNK2B, EIF5, EIF2B4,PRKDC, CK2alpha	This protein functions in the early steps of protein synthesis
H06524 (1)	GSN	Gelsolin (amyloidosis, Finnish type)	GSN, DKFZp313L0718	Interacting partner of cancer gene, Cancer pathway affiliated gene (Regulation of actin cytoskeleton)	**BCAR1,LIMK2, PIK3,** ACTN4, ACTB, Tropomyosin, VCL, TNIK, CLIC5, VDAC1,PXN, VASP, APP, ACTA1, PTK2B, FN1, G-actin, tax, AR	GSN functions as a switch that controls E- and N-cadherin conversion via Snail. Knockdown of GSN leads to EMT in human mammary epithelial cells and possibly to the development of human mammary tumors (313)
M96233 (0)	GSTM4	glutathione S-transferase M4	GSTM4-4, GTM4, MGC131945, MGC9247	implication in cancer	GSTM4	A T2517C polymorphism in the GSTM4 gene is associated with risk of developing lung cancer (314)
Z50753 (6)	GUCA2B	Guanylate cyclase activator 2B (uroguanylin)	GCAP-II, UGN	implication in cancer	GUCY2D	This protein synthesizes cGMP, which concentration of human colon tumors was higher than that of the surrounding mucosa (315)
X12671 (1)	HNRPA1	heterogeneous nuclear ribonucleoprotein A1	HNRNPA1, MGC102835	implication in cancer	TTF2, ELAVL1, PTMA, MAP3K14, RIPK3, FEN1, PTMA, BAT5, PRKCZ, TNPO1, SFRS12, SAFB, UPF3A, BAT2	This protein may contribute to maintenance of telomere repeats in cancer cells with enhanced cell proliferation and the quantitative alteration of this protein could facilitate colon epithelial cell transformation through transcriptional and translational perturbation (316)
T51023 (1)	HSP90AB1	Heat shock protein 90kDa alpha (cytosolic), class B member 1	D6S182, FLJ26984, HSP90-BETA, HSP90B, HSPC2, HSPCB	Cancer related gene	WASL, STARD13, PPID, TRADD, MAP3K7	This protein is important for signaling by types I and II interferons (317)
M22382 (2)	HSPD1	Heat shock 60kDa protein 1 (chaperonin)	CPN60, GROEL, HSP60, HSP65, HuCHA60, SPG13	Interacting partner of cancer gene	**BAK1,** CA2, HSPE1, p21ras, PRNP, HIST2H2BE, CASP9, integrin, DHFR, CASP6, CASP3, PKA C, RASA1, ALDH2	This protein may function as a signaling molecule in the innate immune system
X02492 (0)	IFI6	Interferon, alpha-inducible protein 6	FAM14C, G1P3, IFI-6-16, IFI616	Cancer related gene	p-G1P3, ACTB, HIST4H4, POLR2A	This protein may have function as a cell survival protein by inhibiting mitochondrial-mediated apoptosis (318)
D14812 (1)	MORF4L2	Mortality factor 4 like 2	KIAA0026, MORFL2, MRGX	Interacting partner of cancer gene	**HDAC1, RB1, Sin3A,** TLE, C20orf20, PAM14, MRFAP1	
T71025 (2)	MT1G	Metallothionein 1G	MGC12386, MT1, MT1K	Interacting partner of cancer gene	**HDAC1, RB1, RBL1; RBL2(P107), RBL(P130),** E2F1, E2F2, E2F3, E2F4, E2F5, ECRG2, SPINK7, MTF1	

R87126 (7)	MYH9	Myosin, heavy polypeptide 9, non-muscle	DFNA17, EPSTS, FTNS, MGC104539, MHA, NMHC-II-A, NMMHCA	Cancer related gene, Cancer pathway affiliated gene (Tight junction, Regulation of actin cytoskeleton)	S100A4, CD163, GRIN1	
H20709 (2)	MYL6	Myosin, light polypeptide 6, alkali, smooth muscle and non-muscle	MYL6, ESMLC, LC17-GI, LC17-NM, LC17A, LC17B, MLC1SM, MLC3NM, MLC3SM	Cancer pathway affiliated gene(Focal adhesion, Tight junction)	CHUK, TNFRSF1A, IKBKE, RIPK3, IKBKG, NFKB1, MAP3K3	
X73424 (0)	PCCB	Propionyl Coenzyme A carboxylase, beta polypeptide	PCCB, DKFZp451E113	implication in cancer		This protein is an important source of energy for colonocytes (319)
T94350 (0)	PMP22	Peripheral myelin protein 22	CMT1A, CMT1E, DSS, GAS-3, HMSNIA, HNPP, MGC20769, Sp110	implication in colon cancer	PEX19, CNX, MPZ	The level of this protein was significantly decreased in various types of tumors including the colon carcinoma (320)
U21090 (0)	POLD2	Polymerase (DNA directed), delta 2		Cancer pathway affiliated gene(DNA repair mechanism)	PCNA, POLDIP2, POLDIP3, POLD3, DNA polymerase delta CDKN1A, POLD1, KCTD13	
D13665 (0)	POSTN	Periostin, osteoblast specific factor	MGC119510, MGC119511, OSF-2, PDLPOSTN, PN, RP11-412K4.1	implication in colon cancer		This protein potently promotes metastatic growth of colon cancer by augmenting cell survival via the Akt/PKB pathway (321)
T86444 (0)	PPP1R9B	Protein phosphatase 1, regulatory subunit 9B, spinophilin	FLJ30345, PPP1R6, PPP1R9, SPINO	Interacting partner of cancer gene, Cancer pathway affiliated gene	**CDKN2A**, ACTC, ADRA1A, DRD2, PPP1R2, PPYR1, TGOLN2, TIAM1, DCX, PPP1R9A	This protein is involved in the regulation of a variety of cellular processes, such as cell division, glycogen metabolism, muscle contractility and protein synthesis
D15049 (0)	PTPRH	Protein tyrosine phosphatase, receptor type, H	FLJ39938, MGC133058, MGC133059, SAP-1	Interacting partner of cancer gene	**BCAR1, PTK2, MAPK1,** PXN, DOK1	This protein was found to be expressed in several cancer cell lines, but not in the corresponding normal tissues. This protein induce apoptosis by stomach cancer-associated protein-tyrosine phosphatase-1 (322). Overexpression of this protein can be found in human colorectal cancers (323)
T47377 (3)	S100P	S100 calcium binding protein P	MIG9	Cancer related gene	VIL2, CACYBP, ECD, AGER, S100Z, Melittin	This protein involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation.

T51261 (1)	SERPINE2	Serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2	GDN, PI7, PN1, PNI	implication in cancer	COL4A6, COL4A1, COL4A2, COL4A3, COL4A5, COL4A4	This protein promotes extracellular matrix production and local invasion of pancreatic tumors in vivo (324)
U22055 (1)	SND1	Staphylococcal nuclease domain containing 1	p100	Interacting partner of cancer gene	**RNA pol II, PIM1,STAT6, MYB, MYBL2,** TFIIE, GTF2E2,RBPSUH, POLR2A, GTF2E1	This protein recruits histone acetyltransferase activity to STAT6 and mediates interaction between the CREB-binding protein and STAT6 (325)
U19969 (1)	TCF8	Transcription factor 8 (represses interleukin 2 expression)	AREB6, BZP, MGC133261, NIL-2-A, NIL-2A, ZEB, ZEB1, ZFHEP, ZFHX1A	Interacting partner of cancer gene, Cancer pathway affiliated gene (WNT signaling)	**CDH1, SMAD2, SMAD3,** HTATIP, CTBP2, DRAP1	In colon cancer patients, the correlation between the expression of SNAIL and the downregulation of E-cadherin (CDH1) is lost when ZEB1 is overexpressed (326). This protein plays a role in repressing E-cadherin and MUC1 in epithelial cells (327)
H81068 (0)	WASF2	WAS protein family, member 2, suppressor of cyclic-AMP receptor (WASP-family)	SCAR2, WAVE2, dJ393P12.2	Interacting partner of cancer gene, Cancer pathway affiliated gene (Adherens junction, Regulation of actin cytoskeleton)	**CDC42, GRB2,** Syndapin I, ACTR3, PSTPIP1, DOCK1, BAIAP2, FYN, Actin, ACTR2, BTK	This protein involved in transducing signals that involve changes in cell shape, motility or function
T92451 (3)	TPM2	Tropomyosin 2 (beta)	AMCD1, DA1, TMSB	Interacting partner of cancer gene	**JUN,** TPM1, ACTB, PDLIM7, S100A4, RRAD, S100A2	
X86693 (3)	SPARCL1	SPARC-like 1 (mast9, hevin)	PIG33, SC1	implication in cancer		
R62549 (0) U09564 (0)	SRPK1	SFRS protein kinase 1	SFRSK1	implication in cancer	MBP, SFRS6, B1C8, SFRS1, U2AF1, SFRS3, C20orf42, U2AF2, SFRS4, YWHAG, SFRS12, SAFB, SFRS5, PRM1, SFRS2IP, LBR	SRPK1 gene expression was increased in oxaliplatin-resistant HT29 colon cancer cells.
H64489 (1)	TSPAN1	Tetraspanin 1	NET-1, RP11-322N21	implication in cancer		This proteins mediate signal transduction events that play a role in the regulation of cell development, activation, growth and motility
R55310 (0)	UQCRC1	Ubiquinol-cytochrom E C reductase core protein I	D3S3191	implication in cancer	RTN4, CYCS, UQCRH	This protein was highly expressed in breast (74%) and ovarian tumors (34%) (328)
L32977 (0)	UQCRFS1	ubiquinol-cytochrom E C reductase, Rieske iron-sulfur polypeptide 1	RIS1	implication in cancer		This protein is related to breast cancer

T64012 (0)	WDR7	WD repeat domain 7, TGF-beta resistance associated gene; rabconnectin-3 beta	KIAA0541, TRAG	implication in cancer	RAB3GAP1	This protein involved in a variety of cellular processes, including cell cycle progression, signal transduction, apoptosis, and gene regulation.
T58861 (0)	RPL30	Ribosomal protein L30		Cancer pathway affiliated gene (Ribosome)		
T57619 (0)	RPS6	Ribosomal protein S6		Cancer pathway affiliated gene (Ribosome, mTOR signaling pathway, Insulin signaling pathway)	RPS6KB1	BMK1 pathway is crucial for tumor-associated angiogenesis through its role in the regulation of the RSK-RPS6 signaling module (329)
D31885 (0)	ARL6IP	ADP-ribosylation factor-like 6 interacting protein	AIP1, ARL6IP1, ARMER, KIAA0069		ARL4D, ARL6	This protein may involve in cell survival (330)
R36977 (1)	GTF3A	General transcription factor IIIA	AP2, TFIIIA		OPTN, GTF3C2	
U31525 (0)	GYG1	Glycogenin 1	GYG		GYG2, TRIM7	
X57351 (0)	IFITM2	Interferon induced transmembrane protein 2 (1-8D)	1-8D		UPF3A	
R98842 (1)	DTWD2	DTW domain containing 2	FLJ33977, MGC138579, MGC138580			
R06601 (0)	MT1M	Metallothionein 1M	MGC118949, MGC40498, MT1, MT1K			
J02854 (7)	MYL9	Myosin, light polypeptide 9, regulatory	LC20, MYRL2, MRLC1, MLC2, MGC3505			
T65380 (0)	PRPSAP1	Phosphoribosyl pyrophosphate synthetase-associated protein 1	PAP39		PRPS2, PRPS1	
R44301 (1)	NR3C2	Nuclear receptor subfamily 3, group C, member 2	MCR, MR, MLR, MGC133092		FKBP52, G actin, HSPA4, PIAS1, F actin, FKBP4, HSP90AA1, TRIM24, PTGES3, NR3C1	

T95018 (0)	PCNP	PEST proteolytic signal containing nuclear protein	PCNP, DKFZp781I24156		UHRF2	
T51534 (1)	CST3	Cystatin C	MGC117328		C4A, Actinidin, PDZD2, CTSS, CTSB, Papain	
J05032 (3)	DARS	Aspartyl-tRNA synthetase	DKFZp781B1120 2, MGC111579		EEF1A1, MAP3K7, MARS, EEF1D	
X56597 (1)	FBL	Fibrillarin	FIB, FLRN, RNU3IP1		PRMT1, U3 snoRNA, U3 snoRNP, PSMB6, PIN4, DDX5, RNU3, SNRPN NOP5/NOP58	
M80815 (1)	FUCA1	fucosidase	FUCA1			
T67077 (0)	FXYD1	FXYD domain containing ion transport regulator 1(phospholemman)	MGC44983, PLM		PKA, PPAP2A, Type 1phosphatase, ATP1A1, PPP1CA, ATP1B1	
M82919 (1)	GABRB3	Gamma-aminobutyri c acid (GABA) A receptor	MGC9051			
D42047 (1)	GPD1L	Glycerol-3-phosphate dehydrogenase 1-like	GPD1L, KIAA0089			
H25136 (0)	ITPR3	Inositol 1,4,5-triphosphate receptor, type 3	FLJ36205, IP3R3		TRPC5, TRPC6, TRPC4, TRPC3, TRPC2, OPRS1, TRPM2, CABP1, TRPC1	
U06698 (0)	KIF5A	Kinesin family member 5A	D12S1889, MY050, NKHC, SPG10		KLC3, DTNB, KNS2	
X70326 (1)	MARCKSL1	MARCKS-like 1	F52, MACMARCKS, MLP, MLP1, MRP		DCTN2	
H87135 (0)	MGC22793	Hypothetical protein MGC22793				
D25217 (1)	MLC1	Megalencephalic leukoencephalopathy with subcortical cysts 1	KIAA0027, LVM, MLC, VL			
U17899 (0)	CLNS1A	Chloride channel, nucleotide-sensitive, 1A	CLCI, CLNS1B, ICln		PRMT5, Actin	

D16294 (0)	ACAA2	Acetyl-Coenzyme A acyltransferase 2	DSAEC		SCP2	This protein catalyzes the last step of the mitochondrial fatty acid beta-oxidation spiral
R88740 (3)	ATP5J	ATP synthase, H+ transporting, mitochondrial F0 complex	ATP5, ATP5A, ATPM, CF6, F6			This synthase involved in oxidative phosphorylation
H43887 (1)	CFD	Complement factor D (adipsin)	ADN, DF, PFD		SERPINF2	
H48072 (0)	COX6A1	Cytochrome c oxidase subunit VIa polypeptide 1	COX6A, COX6AL, MGC104500			
T51250 (0)	COX8A	Cytochrome c oxidase subunit 8A (ubiquitous)	COX VIII, COX VIII-L, COX8, COX8-2, COX8L			
X87159 (1)	SCNN1B	Sodium channel, nonvoltage-gated 1, beta (Liddle syndrome)	ENaCb, ENaCbeta, SCNEB		STX1A, SCNN1A, WWP2, CNTN1, SCNN1G, NEDD4, NEDD4L, SCNN1D	
R84411 (1)	SNRPB	Small nuclear ribonucleoprotein polypeptides B and B1	COD, SNRPB1, SmB/SmB', snRNP-B		TOP3B, WBP4, GEMIN5, GEMIN7, STXBP2, STXBP3, SMN1, SMN2, GEMIN4, DDX20, PACT, SNRPD3, GEMIN6, LSM11, WDR77, CTDP1, pICln, COIL, DHX9, RBBP6	
M36634 (3)	VIP	Vasoactive intestinal peptide	MGC13587, PHM27		VIPR1	
H08393 (4)	WDR77	WD repeat domain 77	HKMT1069, MEP50, MGC2722, Nbla10071, RP11-552M11.3		SNRPE, SNRPD1, SNRPD3, YWHAQ, SNRPB, PRMT5, SNRPF, SNRPD2	
H81558 (1)	ZNF358	Zinc finger protein 358	FLJ10390, ZFEND			
H73908 (0)		similar to contains Alu repetitive element; contains THR repetitive element				
H11084 (1)		ym09g08.s1 Soares infant brain 1NIB Homo sapiens cDNA				

		clone				
H40095 (1)		similar to gb:L19686_rna1 macrophage migration inhibitory factor (human)				
H64807 (1)		similar to contains Alu repetitive element				

[1]The signatures selected by other groups was obtained from (119, 124, 127, 147-150), which were used to differentiate colon cancer patients from normal people by using the same microarray dataset as our used dataset (108).

[2]Tumor target information was obtained from therapeutic target database (http://bidd.nus.edu.sg/group/cjttd/ttd.asp) (270, 271). Most of the pathway information was obtained from KEGG database (http://www.genome.jp/kegg/pathway.html) (331), Reactome (http://www.reactome.org/) (332), and review articles (333-338). The information of cancer genes was obtained from the review articles (333-338).

[3]Interacting partner information was obtained from Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) (339).

[4]If no reference was indicated, the information was extract from Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) (339).

Table 3-4 Statistics of the selected colon cancer genes from a colon cancer microarray dataset by class-differentiation systems constructed from 20 different sampling-sets each composed of 500 training-testing sets generated by random sampling. The dataset is from (108).

Sampling Set	No of selected predictor genes in signature	No of predictor-genes also included in N other signatures derived by using different sampling-set																			
		19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	155	104	3	2	5	3	4	3	4	3	3	1	4	2	1	4	2	3	3	1	0
2	135	104	3	2	4	3	4	2	4	2	1	0	1	2	1	0	1	1	0	0	0
3	156	104	3	2	5	3	4	3	3	3	3	1	3	3	0	4	4	4	0	1	3
4	146	104	3	2	5	3	4	3	3	3	2	1	2	2	1	2	2	1	2	0	1
5	116	104	3	1	4	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0
6	112	104	2	0	2	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0
7	119	104	2	2	3	1	3	0	3	0	0	1	0	0	0	0	0	0	0	0	0
8	127	104	3	2	5	3	2	1	2	1	1	1	1	0	0	1	0	0	0	0	0
9	133	104	3	2	5	2	2	2	2	2	2	0	2	1	2	1	0	0	0	0	1
10	156	104	3	2	4	3	4	3	4	3	3	0	3	2	1	2	3	5	3	2	2
11	139	104	3	2	5	3	3	3	3	3	2	1	2	1	1	3	0	0	0	0	0
12	115	104	2	2	3	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
13	144	104	3	2	5	3	4	3	3	2	3	1	2	2	1	3	0	1	2	0	0
14	157	104	3	2	5	3	4	3	4	3	3	0	3	2	2	3	2	3	3	3	2
15	149	104	3	2	4	3	4	3	3	3	3	0	3	2	0	3	1	3	2	0	3
16	136	104	3	2	5	3	4	3	3	2	0	0	3	2	1	0	0	0	1	0	0
17	136	104	3	2	5	3	3	2	3	1	3	0	2	1	1	0	1	1	1	0	0
18	127	104	3	2	4	3	2	3	1	0	1	0	2	0	0	1	1	0	0	0	0
19	146	104	3	2	5	2	4	3	4	3	2	1	2	2	1	1	3	1	1	1	1
20	122	104	3	1	2	2	2	1	1	2	1	1	0	0	1	1	0	0	0	0	0

Table 3-5 Overall accuracies of 500 training-test sets on the optimal SVM parameters

Sampling set	Optimal SVM parameter	Overall performance in 500 training dataset							Overall performance in 500 corresponding test dataset						
		Colon cancer patient			Normal people			Q	Colon cancer patient			Normal people			Q
		TP	FN	SE	TN	FP	SP		TP	FN	SE	TN	FP	SP	
1	18	9448	632	93.7%	5140	280	94.8%	94.1%	9171	749	92.4%	5055	525	90.6%	91.8%
2	17	9543	611	94.0%	4963	383	92.8%	93.6%	9107	739	92.5%	5128	526	90.7%	91.8%
3	18	9463	597	94.1%	5093	347	93.6%	93.9%	9191	749	92.5%	5031	529	90.5%	91.8%
4	18	9425	667	93.4%	5067	341	93.7%	93.5%	9135	773	92.2%	5084	508	90.9%	91.7%
5	17	9447	708	93.0%	4906	439	91.8%	92.6%	9140	705	92.8%	5146	509	91.0%	92.2%
6	17	9416	658	93.5%	4972	454	91.6%	92.8%	9174	752	92.4%	5079	495	91.1%	92.0%
7	17	9434	648	93.6%	5007	411	92.4%	93.2%	9145	773	92.2%	5048	534	90.4%	91.6%
8	17	9433	677	93.3%	5037	353	93.5%	93.4%	9154	736	92.6%	5110	500	91.1%	92.0%
9	17	9453	665	93.4%	4979	403	92.5%	93.1%	9150	732	92.6%	5109	509	90.9%	92.0%
10	17	9638	477	95.3%	5047	338	93.7%	94.7%	9157	728	92.6%	5105	510	90.9%	92.0%
11	17	9434	627	93.8%	5147	292	94.6%	94.1%	9194	745	92.5%	5063	498	91.0%	92.0%
12	17	9470	672	93.4%	4902	456	91.5%	92.7%	9148	710	92.8%	5097	545	90.3%	91.9%
13	18	9463	652	93.6%	5021	364	93.2%	93.4%	9140	745	92.5%	5119	496	91.2%	92.0%
14	17	9528	559	94.5%	5111	302	94.4%	94.4%	9176	737	92.6%	5062	525	90.6%	91.9%
15	17	9515	587	94.2%	5067	331	93.9%	94.1%	9148	750	92.4%	5091	511	90.9%	91.9%
16	17	9554	599	94.1%	5058	289	94.6%	94.3%	9106	741	92.5%	5127	526	90.7%	91.8%
17	17	9512	618	93.9%	5009	361	93.3%	93.7%	9145	725	92.7%	5090	540	90.4%	91.8%
18	17	9483	624	93.8%	5017	376	93.0%	93.5%	9154	739	92.5%	5108	499	91.1%	92.0%
19	18	9472	656	93.5%	5049	323	94.0%	93.7%	9131	741	92.5%	5079	549	90.2%	91.7%
20	18	9386	671	93.3%	4989	454	91.7%	92.7%	9173	770	92.3%	5035	522	90.6%	91.7%

3.3.3 The predictive performance of identified markers in disease differentiation

To further evaluate the predictive capability of our selected genes sets, we collected the gene expression profiles of 34 colon cancer cell line and 8 normal colon tissues from Stanford Microarray Database (SMD) (Appendices Table S1) (284). The predictive capability of our selected and the 10 other previously-derived signatures were evaluated by using the SVM classification system and 500 randomly-generated training-test sets generated from this dataset. The performance was evaluated using the associated test set and are shown in Table 3-6. The overall accuracy for the 104 predictor-genes was 96.8% with the standard deviation of 3.3%. The accuracies for all predictor-genes were in the range of 95.8~97.4% with the standard deviation of 4.7~5.0%. Using genes selected by other methods, the overall accuracies were in the range of 80.5%~94.9%, with the standard deviation of 2.9%~6.6%. These results suggest that the selected signatures using our system can perform well when compared with those selected by other methods.

Table 3-6 Average colon cancer prediction accuracy and standard deviation of 500 SVM class-differentiation systems constructed by 42 samples collected from Stanford Microarray Database (284) and by using signatures derived from this work and 10 previous works. The results were obtained from the overall accuracies of 500 test sets.

Signature (method)	No of selected predictor-genes in signature	Colon cancer tissue			Normal tissue			Q	STDEV
		TP	FN	SE	TN	FP	SP		
1*	155	8235	265	96.9%	1869	131	93.5%	96.2%	4.8%
2*	135	8266	234	97.2%	1840	160	92.0%	96.2%	4.8%
3*	156	8228	272	96.8%	1861	139	93.1%	96.1%	4.8%
4*	146	8342	158	98.1%	1885	115	94.3%	97.4%	4.8%
5*	116	8274	226	97.3%	1931	69	96.6%	97.2%	5.0%
6*	112	8246	254	97.0%	1979	21	99.0%	97.4%	5.0%
7*	119	8268	232	97.3%	1868	132	93.4%	96.5%	4.9%
8*	127	8272	228	97.3%	1875	125	93.8%	96.6%	4.8%

9*	133	8250	250	97.1%	1905	95	95.3%	96.7%	5.0%
10*	156	8206	294	96.5%	1928	72	96.4%	96.5%	5.0%
11*	139	8337	163	98.1%	1770	230	88.5%	96.3%	4.9%
12*	115	8266	234	97.2%	1793	207	89.7%	95.8%	4.6%
13*	144	8366	134	98.4%	1847	153	92.4%	97.3%	4.8%
14*	157	8345	155	98.2%	1839	161	92.0%	97.0%	4.7%
15*	149	8375	125	98.5%	1792	208	89.6%	96.8%	4.8%
16*	136	8401	99	98.8%	1722	278	86.1%	96.4%	4.8%
17*	136	8322	178	97.9%	1896	104	94.8%	97.3%	4.8%
18*	127	8331	169	98.0%	1814	186	90.7%	96.6%	4.9%
19*	146	8343	157	98.2%	1810	190	90.5%	96.7%	4.7%
20*	122	8244	256	97.0%	1855	145	92.8%	96.2%	4.8%
104 genes selected by all of the 20 signatures*	104	8328	172	98.0%	1839	161	92.0%	96.8%	3.3%
Ding and Peng 2005 (NB, SVM, LDA, LR) (127)	60	8193	307	96.4%	1771	229	88.6%	94.9%	5.9%
Huang and Kecman 2005 (SVM3)(149)	10	8194	306	96.4%	1760	240	88.0%	94.8%	6.5%
Bo and Jonassen 2002 (linear discriminant) (148)	50	8145	355	95.8%	1788	212	89.4%	94.6%	6.5%
Huang and Kecman 2005 (SVM2) (149)	10	8130	370	95.6%	1708	292	85.4%	93.7%	6.5%
Huang and Kecman 2005 (SVM1) (149)	10	8207	293	96.6%	1579	421	79.0%	93.2%	6.6%
Liu, Krishnan et al. 2005 (clustering method) (150)	6	7973	527	93.8%	1771	229	88.6%	92.8%	6.7%
Zhou and Mao 2005 (LS-SVM) (147)	15	8345	155	98.2%	989	1011	49.5%	88.9%	6.0%
Inza, Larranaga et al. 2004 (decision tree2) (124)	4	8056	444	94.8%	974	1026	48.7%	86.0%	5.4%
Inza, Larranaga et al. 2004 (decision tree1) (124)	5	7526	974	88.5%	1178	822	58.9%	82.9%	4.3%
Isabelle Guyon 2002 (SVM) (259)	7	7434	1066	87.5%	1019	981	51.0%	80.5%	2.9%

* Gene signatures were selected by using our system.

The predictive capability of our selected and the 10 other previously-derived

signatures (which was generated from the same dataset as ours) were further evaluated by using additional 500 randomly-generated training-test sets generated from the original Alon's colon cancer microarray dataset (108) and contained different combinations of samples from those 10,000 training-test datasets used for gene signatures selection. By using the 500 training sets, we constructed 500 SVM class-differentiation systems, each of which was tested by using the associated test set, and the results are shown in Table 3-7. The average cancer-differentiating accuracies of our signatures over these 500 testing sets are, respectively, 92.0% to 92.3%, 91.3% to 91.7%, 88.5% to 91.2%, 83.8% to 86.9% and 79.2% to 82.8% when all, top-100, top-50, top-30 and top-10 predictor-genes are used. The standard deviations of the individual accuracies are in the range of 3.3% to 3.4%, 3.3% to 3.4%, 3.8% to 4.0%, 4.6% to 5.5% and 5.6% to 6.2% respectively. In contrast, the average accuracies and standard deviations of the 10 previously-derived signatures are in the range of 75.3% to 87.1% and 4.2% to 14.9%. These results further illustrate that the performance of the signatures selected by our system is better and stabler than those of signatures selected by using other strategies. The performances of top-50, top-30 and top-10 predictor-genes are substantially less stable than those of all and top-100 predictor-genes.

Table 3-7 Average colon cancer prediction accuracy and standard deviation of 500 SVM class-differentiation systems constructed by using Alon's colon cancer microarray dataset (108) and by using each of the signatures derived from this work and 10 previous works.

Signature (method)	Number of selected cancer genes in signature	Colon cancer tissue			Normal tissue			Q	STDEV
		TP	FN	SE	TN	FP	SP		
1*	155	9491	683	93.3%	4809	517	90.3%	92.3%	3.3%
2*	135	9491	683	93.3%	4783	543	89.8%	92.1%	3.3%
3*	156	9490	684	93.3%	4798	528	90.1%	92.2%	3.4%
4*	146	9494	680	93.3%	4792	534	90.0%	92.2%	3.4%
5*	116	9492	682	93.3%	4796	530	90.0%	92.2%	3.3%
6*	112	9487	687	93.2%	4798	528	90.1%	92.2%	3.4%
7*	119	9495	679	93.3%	4800	526	90.1%	92.2%	3.4%
8*	127	9494	680	93.3%	4798	528	90.1%	92.2%	3.4%
9*	133	9492	682	93.3%	4787	539	89.9%	92.1%	3.4%
10*	156	9495	679	93.3%	4806	520	90.2%	92.3%	3.3%
11*	139	9486	688	93.2%	4797	529	90.1%	92.1%	3.3%
12*	115	9490	684	93.3%	4791	535	90.0%	92.1%	3.4%
13*	144	9486	688	93.2%	4787	539	89.9%	92.1%	3.4%
14*	157	9488	686	93.3%	4813	513	90.4%	92.3%	3.3%
15*	149	9494	680	93.3%	4800	526	90.1%	92.2%	3.4%
16*	136	9493	681	93.3%	4806	520	90.2%	92.3%	3.4%
17*	136	9482	692	93.2%	4798	528	90.1%	92.1%	3.4%
18*	127	9493	681	93.3%	4814	512	90.4%	92.3%	3.3%
19*	146	9485	689	93.2%	4774	552	89.6%	92.0%	3.3%
20*	122	9490	684	93.3%	4793	533	90.0%	92.1%	3.4%
104 genes selected by all of the 20 signatures*	104	9478	696	93.2%	4779	547	89.7%	92.0%	3.3%
top 100 genes in signature 1*	100	9432	742	92.7%	4751	575	89.2%	91.5%	3.4%
top 100 genes in signature 2*	100	9429	745	92.7%	4773	553	89.6%	91.6%	3.4%
top 100 genes in signature 3*	100	9432	742	92.7%	4765	561	89.5%	91.6%	3.4%
top 100 genes in signature 4*	100	9435	739	92.7%	4754	572	89.3%	91.5%	3.5%
top 100 genes in signature 5*	100	9442	732	92.8%	4755	571	89.3%	91.6%	3.3%
top 100 genes in signature 6*	100	9433	741	92.7%	4767	559	89.5%	91.6%	3.3%
top 100 genes in signature 7*	100	9445	729	92.8%	4768	558	89.5%	91.7%	3.4%
top 100 genes in signature 8*	100	9443	731	92.8%	4753	573	89.2%	91.6%	3.3%
top 100 genes in signature 9*	100	9432	742	92.7%	4749	577	89.2%	91.5%	3.3%
top 100 genes in signature 10*	100	9429	745	92.7%	4776	550	89.7%	91.6%	3.3%
top 100 genes in signature 11*	100	9444	730	92.8%	4766	560	89.5%	91.7%	3.4%
top 100 genes in signature 12*	100	9427	747	92.7%	4743	583	89.1%	91.4%	3.3%
top 100 genes in signature 13*	100	9428	746	92.7%	4760	566	89.4%	91.5%	3.4%

top 100 genes in signature 14*	100	9432	742	92.7%	4779	547	89.7%	91.7%	3.3%
top 100 genes in signature 15*	100	9421	753	92.6%	4767	559	89.5%	91.5%	3.4%
top 100 genes in signature 16*	100	9436	738	92.7%	4753	573	89.2%	91.5%	3.3%
top 100 genes in signature 17*	100	9430	744	92.7%	4789	537	89.9%	91.7%	3.3%
top 100 genes in signature 18*	100	9446	728	92.8%	4771	555	89.6%	91.7%	3.3%
top 100 genes in signature 19*	100	9431	743	92.7%	4726	600	88.7%	91.3%	3.3%
top 100 genes in signature 20*	100	9439	735	92.8%	4747	579	89.1%	91.5%	3.4%
Top 50 genes in signature 1*	50	9308	866	91.5%	4408	918	82.8%	88.5%	4.0%
Top 50 genes in signature 2*	50	9320	854	91.6%	4725	601	88.7%	90.6%	3.8%
Top 50 genes in signature 3*	50	9308	866	91.5%	4592	734	86.2%	89.7%	4.0%
Top 50 genes in signature 4*	50	9300	874	91.4%	4582	744	86.0%	89.6%	4.0%
Top 50 genes in signature 5*	50	9315	859	91.6%	4734	592	88.9%	90.6%	3.8%
Top 50 genes in signature 6*	50	9364	810	92.0%	4655	671	87.4%	90.4%	4.0%
Top 50 genes in signature 7*	50	9275	899	91.2%	4635	691	87.0%	89.7%	4.1%
Top 50 genes in signature 8*	50	9306	868	91.5%	4714	612	88.5%	90.5%	3.8%
Top 50 genes in signature 9*	50	9412	762	92.5%	4726	600	88.7%	91.2%	3.9%
Top 50 genes in signature 10*	50	9412	762	92.5%	4726	600	88.7%	91.2%	3.9%
Top 50 genes in signature 11*	50	9298	876	91.4%	4698	628	88.2%	90.3%	3.8%
Top 50 genes in signature 12*	50	9324	850	91.6%	4719	607	88.6%	90.6%	3.9%
Top 50 genes in signature 13*	50	9300	874	91.4%	4582	744	86.0%	89.6%	4.0%
Top 50 genes in signature 14*	50	9328	846	91.7%	4728	598	88.8%	90.7%	3.9%
Top 50 genes in signature 15*	50	9363	811	92.0%	4635	691	87.0%	90.3%	3.9%
Top 50 genes in signature 16*	50	9391	783	92.3%	4678	648	87.8%	90.8%	3.8%
Top 50 genes in signature 17*	50	9402	772	92.4%	4698	628	88.2%	91.0%	3.8%
Top 50 genes in signature 18*	50	9365	809	92.0%	4641	685	87.1%	90.4%	3.8%
Top 50 genes in signature 19*	50	9304	870	91.4%	4597	729	86.3%	89.7%	3.8%
Top 50 genes in signature 20*	50	9285	889	91.3%	4616	710	86.7%	89.7%	4.0%
Top 30 genes in signature 1*	30	9155	1019	90.0%	4021	1305	75.5%	85.0%	5.2%
Top 30 genes in signature 2*	30	9202	972	90.4%	4080	1246	76.6%	85.7%	4.6%
Top 30 genes in signature 3*	30	9155	1019	90.0%	4021	1305	75.5%	85.0%	5.2%
Top 30 genes in signature 4*	30	9080	1094	89.2%	3909	1417	73.4%	83.8%	5.5%
Top 30 genes in signature 5*	30	9180	994	90.2%	4065	1261	76.3%	85.5%	4.7%
Top 30 genes in	30	9207	967	90.5%	4046	1280	76.0%	85.5%	4.9%

signature 6*									
Top 30 genes in signature 7*	30	9047	1127	88.9%	3904	1422	73.3%	83.6%	5.4%
Top 30 genes in signature 8*	30	9155	1019	90.0%	4092	1234	76.8%	85.5%	4.7%
Top 30 genes in signature 9*	30	9155	1019	90.0%	4021	1305	75.5%	85.0%	5.2%
Top 30 genes in signature 10*	30	9227	947	90.7%	4166	1160	78.2%	86.4%	4.5%
Top 30 genes in signature 11*	30	9110	1064	89.5%	4352	974	81.7%	86.9%	4.4%
Top 30 genes in signature 12*	30	9202	972	90.4%	4080	1246	76.6%	85.7%	4.6%
Top 30 genes in signature 13*	30	9170	1004	90.1%	4051	1275	76.1%	85.3%	5.2%
Top 30 genes in signature 14*	30	9155	1019	90.0%	4021	1305	75.5%	85.0%	5.2%
Top 30 genes in signature 15*	30	9147	1027	89.9%	4052	1274	76.1%	85.2%	4.9%
Top 30 genes in signature 16*	30	9180	994	90.2%	4049	1277	76.0%	85.3%	5.0%
Top 30 genes in signature 17*	30	9162	1012	90.1%	4053	1273	76.1%	85.3%	5.0%
Top 30 genes in signature 18*	30	9180	994	90.2%	4065	1261	76.3%	85.5%	4.7%
Top 30 genes in signature 19*	30	9079	1095	89.2%	4010	1316	75.3%	84.4%	4.9%
Top 30 genes in signature 20*	30	9127	1047	89.7%	4014	1312	75.4%	84.8%	4.4%
Top 10 genes in signature 1*	10	8842	1332	86.9%	3769	1557	70.8%	81.4%	6.0%
Top 10 genes in signature 2*	10	8794	1380	86.4%	3739	1587	70.2%	80.9%	5.9%
Top 10 genes in signature 3*	10	8866	1308	87.1%	3681	1645	69.1%	80.9%	6.1%
Top 10 genes in signature 4*	10	8842	1332	86.9%	3769	1557	70.8%	81.4%	6.0%
Top 10 genes in signature 5*	10	8901	1273	87.5%	3712	1614	69.7%	81.4%	5.9%
Top 10 genes in signature 6*	10	8842	1332	86.9%	3769	1557	70.8%	81.4%	6.0%
Top 10 genes in signature 7*	10	8743	1431	85.9%	3565	1761	66.9%	79.4%	6.2%
Top 10 genes in signature 8*	10	8970	1204	88.2%	3869	1457	72.6%	82.8%	5.6%
Top 10 genes in signature 9*	10	8901	1273	87.5%	3712	1614	69.7%	81.4%	5.9%
Top 10 genes in signature 10*	10	8901	1273	87.5%	3712	1614	69.7%	81.4%	5.9%
Top 10 genes in signature 11*	10	8842	1332	86.9%	3769	1557	70.8%	81.4%	6.0%
Top 10 genes in signature 12*	10	8794	1380	86.4%	3739	1587	70.2%	80.9%	5.9%
Top 10 genes in signature 13*	10	8794	1380	86.4%	3739	1587	70.2%	80.9%	5.9%
Top 10 genes in signature 14*	10	8842	1332	86.9%	3769	1557	70.8%	81.4%	6.0%
Top 10 genes in signature 15*	10	8730	1444	85.8%	3552	1774	66.7%	79.2%	5.9%
Top 10 genes in signature 16*	10	8842	1332	86.9%	3769	1557	70.8%	81.4%	6.0%
Top 10 genes in signature 17*	10	8842	1332	86.9%	3769	1557	70.8%	81.4%	6.0%
Top 10 genes in signature 18*	10	8807	1367	86.6%	3518	1808	66.1%	79.5%	6.0%

Top 10 genes in signature 19*	10	8901	1273	87.5%	3712	1614	69.7%	81.4%	5.9%
Top 10 genes in signature 20*	10	8845	1329	86.9%	3723	1603	69.9%	81.1%	6.2%
Inza, Larranaga et al. 2004 (decision tree2) (124)	4	8239	1935	81.0%	3436	1890	64.5%	75.3%	6.5%
Inza, Larranaga et al. 2004 (decision tree1) (124)	5	8493	1681	83.5%	3615	1711	67.9%	78.1%	7.0%
Liu, Krishnan et al. 2005 (clustering method) (150)	6	8439	1735	82.9%	3792	1534	71.2%	78.9%	6.9%
Huang and Kecman 2005 (SVM1) (149)	10	8736	1438	85.9%	3760	1566	70.6%	80.6%	6.5%
Huang and Kecman 2005 (SVM2) (149)	10	8647	1527	85.0%	3921	1405	73.6%	81.1%	7.3%
Isabelle Guyon 2002 (SVM) (259)	7	8333	1841	81.9%	4250	1076	79.8%	81.2%	13.2%
Huang and Kecman 2005 (SVM3) (149)	10	8818	1356	86.7%	3937	1389	73.9%	82.3%	5.5%
Zhou and Mao 2005 (LS-SVM) (147)	15	8619	1555	84.7%	4601	725	86.4%	85.3%	14.9%
Ding and Peng 2005 (NB, SVM, LDA, LR) (127)	60	8934	1240	87.8%	4288	1038	80.5%	85.3%	4.7%
Bo and Jonassen 2002 (linear discriminant) (148)	50	9207	967	90.5%	4294	1032	80.6%	87.1%	4.2%

* Signatures were selected by using our system

3.3.4 Hierarchical clustering analysis of samples

Figure 3-2 shows the hierarchical cluster analysis of 62 samples based on the gene expression profiles from 104 selected genes. These 62 samples form two groups. One group contains 20 normal samples and three tumor samples T30, T33 and T36. Another group contains 37 tumor samples and 2 normal samples N34 and N36. These results are similar to Alon's results (108), showing that 35 tumor and 3 normal samples form a cluster, and the left 19 normal and 5 tumor samples form another cluster.

3.3.5 Evaluation of sample labels

Using the SVM method with leave-one-out validation, all except 6 samples were correctly classified (120). The 6 samples included three tumor samples (T30, T33, T36) which were more probably to be normal ones, whereas three normal samples (N8, N34, N36) which were more likely to be cancerous. This suggested that some samples in the colon cancer dataset might have been wrongly labeled. In our system, the average training accuracies of the best models, which were evaluated by the predictive accuracies from the test sets, was only in the range of 92.6% to 94.7%. This result from hierarchical clustering analysis also indicated that some samples might be mistakenly labeled. In microarray experiments, mistakenly labeling may happen. For example, a normal ovarian tissue sample was mistakenly labeled as cancerous tissue in an ovarian cancer microarray experiment (120). A robust model should have the ability to identify the mislabeled samples.

Using 10,000 SVM model constructed by 104 genes we selected, 5 samples were misclassified at most occasions. The three confusion tumor samples T33, T36 and T30 were misclassified as normal one in 92.9%, 92.8% and 90.1% of occasions respectively. The normal samples N36 and N34 were misclassified as cancer ones in 96.0% and 73.6% of occasions. Another normal sample N8 was misclassified in 3.4% of occasions. The misclassification rates for other samples were all less than 0.5%. Misclassification of T33, T36 and T30 into their opposite labels was actually consistent with the opinion that these were more likely normal tissues. Likewise, misclassification of N36 and N34 was consistent with the opinion that they were more likely cancerous. These results suggest that our method and derived SVM models are insensitive to incorrect labeling of a small percentage of

samples.

To reduce the effect of mislabeling samples, 6 samples (T33, T36, T30, N8, N34, and N36) was excluded, and the hierarchical clustering analysis was conducted again from the expression profiles of the 104 selected colon cancer genes on the other 56 samples. The result is displayed in Figure 3-3. At this time the samples were separated into two distinct clusters: normal people and colon cancer patients with no error. It further suggests that our selected genes can predict the sample groups accurately by using the unsupervised cluster method.

Figure 3-2 Hierarchical clustering analysis of 62 samples from the gene expression profile of 104 selected genes.

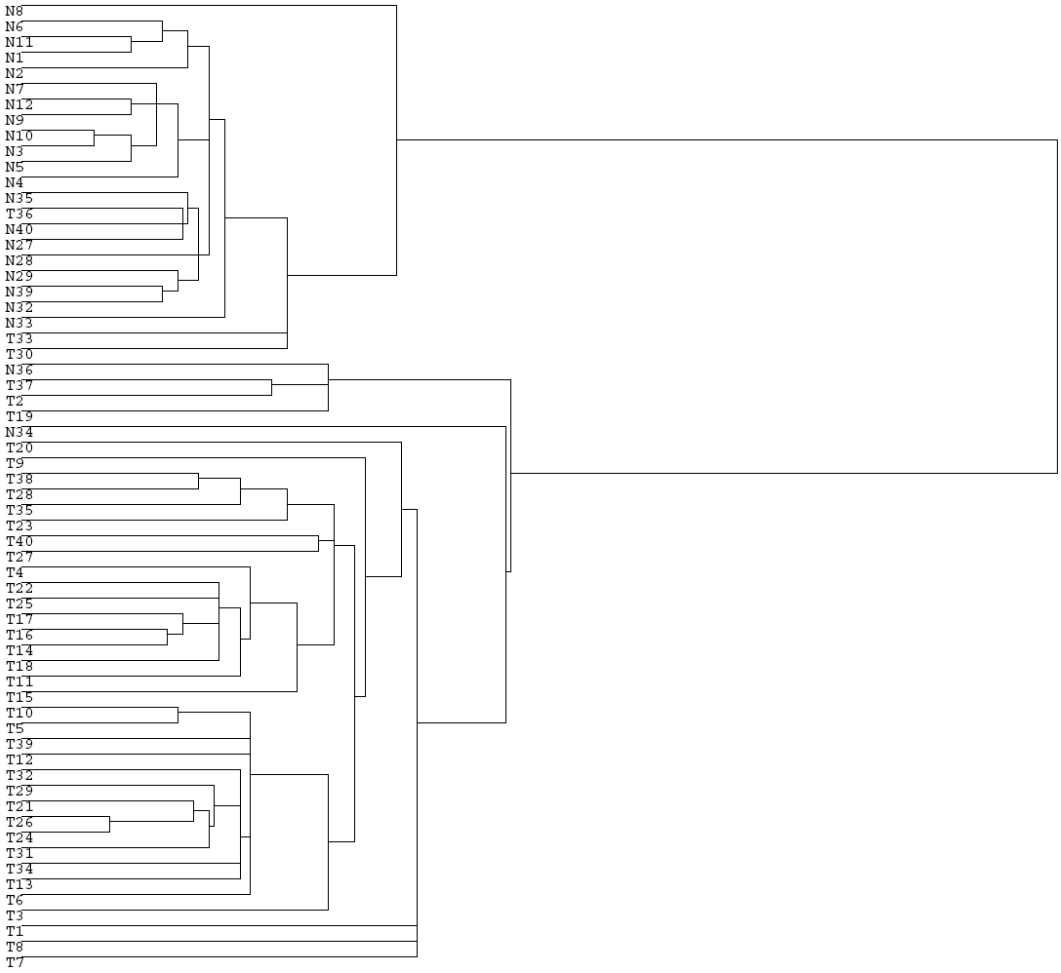
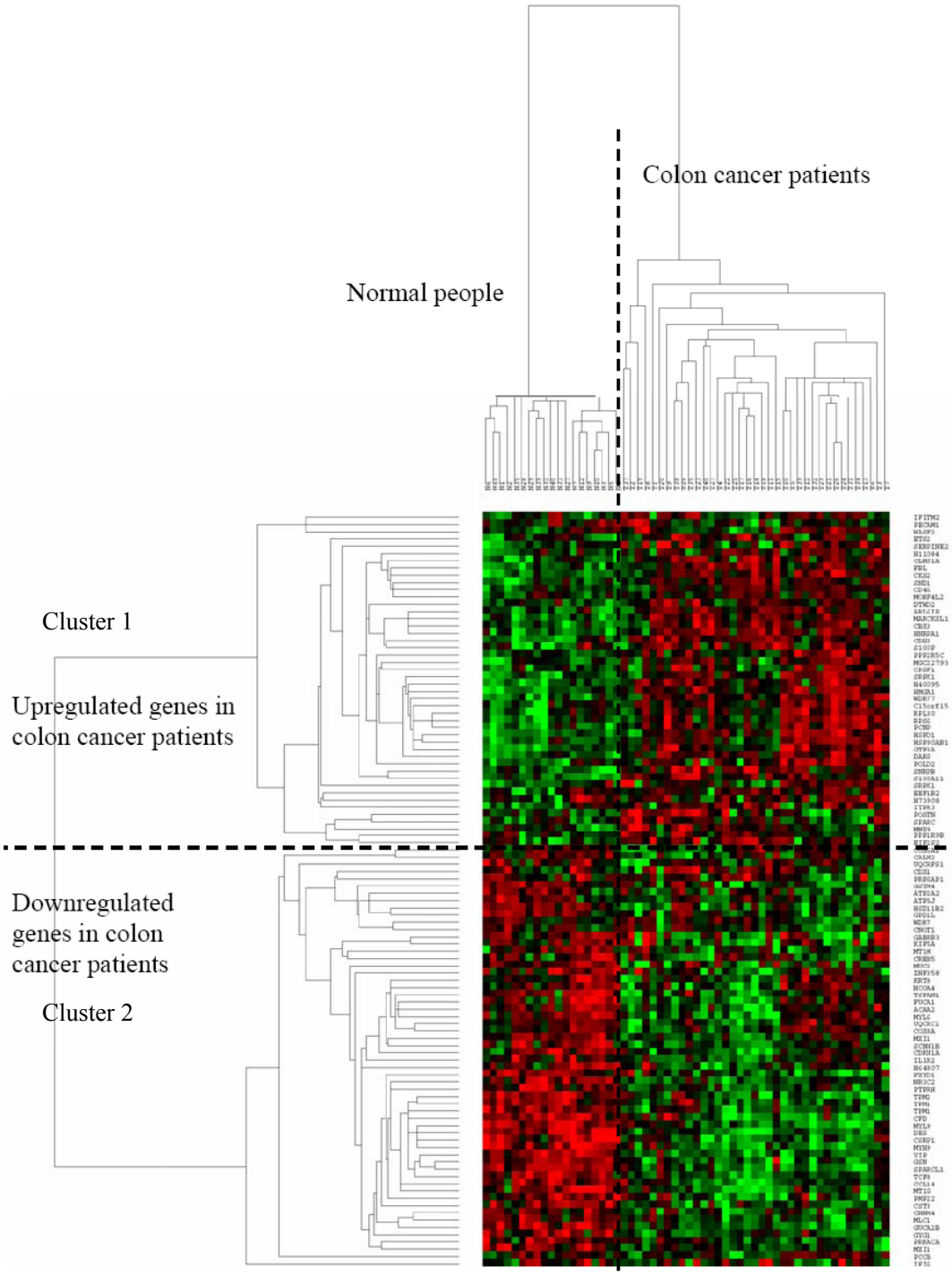


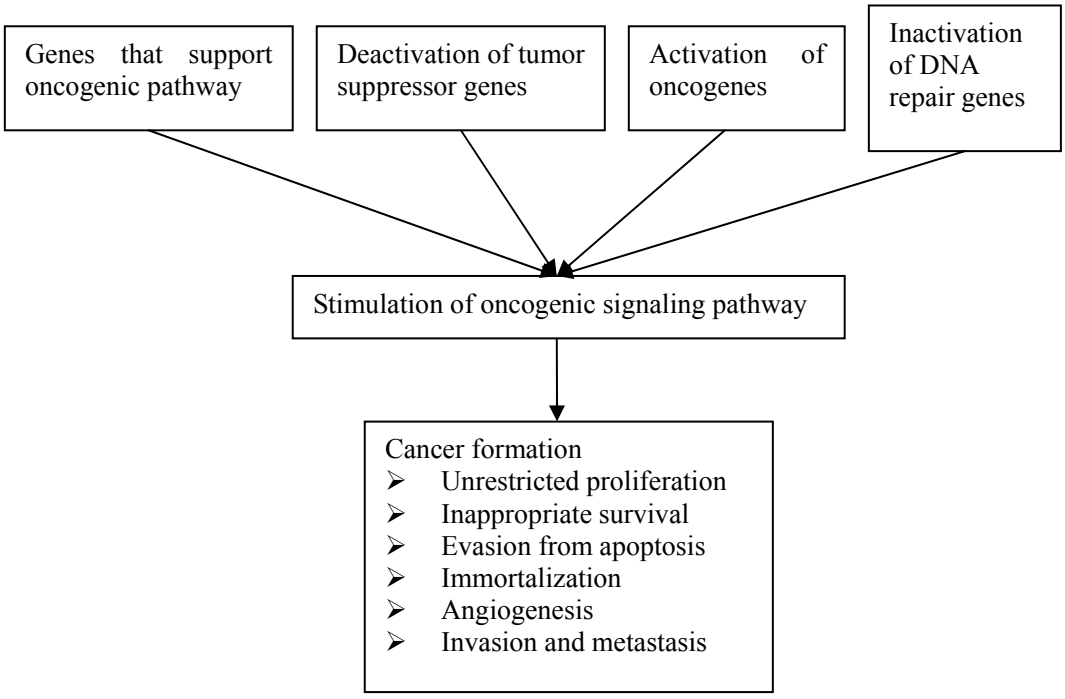
Figure 3-3 Hierarchical clustering analysis of 56 samples and 104 genes on colon cancer microarray. (Red color represents the higher relative expression level and green represents the lower relative expression level)



3.3.6 The function of the identified colon cancer markers

It is now well known that cancer is caused and driven by mutations in DNA that change the signal pathways which normally operate to regulate proliferation and death in normal cell. The activation of oncogene (drive excessive proliferation of cells) and inactivation of tumor suppressor genes (lose the inhibitory effect which is crucial to prevent inappropriate growth) change the normal signal pathway and hence leads to various well-defined phenotypic traits of cancer (Figure 3-4) (15, 16). These traits include proliferation, inappropriate survival, immortalization, invasion, angiogenesis and metastasis (15, 16). Considering such complexity of tumorigenesis, the number of cancer genes in the signatures should not be very few. It was reported that there are 291 known cancer genes (333), 15 cancer-associated pathways (334), and 34 angiogenesis genes (335, 336). Because of biological differences and complex nature of cancers, a signature applicable for many patients is expected to include a substantial percentage of these cancer-related genes, together with some of their interacting-partners and consequence-genes (333). Moreover, because of measurement variability, a certain number of irrelevant genes may be inevitably included in a signature. Therefore, it is not surprising that the number of selected predictor-genes in our signatures ranged from 112 to 157. Moreover, it is probably unrealistic to assume that only a few genes stand out from the thousands of gene with sufficient clarity allowing target selection (62), which is a very important application of gene selection from microarray analysis.

Figure 3-4 Classes of genes involved in oncogenic transformation



The selected 104 predictor-genes shared by all 20 signatures (Table 3-3, Table 3-8) include 48 cancer-related genes (6 anticancer targets, 2 oncogenes, 8 tumor-suppressors, 2 angiogenesis gene, 4 cancer-genes, 3 tumor-marker, 17 cancer-gene interacting genes, and 6 cancer-pathway-affiliated genes). In our analysis, anticancer targets were obtained from the latest version of TTD database (271, 340). Tthe cancer-related genes and cancer-pathways were taken from recent publications (333-338) and references in Table 3-3.

These 104 shared predictor-genes also include 16 genes possibly implicated in cancer (description and references in Table 3-3). They have been reported to be involved in cancer risk (GSTM4), promotion of metastatic growth (POSTN) and tumor invasion (SERPINE2), maintenance of telomere repeats in cancer cells (HNRPA1), energy metabolism in cancer cells (PCCB), regulation of cell cycle (CSRP1, EEF1B2, TSPAN1, WDR7) and gene expression (CNOT1), and

synthesis of signaling molecules with elevated levels in tumors (GUCA2B). Genes reported to have significantly altered expression level in tumors with unclear connection to cancer (PMP22, SRPK1, UQCRC1) were also included here because of their possible roles as cancer consequence-genes.

Table 3-8 List of colon cancer genes shared by all 20 signatures.

Gene Group	Predictor-genes selected by both this work and other studies (number of studies)	Predictor-genes selected by this work only
Anticancer target (successful or research)	CDKN1A(2), HSD11B2(1)	MMP9, SPARC, NCOA4, IL1R2
Cancer gene (Oncogene)	HMGA1(1), ETS2(1)	
Cancer gene (Tumor-suppressor)	MXI1(1), TPM1(2), CDH3(2)	CDX1, PRKACA, CREB5, PPP2R5C, CNNM4
Cancer gene (angiogenesis gene)		PECAM1, CCL14
Cancer gene (other types)	MYH9(7), S100P(3), HSP90AB1(1)	IFI6
Tumor marker	MUC2(1)	S100A11, KRT8
Interacting partner of cancer gene	TPM2(3), CKS2(2), EIF2S2(2), HSPD1(2), MT1G(2), GSN(1), MORF4L2(1), SND1(1), TCF8(1)	ATP2A2, CD46, CPSF1, PTPRH, WASF2, PPP1R9B, CALM2, CBX3
Cancer pathway affiliated gene	DES(7), C15orf15(3), MYL6(2)	RPS6, RPL30, POLD2
Gene having possible implication in cancer	GUCA2B(6), CSRP1(5), SPARCL1(3), HNRPA1(1), SERPINE2(1), TSPAN1(1)	CNOT1, EEF1B2, PCCB, PMP22, POSTN, SRPK1, WDR7, GSTM4, UQCRFS1, UQCRC1
Others	MYL9(7), WDR77(4), DARS(3), VIP(3), ATP5J(3), GTF3A(1), NR3C2, MLC1(1), GPD1L(1), CFD(1), ZNF358(1), FUCA1(1), GABRB3(1), SNRPB(1), DTWD2(1), CST3(1), FBL(1), MARCKSL1(1), SCNN1B(1), H11084(1), H40095(1), H64807(1)	ACAA2, ARL6IP, CLNS1A, COX6A1, COX8A, FXYD1, GYG1, IFITM2, ITPR3, KIF5A, MGC22793, MT1M, PCNP, PRPSAP1, H73908

3.3.7 Hierarchical clustering analysis of the identified markers

In Figure 3-3, the 104 colon cancer genes form 2 distinctive clusters. One is the upregulated genes in the colon cancer patients, whereas another one is the downregulated genes in colon cancer patients.

The upregulated genes included 1 successful tumor marker (S100A11), 2 research tumor targets (SPARC and MMP9), 1 angiogenesis gene (PECAM1), 2 oncogenes (ETS2 and HMGA1) and 2 tumor suppressor genes (PPP2R5C and CDH3). Since the activation of oncogenes, angiogenesis genes, tumor targets and tumor markers can promote tumorigenesis, it is not surprising that S100A11, SPARC, MMP9, PECAM1, ETS2, and HMGA1 are upregulated in colon cancer patients. Although CDH3 as a tumor suppressor gene, it was reported to be present at an increased level in colon cancer cells (300, 301). Therefore the upregulating of these genes which are associated with colon tumorigenesis is actually consistent with the experiments.

Other upregulated genes in colon cancer patients are IFITM2, WASF2, SERPINE2, H11084, CLNS1A, CKS2, SND1, CD46, MORF4L2, DTWD2, ARL6IP, MARCKSL1, CBX3, HNRPA1, MGC22793, CPSF1, SRPK1, H40095, WDR77, C15orf15, RPL30, RPS6, PCNP, HSPD1, HSP90AB1, GTF3A, DARS, POLD2 , SNRPB, EEF1B2, H73908, ITPR3, POSTN, PPP1R9B and EIF2S2.

The downregulated genes in colon cancer patients include 6 tumor suppressor genes (CDX1, CREB5, MXI1, TPM1, CNNM4 and PRKACA), 1 angiogenesis genes (CCL14), 1 successful tumor target (HSD11B2), 2 research tumor target (CDKN1A and NCOA4 and IL1R2), and 2 tumor marker (MUC2 and KRT8). Since the deactivation of tumor suppressor genes promotes tumorigenesis, tumor suppressor genes should down-express in cancer patients. It was also reported that the decrease of tumor target HSD11B2 mRNA abundance and enzyme activity was associated with colorectal cancer (287) and the expression level of tumor target DKN1 was tightly controlled by the tumor suppressor protein p53 and could

mediate the p53-dependent cell cycle G1 phase arrest in response to a variety of stress stimuli (290). Reduction of tumor marker MUC2 expression may be associated with the occurrence and progression of colorectal carcinomas (308). IL1R2 and CCL2 belong to immune tolerance genes. The cancer patients are normally associated with low immune tolerance (338). Therefore the downregulating of these genes which are associated with colon tumorigenesis is actually consistent with the experiments.

Other downregulated genes in colon cancer patients are COX6A1, CALM2, UQCRFS1, PRPSAP1, GSTM4, ATP2A2, ATP5J, GPD1L, WDR7, GABRB3, KIF5A, MT1M, ZNF358, TSPAN1, FUCA1, ACAA2, MYL6, UQCRC1, COX8A, SCNN1B, H64807, FXYD1, NR3C2, PTPRH, TPM2, CFD, MYL9, DES, CSRP1, MYH9, VIP, GSN, SPARCL1, TCF8, MT1G, PMP22, CST3, MLC1, GUCA2B, GYG1 and PCCB.

3.3.8 Therapeutic target prediction

For facilitate the identification of therapeutic targets, a therapeutic target system based on SVM classifier was developed (268), as described in Section 2.2.3 of Chapter 2. SVM classifier separates positive (therapeutic target) and negative (non-therapeutic target) training samples in a multi-dimensional space by constructing a hyper-plane optimally positioned between the positive and negative samples. A testing sample is then projected onto this multi-dimensional space to determine its class affiliation based on its relative position to the hyperplane.

The performance of SVM prediction of therapeutic targets was evaluated based on

a 5-fold cross validation study of 1,484 therapeutic targets and 6,637 non-therapeutic targets. The computed prediction accuracies for therapeutic targets and non-therapeutic targets were in the range of range 64.1–71.0% and 85.0–85.8% respectively (268). This suggested the SVM is capable of predicting therapeutic targets.

Therapeutic target prediction system was utilized to predict therapeutic targets from the identified colon cancer markers. All of the 6 known therapeutic targets (NCOA4, SPAC, MMP9, IL1R2, HSD1B2 and CDKN1A) were predicted correctly by our therapeutic target system (Table 3-9). 18 markers (PRKACA, CDH3, HSP90AB1, PECAM1, SND1, SCNN1B, PCCB, NR3C2, KIF5A, HSPD1, GYG1, GPD1L, CNOT1, CFD, CD46, CALM2, ATP2A2 and ACAA2) were predicted as therapeutic targets.

Table 3-9 Prediction results from therapeutic target prediction system

Gene Name	Gene function	therapeutic target prediction result
NCOA4	oncogene; research tumor target	target
SPARC	research tumor target	target
MMP9	research tumor target	target
IL1R2	research therapeutic target; immune tolerance gene	target
HSD11B2	successful research target	target
CDKN1A	research tumor target	target
PRKACA	tumor suppressor gene	target
CDH3	tumor suppressor gene	target
HSP90AB1	cancer gene	target
PECAM1	angiogenesis gene	target
SND1		target
SCNN1B		target
PCCB		target
NR3C2		target
KIF5A		target
HSPD1		target
GYG1		target
GPD1L		target
CNOT1		target
CFD		target

CD46		target
CALM2		target
ATP2A2		target
ACAA2		target
CDX1	tumor suppressor gene	
TPM1	tumor suppressor gene	
MXI1	tumor suppressor gene	
CREB5	tumor suppressor gene	
CNNM4	tumor suppressor gene	
PPP2R5C	tumor suppressor gene	
KRT8	tumor marker	
MUC2	tumor marker	
S100A11	tumor marker	
CCL14	angiogenesis gene, immune tolerance gene	
HMGA1	oncogene	
ETS2	oncogene	
S100P	cancer gene	
MYH9	cancer gene	
IFI6	cancer gene, immune tolerance gene	
TPM2		
TCF8		
SRPK1		
MYL6		
MGC22793		
GTF3A		
UQCRC1		
FBL		
CST3		
WDR7		
UQCRFS1		
PTPRH		
GSTM4		
GSN		
POSTN		
ARL6IP		
ATP5J		
C15orf15		
CBX3		
CKS2		
CLNS1A		
COX6A1		
COX8A		
CPSF1		
CSRP1		
DARS		
DES		
DTWD2		
EEF1B2		
EIF2S2		
FUCA1		
FXYD1		
GABRB3		

GUCA2B		
HNRPA1		
IFITM2		
ITPR3		
MARCKSL1		
MLC1		
MORF4L2		
MT1G		
MT1M		
MYL9		
PCNP		
PMP22		
POLD2		
PPP1R9B		
PRPSAP1		
RPL30		
RPS6		
SERPINE2		
SNRPB		
SPARCL1		
TSPAN1		
VIP		
WASF2		
WDR77		
ZNF358		

3.4 Summary

This chapter described a system for marker discovery. The system was designed to overcome the unstable signatures from different combination of samples and different classification method. Multiple random sampling method and consistency evaluation strategy were incorporated into the normal RFE gene selection procedure. The system was tested on colon cancer marker discovery. The results show that our selected markers could present both better stability and higher predictive performance on different microarray datasets than other signatures. 104 genes were selected in twenty groups of colon cancer gene signatures, in which the number of genes were in the range of 112 to 157. The results from SVM classification system and hierarchical clustering analysis suggest that our selected genes could perform stable and well with the variation of

samples. Our selected genes contain a significantly higher number of cancer-related genes and half of the genes selected by other groups. Therefore, our signatures tend to more closely reflect the complex nature of cancer known which involves collective actions of many genes of different functions. A therapeutic target prediction system was utilized to identify the possible therapeutic target from the selected markers. 6 known targets and 18 novel targets were identified, indicating that our gene selection system may be used to identify the therapeutic targets.

4 Lung adenocarcinoma survival marker selection

This chapter provides another case study for our gene selection system – lung adenocarcinoma survival marker selection. The predictive ability of these survival markers are evaluated by neural network models, SVM models, and unsupervised hierarchical clustering analysis from different lung adenocarcinoma datasets. Hierarchical clustering analysis and literature search are used to evaluate the expression pattern of the identified markers. A therapeutic target prediction system is applied to identify the potential therapeutic targets from the identified markers.

4.1 Introduction

The fundamental goals of disease subtype and prognostic prediction, which includes the prediction of disease stages, disease recurrence and disease survivability, are quite different from the goals of disease diagnosis. Disease subtype and prognostic prediction either predict the likelihood of disease redeveloping following an apparent resolution of a disease or to predict outcomes such as life expectancy, survivability, progression, drug sensitivity after the diagnosis of a disease (341). In order to apply proper treatment regime and ultimately extend the survival of the patients, the accurate identification of disease subtype and prognosis effect is crucial (12). However, the disease subtype differentiation and prognostic prediction towards treatment is difficult and expensive. Furthermore, the successful rate of prediction is low, and a collective expertise of professionals is demanded (29). Taking cancer as an example, they

are complex and very heterogeneous. Even for a specific cancer type, such as leukemia, a few subclasses with various phenotypes exist (5, 6). By use of traditional diagnostic technologies, the different subtypes of a specific disease are reluctant to be figured out because they tend to look alike from microscopic analysis, and share the same symptoms or markers used in the diagnostic, or simply because one or more subclasses of the disease are unknown (12). As a result, such similarity leads to misdiagnosis and improper treatment (27). Fortunately, recently developed microarray technology provides the opportunity of subtype discovery and prognostic prediction based on disease and patients molecular details (5, 6, 26-29).

Since disease is a kind of broad class, in order to be specific, lung adenocarcinoma was chosen in this study. Our rational is that the study on lung adenocarcinoma survival marker discovery and prognosis prediction will provide a platform to study other diseases in the same way.

Currently cancer accounts for about 23% of death on human (281). Among different sites of cancer origin, lung cancer is the most dangerous one. The death rate of lung cancer is 31% for men and 26% for women (282). Lung cancers can be classified as non-small cell lung cancer (NSCLC) and small-cell lung cancer (SCLC) based on simple pathomorphological criteria. The prognosis for both type of lung cancers are poor. 80% of lung cancers are diagnosed as NSCLC (342). For advanced NSCLC, average survival time is 6 months for untreated patients, and 9 months for patients treated with chemotherapy (342). 5-year survival rate is 60~70% for patients with stage I disease and zero for patients with stage IV disease (342). The NSCLC can be further classified as squamous cell carcinoma

and adenocarcinoma. The proportional of squamous cell carcinoma is around 20~25%, and The proportional of adenocarcinoma is 50~60% (342).

The lung cancer patients can be roughly stratified from the morphological assessment based on conventional sputum cytology and chest radiography. These techniques have so far not demonstrated an impact on decreasing lung-cancer mortality (343). In one study, only 41% of cases that independent lung pathologists agreed on lung adenocarcinoma subclassification (344). Recently some specific indicators, including tumor size, poor differentiation and high tumor-proliferative index, have been identified to predict the survival of lung cancer patients (345-348), However, these indicators have only limited power in survival prediction.

The development of microarray technology makes it possible to find molecular markers of lung cancer subtype and outcome prediction systematically (349-352). These markers allow new insights in the process of lung carcinogenesis, and they may provide new tools for determination of prognosis and identification of innovative treatments. However, the molecule marker selection is strongly depended on the patient samples, causing the significantly different marker signatures in different groups for lung adenocarcinoma prognosis (Table 4-1) (350, 351) and diminishing their application potential for prognosis (70). Moreover, the prognostic power of previous selected survival genes for individual patients was seldom reported in their studies (350, 351). Guo et al. provided the prediction accuracy for their selected survival genes. However, their selected survival genes were only applicable to one dataset, and the predictive power to the independent dataset was very limited (349).

Table 4-1 Statistics of lung adenocarcinoma survival marker signatures from references

Reference	Number of selected survival genes in signature	Number of genes selected by other N studies				
		4	3	2	1	0
Lu et al (353)	125	0	0	0	8	116
Chen et al (354)	16	0	0	0	0	16
Xu et al (355)	5	0	0	0	2	3
Beer et al (350)	100	0	0	0	8	92
Guo et al (349)	37	0	0	0	4	34

In this chapter, we present our proposal to identify important marker genes, which can be used to predict the survivability of individual patients with lung adenocarcinoma. Employing similar method as colon cancer gene discovery, totally 21 genes were selected after 10 times of experiments. Results show that the prediction models can accurately predict the clinical outcome for individual patients with lung adenocarcinoma by use of independent datasets. The differential expression analysis, function prediction, and literature searches of the identified marker genes implies that the 21 survival markers should play important roles in lung adenocarcinoma progress and may contain novel therapeutic targets.

4.2 Materials and Methods

4.2.1 Lung adenocarcinoma microarray datasets and data preprocess

Two independent datasets of clinical samples were used for lung adenocarcinoma survival marker gene selection and validation of the effect of our selected genes. The original gene expression profiles of patient samples have been reported in previous publications (350, 351).

The dataset for survival marker gene selection contained the gene expression

profiles from 86 primary lung adenocarcinomas (Beer's dataset) (350, 356), including 67 stage I and 19 stage III tumor, from oligonucleotide arrays seen at the University of Michigan Hospital between May 1994 and July 2000. This gene expression profile, containing 7129 gene expression levels, was obtained before surgery. 62 patients survived (survivable patients) whereas 24 patients died at last follow-up (non-survivable patients). The detailed clinical information of samples is listed in Appendix Table S3. For preprocessing, those genes with little variation (less than 2) were removed, and 6009 genes were used for survival gene selection (70, 259).

The robustness of our selected signatures in predicting survivability in lung adenocarcinomas was tested using oligonucleotide gene-expression data obtained from a completely independent lung adenocarcinoma dataset (Bhattacharjee's dataset) (351, 357). To ensure equivalent testing power and comparability of samples, 84 primary lung tumor samples of which at least 40% samples being cancer cells were selected (350). In these 84 samples, 41 patients were alive at last follow-up (survivable patients), whereas 43 died (non-survivable patients). The detailed sample clinical information is listed in Appendix Table S4.

4.2.2 Survival marker selection procedure

In order to present a statistically meaningful evaluation, signature selection was conducted based on multiple random sampling on the Beer's dataset (350). In multiple random sampling, this dataset was randomly divided into a training set containing 43 samples (including 12 poor outcome samples and 31 good outcome samples) and an associated test set containing the other 43 samples (including the

other 12 poor outcome samples and 31 good outcome samples). To reduce computational cost, 5,000 training-test sets, each containing a unique combination of samples, were generated. These 5,000 training-test sets were randomly placed into 10 sampling groups; each containing 500 training-test sets. Every sampling group was then used to derive a signature by using the similar way as colon cancer marker discovery. Finally, the 10 different signatures derived from these sampling groups were compared in order to test the level of stability of selected predictor-genes.

4.2.3 Performance evaluation of survival marker signatures

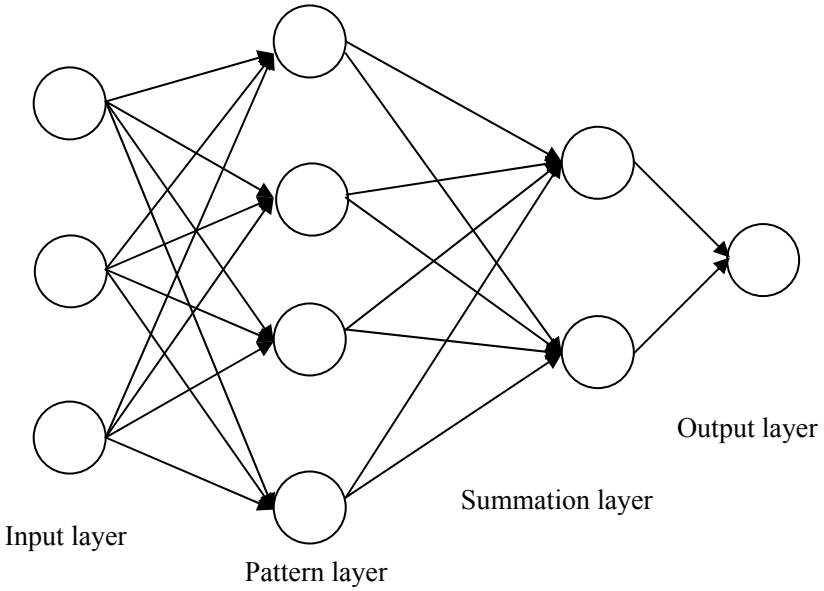
The predictive capability of survival marker signatures were evaluated by using the SVM and PNN classification system on 500 randomly-generated training-testing sets generated from the Bhattacharjee's dataset (351) and the Beer's dataset (350). For each training-testing set, the training data was used to construct a classifier model, whereas the testing data was used to evaluate the performance of the model. The predictive performance of selected signatures was evaluated by the overall accuracies (Q) of the 500 models. Besides the evaluation by using supervised classifiers, unsupervised hierarchical clustering analysis was also applied to evaluate the performance of signatures.

4.2.3.1 Neural Networks

Neural networks are another important machine learning method for microarray data analysis. This method employs a multilayered approach to approximate complex mathematical functions to process data. Probabilistic Neural Networks (PNN) is a specific form of neural networks which has 4 layers (Figure

4-1). The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of examples in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron. The summation layer has a neuron for each class and the neurons sum all the pattern neurons' output corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. The single neuron in the output layer then estimates the class of the unknown vector x by comparing all the probability density function from the summation neurons and by choosing the class with the highest probability density function. The PNN programme is provided by our group.

Figure 4-1 Architecture of neural networks



4.2.3.2 Hierarchical cluster analysis and Kaplan-Meier survival analysis

Hierarchical cluster analysis was conducted by using the selected survival on the 86 samples from Beer et al. (350) and the independent dataset consisting of 84 samples from Bhattacharjee et al. (351).

Kaplan-Meier survival analysis, often referred as survival analysis, was used in this study together with hierarchical cluster analysis. This analysis is popularly employed in medical research to estimate the percentage of patients living for a certain amount of time after surgery. It allows the estimation of survival over time, even when patients drop out or are studied for different lengths of time. A typical application of Kaplan-Meier analysis involves (1) grouping patients into different categories, and (2) comparing the survival curves from those categories by the log-rank test to assess the statistical significance of the difference among the survival curves for the categories. The Kaplan-Meier analysis was performed by using XLSTAT software (358).

4.3 Results and discussion

4.3.1 System of the lung adenocarcinoma survival marker selection

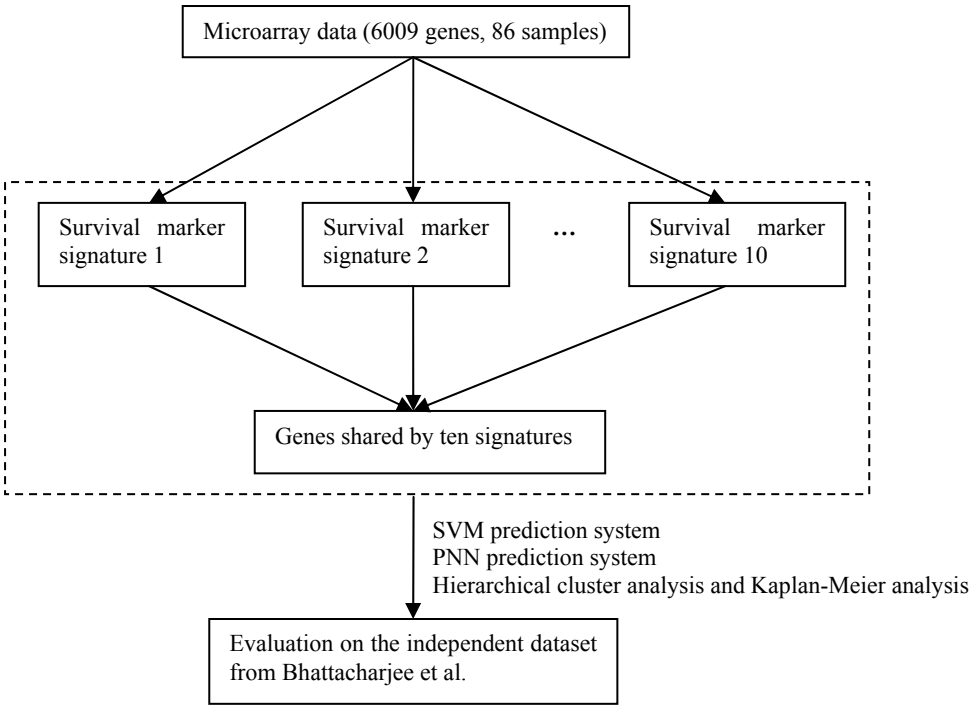
10 sets of survival marker signatures were obtained. PNN and SVM classifiers were used to evaluate the survivability prediction system constructed from selected signatures. Hierarchical cluster analysis and Kaplan-Meier analysis were used for further evaluating survival marker gene signatures, as shown in Fig. 4-2.

4.3.2 Consistency analysis of the identified markers

The stability levels of the 10 derived signatures could be estimated from the percentage of predictor-genes shared by them. The number or predictor-genes in the signatures ranged from 34 to 57 (Table 4-2, Appendices Table S5). A total of

21 predictor-genes were shared by all the 10 signatures, as shown in Table 4-2 and Table 4-3. The signature we generated had a certain level of stability when comparing to the results from 5 previous studies (Table 4-1), which shows that 5~125 selected predictor-genes in each of the 5 previous studies were seldom presented in the other 4 studies.

Figure 4-2 System for lung adenocarcinoma survival marker derivation and survivability prediction



It is noted that the numbers of selected genes in the lung adenocarcinoma dataset (21 genes) is significantly less than the number of genes from the colon cancer dataset (104 genes). One possible reason for this difference is that the expression profiles of some cancer genes important for differentiating cancer and non-cancer patients may not be significantly different in cancer patients of different survival groups. As a result, higher number of cancer genes is expected to be selected in the signatures of the colon cancer dataset than that of the lung adenocarcinoma datasets.

Table 4-2 Statistics of the lung adenocarcinoma survival markers by class-differentiation systems constructed from 10 different sampling-sets, each composed of 500 training-testing sets generated by random sampling.

Signature (method)	Number of selected survival genes in signature	Number of survival-genes also included in N other signatures derived by using different sampling-set									
		9	8	7	6	5	4	3	2	1	0
1	51	21	4	1	7	5	3	3	2	1	4
2	54	21	6	1	6	3	2	5	5	2	3
3	42	21	6	2	4	3	2	2	2	0	0
4	34	21	3	2	1	2	2	1	2	0	0
5	46	21	6	2	7	5	1	2	2	0	0
6	54	21	6	2	8	5	3	2	2	2	3
7	57	21	5	1	7	2	1	3	5	2	10
8	50	21	6	2	6	2	1	4	5	2	1
9	53	21	6	1	5	5	1	4	3	4	3
10	47	21	6	2	5	4	1	2	3	1	2

The optimal SVM parameters for the 10 sample sets were in the range of 41 to 46, and the highest average accuracies across the 10 sampling-sets were 84.1%~88.4% for non-survivable (those died at last follow-up) and 100% for survivable patients (those alive at last follow-up) respectively (Table 4-4). The accuracies for the 5,000 individual testing-sets ranged from 63.6%~100% for non-survivable and 100% for survivable patients respectively. The relatively small variations of optimal SVM parameters and prediction accuracies across the 10 sampling-sets suggest that the performance of the SVM class-differentiation systems constructed by using globally optimized parameters and RFE iteration steps are fairly stable across different sampling combinations.

Table 4-3 Gene information for lung adenocarcinoma survival markers shared by all of 10 signatures.

Gene Name	Gene description	Chromosome Location	Type	Family	Function in metagenesis	Gene Ontology: Function	Gene Ontology: Process	Pathway (from KEGG, Reactome, proteinlounge)	References
VEGF	vascular endothelial growth factor	6p12	Growth Factor	PDGF/VEGF Family of Growth Factors	Angiogenesis, therapeutic target for lung cancer therapy	extracellular matrix binding; growth factor activity; growth factor activity; heparin binding; protein binding; protein homodimerization activity; vascular endothelial growth factor receptor binding	angiogenesis; anti-apoptosis; cell migration; cell proliferation; epithelial cell differentiation; eye photoreceptor cell development; induction of positive chemotaxis; lung development; mesoderm development; multicellular organismal development; nervous system development; nervous system development; positive regulation of epithelial cell proliferation; positive regulation of vascular endothelial growth factor receptor signaling pathway; regulation of progression through cell cycle; response to hypoxia; signal transduction; vasculogenesis	VEGF Pathway; Inhibition of Angiogenesis by TSP1; eNOS Signaling; Relaxin Pathway; Phospholipase-C Pathway; CRHR Pathway; mTOR Pathway; Paxillin Interactions; PAK Pathway; Ras Pathway; Cellular Apoptosis Pathway; Rap1 Pathway; GPCR Pathway; TGF-Beta Pathway; MAPK Family Pathway; P2Y Receptor Signaling; RhoGDI Pathway ; NF-KappaB Family Pathway; FGF Pathway; HIF1Alpha Pathway; Rac1 Pathway; JAK/STAT Pathway; Renin-Angiotensin Pathway; Mitochondrial Apoptosis; NF-KappaB (p50/p65) Pathway; Telomerase Components in Cell Signaling; Rho Family GTPases	(359-364)
BSG	basigin	19p13.3			Tumor marker, angiogenesis, immunoangiostasis	mannose binding; signal transducer activity; sugar binding	cell surface receptor linked signal transduction		(365-369)
CXCL3	chemokine (C-X-C motif) ligand 3	4q21	Cytokine	Intercrine Alpha (Chemokine CXC) Family	Oncogene, immune tolerance gene, angiogenesis, organ-specific metastases	chemokine activity	G-protein coupled receptor protein signaling pathway; chemotaxis; immune response; inflammatory response	Rho Family GTPases	(370-373)
CHRNA2	cholinergic receptor, nicotinic, alpha 2 (neuronal)	8p21	Receptor, Transporter, Neurotransmitter	Ligand-Gated Ionic Channel (TC 1.A.9) Family; autocrine growth factors	therapeutic target for lung cancer therapy	acetylcholine receptor activity; extracellular ligand-gated ion channel activity; ion channel activity; nicotinic acetylcholine-activated cation-selective	ion transport; signal transduction; synaptic transmission		(364, 374, 375)

						channel activity			
FUT3	fucosyltransferase 3	19p13.3				transferase activity, transferring glycosyl groups	carbohydrate metabolic process; protein amino acid glycosylation		(376-379)
FXYD3	FXYD domain containing ion transport regulator 3	19q13.11-q13.12	ion channel activity, chloride channel activity			chloride channel activity; chloride ion binding; ion channel activity	chloride transport; ion transport		(380, 381)
PLD1	phospholipase D1	3q26	Signal Transduction	PLD Family		hydrolase activity; phosphoinositide binding; phospholipase D activity; protein binding	Ras protein signal transduction; cell communication; chemotaxis; lipid catabolic process; metabolic process; phospholipid metabolic process	Ras pathway; Rho Family GTPases; RhoA Pathway ;Rac1 Pathway; Endothelin-1 Signaling Pathway	(382, 383)
POLD3	polymerase (DNA-directed), delta 3, accessory subunit	11q14				DNA binding; delta DNA polymerase activity; transferase activity	DNA synthesis during DNA repair; mismatch repair	DNA polymerase; Purine metabolism; Pyrimidine metabolism; Cell Cycle (Mitotic); DNA Repair; DNA Replication; Maintenance of Telomeres	(384)
PRKACB	protein kinase, cAMP-dependent, catalytic, beta	1p36.1	Kinase	Ser/Thr Family of Protein Kinases (cAMP Subfamily)		ATP binding; cAMP-dependent protein kinase activity; magnesium ion binding; nucleotide binding; protein kinase activity; protein serine/threonine kinase activity; transferase activity	G-protein signaling, coupled to cAMP nucleotide second messenger; protein amino acid phosphorylation; signal transduction	Apoptosis; Calcium signaling pathway; Gap junction; GnRH signaling pathway; Hedgehog signaling pathway; Insulin signaling pathway; Long-term potentiation; MAPK signaling pathway; Olfactory transduction; Taste transduction; Wnt signaling pathway; PKA pathway(17333334)	(385)
CXCR7	chemokine (C-X-C motif) receptor 7	2q37.3			Immune tolerance gene, therapeutic target for lung cancer therapy, organ-specific metastases	receptor activity; rhodopsin-like receptor activity	G-protein coupled receptor protein signaling pathway; biological_process; signal transduction		(364, 370, 373, 386, 387)
REG1A	regenerating islet-derived 1 alpha	2p12				sugar binding	positive regulation of cell proliferation		(388, 389)

RPS3	ribosomal protein S3	11q13.3-q13.5	Structural Protein	S3P Family of Ribosomal Proteins.	involved in DNA repair pathway and apoptosis pathway, interacted with metastasis suppressor nm23	RNA binding; structural constituent of ribosome	translation	DNA repair pathway and apoptosis pathway	(390, 391)
SERPINE1	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	7q21.3-q22	Metabolic	SERPIN Family	angiogenesis				(392-395)
SLC2A1	solute carrier family 2 (facilitated glucose transporter)	1p35-p31.3	Transport	Sugar Transporter (Subfamily-Glucose Transporter)	providing energy to rapidly dividing tumor cells,	glucose transporter activity; protein binding; sugar porter activity; transporter activity	carbohydrate transport; glucose transport		(396)
SPRR1B	small proline-rich protein 1B	1q21-q22	structural molecule activity			protein binding, bridging; structural molecule activity	epidermis development; keratinization; keratinocyte differentiation; peptide cross-linking		(397, 398)
TUBA4A	tubulin, alpha 4a	2q35	Structural	Tubulin Superfamily	angiogenesis		GTP binding; GTPase activity; nucleotide binding; protein binding; structural molecule activity		(399-401)
VDR	vitamin D (1,25-dihydroxyvitamin D3) receptor	12q13.11	Transcription Factor	Nuclear Hormone Receptor Family (NR1 Subfamily)	Research tumor target	metal ion binding; protein binding; sequence-specific DNA binding; steroid hormone receptor activity; transcription factor activity; vitamin D3 receptor activity; zinc ion binding	calcium ion homeostasis; calcium ion transport; intestinal absorption; multicellular organismal development; negative regulation of transcription; organ morphogenesis; regulation of transcription, DNA-dependent; signal transduction; skeletal development; transcription	MAPK	(402)

ADFP	Adipose differentiation-related protein	9p22.1							
ANXA8	annexin A8	10q11.2				calcium ion binding; calcium-dependent phospholipid binding	blood coagulation		
GALNT4	GalNAc transferase 4	12q21.3-q22		GalNAc-transferase family		calcium ion binding; manganese ion binding; sugar binding; transferase activity, transferring glycosyl groups	carbohydrate metabolic process		
LDHB	lactate dehydrogenase B	12p12.2-p12.1	Metabolic	Ldh Family		L-lactate dehydrogenase activity; oxidoreductase activity	anaerobic glycolysis; tricarboxylic acid cycle intermediate metabolic process		

Table 4-4 Average survivability prediction accuracy of 500 SVM class-differentiation systems on the optimal SVM parameters for lung adenocarcinoma prediction. The sigma is the optimal SVM parameter which gives the best average class-differentiation accuracy over the 500 testing-sets. The accuracies are obtained from 500 testing-sets.

Signature (method)	Optimal SVM parameter	Number of selected survival genes in signature	Non-survivable patients			Survivable patients			Q
			TP	FN	SE	TN	FP	SP	
1	45	51	5275	814	86.6%	14908	3	100%	96.1%
2	44	54	5175	939	84.6%	14886	0	100.%	95.5%
3	43	42	5173	909	85.1%	14918	0	100%	95.7%
4	41	34	5347	802	87.0%	14845	6	100%	96.2%
5	43	46	5340	703	88.4%	14956	1	100%	96.6%
6	45	54	5230	865	85.8%	14905	0	100%	95.9%
7	45	57	5139	972	84.1%	14889	0	100%	95.4%
8	46	50	5201	949	84.6%	14850	0	100%	95.5%
9	43	53	5313	801	86.9%	14886	0	100%	96.2%
10	46	47	5333	757	87.6%	14910	0	100%	96.4%

4.3.3 The predictive ability of identified markers

The predictive capabilities of our selected and 10 previously-derived signatures were evaluated by using the SVM and PNN classification system on 500 randomly-generated training-testing sets generated from the Bhattacharjee's dataset (351) and the Beer's dataset (350). For each training-test set, the training data was used to construct a classifier model, whereas the test data was used to evaluate the performance of the model. The predictive performance of selected signatures was evaluated by the overall accuracies (Q) of the 500 models.

Table 4-5 gives the prediction accuracies from the SVM prediction system constructed by independent samples from Bhattacharjee's dataset (351) using our selected gene signatures and 9 other previous-derived signatures. The accuracies for non-survival patients, survival patients and all patients of the our selected 10 gene signatures over the 500 test sets were in the range of 77.8% to 81.2%, 74.3%

to 80.1% and 77.6% to 80.2% correspondingly, while the standard deviations of the accuracy of all patients were in the range of 4.7% to 4.9%. The accuracies for non-survival patients, survival patients and overall accuracies for all patients of the 21 survival genes shared by all of the 10 signatures over the 500 test sets were 78.9%, 76.8% and 77.9% respectively, while the standard deviation of the accuracy of all patients was 4.8%. In contrast, the accuracies for non-survivable patients, survival patients and all patients of the 9 previous-derived signatures were in the range of 70.1%~77.2%, 56.4% to 78.0% and 66.0% to 77.6% respectively, and the standard deviations of the accuracy of all patients were in the range of 5.5%~6.9%. These results suggest that the performance of our signatures is stabler than those of other signatures by using independent database and by applying the SVM models.

Table 4-6 illustrates the predictive performance of the 500 testing sets by using the PNN classification system and the 500 randomly generated training-testing dataset generated from the independent Bhattacharjee's dataset (351) using our selected genes. The accuracies for non-survivable patients, survival patients and all patients of our selected 10 signatures over the 500 test sets were, respectively, in the range of 69.3% to 80.2%, 64.5% to 78.0% and 69.1% to 76.6%, while the standard deviations of the accuracy of all patients were in the range of 4.2% to 4.9%. The accuracies for non-survivable patients, survivable patients and all patients of the 21 survival genes shared by all the 10 signatures over the 500 test sets were 75.2%, 62.6% and 69.2% respectively, while the standard deviation of the overall accuracy was 4.4%. The accuracies for non-survivable patients, survival patients and patients of the 9 previous-derived signatures were in the range of 53.5%~95.6%, 45.7% to 76.8% and 64.0% to 71.7% respectively, while

the standard deviations of accuracy of all patients were in the range of 4.7%~6.1%. These results recommend that the survival genes we selected have a slightly better consistency and stabler predictive performance than those of the signatures selected by other studies with the PNN classification system.

Table 4-5 Average survivability prediction accuracy of the 500 SVM class-differentiation systems constructed by 84 samples from independent Bhattacharjee's lung adenocarcinoma dataset (351) using each of the signatures derived from this study and 9 previous studies. The accuracies were obtained from the 500 testing-sets.

Signature (method)	Number of selected survival genes in signature	Non-survivable patients			Survivable patients			Q	STDEV
		TP	FN	SE	TN	FP	SP		
1*	51	8495	2369	78.2%	7864	2272	77.6%	77.9%	4.8%
2*	54	8602	2262	79.2%	7783	2353	76.8%	78.0%	4.7%
3*	42	8745	2119	80.5%	8014	2122	79.1%	79.8%	4.8%
4*	34	8452	2412	77.8%	7837	2299	77.3%	77.6%	4.8%
5*	46	8723	2141	80.3%	8117	2019	80.1%	80.2%	4.9%
6*	54	8600	2264	79.2%	7731	2405	76.3%	77.8%	4.7%
7*	57	8802	2062	81.0%	7807	2329	77.0%	79.1%	4.8%
8*	50	8414	2450	77.4%	7533	2603	74.3%	75.9%	4.8%
9*	53	8655	2209	79.7%	7992	2144	78.8%	79.3%	4.7%
10*	47	8823	2041	81.2%	7899	2237	77.9%	79.6%	4.8%
Genes selected by all sampling sets*	21	8571	2293	78.9%	7788	2348	76.8%	77.9%	4.8%
Beer et al (350)	100	8287	2577	76.3%	7540	2596	74.4%	75.4%	6.2%
Beer et al (350)	50	7616	3248	70.1%	7407	2729	73.1%	71.5%	6.3%
Chen et al (354)	16	7755	3109	71.4%	7255	2881	71.6%	71.5%	6.6%
Chen et al (354)	5	7684	3180	70.7%	6820	3316	67.3%	69.1%	6.4%
Guo et al (349)	37	8088	2776	74.4%	7443	2693	73.4%	74.0%	6.4%
Guo et al (349)	8	8386	2478	77.2%	7904	2232	78.0%	77.6%	6.6%
Lu et al (353)	125	8348	2516	76.8%	7588	2548	74.9%	75.9%	5.8%
Lu et al (353)	64	8237	2627	75.8%	7612	2524	75.1%	75.5%	5.5%
Xu et al (355)	5	8141	2723	74.9%	5720	4416	56.4%	66.0%	6.9%

* Data from this study

Table 4-6 Average survivability prediction accuracies of the 500 PNN class-differentiation systems constructed by 84 samples from independent Bhattacharjee's lung adenocarcinoma dataset (351) using each of the signatures derived from this study and 9 previous works.

Signature (method)	Number of selected survival genes in signature	Non-survivable patients			Survivable patients			Q	STDEV
		TP	FN	SP	TN	FP	SE		
1*	51	7769	3156	71.1%	7270	2805	72.2%	71.6%	4.5%
2*	54	7837	3088	71.7%	7478	2597	74.2%	72.9%	4.9%
3*	42	8762	2163	80.2%	7333	2742	72.8%	76.6%	4.6%
4*	34	8656	2269	79.2%	6810	3265	67.6%	73.6%	4.3%
5*	46	7995	2930	73.2%	7863	2212	78.0%	75.5%	4.6%
6*	54	8019	2906	73.4%	6502	3573	64.5%	69.1%	4.5%
7*	57	8177	2748	74.8%	7518	2557	74.6%	74.7%	4.4%
8*	50	8000	2925	73.2%	7514	2561	74.6%	73.9%	4.2%
9*	53	7575	3350	69.3%	7140	2935	70.9%	70.1%	4.6%
10*	47	8379	2546	76.7%	7413	2662	73.6%	75.2%	4.7%
Genes selected by all sampling sets*	21	8217	2708	75.2%	6305	3770	62.6%	69.2%	4.4%
Beer et al (350)	100	7537	3388	69.0%	7515	2560	74.6%	71.7%	5.5%
Chen et al (354)	5	10446	479	95.6%	4600	5475	45.7%	71.6%	4.7%
Guo et al (349)	8	7752	3173	71.0%	7189	2886	71.4%	71.1%	5.2%
Guo et al (349)	37	7537	3388	69.0%	7284	2791	72.3%	70.6%	5.5%
Xu et al (355)	5	7884	3041	72.2%	6844	3231	67.9%	70.1%	5.6%
Beer et al (350)	50	9220	1705	84.4%	5310	4765	52.7%	69.2%	4.9%
Chen et al (354)	16	6780	4145	62.1%	7734	2341	76.8%	69.1%	5.7%
Lu et al (353)	125	6874	4051	62.9%	7591	2484	75.3%	68.9%	6.1%
Lu et al (353)	64	5845	5080	53.5%	7591	2484	75.3%	64.0%	6.1%

* Data from this study

The predictive accuracies of the 500 SVM survivability prediction systems from the original Beer' dataset (350) are shown in Table 4-7. These 500 training sets and 500 test sets were different from those used for survivability gene signatures selection. The accuracies for non-survivable patients, survival patients and all patients of the 10 survival gene signatures over the 500 test sets were in the range of 94.2% to 96.1%, 99.8 to 100% and 98.3% to 98.9% respectively, and the

standard deviations of accuracy of all patients were in the range of 3.2~3.7%. The accuracies for non-survival patients, survival patients and all patients of the 21 survival genes shared by all the 10 signatures over the 500 test sets were 90.5%, 99.5% and 96.9% respectively, and the standard deviation of the accuracy of all patients was 4.0%. The performances of our selected genes were both higher and stabler than those of the other 9 studies, in which the accuracies for non-survivable patients, survival patients and all patients were in the range of 52.5% to 66.6%, 81.8% to 96.8% and 75.6% to 88.3% respectively, and the standard deviations of accuracy of all patients were in the range of 5.8% to 8.0%. These results suggest that the survival gene signatures we selected can perform better than those selected by other studies by using the SVM classification system. Furthermore, our selected genes give stabler predictive performance demonstrated by low standard deviation values.

The predictive accuracies of the 500 PNN classification systems for survivability prediction from the original Beer' dataset (350) are shown in Table 4-8. The accuracies for non-survivable patients, survival patients and all patients of the 10 survival gene signatures over the 500 test sets were in the range of 79.6% to 89.8%, 95.9% to 98.9% and 93.4% to 95.5% respectively, and the standard deviations (STDEV) were in the range of 4.3% to 5.2%. The accuracies for non-survivable patients, survival patients and all patients of the 21 survival genes shared by all the 10 signatures over the 500 test sets were 75.1%, 96.2% and 90.2% respectively, and the standard deviations of the overall accuracy was 5.7%. In contrast, the accuracies for non-survivable patients, survival patients and all patients of the 9 gene signatures from other studies over the 500 test sets were in the range of 57.2% to 76.1%, 73.5% to 89.7% and 72.1% to 80.6% respectively,

and the standard deviation were in the range of 7.5% to 11.0%. This comparison indicated that the performance of our selected gene signatures is better and stabler than those of other studies using the PNN classification methods for survivability prediction.

Table 4-7 Average survivability prediction accuracy of 500 SVM class-differentiation systems constructed by 86 samples from Beer's lung adenocarcinoma dataset (350).

Signature (method)	Number of selected survival genes in signature	Non-survival patients			Survivable patients			Q	STDEV
		TP	FN	SE	TN	FP	SP		
1*	51	5589	342	94.2%	15047	22	99.9%	98.3%	3.4%
2*	54	5671	260	95.6%	15043	26	99.8%	98.6%	3.2%
3*	42	5622	309	94.8%	15061	8	99.9%	98.5%	3.5%
4*	34	5630	301	94.9%	15037	32	99.8%	98.4%	3.3%
5*	46	5679	252	95.8%	15039	30	99.8%	98.7%	3.5%
6*	54	5664	267	95.5%	15054	15	99.9%	98.7%	3.7%
7*	57	5678	253	95.7%	15059	10	99.9%	98.7%	3.4%
8*	50	5694	237	96.0%	15069	0	100%	98.9%	3.3%
9*	53	5702	229	96.1%	15047	22	99.9%	98.8%	3.3%
10*	47	5686	245	95.9%	15052	17	99.9%	98.8%	3.3%
Genes selected by all sampling sets *	21	5369	562	90.5%	14987	82	99.5%	96.9%	4.0%
Beer et al (350)	100	3951	1980	66.6%	14589	480	96.8%	88.3%	5.8%
Beer et al (350)	50	3302	2629	55.7%	14134	935	93.8%	83.0%	6.7%
Lu et al (353)	64	3526	2405	59.5%	13658	1411	90.6%	81.8%	6.4%
Lu et al (353)	125	3467	2464	58.5%	13570	1499	90.0%	81.1%	6.2%
Guo et al (349)	37	2760	3171	46.5%	13974	1095	92.7%	79.7%	7.0%
Chen et al (354)	16	2925	3006	49.3%	13702	1367	90.9%	79.2%	7.0%
Xu et al (355)	5	3696	2235	62.3%	12432	2637	82.5%	76.8%	7.5%
Chen et al (354)	5	3577	2354	60.3%	12325	2744	81.8%	75.7%	8.0%
Guo et al (349)	8	3113	2818	52.5%	12760	2309	84.6%	75.6%	7.3%

* Data from this study

Table 4-8 Average survivability prediction accuracies of the 500 PNN class-differentiation systems constructed by 86 samples from Beer's lung adenocarcinoma dataset (350).

Signature (Method)	No. of selected predictor genes in signature	Non-survivable patients			Survivable patients			Q	STDEV
		TP	FN	SE	TN	FP	QN		
1*	51	5069	862	85.5%	14635	434	97.1%	93.8%	4.8%
2*	54	5062	869	85.3%	14726	343	97.7%	94.2%	4.6%
3*	42	4939	992	83.3%	14715	354	97.7%	93.6%	4.7%
4*	34	4719	1212	79.6%	14904	165	98.9%	93.4%	5.2%
5*	46	5210	721	87.8%	14798	271	98.2%	95.3%	4.5%
6*	54	5326	605	89.8%	14730	339	97.8%	95.5%	4.3%
7*	57	5214	717	87.9%	14533	536	96.4%	94.0%	4.9%
8*	50	5089	842	85.8%	14707	362	97.6%	94.3%	4.5%
9*	53	5319	612	89.7%	14450	619	95.9%	94.1%	4.4%
10*	47	5100	831	86.0%	14571	498	96.7%	93.7%	4.8%
Genes selected by all sampling sets*	21	4454	1477	75.1%	14495	574	96.2%	90.2%	5.7%
Beer et al (350)	50	3393	2538	57.2%	13523	1546	89.7%	80.6%	7.5%
Beer et al (350)	100	4183	1748	70.5%	12648	2421	83.9%	80.1%	9.0%
Lu et al (353)	64	4515	1416	76.1%	11700	3369	77.6%	77.2%	10.0%
Xu et al (355)	5	4205	1726	70.9%	11960	3109	79.4%	77.0%	7.5%
Chen et al (354)	5	3601	2330	60.7%	11985	3084	79.5%	74.2%	7.9%
Guo et al (349)	8	3743	2188	63.1%	11768	3301	78.1%	73.9%	8.2%
Chen et al (354)	16	3569	2362	60.2%	11936	3133	79.2%	73.8%	7.8%
Lu et al (353)	125	4310	1621	72.7%	11078	3991	73.5%	73.3%	12.5%
Guo et al (349)	37	3903	2028	65.8%	11232	3837	74.5%	72.1%	11.0%
* Data from this study									

4.3.4 Patient survival analysis using survival markers

Hierarchical cluster analysis can cluster the samples according to their expression profiles across the gene we selected. The comparison of the survival curves from these clusters can be used to assess the statistical significance of the survivability difference among the clusters.

By using 21 identified markers, hierarchical cluster analysis grouped 86 lung adenocarcinoma patients in the Beer's dataset (350) into three clusters (Figure

4-3). Kaplan-Meier survival analysis showed that the survival time after therapy was significantly different in the three patient clusters ($P<0.0001$, log-rank test, Figure 4-4). Cluster 1 was the poor prognosis group. The average survival time of patients in this cluster was 50.6 months. In this cluster, the numbers of survivable patients (SP) and non-survivable patients (NSP) were 12 and 14 respectively (Table 4-9). The survival percentage, which defined by SP/(SP+NSP), were 46%. Cluster 2 was the good prognosis groups with average survival time of 82.2 months. The SP, NSP and survival percentage were 26, 1 and 96% respectively. Cluster 3 was the moderate prognosis group with average survival time of 74.8 months. The SP, NSP and survival percentage were 22, 9 and 72% respectively. By using the similar way, Guo et al (349) clustered these samples (350) into three clusters by using 37 genes and the survival percentages were 69%, 72% and 75% for poor, moderate and good prognosis clusters, respectively (Table 4-9). The survival percentage for three clusters generated by 100 genes in Beer et al (350) are 43%, 57% and 88% for poor, moderate and good prognosis clusters, respectively (Table 4-9). These results indicate that the 21 genes selected by using our method can be classified into better clinically meaningful groups for further prognosis than the genes selected by other group.

Hierarchical clustering of the 21 genes on the independent validation dataset - Bhattacharjee's dataset (351) showed the similar results (Figure 4-5). Three clusters had significant difference by using Kaplan-Meier analysis with $P<0.001$ from log-rank test (Figure 4-6). The average survival time for cluster 1, which was poor prognosis group, was 35.7 months. The average survival time for cluster 2, which was moderate prognosis group, was 32.0 months. The average survival time for cluster 3, which was good prognosis group, was 78.3 months. The survival

percentages of the three clusters were 30%, 43 % and 73% for poor, moderate and good prognosis clusters, as shown in Table 4-10. By using the similar strategies, Guo et al (349) clustered the sample into three clusters. However, the survivability percentages among the clusters were 45%, 46% and 51% for three clusters by using the Kaplan-Meier analysis, showing little statistically different among the clusters (Table 4-9). The survival percentage of three clusters formed by 21 genes we selected were more spread out than those formed by the genes selected by other researchers, further suggesting that 21 genes we selected have robust behavior for prognosis prediction.

Table 4-9 Comparison of the survival rate in clusters with other groups, by using different signatures and Beer's microarray dataset (350).

Study	Gene number in signatures	Poor prognosis cluster			Moderate prognosis cluster			Good prognosis cluster		
		SP[1]	NSP[2]	Survival rate[3]	SP	NSP	Survival rate	SP	NSP	Survival rate
This study	21	12	14	46%	22	9	72%	26	1	96%
Guo's group (349)	37	25	11	69%	15	6	71%	20	7	74%
Beer's group (350)[4]	100	25	19	43%	23	19	57%	37	5	88%

[1] SP: the number of survivable patients
[2] NSP: the number of non-survivable patients
[3] Survival rate= SP/(SP+NSP)
[4] The cluster analysis was done on 128 lung cancer samples

Figure 4-3 Hierarchical clustering analysis of the 21 lung adenocarcinoma survival markers from Beer's microarray dataset (350). The tumor samples were aggregated into three clusters. Substantially elevated (red) and decreased (green) expression of the genes is observed in individual tumors.

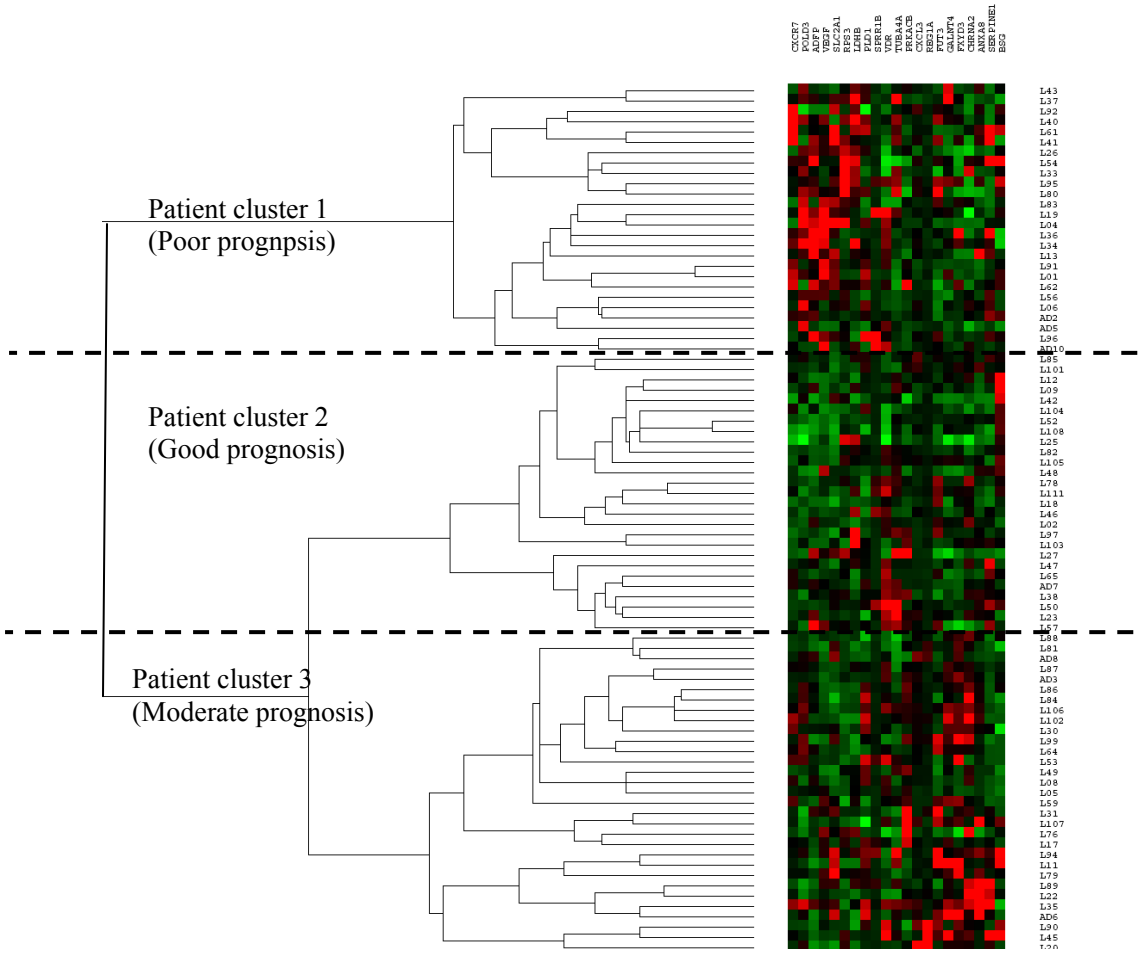


Figure 4-4 Kaplan-Meier survival analysis of the three clusters of patients from Figure 4-3. Average survival time of patients in cluster 1 is 50.6 months; average survival time of patients in cluster 2 is 82.2 months; average survival time of patients in cluster 3 is 74.8 months (P<0.0001, log-rank test).

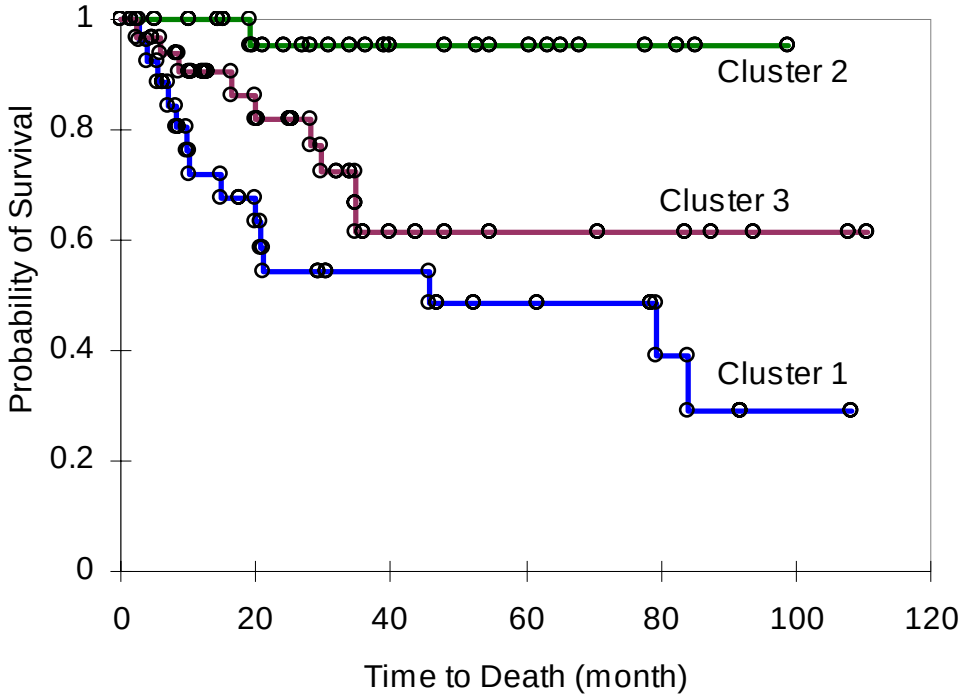


Figure 4-5 Hierarchical clustering analysis of the 21 lung adenocarcinoma markers from Bhattacharjee's microarray dataset (351). The tumor samples were aggregated into three clusters. This 21-gene signature are shared by 10 survival genes sets of lung adenocarcinoma derived by using datasets from Beer et al (350) and by using multiple random sampling method.

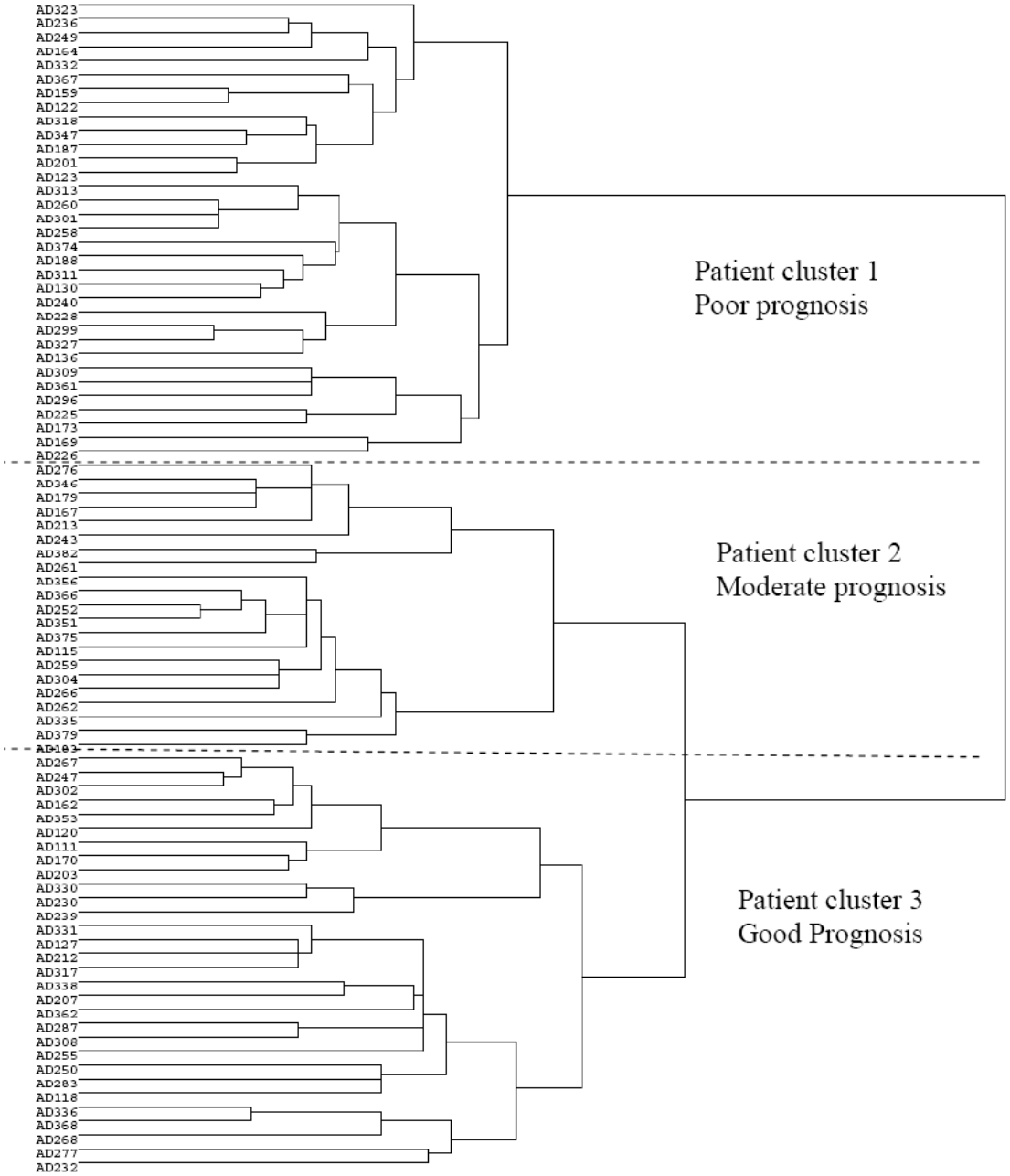


Figure 4-6 Kaplan-Meier survival analysis of the three clusters of patients from Figure 4-5. Average survival time of patients in cluster 1 is 35.7 months; average survival time of patients in cluster 2 is 32.0 months; average survival time of patients in cluster 3 is 78.3 months (P<0.001, log-rank test).

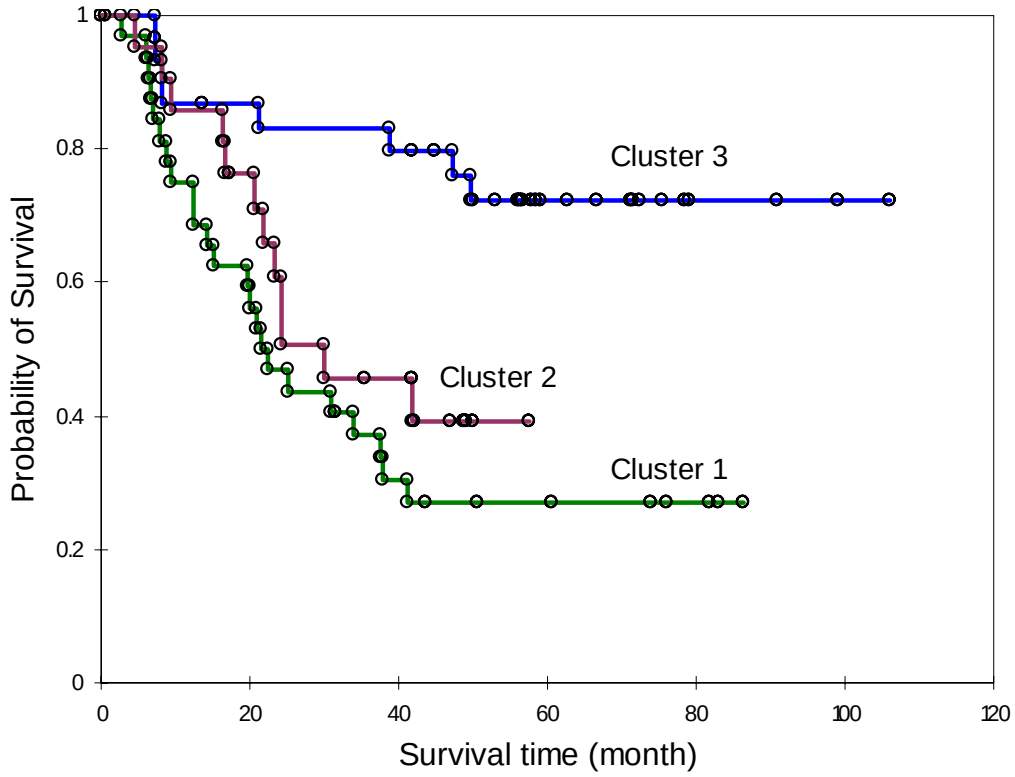


Table 4-10 Comparison of the survival rate in clusters with other groups, by using different signatures and Bhattacharjee's microarray dataset (351).

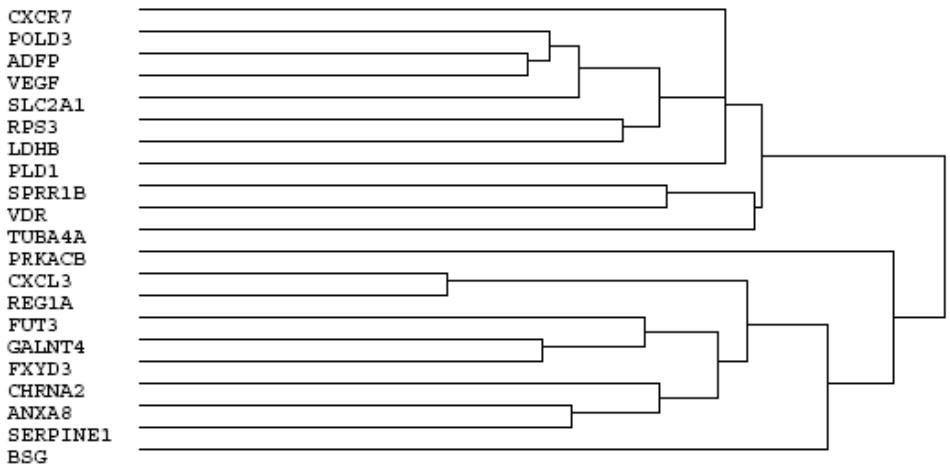
Study	Gene number in signatures	Poor prognosis cluster			Moderate prognosis cluster			Good prognosis cluster		
		SP	NSP	Survival rate	SP	NSP	Survival rate	SP	NSP	Survival rate
This study	21	10	23	30%	9	12	43%	22	8	73%
Guo's group (349)	37	9	11	45%	11	13	46%	20	19	51%

4.3.5 Hierarchical clustering analysis of the survival markers

In the hierarchical analysis for 86 lung adenocarcinoma patients in the Beer's dataset (350), 21 survival genes were formed into two clusters (Figure 4-7). Genes in gene cluster 1 are correlated with a poor prognosis of lung adenocarcinoma,

whereas genes in cluster 2 are correlated with a moderate prognosis of lung adenocarcinoma. Figure 4-3 shows that higher expression of the genes in cluster 1 is associated with poor prognosis in patients in lung adenocarcinoma, and higher expression of the genes in cluster 2 is associated with moderate prognosis in patients in lung adenocarcinoma. On the other hand, the lower expression of all these 21 genes in both cluster 1 and 2 is associated with good prognosis. The 11 poor-prognosis genes in cluster 1 are CXCR7, POLD3, ADFP, VEGF, SLC2A1, RPS3, LDHB, PLD1, SPRR1B, VDR, and TUBA4A, of which four genes, VEGF, CXCR7, TUBA4A and VDR, are therapeutic tumor targets. The 10 moderate-prognosis genes in cluster 2 consist of PRKACB, CXCL3, REG1A, FUT3, GALNT4, FXYD3, CHRNA2, ANXA8, SERPINE1 and BSG. CHRNA2 is a molecular target for lung cancer therapy. The target information was obtained from the latest version of therapeutic target database (270, 271),

Figure 4-7 Hierarchical clustering analysis of the 21 lung adenocarcinoma survival markers from Beer's microarray dataset (350)



Most of the selected genes were experimentally proved that high expression of these genes was related to adverse survivability of patients. High level of serum

VEGF (vascular endothelial growth factor) in the NSCLC may directly predict worse survival (403, 404), and acts as a crucial parameter in lung cancer, especially associated with NSCLC (403, 404). The expression of VDR (Vitamin D receptor) was observed in lung adenocarcinoma (405). Increased SLC2A1 (solute carrier family 2) expression in tumors was identified as an adverse prognostic factor and a predictive prognosis marker (406). Elevated PLD1 (phospholipase D1) activity could promote tumor progress and show high invasive potential (383, 407-409). Up-regulated expression of FXYD3 (FXYD domain containing ion transport regulator 3) in cancer indicated that FXYD3 might contribute to the proliferative activity of malignancy (380). In vivo experiments demonstrated that BSG (basigin; CD147) overexpression stimulated tumor angiogenesis and growth (367). Higher expression of FUT3 (fucosyltransferase 3) was often observed in high grade and poor prognosis tumors (410). The expression level of SERPINE1 (plasminogen activator inhibitor-1) in tissue was significantly and positively correlated with tumor severity and tumor size (411), and high level of SERPINE1 could indicate an aggressive phonotype of carcinomas (412, 413), serving as an indicator of poor prognosis in adenocarcinomas of the lung (414). REG1A (regenerating islet-derived 1 alpha) expression was reported to be closely related to the carcinoma invasiveness of neoplasm (415), and to be an independent predictor of overall cancer patient survival as well (416). The over-expression of SPRR1B (small proline-rich protein 1B (cornifin)) for prolonged periods might disrupt normal progression of mitosis (397). Therefore, the expression of most of our selected survival marker genes has been validated as either directly or closely related to cancer metastasis and prognosis in the literatures.

4.3.6 Therapeutic target prediction of survival markers

4.3.6.1 The prediction results

Therapeutic target prediction system (268), which was developed by Han et al, can be applied to predict the function of the survival markers. The detailed information and performance evaluation of this system were introduced in Chapter 2 (Section 2.2.3) and Chapter 3 (Section 3.3.8). Table 4-11 shows the prediction results. All of the five known therapeutic targets (VEGF, CXCR7, CHRNA2, TUBA4A and VDR) were predicted correctly. 7 markers (SERPINE1, PLD1, RPS3, BSG, ANXA8, LDHB and SLC2A1) were predicted as novel therapeutic targets.

4.3.6.2 The function of the known targets

The five known targets play important roles for lung cancer survivability from the literature searches. It was reported that VEGF induction might promote angiogenesis in lung adenocarcinoma (417), and genetic variations in VEGF might predict both carcinoma risk and tumor aggressiveness (418). CXCR7 (chemokine receptor 7) (364, 370, 386, 387) and CHRNA2 (cholinergic receptor, nicotinic, alpha 2) (364, 374, 375) are molecular targets for lung cancer therapy. CXCR7 (chemokine (C-X-C motif) receptor 7) has properties that affect a spectrum of biological and pathological processes, including cell growth/survival and adhesion, as well as promotion of tumor growth (386). TUBA4A (Tubulin 4A), as a successful anticancer target (419), can moderate drug resistance in lung carcinoma (400). VDR (Vitamin D receptor) polymorphisms may be associated with improved survival among SCC patients of early-stage NSCLC (420).

Table 4- 11Prediction result from therapeutic target prediction system

Gene name	Gene function	Prediction status
VEGF	therapeutic target for lung cancer therapy, angiogenesis gene	target
CXCR7	therapeutic target for lung cancer therapy, immune tolerance gene	target
CHRNA2	therapeutic target for lung cancer therapy	target
TUBA4A	successful tumor target, angiogenesis gene	target
VDR	research tumor target, cancer gene	target
SERPINE1	angiogenesis gene	target
PLD1	cancer gene	target
RPS3	cancer gene	target
BSG	tumor marker, angiogenesis (stimulate angiogenesis genes), immunoangiostasis	target
ANXA8		target
LDHB		target
SLC2A1		target
SPRR1B	cancer gene	
CXCL3	oncogene, immune tolerance gene, angiogenesis gene	
PRKACB		
ADFP		
FUT3		
FXYD3		
GALNT4		
POLD3		
REG1A		

4.3.6.3 The function of the predicted novel targets

Literatures also indicated the important roles of the 4 of the 7 novel targets (SERPINE1, PLD1, RPS3 and BSG) in lung adenocarcinoma progression. Currentky there is no biological evidence that the other 3 novel targets (ANXA8, LDHB and SLC2A1) involve in lung cancer progress .

SERPINE (plasminogen activator inhibitor-1), a component of plasminogen/plasmin system, is an important player in tumor growth, invasion and metastasis, through the regulation of cellular proteolysis, adhesion, proliferation, migration, and processes closely related to the pathogenesis of lung

injury and neoplasia (421-423). Therefore, SERPINE have previously been suggested as prognostic markers in NSCLC (414, 423).

PLD1 (Phospholipase D1), which is recognized as a receptor-regulated signaling enzyme that can modulate many cellular functions, plays an important role in signal transduction (424). Endogenous PLD1 is a critical factor in the organization of the actin-based cytoskeleton, with regard to cell adhesion and migration (382) and a critical downstream mediator of H-Ras-induced tumor formation. PLD1 is critical in the oncogenic ability of Ras (425). Elevated PLD activity generates survival signals allowing cells to overcome default apoptosis programs (383) and contributes to the cell's high invasive potential in a protein phosphorylation-independent manner (409).

RPS3 (ribosomal protein S3), a component of the 40S ribosomal subunit of eukaryotes, plays a role as a base damage endonuclease. It induces apoptosis whose signal is executed through the activation of caspase-8 followed by caspase-3 activation, and increases the proapoptotic potential of cytokines(390). It was suggested that DNA repair pathway and apoptosis pathway might cross-talk via RPS3 (390). RPS3 inhibits tumor invasion via blocking the ERK pathway and MMP-9 secretion. The interaction of RPS3 and nm23-H1, a metastasis suppressor, may be critical in this inhibition (391).

BSG (Basigin) mediates tumor-stroma interactions and directly contributes to tumor invasion, metastasis, tumor angiogenesis and growth by stimulating extracellular matrix remodeling around tumor cell clusters, stroma, and blood vessels (366) and by stimulating VEGF and MMP expression (367).

The literature searching indicated that our gene selection method could identify therapeutic targets. The therapeutic target prediction system can be incorporated into this system to predict the novel therapeutic targets.

4.3.6.4 The function of other survival markers

The other identified survival markers may also involve in cancer progress. FUT3 plays an important role in organ-specific metastasis (426). FXYD3 plays an important role in cellular growth of carcinomas (380). PRKACB (protein kinase, cAMP-dependent) plays different roles in proliferation and differentiation and could be potential markers for cancer progression (385). SPRR1B is likely coupled to signals responsible for withdrawal from the proliferative state (397). The analysis showed that the function of the identified survival markers might have direct impact on cancer development in the literatures.

4.4 Summary

In this chapter, the comprehensive gene selection system was further evaluated on the selection of survival marker for lung adenocarcinoma. By way of multiple random sampling, 21 genes were selected by all of ten sets of lung adenocarcinoma survival marker signatures, in which 34 to 53 genes were selected. These 21 markers were then used to develop PNN and SVM prediction models to predict prognosis for lung adenocarcinoma patients from different datasets. The survivability analysis by hierarchical clustering analysis and Kaplan-Meier survival analysis further suggested that the derived signatures from our system

could provide better performance when comparing with other signatures. Most of the selected genes have been experimentally proved that high expression of the genes is relevant to adverse survivability of patients. 12 markers, including 5 known targets and 7 novel targets, were successfully predicted as therapeutic targets by using a therapeutic target prediction system.

5 The development of bioinformatics tools for disease targeting antibody prediction

An important application for gene selection from microarray data is to discover potential disease targets which can be used for therapeutic molecule design and achieve the goal of treatment and prevention of disease. Antibody, as a very effective therapeutic molecule, was chosen in this chapter as an example of therapeutic molecular. A computational tool for therapeutic antibody prediction was developed in this chapter. First, we developed an immunoinformatics database - antibody-antigen information resource (AAIR) (Section 5.2). This database provides information about the known antibody and its corresponding antigen together with the targeted disease and the diagnosis and therapeutic indications (Section 5.3). Then, a statistical analysis of this database is presented (Section 5.4), which helps to explain the trends of antibody therapeutic development. Finally, a preliminary SVM prediction model is built for therapeutic antibody design (Section 5.5). Such information may provide useful hints about the current trends for the exploration of the antibodies and for disease treatment.

5.1 Introduction

Targeted therapy is a type of treatment which is based on the idea that therapeutic molecules will attack their specific molecular targets involved in pathogenesis and disease progress without damaging other tissue (155, 156). Currently, antibodies, as a frequently used form of therapeutic molecule, can specifically act on the disease-causing targets (antigens) (15) on many diseases

such as cancer (16), heart disease (160) and rheumatic diseases (161).

The discovery of potential target genes is greatly facilitated by microarray technology (30), which was also shown in Chapters 3 and 4. Microarray can be used to discover the upregulated genes in disease status or bad prognosis. Those upregulated genes with important biological clues give a strong indication to act as potential targets for therapeutic molecular design, including antibody design (427).

The research and application of therapeutic antibodies grow very fast, and antibody is the second largest class of drugs (167). This chapter covers the usage of antibody as an example for therapeutic molecular design. As a well-established drug class, the successful rate of antibody therapeutics is 18−29% from the first use in humans to regulatory approval (162), much higher than 11% of successful rate for small-molecule drugs (159).

However, the development of antibody has not been an easy task because the behavior of antibodies seems to vary, even though they have similar structures (177). The explosive growth in biotechnology combined with major advances in informatics technology has the potential to radically transform immunology (196). Publicly accessible resources, which include the rapidly increasing number of databases of immunoinformatics and computational tools, can be used for antibody design.

All therapeutic applications of antibodies are based on their ability to recognize specific target molecules – antigens. Much effort has been done to understand the

antibody-antigen interaction for generation and optimization of antibodies to improve their potential in the prevention and treatment of disease. The rapid advances in informatics technology and computational technology may be helpful to understand the antibody-antigen interaction (428, 429). Since the interaction between targets (antigens) and their corresponding therapeutic molecules (antibodies) are very important for targeted therapy, in this chapter we constructed a bioinformatics database emphasizing on the sequence-recognition of antibody and antigens and their disease indication. A computational tool can be developed based on the information from the database.

5.2 The development of antibody information database

5.2.1 The objective of the AAIR development

Molecular-level information about antibody-antigen (Ab-Ag) recognition is critically important because it can help to understand the mechanism of immune responses, to discover new vaccines, antibody-based drugs and diagnostic tools (167). A number of immunological databases have been developed (KABAT (197), IMGT (200), IEDB (205), FIMM (202), MMDB (203), JenPep (204), SACS (209), BCIPEP (214), VBASE2 (215), TumorAntigen database (216), AntiJen (220), HaptenDB (206), Epitome (207) and CED (221) etc). These databases provide valuable information about the antibodies and antigens, such as sequences (IMGT, KABAT, FIMM, BCIPEP), structures (IMGT, IEDB, MMDB, SACS), epitope information (IEDB, FIMM, Epitome, CED), binding information (IEDB, JenPep, AntiJen, Epitome) and disease implication (IMGT, FIMM). However, it tends to be difficult to extract the information of targeted diseases, the therapeutic

indications and sequence-level recognition data (i.e. which antibody sequence recognizes which antigen sequence or sequences) from these databases. Although other database such as the epitome database (207) contains sequence-specific information about antibody and antigen interactions, it only covers a limited number of antigen-antibody pairs obtained from protein Databank (194). As a result, there is a need to develop a database capable of providing both easily accessible information and more comprehensive coverage of sequence-specific Ab-Ag recognition to complement existing databases.

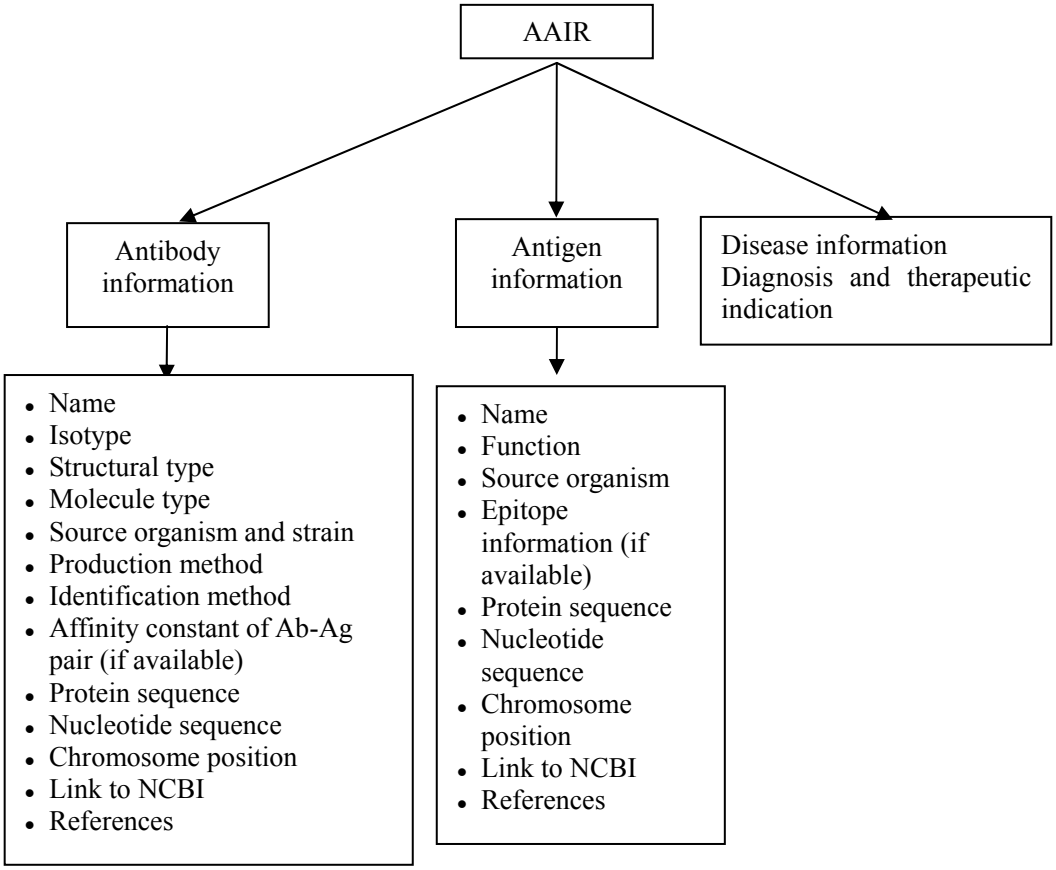
5.2.2 The collection of related information

Three classes of information should be included: The first class is antigen information that includes antigen name sequences, function and source organism. The second is antibody information that contains antibody isotype, source organism, molecular and structural type of antibody. The third is disease and therapeutic information that describes disease class, targeted disease, diagnosis and therapeutic indication. Figure 5-1 shows the detailed information included in AAIR.

The entries of the database were obtained by a comprehensive search of published literatures via Medline (277). Specifically, in order to collect the data for AAIR development, keywords such as “antibody”, “antibodies”, and “immunoglobulin” were used to search those literatures containing both the antibody sequence and the corresponding antigen or antigens. Meanwhile, the sequence of the corresponding antigen or antigens were obtained either from the respective literatures (if they are described in that literature) or from the protein sequence

Figure 5-1 Structure of AAIR

(http://bidd.nus.edu.sg/group/antibody/antibody.asp)



database Swiss-Prot based on the name and host species of the antigen or antigens. Other information such as disease and therapeutic indications was retrieved by using keywords either as a combination of "disease" and the name of antibody (if available) or as a combination of "disease", "antibody" and antigen name.

5.2.3 The construction of AAIR database

AAIR is a relational database, which represents the antibody-antigen interaction database in the form of two-dimension tables. The two-dimensional tables include antibody-antigen pair ID table (Table 5-1), antibody-antigen pair

main information table (Table 5-2), antibody-antigen pair data type table (Table 5-3), antibody or antigen information table (Table 5-4), antibody or antigen data type table (Table 5-5), and reference information table (Table 5-6). In these tables, antibody-antigen pair ID serves as primary key; antibody-antigen pair data type ID, antibody or antigen ID, antibody or antigen Data type ID and reference ID are considered as foreign keys.

Table 5-1 Antibody-antigen pair ID table

Antibody-antigen pair ID	The antigen name of this pair
AAIR0001	HEL
AAIR0002	AahII

Table 5-2 Antibody-antigen pair main information table

Antibody-antigen pair ID	Antibody-antigen pair data type ID	Antibody-antigen pair data content
AAIR0001	101	AA5523
AAIR0001	102	AA5524
AAIR0001	103	AA1396
AAIR0001	104	6806606
AAIR0001	106	Food Allergy

Table 5-3 Antibody-antigen pair data type table

Antibody-antigen pair data type ID	Data type
101	Antibody heavy chain ID
102	Antibody light chain ID
103	Antigen ID
104	Reference
105	Disease indication

Table 5-4 Protein information table

Antibody or antigen ID	Antibody or antigen Data type ID	Data content
AA5523	101	AAA39270
AA5523	102	Mus musculus (house mouse)

Table 5-5 Protein data type table

Antibody or antigen Data type ID	Data content
101	Locus
102	Source Organism

Table 5-6 Reference information table

Reference ID	Reference
6806606	Kobayashi T, Fujio H et al, A monoclonal antibody specific for a distinct region of hen egg-white lysozyme. Mol Immunol. 1982 Apr;19(4):619-30

5.2.4 The interface of the AAIR database

Basically AAIR web interface comprise four layers, and the top layer is the main graphical user interface with a querying tool for finding specific entries. The searching results followed by some specific matching rules will be listed in the second layer. By clicking into each pair entry, the browser can access the detailed information for this interacted pairs, which is displayed in the third layer. More information about antibody and antigen is given in the fourth layer.

The AAIR database can be found at the website http://bidd.nus.edu.sg/group/antibody/antibody.asp. Entries of this database are searchable by several methods. These methods include the search of antigen information (names or source organisms), antibody information (isotype, source organism, molecular type or structure type), disease and therapeutic information (disease classes, disease names, or diagnosis and therapeutic indications). The disease classes are defined by the international classification of diseases from World Health Organization (430). The diseases names and diagnosis and therapeutic indications are derived from the related terms described in the relevant

publications. Full list of antigen names, antibody information, disease names, disease classes, and diagnosis and therapeutic indications are provided on the web page for facilitating the search of particular entries.

Moreover, case-insensitive keyword-based text search and wildcards are also supported. In a query, one can specify full name or part of the name in a text field. For instance, wild characters of '*' and '?' are allowed in the text field. In this case, '?' represents any single character, and '*' represents a string of characters of any length. As an example, input of ‘phosphatase’ in the field of antigen name enables the search of all entries containing the antigen name of ‘phosphatase’ such as heat stable alkaline phosphatase, acid phosphatase, anti-placental alkaline phosphatase antibody, small guanosine triphosphatase (GTPase) Rab6 etc. As another example, input of ‘heat*phosphatase’ enables the finding of all phosphatases whose names start with ‘heat’.

The outcome of a typical search result is illustrated in Figure 5-2. In this interface, all entries that satisfy search criteria are listed along with AAIR antibody-antigen pair entry ID, antibody name (extracted from the original publications), antigen name, targeted disease and diagnosis/therapeutic indication. More detailed information of an antibody-antigen pair entry can be obtained by clicking the corresponding pair ID. The result is displayed in an interface shown in Figure 5-3, from which one may find the three classes of information. The first class is antibody information which includes antibody sequence, isotype, structure and biochemical type, source organism/strain, production and identification method, and affinity constant with relevant antigen. The second is antibody-targeted antigen information which consists of antigen name, function, source organism

and epitope position or sequence. The third is disease and therapeutic information of the retrieved antibody which includes targeted disease, and diagnosis and therapeutic indications. For completeness, the relevant references are provided in the interface.

Further information about each antibody or antigen entry can be retrieved from an interface by clicking the corresponding entry ID, as illustrated in Figure 5-4. From this interface, one can locate antibody or antigen name, entry ID, source organism and strain/isolate, protein sequence and NCBI Entrez protein (277) ID, related DNA sequences and Entrez nucleotide (277) ID. Antibody entry provides additional information such as tissue type, development stage, cell line, cell type, clone, the express system, cell line/strain and host plasmid/vector, while antigen entry provides functional information. Similarly, related references and links to Entrez protein and nucleotide are provided in the interface.

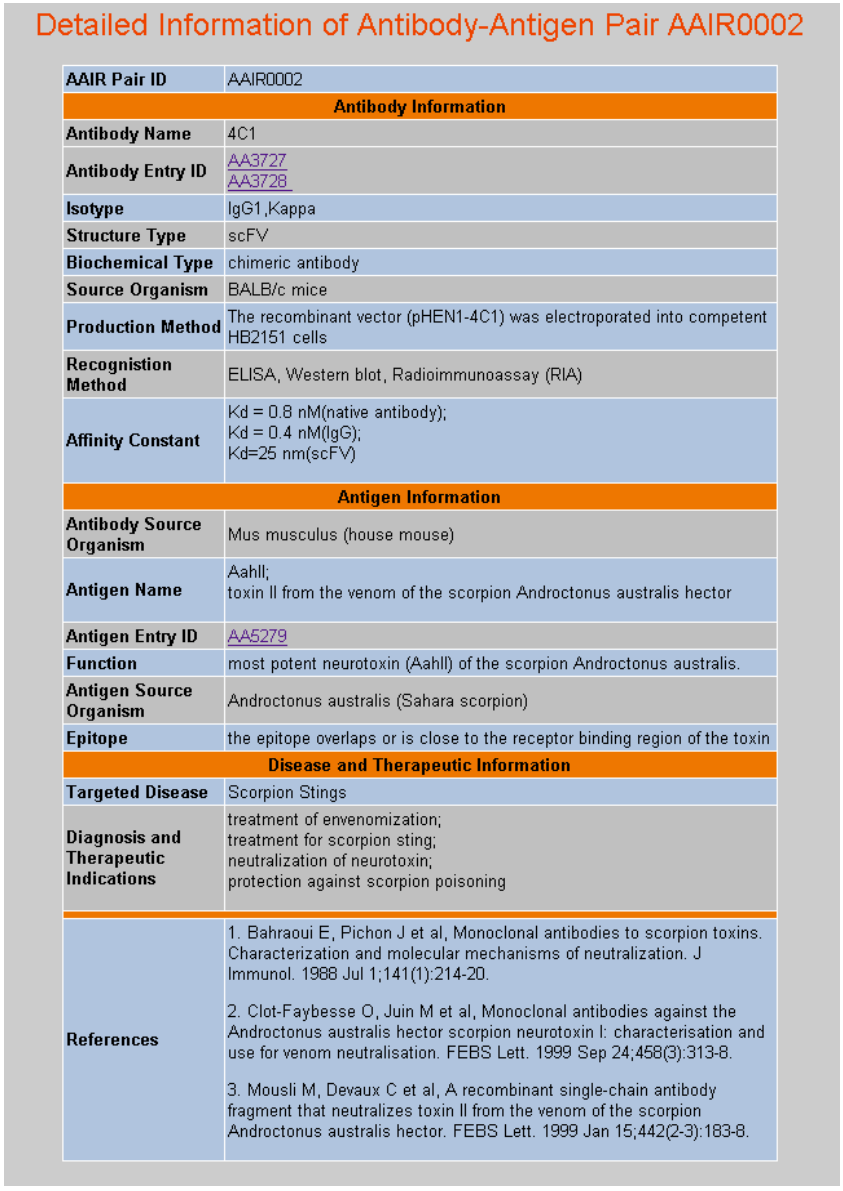
Figure 5-2 The interface displaying a research result on AAIR. All entries that satisfy the specified search criteria are listed along with the Ab-Ag pair ID, antibody name, antigen name, targeted disease type and diagnosis and therapeutic indications in this database.

You searched for: Apolipoprotein

<<First <Previous Page 1 of 2 Next> Last>>

Pair ID	Antibody Name	Antigen Name	Targeted Disease Type	Diagnosis and Therapeutic Indication
AAIR 0926	mAb(a)23	apolipoprotein A	Myocardial Infarction; Atherosclerotic Vascular Disease; Lupus; Stroke; Alzheimer's Disease	diagnosis of Systemic Lupus Erythematosus
AAIR 0927	mAb(a)20	apolipoprotein A	Myocardial Infarction; Atherosclerotic Vascular Disease; Lupus; Stroke; Alzheimer's Disease	diagnosis of Systemic Lupus Erythematosus
AAIR 1001	B55	apolipoprotein B-100	Myocardial Infarction; Atherosclerotic Vascular Disease; Stroke; Alzheimer's Disease	clinical assay for atherogenesis
AAIR 1057	B9	apolipoprotein B-100	Myocardial Infarction; Atherosclerotic Vascular Disease; Stroke; Alzheimer's Disease	clinical assay for atherogenesis
AAIR 1263	2e8Fab	apolipoprotein H	Myocardial Infarction; Atherosclerotic Vascular Disease; Stroke; Alzheimer's Disease; Amyloidosis	diagnosis of systemic amyloidosis; immunohistochemical studies of systemic amyloidosis
AAIR 2487	unknown	apolipoprotein E	Myocardial Infarction; Atherosclerotic Vascular Disease; Stroke; Alzheimer's Disease; Amyloidosis	immunohistochemical studies of systemic amyloidosis
AAIR 3956	CAR	apolipoprotein H	Myocardial Infarction; Atherosclerotic Vascular Disease; Stroke; Alzheimer's Disease; Amyloidosis	diagnosis systemic amyloidosis; immunohistochemical studies of systemic amyloidosis
AAIR 3957	CAM	apolipoprotein H	Myocardial Infarction; Atherosclerotic Vascular Disease; Stroke; Alzheimer's Disease; Amyloidosis	diagnosis systemic amyloidosis; immunohistochemical studies of systemic amyloidosis
			Myocardial Infarction;	

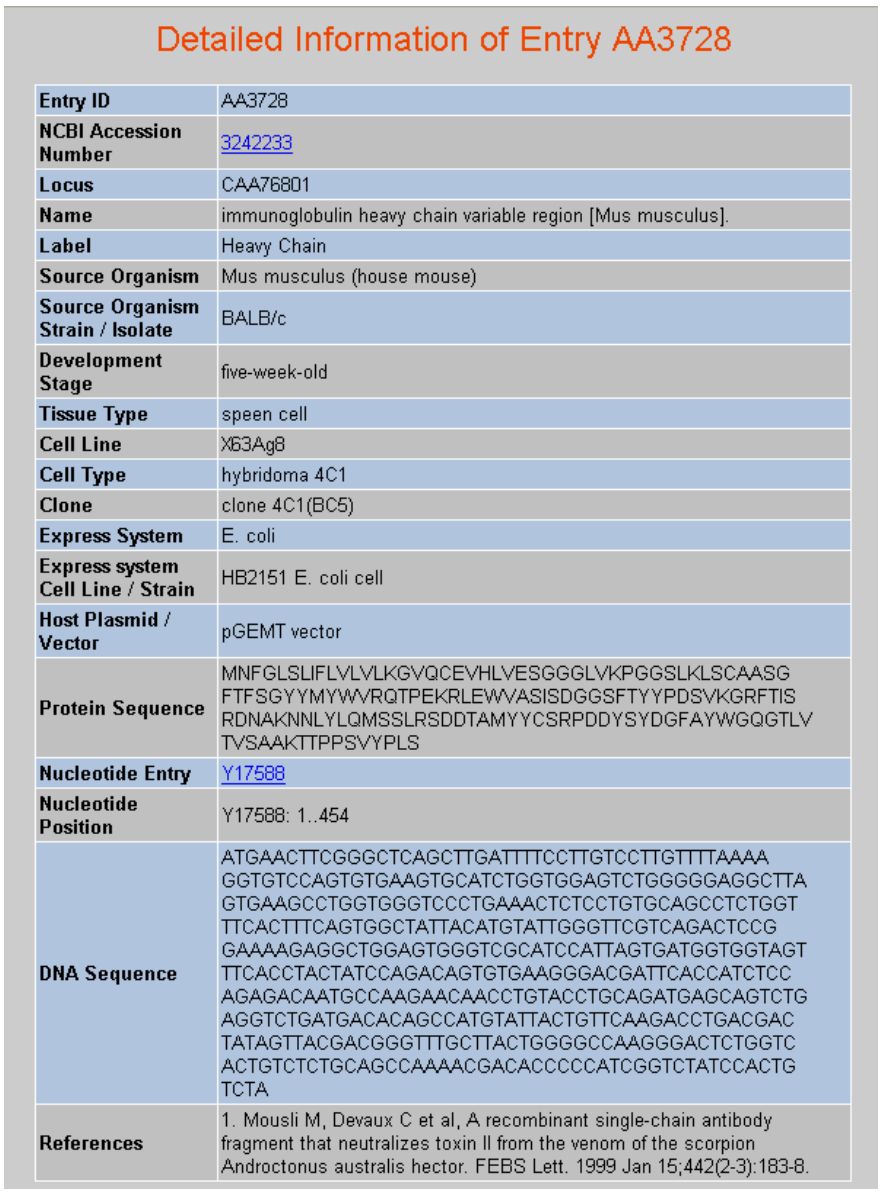
Figure 5-3 Interface displaying the detailed information of an antibody-antigen pair in the AAIR



Detailed Information of Antibody-Antigen Pair AAIR0002

AAIR Pair ID	AAIR0002
Antibody Information	
Antibody Name	4C1
Antibody Entry ID	AA3727 AA3728
Isotype	IgG1,Kappa
Structure Type	scFV
Biochemical Type	chimeric antibody
Source Organism	BALB/c mice
Production Method	The recombinant vector (pHEN1-4C1) was electroporated into competent HB2151 cells
Recognistion Method	ELISA, Western blot, Radioimmunoassay (RIA)
Affinity Constant	Kd = 0.8 nM(native antibody); Kd = 0.4 nM(IgG); Kd=25 nm(scFV)
Antigen Information	
Antibody Source Organism	Mus musculus (house mouse)
Antigen Name	AahII; toxin II from the venom of the scorpion Androctonus australis hector
Antigen Entry ID	AA5279
Function	most potent neurotoxin (AahII) of the scorpion Androctonus australis.
Antigen Source Organism	Androctonus australis (Sahara scorpion)
Epitope	the epitope overlaps or is close to the receptor binding region of the toxin
Disease and Therapeutic Information	
Targeted Disease	Scorpion Stings
Diagnosis and Therapeutic Indications	treatment of envenomization; treatment for scorpion sting; neutralization of neurotoxin; protection against scorpion poisoning
References	1. Bahraoui E, Pichon J et al, Monoclonal antibodies to scorpion toxins. Characterization and molecular mechanisms of neutralization. J Immunol. 1988 Jul 1;141(1):214-20. 2. Clot-Faybesse O, Juin M et al, Monoclonal antibodies against the Androctonus australis hector scorpion neurotoxin I: characterisation and use for venom neutralisation. FEBS Lett. 1999 Sep 24;458(3):313-8. 3. Mousli M, Devaux C et al, A recombinant single-chain antibody fragment that neutralizes toxin II from the venom of the scorpion Androctonus australis hector. FEBS Lett. 1999 Jan 15;442(2-3):183-8.

Figure 5-4 Interface displaying the detailed information of an antibody entry in AAIR.



Detailed Information of Entry AA3728

Entry ID	AA3728
NCBI Accession Number	3242233
Locus	CAA76801
Name	immunoglobulin heavy chain variable region [Mus musculus].
Label	Heavy Chain
Source Organism	Mus musculus (house mouse)
Source Organism Strain / Isolate	BALB/c
Development Stage	five-week-old
Tissue Type	speen cell
Cell Line	X63Ag8
Cell Type	hybridoma 4C1
Clone	clone 4C1(BC5)
Express System	E. coli
Express system Cell Line / Strain	HB2151 E. coli cell
Host Plasmid / Vector	pGEMT vector
Protein Sequence	MNFGLSLIFLVLVLKGVQCEVHLVESGGGLVKPGGSLKLSCAASG FTFSGYYMYWVRQTPEKRLEWVASISDGGSFTYYPDSVKGRFTIS RDNAKNNLYLQMSSLRSDDTAMYYCSRPDDYSYDGFAYWGQGTLV TVSAAKTTPPSVYPLS
Nucleotide Entry	Y17588
Nucleotide Position	Y17588: 1..454
DNA Sequence	ATGAACTTCGGGCTCAGCTTGATTTTCCTTGTCCTTGTTTTAAAA GGTGTCCAGTGTGAAGTGCATCTGGTGGAGTCTGGGGGAGGCTTA GTGAAGCCTGGTGGGTCCCTGAAACTCTCCTGTGCAGCCTCTGGT TTCACTTTCAGTGGCTATTACATGTATTGGGTTCGTCAGACTCCG GAAAAGAGGCTGGAGTGGGTCGCATCCATTAGTGATGGTGGTAGT TTCACCTACTATCCAGACAGTGTGAAGGGACGATTCACCATCTCC AGAGACAATGCCAAGAACAACCTGTACCTGCAGATGAGCAGTCTG AGGTCTGATGACACAGCCATGTATTACTGTTCAAGACCTGACGAC TATAGTTACGACGGGTTTGCTTACTGGGGCCAAGGGACTCTGGTC ACTGTCTCTGCAGCCAAAACGACACCCCCATCGGTCTATCCACTG TCTA
References	1. Mousli M, Devaux C et al, A recombinant single-chain antibody fragment that neutralizes toxin II from the venom of the scorpion Androctonus australis hector. FEBS Lett. 1999 Jan 15;442(2-3):183-8.

5.3 Statistic analysis of disease targeting antibody information database

The disease targeting antibody information database currently contains 2,777 antibody-antigen pairs covering 159 disease conditions, 2,035 heavy chain sequences (535 IgG (232 IgG1, 37 IgG2a, 14 IgG2b, 6 IgG3 and 4 IgG4), 52 IgM and 23 IgE), and 1,701 light chain sequences (693 kappa and 113 lambda), 619 distinct antigen sequences (584 proteins and 35 other type of molecules), 254 antigen epitope sequences and 157 binding affinity constants for Ab-Ag pairs, from various viruses, bacteria, tumor types, and autoimmune responses. With the rapid advances in genomics (431), proteomics (431, 432), immunology (433) and biotechnology (434), new entries about disease targeted antibody can be incorporated or the corresponding database can be cross-linked to AAIR database to provide more comprehensive information about the disease targeted antibody, corresponding antigen, and related disease and therapeutic information. It is expected that a significantly higher number of entries of naturally occurring antibodies can be included in this and other databases as the relevant information are being made available from the vast number of medical studies and therapeutic explorations.

5.3.1 Distribution pattern of antibody-antigen pairs

5.3.1.1 Distribution pattern of antibody-antigen pairs with respective disease classes

Ab-Ag recognition has been widely explored for disease treatment (167,

435-437) . Table 5-7 lists the distribution pattern of Ab-Ag pairs involved in disease classes. The Ab-Ag pairs involved in infectious and parasites diseases, immunity disorders, and neoplasms contain 1123, 587, and 398 pairs, respectively. They constitute the group which has the largest number of Ab-Ag pairs. Other disease types composed of a substantial number of Ab-Ag pairs (indicated by the numerical number in brackets) are skin and subcutaneous tissue disease (143), circulatory system disease (119), musculoskeletal system and connective tissue diseases (108), blood and blood-forming organs diseases (92), nervous system and sense organs diseases (87), injury and poisoning (67), endocrine disorders (52), congenital anomalies (47), and digestive system diseases (45).

Table 5-7 Distribution pattern of antibody-antigen pairs involved in different disease classes

Disease Class	Number of Ab-Ag pairs	Percentage in the AAIR database
Infectious and parasitic diseases	1123	40.4%
Immunity disorders	587	21.1%
Neoplasms	398	14.3%
Skin and subcutaneous tissue diseases	143	5.1%
Circulatory system diseases	119	4.3%
Musculoskeletal system and connective tissue diseases	108	3.9%
Blood and blood-forming organs diseases	92	3.3%
Nervous system and sense organs diseases	87	3.1%
Injury and poisoning	67	2.4%
Endocrine disorders	52	1.9%
Congenital anomalies	47	1.7%
Digestive system diseases	45	1.6%
Inflammation	27	1.0%
Respiratory system diseases	21	0.8%
Nutritional and metabolic diseases	17	0.6%
Genitourinary system diseases	13	0.5%
Mental disorders	6	0.2%

The distribution pattern of Ab-Ag involved in disease types are listed in Table 5-8. Antibodies targeting cancer, influenza, HIV infection, allergy, rabies and hepatitis,

which contain 398, 220, 167, 158, 150, and 116 antibodies respectively, constitute the group with the largest number of antibodies.

Table 5-8 Distribution pattern of antibody-antigen pairs involved in different disease types (only those disease types with more than 25 antibodies are listed here)

Disease Type	Numbers of Ab-Ag pairs	Percentage in the AAIR database
Cancer	398	14.4%
Influenza	220	8.0%
HIV Infection	167	6.1%
Allergy	158	5.7%
Rabies	150	5.4%
Hepatitis	116	4.2%
Lupus	88	3.2%
Organ Transplantation	78	2.8%
Cytomegalovirus Infection	58	2.1%
Prion Disease	55	2.0%
Rheumatoid Arthritis	55	2.0%
Sjogren's Syndrome	47	1.7%
Botulinum Intoxication	42	1.5%
Graves' Disease	42	1.5%
Staphylococcal Infection	42	1.5%
Multiple Sclerosis	41	1.5%
Rotavirus Infection	37	1.3%
Anthrax Infection	36	1.3%
Thyroid Disease	35	1.3%
Thyroiditis	32	1.2%
Vitiligo	32	1.2%
Injury	31	1.1%
Heart Block	28	1.0%
Thrombocytopenia	28	1.0%
Alzheimer's Disease	26	0.9%
Rheumatic Heart Disease	26	0.9%
Rheumatic Disease	26	0.9%
Myocarditis	25	0.9%

The most common and earliest application of antibodies is the detection and treatment of viral, bacterial and other types of infection (166). Our database includes 1123 entries related to infectious and parasites diseases, such as influenza,

rabies, hepatitis, HIV infection, prion disease, staphylococcal infection, rotavirus infection and anthrax infection. For examples, 80 antibodies targeted on HIV glycoprotein gp120, 31 antibodies targeted on HIV glycoprotein gp41, 147 antibodies on rabies virus glycoprotein, and 36 antibodies on cytomegalovirus glycoprotein B are included in AAIR database.

Cancer treatment is another crucially important application for antibody (438). In the development of our database, one of the major focuses is to collect as much as possible numbers of cancer-related antibody and its corresponding antigen. AAIR database includes 398 Ab-Ag pairs involved in breast cancer, colon cancer, non-Hodgkin's lymphoma, leukemia and other cancers. The most common cancer targets are epidermal growth factor receptor (EGFR), ERBB2 (also known as HER2/neu, associated with lung and breast cancer), CD52 on lymphocytes, CD20 on B cells (a marker for non-Hodgkin's lymphoma (NHL)) and vascular endothelial growth factor (VEGF). In our database we include 5, 15, 5, 6, 5 antibody-antigen pairs for EGFR, ERBB2, CD52, CD20 and VEGF respectively. These cancer targets are the most popular antigens of the FDA approved anticancer antibodies, for example, Erbitux (Cetuximab) and Vectibix (Panitumumab) which target EGFR for Metastatic colorectal cancer treatment and head and neck cancer treatment, herceptin (Trastuzumab) which targets ERBB2 for metastatic breast cancer treatment (167), campath (Alemtuzumab) which targets CD52 for B-cell chronic lymphocytic leukemia treatment, bexxar (Tositumomab), rituxan (Rituximab) and zevalin (Ibritumomab tiuxetan) which target CD20 for non-Hodgkin's lymphoma treatment (166, 167), and Avastin (Bevacizumab) and Lucentis (Ranibizumab) which target VEGF so as to inhibit angiogenesis for the treatment of breast cancer and colorectal cancer.

Our database also includes a substantially number of antibodies targeted on cytokines and cytokine receptors which are associated with inflammation and autoimmunity for the treatment of inflammatory diseases such as rheumatoid arthritis and nephritis (166) and transplant rejection. These cytokines and cytokine receptors include tumor-necrosis factor-alpha (TNFalpha), complement proteins, interleukins (IL), and interleukin receptors. There are 16 antibodies targeted for TNFalpha, 6 antibodies for complement proteins, 11 antibodies for interleukin 2 and 6 antibodies for interleukin 2 receptor in AAIR database. These targets are the most popular antigens of the FDA approved antibodies for the treatment of inflammatory diseases and transplant rejection. TNFalpha was targeted by Humira (Adalimumab) and Remicade (Infliximab) in the treatment of Inflammatory diseases (mostly auto-immune disorders), interleukin 2 receptor was targeted by Simulect (Basiliximab) and Zenapax(Daclizumab) in the treatment of transplant rejection, complement protein C5 was targeted by Soliris (Eculizumab) in the treatment of Inflammatory diseases such as paroxysmal nocturnal hemoglobinuria.

5.3.1.2 Distribution pattern of antibody-antigen pairs with respective Pfam of antigen

We also investigated the distribution pattern of protein families of antigens in the database. Pfam, a comprehensive database of protein families, which contains over 8957 protein families in the current release (version 21.0) (272), was employed herein. It has been widely used for protein function prediction (236) and protein structure prediction (439). The family that Pfam provides is the domain family, and was widely used in protein function prediction (236) and protein structure prediction (439). The disease related antibody can be classified recording

to the Pfam of corresponding antigen. Table 5-9 shows the distribution pattern of Ab-Ag pairs with respect to the Pfam of corresponding antigen. Most of the antigens belong to PF01500 (Keratin B2 protein), PF02440 (Adenovirus E3 region protein CR1) and PF00509 (Hemagglutinin).

Table 5-9 Distribution pattern of antigen in different Pfam (only those Pfams with more than 50 antigens are listed here)

Pfam	Description	Number of antigens in AAIR database
PF01500	Keratin B2 protein	124
PF02440	Adenovirus E3 region protein CR1	92
PF00509	Hemagglutinin	89
PF01464	Transglycosylase SLT domain	81
PF01686	Adenovirus penton base protein	80
PF00516	Envelope glycoprotein GP120	80
PF00517	Envelope Polyprotein GP41	75
PF07654	Immunoglobulin C1-set domain	75
PF08791	Viral envelope protein	72
PF00062	C-type lysozyme/alpha-lactalbumin family	72
PF06737	Transglycosylase-like domain	72
PF08205	CD80-like C2-set immunoglobulin domain	72
PF08475	Viral capsid protein 91 N-terminal	69
PF03236	Domain of unknown function DUF263	69
PF04110	Ubiquitin-like autophagy protein Apg12	69
PF00431	CUB domain	69
PF02525	Flavodoxin-like fold	68
PF04839	Plastid and cyanobacterial ribosomal protein (PSRP-3 / Ycf65)	67
PF03056	Env gp36 protein (HERV/MMTV type)	63
PF03151	Triose-phosphate Transporter family	60
PF03217	Bacterial surface layer protein	57
PF07690	Major Facilitator Superfamily	56
PF07951	Clostridium neurotoxin, C-terminal receptor binding	56
PF03595	C4-dicarboxylate transporter/malic acid transport protein	55
PF05316	Mitochondrial ribosomal protein (VAR1)	55
PF03938	Outer membrane protein (OmpH-like)	54
PF02489	Herpesvirus glycoprotein H	52
PF03600	Citrate transporter	50
PF07554	Uncharacterised Sugar-binding Domain	50

5.3.2 Statistical analysis of sequence specificity of antibody-antigen recognition

While there are many instances of antibodies interacting with multiple antigens and antigens interacting with multiple antibodies, antibodies typically recognize their target antigen selectively. This indicates that each antibody binds specifically to a particular antigen sequence, which is a key feature for therapeutic applications of antibodies (167). Furthermore, it is uncommon for antibodies to bind the corresponding antigen from different species. Therefore, the level of sequence selectivity can be analyzed from the sequence data of AAIR. As an example, the level of selectivity of antigen recognition can be characterized by the extent of sequence variation of the antigens recognizable by antibodies with different sequence variations, which can be measured by the sequence variations of VH-VL of the antibodies. As a result, this method can be used to address the questions such as whether antibodies differing by a few amino acids are able to selectively recognize antigens differing by both lower (close homologues) and higher number of amino acids (remote homologues).

A statistical picture about sequence selectivity of antigen recognition in the known Ab-Ag pairs was obtained by the following study: First, all antibody pairs generated from antibodies of the known Ab-Ag pairs were grouped into classes that differ by one, two, …, and *n* number of amino acids of VH-VL of the corresponding antibodies. The sequence variation among the corresponding antigen pairs for the antibody pairs in each antibody group was then analyzed to determine to what extent the VH-VL of antibodies that differ by one to n number of amino acids were able to selectively recognize antigens that differ by one to m number of amino acids.

Table 5-10 summarizes the distribution of antigen pairs differing by one to 2815 amino acids selectively recognized by antibodies in which VH-VL differ by one to 208 amino acids. It can be found that a substantial number of antibodies, in which VH-VL differ by a few amino acids, are capable of specifically recognizing antigens that differ by both a few amino acids and up to 208 amino acids. One example of such antigen pairs is horse cytochrome C and mouse cytochrome C, which can be recognized by antibody E8 and CA4-1 respectively. The sequences of the VH-VL in the two antibodies differ by 1 amino acid only, compared to the difference of 93 amino acids between horse cytochrome C and mouse cytochrome C. Another example of such antigen pairs is Der p I allergen and Human Heat Shock Protein 70. The sequence variation of these two antigens is 228 amino acids, compared to the difference of 2 amino acids between the VH-VL of their corresponding antibodies. It can be inferred that antibodies in which the VH-VL showing small variations are able to selectively recognize antigens of both close and remote homologues that differ by both lower and higher number of amino acids.

Table 5-10 Distribution of antigens of different sequence variations that can be selectively recognized by antibodies in which the VH-VL differ by one to 208 amino acids

Sequence variation between antibodies (VH-VL)	Number of antibody pairs with this sequence variation	Number of corresponding antigen pairs with different sequence variation									
		variation by 1 amino acid	variation by 2~5 amino acids	variation by 6~10 amino acids	variation by 11~20 amino acids	variation by 21~50 amino acids	variation by 51~100 amino acids	variation by 101~150 amino acids	variation by 151~200 amino acids	variation by 201~250 amino acids	variation by >250 amino acids
1 amino acid	385	120	60	0	1	0	2	195	7	0	0
2~5 amino acids	422	135	31	11	8	1	7	218	4	0	7
6~10 amino acids	73	5	0	0	2	3	0	43	3	0	17
11~20 amino acids	499	61	5	15	0	11	2	369	2	0	34
21~50 amino acids	486	8	2	9	13	50	62	93	60	40	149
51~100 amino acids	9728	44	38	126	266	348	1798	1956	1135	914	3103
101~150 amino acids	367506	2946	1175	3730	5658	11998	49914	78513	61655	27948	123969
151~200 amino acids	1175133	7442	5993	19033	18591	40122	148014	259527	199895	95245	381271
>201 amino acids	80842	346	238	1175	1374	2757	9193	13481	14861	9160	28257

5.4 Prediction performance of disease targeting antibody prediction system

5.4.1 Overview of the prediction system

The antibody-antigen sequence-recognition information in the AAIR database is very important to understand the mechanism of antibody-antigen interaction and helpful for vaccine design and antibody design. Therefore, we developed some prototype prediction systems based on the sequence information of antibodies and antigens which either makes use of all data, or is restricted to a certain disease or Pfam. It was expected that this prototype system can be further extended for antibody design and vaccine design.

5.4.2 Prediction performance

5.4.2.1 Prediction performance of machine learning model for prediction antibody-antigen pairs involved in certain diseases.

Figure 5-11 give the performance of SVM prediction of Ab-Ag pairs for some datasets generated from AAIR using the 5-fold cross validation method described above. The four datasets in Figure 5-11 contain Ab-Ag pairs involved in four disease types, which are cancer, influenza, HIV infection, and allergy. The corresponding non-Ab-Ag pairs in each dataset contain two parts. For those antibodies and antigens in Ab-Ag pairs, random pairing of antibodies with antigens of other antibodies forms one part. Another part is selected as a representative Ab-Ag pairs which are not involved in this disease types.

The prediction accuracies for antibody-antigen and non-antibody-antigen pairs involved in cancer, influenza, HIV infection, and allergy were in the range of 96.9% to 97.7%, 95.5% to 97.9%, 94.1% to 96.7%, and 92.6% to 95.4% respectively, suggesting that SVM is potentially useful for predicting antibody-antigen pairs.

Table 5-11 Performance evaluation of SVM prediction system of antibody-antigen pairs involved in cancer, influenza, HIV infection and allergy by using five-fold cross validation.

Disease type in which Ab-Ag pairs involved	Cross validation	Training data		Test data								
		Ab-Ag pair	non-Ab-Ag pair	Prediction accuracy for Ab-Ag pairs			Prediction accuracy for non-Ab-Ag pairs			PPV (%)	NPV (%)	Q (%)
				TP	FN	SE (%)	TN	FP	SP (%)			
Cancer	1	264	5296	52	15	77.6	1307	17	98.7	75.4	98.9	97.7
	2	265	5296	46	20	69.7	1301	23	98.3	66.7	98.5	96.9
	3	265	5296	49	17	74.2	1309	15	98.9	76.6	98.7	97.7
	4	265	5296	52	14	78.8	1297	27	98.0	65.8	98.9	97.1
	5	265	5296	48	18	72.7	1299	25	98.1	65.8	98.6	96.9
Average						74.6			98.4	70.0	98.7	97.3
Standard deviation						±3.3			±0.3	±4.9	±0.2	±0.4
Influenza	1	92	1856	15	9	62.5	457	7	98.5	68.2	98.1	96.7
	2	93	1856	15	8	65.2	455	9	98.1	62.5	98.3	96.5
	3	93	1856	18	5	78.3	456	8	98.3	69.2	98.9	97.3
	4	93	1856	18	5	78.3	447	17	96.3	51.4	98.9	95.5
	5	93	1856	17	6	73.9	460	4	99.1	81.0	98.7	97.9
Average						71.6			98.1	66.5	98.6	96.8
Standard deviation						±6.6			±0.9	±9.6	±0.3	±0.8
HIV infection	1	115	2209	14	15	48.3	538	14	97.5	50.0	97.3	95.0
	2	115	2209	18	11	62.1	538	14	97.5	56.3	98.0	95.7
	3	115	2209	18	11	62.1	540	12	97.8	60.0	98.0	96.0
	4	115	2209	15	14	51.7	532	20	96.4	42.9	97.4	94.1
	5	116	2208	15	13	53.6	547	6	98.9	71.4	97.7	96.7
Average						55.5			97.6	56.1	97.7	95.5
Standard deviation						±5.6			±0.8	±9.6	±0.3	±0.9
Allergy	1	105	1035	17	10	63.0	255	3	98.8	85.0	96.2	95.4
	2	105	1035	18	9	66.7	252	6	97.7	75.0	96.6	94.7
	3	106	1034	13	13	50.0	251	8	96.9	61.9	95.1	92.6
	4	106	1034	16	10	61.5	248	11	95.8	59.3	96.1	92.6
	5	106	1034	16	10	61.5	248	11	95.8	59.3	96.1	92.6
Average						60.5			97.0	68.1	96.0	93.6
Standard deviation						±5.6			±1.2	±10.3	±0.5	±1.2

5.4.2.2 Prediction performance of machine learning model for prediction antibody-antigen pairs with respect to Pfam of antigen

The four datasets in Figure 5-11 contains Ab-Ag pairs of antigens that belong to the Pfam (440) protein family PF01500 (Keratin B2 protein), PF02440 (Adenovirus E3 region protein CR1), PF00509 (Hemagglutinin), and PF01464 (Transglycosylase SLT domain). The generation of corresponding non-Ab-Ag is similar as previous description in generating non-Ab-Ag pairs for disease types. The computed prediction accuracies for antibody-antigen and non-antibody-antigen pairs of which antigens belong to PF01500, PF02440, PF00509, and PF01464 were in the range of 96.3% to 97.1%, 96.6% to 98.2%, 95.2% to 98.9%, and 93.5% to 95.3% respectively.

5.4.2.3 Prediction performance of machine learning model for prediction antibody-antigen pairs

Table 5-13 shows the performances of SVM prediction of antibody-antigen pairs which have certain therapeutic applications based on a 5-fold cross validation study. The antibody-antigen pairs were chosen if they have certain diagnosis or therapeutic application. The non-antibody-antigen pairs were generated by random pairing of antibodies with antigens of other antibodies.

The computed prediction accuracies for antibody-antigen pairs and non-antibody-antigen pairs were in the range of 61.6% to 65.4% and 99.9% to 100%. The positive predictive values were in the range of 94.1% to 98.1%. The overall accuracy for all data was in 99.5%. Therefore, our SVM prediction system appears to show reasonably good capability for prediction the antibody-antigen

interaction based on the data in the AAIR database.

Table 5-12 Performance evaluation of SVM prediction system of antibody-antigen pairs for antigens from four different protein domain families, Keratin high sulfur B2 protein, Adenovirus E3 region protein CR1, Hemagglutinin and Transglycosylase SLT domain by using five-fold cross validation.

Pfam in which antigens belong to	Cross validation	Training data		Test data								
		Ab-Ag pair	non-Ab-Ag pair	Prediction accuracy for Ab-Ag pairs			Prediction accuracy for non-Ab-Ag pairs			PPV (%)	NPV (%)	Q (%)
				TP	FN	SE (%)	TN	FP	SP (%)			
Keratin, high sulfur B2 protein (PF01500)	1	99	992	20	5	80.0	243	5	98.0	80.0	98.0	96.3
	2	99	992	20	5	80.0	244	4	98.4	83.3	98.0	96.7
	3	99	992	24	1	96.0	241	7	97.2	77.4	99.6	97.1
	4	99	992	16	9	64.0	247	1	99.6	94.1	96.5	96.3
	5	100	992	19	5	79.2	245	3	98.8	86.4	98.0	97.1
Average						79.8			98.4	84.2	98.0	96.7
Standard deviation						±10.1			±0.8	±5.8	±1.0	±0.3
Adenovirus E3 region protein CR1 (PF02440)	1	73	1472	17	2	89.5	360	8	97.8	68.0	99.4	97.4
	2	73	1472	10	9	52.6	364	4	98.9	71.4	97.6	96.6
	3	74	1472	12	6	66.7	361	7	98.1	63.2	98.4	96.6
	4	74	1472	10	8	55.6	363	5	98.6	66.7	97.8	96.6
	5	74	1472	17	1	94.4	362	6	98.4	73.9	99.7	98.2
Average						71.8			98.4	68.6	98.6	97.1
Standard deviation						±17.2			±0.4	±3.7	±0.9	±0.6
Hemagglutinin (PF00509)	1	71	679	14	4	77.8	168	2	98.8	87.5	97.7	96.8
	2	71	679	14	4	77.8	165	5	97.1	73.7	97.6	95.2
	3	71	679	15	3	83.3	167	3	98.2	83.3	98.2	96.8
	4	71	680	16	2	88.9	169	0	100.0	100.0	98.8	98.9
	5	72	679	15	2	88.2	169	1	99.4	93.8	98.8	98.4
Average						83.2			98.7	87.7	98.2	97.2
Standard deviation						±4.8			±1.0	±9.0	±0.5	±1.3
Transglycosylase SLT domain (PF01464)	1	64	1296	13	4	76.5	307	17	94.8	43.3	98.7	93.8
	2	65	1296	13	3	81.3	311	13	96.0	50.0	99.0	95.3
	3	65	1296	12	4	75.0	306	18	94.4	40.0	98.7	93.5
	4	65	1296	11	5	68.8	310	14	95.7	44.0	98.4	94.4
	5	65	1296	10	6	62.5	311	13	96.0	43.5	98.1	94.4
Average						72.8			95.4	44.2	98.6	94.3
Standard deviation						±6.5			±0.6	±3.2	±0.3	±0.6

Table 5-13 Performance evaluation of SVM prediction system of antibody-antigen pairs

Cross validation	training data		Test data								
	Ab-Ag pair	non-Ab-Ag pair	Prediction accuracy for Ab-Ag pairs			Prediction accuracy for non-Ab-Ag pairs			PPV (%)	NPV (%)	Q (%)
			TP	FN	SE (%)	TN	FP	SP (%)			
1	1376	110176	217	128	62.9	27539	5	100.0	97.7	99.5	99.5
2	1377	110176	225	119	65.4	27530	14	99.9	94.1	99.6	99.5
3	1377	110176	221	123	64.2	27538	6	100.0	97.4	99.6	99.5
4	1377	110176	223	121	64.8	27535	9	100.0	96.1	99.6	99.5
5	1377	110176	212	132	61.6	27540	4	100.0	98.1	99.5	99.5
Average					63.8			100.0			99.5
Standard deviation					±1.5			±0.0			±0.0

Based on the sequence information of antibodies and antigens which either makes use of all data, or is restricted to a certain disease or is restrict to a certain Pfam, the SVM prediction results suggest that SVM is potentially useful for predicting Ab-Ag pairs, which give insight for developing systems for predicting antibody from antigen sequence or antigen sequence from antibody sequence in our future study. The prediction accuracies for non-Ab-Ag pairs are better than that of Ab-Ag pairs in all of these datasets. This probably results from the more diverse set of non-Ab-Ag pairs compared to that of Ab-Ag pairs which enables SVM to better recognize non-Ab-Ag pairs.

5.5 Conclusion

The AAIR database is intended to provide comprehensive information about Ab-Ag pair sequence, function, diagnosis and therapeutic indications, and other information from a single source. Through a few illustrative case studies, data from the database are proved to be highly effective in facilitating immunological research and development tasks such as the development of antibody discovery tools and the analysis of selective antigen recognition. Preliminary results of the

machine learning models suggest that the information in AAIR is useful to characterize pair-wise antibody-antigen interactions. Further investigations are being made to collect more Ab-Ag data via improved searching algorithms and from more recently published papers. Rapid advances in the technologies for identifying Ab-Ag recognitions will enable the generation of more comprehensive and useful molecular level data that can be incorporated into our database.

6 Conclusion and future works

In drug discovery, disease targets and therapeutic molecules are two important molecules. This work developed a computational system to discover potential disease targets from microarray data, and implement a bioinformatics tool for therapeutic molecular discovery.

In this study, a robust computational system for gene signature derivation from microarray data was developed. A popular and accurate machine learning method, support vector machines, was applied to classify the samples. Recursive feature selection incorporating with multiple random sampling method and gene consistency evaluation strategies was used in gene selection procedure. This system was used to select colon cancer markers and lung adenocarcinoma survivability markers. For both cases, the markers were consistent with the variation of the samples, and present good predictive performances. 104 colon cancer markers were shared by all of the 20 signatures which were selected from different combination of samples. By applying a therapeutic target prediction system, all of the 6 known therapeutic targets were predicted correctly and 18 markers were predicted as potential therapeutic targets. 21 lung adenocarcinoma survivability markers were shared by all of the 10 signatures. All of the 5 know therapeutic targets were predicted correctly by therapeutic target prediction system. In addition, 7 markers were identified as potential therapeutic targets.

The results from the two demonstrative examples – colon cancer gene selection and lung adenocarcinoma survival gene selection suggest that, our system can derive stable and good predictive marker signatures. Since the cost for microarray

experiments is very high, the sample size is much smaller than what is required for a satisfactory diagnosis and prognosis of a certain disease such as cancer. In such situations, our system is particular useful to get real important markers for disease diagnosis, patient survival prediction and therapeutic target discovery. The use of consensus scoring for multiple random sampling and evaluation of gene-ranking consistency seem to have impressive capability in avoiding erroneous elimination of predictor-genes due to such noise as measurement variability and biological differences. Further improvement in measurement quality, annotation accuracy and coverage, and signature-selection will enable the derivation of more accurate signatures for facilitating biomarker and target discovery. The currently available platforms for microarray data are different. Therefore if we could synchronize the platform and provide more samples, we could further improve the accuracy of our system and reduce the computational time. The gene ontology information also could be integrated into the system and the selected genes would be given a biological meaning directly.

Another aspect of this work was to develop a bioinformatics tools for disease targeting antibody prediction. An antibody-antigen interaction database (AAIR) was developed. The information from the AAIR database was used to develop a prediction system for antibody design and vaccine design. The accuracy of the system is in the range of 92.6% to 99.5% from the five-fold cross validation. The information of this database was also used to analyze the recognition variation between antibody and antigen. It was shown that small variation of antibodies can recognize both close homologues and remote homologues.

With the development of immunology and biotechnology, it is expected that a

significant higher number of entries of antibodies can be incorporated in this database. Structural information of antibody and antigen may include in this database and develop a prediction system not only based on sequence information but also on the structural information in order to get a better performance of the system. A more comprehensive database system could make it possible to screen the genome to find out the possible antibody-antigen pairs for the purpose of antibody design or vaccine design.

BIOBLIOGRAPHY

1. Larsson TP, Murray CG, Hill T, Fredriksson R, and Schioth HB. Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery. FEBS Lett, 579*:* 690-8, 2005.
2. Sandberg AA and Chen Z. Cancer cytogenetics and molecular genetics: detection and therapeutic strategy. In Vivo, 8*:* 807-18, 1994.
3. Chen Z and Sandberg AA. Molecular cytogenetic aspects of hematological malignancies: clinical implications. Am J Med Genet, 115*:* 130-41, 2002.
4. Mrozek K, Heerema NA, and Bloomfield CD. Cytogenetics in acute leukemia. Blood Rev, 18*:* 115-36, 2004.
5. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286*:* 531-7, 1999.
6. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A, 98*:* 15149-54, 2001.
7. Robinson S and Kessling A. Diabetes secondary to genetic disorders. Baillieres Clin Endocrinol Metab, 6*:* 867-98, 1992.
8. Clee SM, Yandell BS, Schueler KM, et al. Positional cloning of Sorcs1, a type 2 diabetes quantitative trait locus. Nat Genet, 38*:* 688-93, 2006.
9. Li D. Positional cloning: single-gene cardiovascular disorders. Methods Mol Med, 128*:* 125-36, 2006.
10. Gulcher J and Stefansson K. Positional cloning: complex cardiovascular traits. Methods Mol Med, 128*:* 137-52, 2006.
11. Hotta K. [Genetic testing and gene-based testing for obesity]. Nippon Rinsho, 63 Suppl 12*:* 280-4, 2005.
12. Zhang W, Rekaya R, and Bertrand K. A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. Bioinformatics, 22*:* 317-25, 2006.
13. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409*:* 928-33, 2001.
14. Knowles J and Gromo G. A guide to drug discovery: Target selection in drug discovery. Nat Rev Drug Discov, 2*:* 63-9, 2003.
15. Collins I and Workman P. New approaches to molecular cancer therapeutics. Nat Chem Biol, 2*:* 689-700, 2006.
16. Workman P. Genomics and the second golden era of cancer drug development. Mol Biosyst, 1*:* 17-26, 2005.
17. Shimoji T, Miki Y, and Nagasaki K. [Gene expression profiling for prediction of response to chemotherapy]. Gan To Kagaku Ryoho, 33*:* 1-5, 2006.
18. Boiesen P, Bendahl PO, Anagnostaki L, et al. Histologic grading in breast cancer--reproducibility between seven pathologic departments. South Sweden Breast Cancer Group. Acta Oncol, 39*:* 41-5, 2000.
19. Dietel M and Sers C. Personalized medicine and development of targeted therapies: The upcoming challenge for diagnostic molecular pathology. A

review. Virchows Arch, 448*:* 744-55, 2006.
20. Isaacs C, Stearns V, and Hayes DF. New prognostic factors for breast cancer recurrence. Semin Oncol, 28*:* 53-67, 2001.
21. Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet, 24*:* 227-35, 2000.
22. Yeang CH, Ramaswamy S, Tamayo P, et al. Molecular classification of multiple tumor types. Bioinformatics, 17 Suppl 1*:* S316-22, 2001.
23. Ooi CH and Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics, 19*:* 37-44, 2003.
24. Peng S, Xu Q, Ling XB, et al. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Lett, 555*:* 358-62, 2003.
25. Massion PP and Carbone DP. The molecular basis of lung cancer: molecular abnormalities and therapeutic implications. Respir Res, 4*:* 12, 2003.
26. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403*:* 503-11, 2000.
27. Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med, 7*:* 673-9, 2001.
28. Ross ME, Zhou X, Song G, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood, 102*:* 2951-9, 2003.
29. Yeoh EJ, Ross ME, Shurtleff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell, 1*:* 133-43, 2002.
30. Giallourakis C, Henson C, Reich M, Xie X, and Mootha VK. Disease gene discovery through integrative genomics. Annu Rev Genomics Hum Genet, 6*:* 381-406, 2005.
31. Kim HL and Steinberg GD. New insights and candidate genes and their implications for care of patients with hereditary prostate cancer. Curr Urol Rep, 1*:* 9-14, 2000.
32. Tanigawa G, Jarcho JA, Kass S, et al. A molecular basis for familial hypertrophic cardiomyopathy: an alpha/beta cardiac myosin heavy chain hybrid gene. Cell, 62*:* 991-8, 1990.
33. Malkin D, Li FP, Strong LC, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science, 250*:* 1233-8, 1990.
34. Dryja TP, McGee TL, Reichel E, et al. A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. Nature, 343*:* 364-6, 1990.
35. Farrar GJ, Kenna P, Jordan SA, et al. A three-base-pair deletion in the peripherin-RDS gene in one form of retinitis pigmentosa. Nature, 354*:* 478-80, 1991.
36. Cui JF, Liu YK, Zhang LJ, et al. Identification of metastasis candidate proteins among HCC cell lines by comparative proteome and biological function analysis of S100A4 in metastasis in vitro. Proteomics, 6*:* 5953-61, 2006.
37. Pharoah PD, Tyrer J, Dunning AM, Easton DF, and Ponder BA. Association between Common Variation in 120 Candidate Genes and

Breast Cancer Risk. PLoS Genet, 3*:* e42, 2007.
38. Smith AK and Meyers DA. Family studies and positional cloning of genes for asthma and related phenotypes. Immunol Allergy Clin North Am, 25*:* 641-54, 2005.
39. Cho WC. Contribution of oncoproteomics to cancer biomarker discovery. Mol Cancer, 6*:* 25, 2007.
40. Bharti A, Ma PC, and Salgia R. Biomarker discovery in lung cancer-promises and challenges of clinical proteomics. Mass Spectrom Rev, 2007.
41. Brusic V, Marina O, Wu CJ, and Reinherz EL. Proteome informatics for cancer research: From molecules to clinic. Proteomics, 7*:* 976-91, 2007.
42. de Hoog CL and Mann M. Proteomics. Annu Rev Genomics Hum Genet, 5*:* 267-93, 2004.
43. Schonberger J and Seidman CE. Many roads lead to a broken heart: the genetics of dilated cardiomyopathy. Am J Hum Genet, 69*:* 249-60, 2001.
44. Teitelbaum SL and Ross FP. Genetic regulation of osteoclast development and function. Nat Rev Genet, 4*:* 638-49, 2003.
45. Chalhoub N, Benachenhou N, Rajapurohitam V, et al. Grey-lethal mutation induces severe malignant autosomal recessive osteopetrosis in mouse and human. Nat Med, 9*:* 399-406, 2003.
46. Bagshaw RD, Mahuran DJ, and Callahan JW. A proteomic analysis of lysosomal integral membrane proteins reveals the diverse composition of the organelle. Mol Cell Proteomics, 4*:* 133-43, 2005.
47. Hellstrom M, Lexander H, Franzen B, and Egevad L. Proteomics in prostate cancer research. Anal Quant Cytol Histol, 29*:* 32-40, 2007.
48. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A, 96*:* 4285-8, 1999.
49. Tatusov RL, Koonin EV, and Lipman DJ. A genomic perspective on protein families. Science, 278*:* 631-7, 1997.
50. Wu J, Kasif S, and DeLisi C. Identification of functional links between genes using phylogenetic profiles. Bioinformatics, 19*:* 1524-30, 2003.
51. Warren RM, Richardson M, Sampson SL, et al. Molecular evolution of Mycobacterium tuberculosis: phylogenetic reconstruction of clonal expansion. Tuberculosis (Edinb), 81*:* 291-302, 2001.
52. Li JB, Gerdes JM, Haycraft CJ, et al. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell, 117*:* 541-52, 2004.
53. Chiang AP, Nishimura D, Searby C, et al. Comparative genomic analysis identifies an ADP-ribosylation factor-like gene as the cause of Bardet-Biedl syndrome (BBS3). Am J Hum Genet, 75*:* 475-84, 2004.
54. Velculescu VE, Zhang L, Vogelstein B, and Kinzler KW. Serial analysis of gene expression. Science, 270*:* 484-7, 1995.
55. Winegarden N. Microarrays in cancer: moving from hype to clinical reality. Lancet, 362*:* 1428, 2003.
56. Ramaswamy S, Ross KN, Lander ES, and Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet, 33*:* 49-54, 2003.
57. Staudt LM. Molecular diagnosis of the hematologic cancers. N Engl J Med, 348*:* 1777-85, 2003.
58. Bullinger L, Dohner K, Bair E, et al. Use of Gene-Expression Profiling to

Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. N Engl J Med, 350*:* 1605-16, 2004.
59. Valk PJM, Verhaak RGW, Beijen MA, et al. Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. N Engl J Med, 350*:* 1617-28, 2004.
60. Garaizar J, Brena S, Bikandi J, Rementeria A, and Ponton J. Use of DNA microarray technology and gene expression profiles to investigate the pathogenesis, cell biology, antifungal susceptibility and diagnosis of Candida albicans. FEMS Yeast Res, 6*:* 987-98, 2006.
61. Nguyen DM and Schrump DS. Lung cancer staging in the genomics era. Thorac Surg Clin, 16*:* 329-37, 2006.
62. Meltzer PS. Spotting the target: microarrays for disease gene discovery. Curr Opin Genet Dev, 11*:* 258-63, 2001.
63. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet, 34*:* 267-73, 2003.
64. Tanaka F, Niwa J, Ishigaki S, et al. Gene expression profiling toward understanding of ALS pathogenesis. Ann N Y Acad Sci, 1086*:* 1-10, 2006.
65. Schlee M, Holzel M, Bernard S, et al. C-myc activation impairs the NF-kappaB and the interferon response: implications for the pathogenesis of Burkitt's lymphoma. Int J Cancer, 120*:* 1387-95, 2007.
66. Garber K. Genomic medicine. Gene expression tests foretell breast cancer's future. Science, 303*:* 1754-5, 2004.
67. Hughes TR, Marton MJ, Jones AR, et al. Functional discovery via a compendium of expression profiles. Cell, 102*:* 109-26, 2000.
68. Inoue R, Matsuyama H, Yano S, et al. Gefitinib-related gene signature in bladder cancer cells identified by a cDNA microarray. Anticancer Res, 26*:* 4195-202, 2006.
69. Narayanan BA. Chemopreventive agents alters global gene expression pattern: predicting their mode of action and targets. Curr Cancer Drug Targets, 6*:* 711-27, 2006.
70. Michiels S, Koscielny S, and Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet, 365*:* 488-92, 2005.
71. Caldas C and Aparicio SA. The molecular outlook. Nature, 415*:* 484-5, 2002.
72. Cho RJ, Huang M, Campbell MJ, et al. Transcriptional regulation and function during the human cell cycle. Nat Genet, 27*:* 48-54, 2001.
73. Iyer VR, Eisen MB, Ross DT, et al. The transcriptional program in the response of human fibroblasts to serum. Science, 283*:* 83-7, 1999.
74. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, 9*:* 3273-97, 1998.
75. Kirmizis A and Farnham PJ. Genomic approaches that aid in the identification of transcription factor target genes. Exp Biol Med (Maywood), 229*:* 705-21, 2004.
76. Lind GE, Kleivi K, Meling GI, et al. ADAMTS1, CRABP1, and NR3C1 identified as epigenetically deregulated genes in colorectal tumorigenesis. Cell Oncol, 28*:* 259-72, 2006.
77. Shi H, Guo J, Duff DJ, et al. Discovery of novel epigenetic markers in non-Hodgkin's lymphoma. Carcinogenesis, 28*:* 60-70, 2007.

78. Mrozek K, Dohner H, and Bloomfield CD. Influence of new molecular prognostic markers in patients with karyotypically normal acute myeloid leukemia: recent advances. Curr Opin Hematol, 14*:* 106-14, 2007.
79. Babu MM. An Introduction to Microarray Data Analysis: Horizon Bioscience, 2004.
80. Leung YF and Cavalieri D. Fundamentals of cDNA microarray data analysis. Trends Genet, 19*:* 649-59, 2003.
81. Pinkel D, Segraves R, Sudar D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet, 20*:* 207-11, 1998.
82. Hughes TR, Mao M, Jones AR, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol, 19*:* 342-7, 2001.
83. Lockhart DJ, Dong H, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol, 14*:* 1675-80, 1996.
84. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, and Miyada CG. The affymetrix GeneChip platform: an overview. Methods Enzymol, 410*:* 3-28, 2006.
85. Demeter J, Beauheim C, Gollub J, et al. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res, 35*:* D766-70, 2007.
86. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res, 35*:* D760-5, 2007.
87. Parkinson H, Kapushesky M, Shojatalab M, et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res, 35*:* D747-50, 2007.
88. http://www.mged.org/Workgroups/MIAME/miame.html.
89. Allison DB, Cui X, Page GP, and Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet, 7*:* 55-65, 2006.
90. Harrison R and DeLisi C. Condition specific transcription factor binding site characterization in Saccharomyces cerevisiae. Bioinformatics, 18*:* 1289-96, 2002.
91. Aach J, Rindone W, and Church GM. Systematic management and analysis of yeast gene expression data. Genome Res, 10*:* 431-45, 2000.
92. Su AI, Cooke MP, Ching KA, et al. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A, 99*:* 4465-70, 2002.
93. Smith CM, Finger JH, Hayamizu TF, et al. The mouse Gene Expression Database (GXD): 2007 update. Nucleic Acids Res, 35*:* D618-23, 2007.
94. Wiederkehr C, Basavaraj R, Sarrauste de Menthiere C, et al. GermOnline, a cross-species community knowledgebase on germ cell differentiation. Nucleic Acids Res, 32*:* D560-7, 2004.
95. Haverty PM, Weng Z, Best NL, et al. HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. Nucleic Acids Res, 30*:* 214-7, 2002.
96. Argraves GL, Barth JL, and Argraves WS. The MUSC DNA Microarray Database. Bioinformatics, 19*:* 2473-4, 2003.
97. Bono H, Kasukawa T, Hayashizaki Y, and Okazaki Y. READ: RIKEN Expression Array Database. Nucleic Acids Res, 30*:* 211-3, 2002.

98. Yazaki J KN, Ishikawa M, Endo D, Kojima K, MicroArray Center, Kikuchi S. The Rice Expression Database (RED): gateway to rice functional genomics. Trends in Plant Science, 12*:* 563-4, 2002.
99. Manduchi E, Pizarro, A., Stoeckert, C. RAD (RNA Abundance Database): an infrastructure for array data analysis. Proc. SPIE, 4266*:* 68-78, 2001.
100. Dwight SS, Harris MA, Dolinski K, et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res, 30*:* 69-72, 2002.
101. Cheung KH, White K, Hager J, et al. YMD: a microarray database for large-scale gene expression analysis. Proc AMIA Symp 140-4, 2002.
102. Lelandais G, Le Crom S, Devaux F, et al. yMGV: a cross-species expression data mining tool. Nucleic Acids Res, 32*:* D323-5, 2004.
103. Schoch C DM, Kern W, Kohlmann A, Schnittger S, Haferlach T. "Deep insight" into microarray technology. Atlas Genet Cytogenet Oncol Haematol, 2004.
104. DeRisi JL, Iyer VR, and Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 278*:* 680-6, 1997.
105. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM. Systematic determination of genetic network architecture. Nat Genet, 22*:* 281-5, 1999.
106. Jansen R, Greenbaum D, and Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res, 12*:* 37-46, 2002.
107. Ramirez-Benitez Mdel C, Moreno-Hagelsieb G, and Almagro JC. VIR.II: a new interface with the antibody sequences in the Kabat database. Biosystems, 61*:* 125-31, 2001.
108. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A, 96*:* 6745-50, 1999.
109. Eisen MB, Spellman PT, Brown PO, and Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A, 95*:* 14863-8, 1998.
110. van der Pouw Kraan TC, van Gaalen FA, Huizinga TW, et al. Discovery of distinctive gene expression profiles in rheumatoid synovium using cDNA microarray technology: evidence for the existence of multiple pathways of tissue destruction and repair. Genes Immun, 4*:* 187-96, 2003.
111. Ma SF, Grigoryev DN, Taylor AD, et al. Bioinformatic identification of novel early stress response genes in rodent models of lung injury. Am J Physiol Lung Cell Mol Physiol, 289*:* L468-77, 2005.
112. Tiranti V, D'Adamo P, Briem E, et al. Ethylmalonic encephalopathy is caused by mutations in ETHE1, a gene encoding a mitochondrial matrix protein. Am J Hum Genet, 74*:* 239-52, 2004.
113. Sherlock G. Analysis of large-scale gene expression data. Curr Opin Immunol, 12*:* 201-5, 2000.
114. Vapnik V. Statistical Learning Theory. 1998.
115. Bishop C. neural networks for pattern recognition. 1995.
116. Qiu P, Wang ZJ, and Liu KJ. Ensemble dependence model for classification and prediction of cancer and normal gene expression data. Bioinformatics, 21*:* 3114-21, 2005.
117. Li F and Yang Y. Analysis of recursive gene selection approaches from microarray data. Bioinformatics, 21*:* 3741-7, 2005.

118. Pochet N, De Smet F, Suykens JA, and De Moor BL. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics, 20*:* 3185-95, 2004.
119. Isabelle Guyon JW, Stephen Barnhill, Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46*:* 389-422, 2002.
120. Furey TS, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16*:* 906-14, 2000.
121. Brown MP, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A, 97*:* 262-7, 2000.
122. Bellman. RE. Adaptive Control Processes. 1961.
123. Koeppen M. The Curse of Dimensionality. 5th Online World Conference on Soft Computing in Industrial Applications (WSC5), 2000.
124. Inza I, Larranaga P, Blanco R, and Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med, 31*:* 91-103, 2004.
125. Model F, Adorjan P, Olek A, and Piepenbrock C. Feature selection for DNA methylation based cancer classification. Bioinformatics, 17 Suppl 1*:* S157-64, 2001.
126. Robnik-Šikonja M and Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning, 53*:* 23-69, 2003.
127. Ding C and Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol, 3*:* 185-205, 2005.
128. Ben-Bassat M. Pattern recognition and reduction of dimensionality. Handbook of statistics II p. 773—91., 1982.
129. Cheng J and Greiner R. Comparing Bayesian Network Classifiers. Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99) 101-10, 1999.
130. Aris V RM. A method to improve detection of disease using selectively expressed genes in microarray data. Methods of Microarray Data Analysis. Proceedings of the First Conference on Critical Assessment of Microarray Data Analysis, CAMDA'00., p. 69—80., 2002.
131. Beibel M. Selection of informative genes in gene expression based diagnosis: a nonparametric approach. Lecture Notes in Computer Sciences. Proceedings of the First International Symposium in Medical Data Analysis, ISMDA'00, 1933*:* p. 300-7, 2000.
132. Ding C. Analysis of gene expression profiles: class discovery and leaf ordering. Proceedings of the Sixth International Conference on Research in Computational Molecular Biology p. 127-36, 2002.
133. Baker SG and Kramer BS. Identifying genes that contribute most to good classification in microarrays. BMC Bioinformatics, 7*:* 407, 2006.
134. Kohavi R and John GH. Wrappers for feature subset selection. Artificial Intelligence, Special issue on relevance*:* 273 - 324, 97.
135. Xiong M, Fang X, and Zhao J. Biomarker identification by feature wrappers. Genome Res, 11*:* 1878-87, 2001.
136. Kohavi R and John GH. Wrappers for feature subset selection. Artificial Intelligence, Special issue on relevance*:* 273 - 324, 1997.

137. Talvinen K, Tuikkala J, Gronroos J, et al. Biochemical and clinical approaches in evaluating the prognosis of colon cancer. Anticancer Res, 26*:* 4745-51, 2006.
138. Ancona N, Maglietta R, Piepoli A, et al. On the statistical assessment of classifiers using DNA microarray data. BMC Bioinformatics, 7*:* 387, 2006.
139. Zhang XW, Yap YL, Wei D, Chen F, and Danchin A. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. Eur J Hum Genet, 13*:* 1303-11, 2005.
140. Li W and Yang Y. How Many Genes Are Needed for a Discriminant Microarray Data Analysis? Methods of Microarray Data Analysis. Proceedings of the First Conference on Critical Assessment of Microarray Data Analysis, CAMDA'00. 137-50, 2002.
141. Grate LR. Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery. BMC Bioinformatics, 6*:* 97, 2005.
142. Slonim DK, Tamayo P, Masiar J, Golub T, and Lander E. Class prediction and discovery using gene expression data. Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB), 2000.
143. Ahmed AA and Brenton JD. Microarrays and breast cancer clinical studies: forgetting what we have not yet learnt. Breast Cancer Res, 7*:* 96-9, 2005.
144. Brenton JD, Carey LA, Ahmed AA, and Caldas C. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J Clin Oncol, 23*:* 7350-60, 2005.
145. Bullinger L and Valk PJ. Gene expression profiling in acute myeloid leukemia. J Clin Oncol, 23*:* 6296-305, 2005.
146. Ntzani EE and Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet, 362*:* 1439-44, 2003.
147. Zhou X and Mao KZ. LS Bound based gene selection for DNA microarray data. Bioinformatics, 21*:* 1559-64, 2005.
148. Bo T and Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biol, 3*:* RESEARCH0017, 2002.
149. Huang TM and Kecman V. Gene extraction for cancer diagnosis by support vector machines--an improvement. Artif Intell Med, 35*:* 185-94, 2005.
150. Liu X, Krishnan A, and Mondry A. An entropy-based gene selection method for cancer classification using microarray data. BMC Bioinformatics, 6*:* 76, 2005.
151. Draghici S, Khatri P, Eklund AC, and Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet, 22*:* 101-9, 2006.
152. Ioannidis JP. Microarrays and molecular research: noise discovery? Lancet, 365*:* 454-5, 2005.
153. Gardner SN and Fernandes M. Prediction of cancer outcome with microarrays. Lancet, 365*:* 1685, 2005.
154. Biganzoli E, Lama N, Ambrogi F, Antolini L, and Boracchi P. Prediction of cancer outcome with microarrays. Lancet, 365*:* 1683; author reply 4-5, 2005.

155. Sledge GW, Jr. What is targeted therapy? J Clin Oncol, 23*:* 1614-5, 2005.
156. Takimoto C and Kruzelock R. Noval Agents and New Paradigms for colorectal cancers: beyond EGFR and VEGF. Ney York: Human Press Inc, 2007.
157. Tuhackova Z. Molecular therapeutics--lessons from the role of Src in cellular signalling. Folia Biol (Praha), 51*:* 114-20, 2005.
158. Rockey WM and Elcock AH. Rapid computational identification of the targets of protein kinase inhibitors. Curr Opin Drug Discov Devel, 9*:* 326-31, 2006.
159. Kola I and Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov, 3*:* 711-5, 2004.
160. Steffens S, Burger F, Pelli G, et al. Short-term treatment with anti-CD3 antibody reduces the development and progression of atherosclerosis in mice. Circulation, 114*:* 1977-84, 2006.
161. Bayry J, Lacroix-Desmazes S, Kazatchkine MD, and Kaveri SV. Monoclonal antibody and intravenous immunoglobulin therapy for rheumatic diseases: rationale and mechanisms of action. Nat Clin Pract Rheumatol, 3*:* 262-72, 2007.
162. Reichert JM, Rosensweig CJ, Faden LB, and Dewitz MC. Monoclonal antibody successes in the clinic. Nat Biotechnol, 23*:* 1073-8, 2005.
163. http://www.centerwatch.com/patient/drugs/druglist.html.
164. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res, 34*:* D668-72, 2006.
165. Gandhi V, Keating MJ, Bate G, and Kirkpatrick P. Nelarabine. Nat Rev Drug Discov, 5*:* 17-8, 2006.
166. Brekke OH and Sandlie I. Therapeutic antibodies for human diseases at the dawn of the twenty-first century. Nat Rev Drug Discov, 2*:* 52-62, 2003.
167. Carter PJ. Potent antibody therapeutics by design. Nat Rev Immunol, 6*:* 343-57, 2006.
168. Kohler G and Milstein C. Continuous cultures of fused cells secreting antibody of predefined specificity. Nature, 256*:* 495-7, 1975.
169. Miller RA, Maloney DG, Warnke R, and Levy R. Treatment of B-cell lymphoma with monoclonal anti-idiotype antibody. N Engl J Med, 306*:* 517-22, 1982.
170. Oh CS, Stratta RJ, Fox BC, et al. Increased infections associated with the use of OKT3 for treatment of steroid-resistant rejection in renal transplantation. Transplantation, 45*:* 68-73, 1988.
171. Sinnott JTt, Cullison JP, Sweeney MS, and Weinstein SS. Infections in patients receiving OKT3 monoclonal antibody for cardiac rejection: results of a small clinical trial. Tex Heart Inst J, 15*:* 102-6, 1988.
172. D'Alessandro AM, Pirsch JD, Stratta RJ, et al. OKT3 salvage therapy in a quadruple immunosuppressive protocol in cadaveric renal transplantation. Transplantation, 47*:* 297-300, 1989.
173. Jones PT, Dear PH, Foote J, Neuberger MS, and Winter G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. Nature, 321*:* 522-5, 1986.
174. group OMts. A randomized clinical trial of OKT3 monoclonal antibody for acute rejection of cadaveric renal transplants. Ortho Multicenter Transplant Study Group. N Engl J Med, 313*:* 337-42, 1985.
175. Nightingale SL. From the Food and Drug Administration. Jama, 273*:* 982,

1995.
176. Waldmann TA. Immunotherapy: past, present and future. Nat Med, 9*:* 269-77, 2003.
177. Wang W, Singh S, Zeng DL, King K, and Nema S. Antibody structure, instability, and formulation. J Pharm Sci, 96*:* 1-26, 2007.
178. Pavlou AK and Belsey MJ. The therapeutic antibodies market to 2008. Eur J Pharm Biopharm, 59*:* 389-96, 2005.
179. Reichert J and Pavolu A. Monoclonal antibodies market. Nat Rev Drug Discov, 3*:* 383-4, 2004.
180. Kurup S, Wijnhoven TJ, Jenniskens GJ, et al. Characterization of anti-heparan sulfate phage-display antibodies AO4B08 and HS4E4. J Biol Chem, 2007.
181. Floss DM, Falkenburg D, and Conrad U. Production of vaccines and therapeutic antibodies for veterinary applications in transgenic plants: an overview. Transgenic Res, 16*:* 315-32, 2007.
182. Saltz L, Easley C, and Kirkpatrick P. Panitumumab. Nat Rev Drug Discov, 5*:* 987-8, 2006.
183. Yang XD, Jia XC, Corvalan JR, et al. Eradication of established tumors by a fully human monoclonal antibody to the epidermal growth factor receptor without concomitant chemotherapy. Cancer Res, 59*:* 1236-43, 1999.
184. Carter P, Presta L, Gorman CM, et al. Humanization of an anti-p185HER2 antibody for human cancer therapy. Proc Natl Acad Sci U S A, 89*:* 4285-9, 1992.
185. Goldenberg MM. Trastuzumab, a recombinant DNA-derived humanized monoclonal antibody, a novel agent for the treatment of metastatic breast cancer. Clin Ther, 21*:* 309-18, 1999.
186. Rosenfeld PJ, Brown DM, Heier JS, et al. Ranibizumab for neovascular age-related macular degeneration. N Engl J Med, 355*:* 1419-31, 2006.
187. Padlan EA and Helm BA. Modelling study of IgE/receptor interactions. Biochem Soc Trans, 21*:* 963-7, 1993.
188. Garman SC, Wurzburg BA, Tarchevskaya SS, Kinet JP, and Jardetzky TS. Structure of the Fc fragment of human IgE bound to its high-affinity receptor Fc epsilonRI alpha. Nature, 406*:* 259-66, 2000.
189. Tormo J, Blaas D, Parry NR, et al. Crystal structure of a human rhinovirus neutralizing antibody complexed with a peptide derived from viral capsid protein VP2. Embo J, 13*:* 2247-56, 1994.
190. Friedman AR, Roberts VA, and Tainer JA. Predicting molecular interactions and inducible complementarity: fragment docking of Fab-peptide complexes. Proteins, 20*:* 15-24, 1994.
191. Irnaten M, Gallet X, Festy F, et al. Prediction of epitopes and production of monoclonal antibodies against gastric H,K-ATPase. Protein Eng, 11*:* 949-55, 1998.
192. Choulier L, Andersson K, Hamalainen MD, et al. QSAR studies applied to the prediction of antigen-antibody interaction kinetics as measured by BIACORE. Protein Eng, 15*:* 373-82, 2002.
193. Ohtsuka K, Kuroki M, Nojima T, Waki M, and Takenaka S. Interaction analysis of the carcinoembryonic antigen (CEA) with its monoclonal antibody immobilized on a gold surface using Fourier transform infrared reflection-absorption spectroscopy (FT-IR RAS). Anal Sci, 21*:* 215-8, 2005.

194. Berman H, Henrick K, and Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol, 10*:* 980, 2003.
195. Shomer B. Seqalert--a daily sequence alertness server for the EMBL and SWISSPROT databases. Comput Appl Biosci, 13*:* 545-7, 1997.
196. Petrovsky N and Brusic V. Computational immunology: The coming of age. Immunol Cell Biol, 80*:* 248-54, 2002.
197. Johnson G and Wu TT. Kabat Database and its applications: future directions. Nucleic Acids Res, 29*:* 205-6, 2001.
198. Wu TT and Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med, 132*:* 211-50, 1970.
199. Lefranc MP. IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev Comp Immunol, 26*:* 697-705, 2002.
200. Lefranc MP, Giudicelli V, Kaas Q, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res, 33*:* D593-7, 2005.
201. Giudicelli V, Duroux P, Ginestoux C, et al. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res, 34*:* D781-4, 2006.
202. Schonbach C, Koh JL, Flower DR, Wong L, and Brusic V. FIMM, a database of functional molecular immunology: update 2002. Nucleic Acids Res, 30*:* 226-9, 2002.
203. Chen J, Anderson JB, DeWeese-Scott C, et al. MMDB: Entrez's 3D-structure database. Nucleic Acids Res, 31*:* 474-7, 2003.
204. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, and Flower DR. JenPep: a novel computational information resource for immunobiology and vaccinology. J Chem Inf Comput Sci, 43*:* 1276-87, 2003.
205. Peters B, Sidney J, Bourne P, et al. The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol, 3*:* e91, 2005.
206. Singh MK, Srivastava S, Raghava GP, and Varshney GC. HaptenDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies. Bioinformatics, 22*:* 253-5, 2006.
207. Schlessinger A, Ofran Y, Yachdav G, and Rost B. Epitome: database of structure-inferred antigenic epitopes. Nucleic Acids Res, 34*:* D777-80, 2006.
208. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, and Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics, 50*:* 213-9, 1999.
209. Allcorn LC and Martin AC. SACS--self-maintaining database of antibody crystal structure information. Bioinformatics, 18*:* 175-81, 2002.
210. Robinson J, Waller MJ, Parham P, et al. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res, Vol. 31, pp. 311-4, 2003.
211. Bhasin M, Singh H, and Raghava GP. MHCBN: a comprehensive database of MHC binding and non-binding peptides. Bioinformatics, 19*:* 665-6, 2003.
212. Helmberg W, Dunivin R, and Feolo M. The reagent database at dbMHC. Tissue Antigens, 63*:* 142-8, 2004.
213. Blumenfeld OO and Patnaik SK. Allelic genes of blood group antigens: a source of human mutations and cSNPs documented in the Blood Group

Antigen Gene Mutation Database. Hum Mutat, 23*:* 8-16, 2004.
214. Saha S, Bhasin M, and Raghava GP. Bcipep: a database of B-cell epitopes. BMC Genomics, 6*:* 79, 2005.
215. Retter I, Althaus HH, Munch R, and Muller W. VBASE2, an integrative V gene database. Nucleic Acids Res, 33*:* D671-4, 2005.
216. Novellino L, Castelli C, and Parmiani G. A listing of human tumor antigens recognized by T cells: March 2004 update. Cancer Immunol Immunother, 54*:* 187-207, 2005.
217. Bette T. M. Korber CB, Barton F. Haynes, Richard Koup, John P. Moore, Bruce D. Walker, and David I. Watkins. HIV Molecular Immunology 2005. Los Alamos, New Mexico: Los Alamos National Laboratory, Theoretical Biology and Biophysics, 2005.
218. Yusim K, Richardson R, Tao N, et al. Los alamos hepatitis C immunology database. Appl Bioinformatics, 4*:* 217-25, 2005.
219. Reche PA, Zhang H, Glutting JP, and Reinherz EL. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics, 21*:* 2140-1, 2005.
220. Toseland CP, Clayton DJ, McSparron H, et al. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res, 1*:* 4, 2005.
221. Huang J and Honda W. CED: a conformational epitope database. BMC Immunol, 7*:* 7, 2006.
222. Tenette-Souaille C and Smith JC. Structure of the Malpha2-3 toxin alpha antibody-antigen complex: combination of modelling with functional mapping experimental results. Protein Eng, 13*:* 345-51, 2000.
223. Chen J, Liu H, Yang J, and Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids, 2007.
224. Bock JR and Gough DA. Predicting protein--protein interactions from primary structure. Bioinformatics, 17*:* 455-60, 2001.
225. Lo SL, Cai CZ, Chen YZ, and Chung MC. Effect of training datasets on support vector machine prediction of protein-protein interactions. Proteomics, 5*:* 876-84, 2005.
226. Chou KC and Cai YD. Predicting protein-protein interactions from sequences in a hybridization space. J Proteome Res, 5*:* 316-22, 2006.
227. Cui J, Han LY, Lin HH, et al. Advances in exploration of machine learning methods for predicting functional class and interaction profiles of proteins and peptides irrespective of sequence homology. Curr. Bioinformatics accepted, 2007.
228. Vapnik V. Estimation of dependences based on empirical data [in Russian]. [English tanslation: Spring Verlag, New York, 1982], 1979.
229. Vapnik V. The nature of statistical learning theory. New York: Springer, 1995.
230. Souheil Ben-Yacoub YA, and Eddy Mayoraz. Fusion of Face and Speech Data for Person Identity Verification. IEEE transactions on neural networks, 10*:* 1065-74, 1999.
231. Karlsen REG, David J.; Gerhart, Grant R. Target classification via support vector machines. Optical Engineering, 39*:* 704-11, 2000.
232. Shin CSK, K.I. Park, M.H. Kim, H.J. Support vector machine-based text detection in digital video. Pattern recognition, 34*:* 527-9, 2001.
233. Yuan Z, Burrage K, and Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins, 48*:* 566-70, 2002.

234. Ding CH and Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 17*:* 349-58, 2001.
235. Hua S and Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol, 308*:* 397-407, 2001.
236. Cai CZ, Han LY, Ji ZL, Chen X, and Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res, 31*:* 3692-7, 2003.
237. Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2*:* 121-67, 1998.
238. Platt J. Fast Training of Support Vector Machines using Sequential Minimal Optimization: MIT Press, 1998.
239. Keerthi SS and Lin CJ. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput, 15*:* 1667-89, 2003.
240. Lin H-T, C.-J. Lin. A study on sigmoid kernels for SVM and the trainingof non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University., 2003.
241. Gunnarsson RK and Lanke J. The predictive value of microbiologic diagnostic tests if asymptomatic carriers are present. Stat Med, 21*:* 1773-85, 2002.
242. Schuchhardt J, Beule D, Malik A, et al. Normalization strategies for cDNA microarrays. Nucleic Acids Res, 28*:* E47, 2000.
243. Tu Y, Stolovitzky G, and Klein U. Quantitative noise analysis for gene expression microarray experiments. Proc Natl Acad Sci U S A, 99*:* 14031-6, 2002.
244. Bo TH, Dysvik B, and Jonassen I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res, 32*:* e34, 2004.
245. de Brevern AG, Hazout S, and Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. BMC Bioinformatics, 5*:* 114, 2004.
246. Hu J, Li H, Waterman MS, and Zhou XJ. Integrative missing value estimation for microarray data. BMC Bioinformatics, 7*:* 449, 2006.
247. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics, 17*:* 520-5, 2001.
248. Kim H, Golub GH, and Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics, 21*:* 187-98, 2005.
249. Oba S, Sato MA, Takemasa I, et al. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics, 19*:* 2088-96, 2003.
250. Scholz M, Kaplan F, Guy CL, Kopka J, and Selbig J. Non-linear PCA: a missing data approach. Bioinformatics, 21*:* 3887-95, 2005.
251. Tusher VG, Tibshirani R, and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, 98*:* 5116-21, 2001.
252. Bair E and Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol, 2*:* E108, 2004.
253. Scheel I, Aldrin M, Glad IK, et al. The influence of missing value imputation on detection of differentially expressed genes from microarray

data. Bioinformatics, 21*:* 4272-9, 2005.
254. http://helix-web.stanford.edu/pubs/impute/.
255. Lee PD, Sladek R, Greenwood CM, and Hudson TJ. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res, 12*:* 292-7, 2002.
256. Norman Morrison MR, Martin Brutsche, Stephen G. Oliver, Andrew Hayes, Nianshu Zhang, Chris Penkett, Jacqui Lockey, Sudha Rao, Ian Hayes, Ray Jupp, Andy Brass. Robust normalization of microarray data over multiple experiments. Nature Genetics, 23*:* 64, 1999.
257. Chu W, Ghahramani Z, Falciani F, and Wild DL. Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinformatics, 21*:* 3385-93, 2005.
258. Michael E. Wall AR, Luis M. Rocha. Microarray analysis techniques:Singular value decomposition and principal component analysis: Kluwer Academic Press, 2002.
259. Guyon I, Weston J, Barnhill S, and Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46*:* 389-422, 2002.
260. Sima C, Braga-Neto U, and Dougherty ER. Superior feature-set ranking for small samples using bolstered error estimation. Bioinformatics, 21*:* 1046-54, 2005.
261. Fu WJ, Carroll RJ, and Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics, 21*:* 1979-86, 2005.
262. Ohlstein EH, Ruffolo RR, Jr., and Elliott JD. Drug discovery in the next millennium. Annu Rev Pharmacol Toxicol, 40*:* 177-91, 2000.
263. Drews J. Strategic choices facing the pharmaceutical industry: a case for innovation. Drug Discov. Today., 2*:* 72-8, 1997.
264. Hopkins AL and Groom CR. The druggable genome. Nat Rev Drug Discov, 1*:* 727-30, 2002.
265. Wang S, Sim TB, Kim YS, and Chang YT. Tools for target identification and validation. Curr Opin Chem Biol, 8*:* 371-7, 2004.
266. Hajduk PJ, Huth JR, and Fesik SW. Druggability indices for protein targets derived from NMR-based screening data. J Med Chem, 48*:* 2518-25, 2005.
267. Hajduk PJ, Huth JR, and Tse C. Predicting protein druggability. Drug Discov Today, 10*:* 1675-82, 2005.
268. Han LY, Zheng CJ, Xie B, et al. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Discov Today, 12*:* 304-13, 2007.
269. Baldi P, Brunak S, Chauvin Y, Andersen CA, and Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 16*:* 412-24, 2000.
270. Zheng CJ, Han LY, Yap CW, et al. Therapeutic targets: progress of their exploration and investigation of their characteristics. Pharmacol Rev, 58*:* 259-79, 2006.
271. Chen X, Ji ZL, and Chen YZ. TTD: Therapeutic Target Database. Nucleic Acids Res, 30*:* 412-5, 2002.
272. Finn RD, Mistry J, Schuster-Bockler B, et al. Pfam: clans, web tools and services. Nucleic Acids Res, 34*:* D247-51, 2006.
273. Han L, Cui J, Lin H, et al. Recent progresses in the application of machine learning approach for predicting protein functional class independent of

sequence similarity. Proteomics, 6*:* 4023-37, 2006.
274. Li ZR, Lin HH, Han LY, et al. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res, 34*:* W32-7, 2006.
275. Gasteiger E, Hoogland C, Gattiker A, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: MW John (ed.), The Proteomics Protocols Handbook, pp. 571-607: Humana Press, 2005.
276. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 34*:* D173-80, 2006.
277. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 35*:* D5-12, 2007.
278. Bock JR and Gough DA. Whole-proteome interaction mining. Bioinformatics, 19*:* 125-34, 2003.
279. Martin S, Roe D, and Faulon JL. Predicting protein-protein interactions using signature products. Bioinformatics, 21*:* 218-26, 2005.
280. Cui J, Han LY, Lin HH, et al. Advances in exploration of machine learning methods for predicting functional class and interaction profiles of proteins and peptides irrespective of sequence homology. Curr. Bioinformatics 95-112, 2007.
281. Cotran RS, Kumar V, and Collins T. Robbins pathologic basis of disease, 6th edition edition, p. 260. Philsdelphia London Toronto Montreal Sydney Tokyo: W.B.Saunders Company, 1999.
282. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2007. CA Cancer J Clin, 57*:* 43-66, 2007.
283. Kaz AM and Brentnall TA. Genetic testing for colon cancer. Nat Clin Pract Gastroenterol Hepatol, 3*:* 670-9, 2006.
284. Gollub J, Ball CA, Binkley G, et al. The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res, 31*:* 94-6, 2003.
285. http://microarray.princeton.edu/oncology/.
286. http://rana.lbl.gov/EisenSoftware.htm.
287. Zbankova S, Bryndova J, Kment M, and Pacha J. Expression of 11beta-hydroxysteroid dehydrogenase types 1 and 2 in colorectal cancer. Cancer Lett, 210*:* 95-100, 2004.
288. Heron-Milhavet L, Franckhauser C, Rana V, et al. Only Akt1 is required for proliferation, while Akt2 promotes cell cycle exit through p21 binding. Mol Cell Biol, 26*:* 8267-80, 2006.
289. Fan Y, Chen H, Qiao B, et al. c-Jun NH2-terminal kinase decreases ubiquitination and promotes stabilization of p21(WAF1/CIP1) in K562 cell. Biochem Biophys Res Commun, 355*:* 263-8, 2007.
290. Ukomadu C and Dutta A. p21-dependent inhibition of colon cancer cell growth by mevastatin is independent of inhibition of G1 cyclin-dependent kinases. J Biol Chem, 278*:* 43586-94, 2003.
291. Chirco R, Liu XW, Jung KK, and Kim HR. Novel functions of TIMPs in cell signaling. Cancer Metastasis Rev, 25*:* 99-113, 2006.
292. Kato Y, Lewalle JM, Baba Y, et al. Induction of SPARC by VEGF in human vascular endothelial cells. Biochem Biophys Res Commun, 287*:* 422-6, 2001.
293. Giannini G, Cerignoli F, Mellone M, et al. High mobility group A1 is a

molecular target for MYCN in human neuroblastoma. Cancer Res, 65*:* 8308-16, 2005.
294. Russell JP, Shinohara S, Melillo RM, et al. Tyrosine kinase oncoprotein, RET/PTC3, induces the secretion of myeloid growth and chemotactic factors. Oncogene, 22*:* 4569-77, 2003.
295. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell, 4th edition edition. New York: Garland Science, 2002.
296. Varga AE, Stourman NV, Zheng Q, et al. Silencing of the Tropomyosin-1 gene by DNA methylation alters tumor suppressor function of TGF-beta. Oncogene, 24*:* 5043-52, 2005.
297. Raval GN, Bharadwaj S, Levine EA, et al. Loss of expression of tropomyosin-1, a novel class II tumor suppressor that induces anoikis, in primary breast tumors. Oncogene, 22*:* 6194-203, 2003.
298. Suh ER, Ha CS, Rankin EB, Toyota M, and Traber PG. DNA methylation down-regulates CDX1 gene expression in colorectal cancer cell lines. J Biol Chem, 277*:* 35795-800, 2002.
299. Muller N, Reinacher-Schick A, Baldus S, et al. Smad4 induces the tumor suppressor E-cadherin and P-cadherin in colon carcinoma cells. Oncogene, 21*:* 6049-58, 2002.
300. Hardy RG, Tselepis C, Hoyland J, et al. Aberrant P-cadherin expression is an early event in hyperplastic and dysplastic transformation in the colon. Gut, 50*:* 513-9, 2002.
301. Bernstein H, Payne CM, Kunke K, et al. A proteomic study of resistance to deoxycholate-induced apoptosis. Carcinogenesis, 25*:* 681-92, 2004.
302. Ward R, Johnson M, Shridhar V, van Deursen J, and Couch FJ. CBP truncating mutations in ovarian cancer. J Med Genet, 42*:* 514-8, 2005.
303. Neviani P, Santhanam R, Trotta R, et al. The tumor suppressor PP2A is functionally inactivated in blast crisis CML through the inhibitory activity of the BCR/ABL-regulated SET protein. Cancer Cell, 8*:* 355-68, 2005.
304. Struyf S, Schutyser E, Gouwy M, et al. PARC/CCL18 is a plasma CC chemokine with increased levels in childhood acute lymphoblastic leukemia. Am J Pathol, 163*:* 2065-75, 2003.
305. Cao G, O'Brien CD, Zhou Z, et al. Involvement of human PECAM-1 in angiogenesis and in vitro endothelial cell migration. Am J Physiol Cell Physiol, 282*:* C1181-90, 2002.
306. Casanova ML, Bravo A, Martinez-Palacio J, et al. Epidermal abnormalities and increased malignancy of skin tumors in human epidermal keratin 8-expressing transgenic mice. Faseb J, 18*:* 1556-8, 2004.
307. Song S, Byrd JC, Mazurek N, et al. Galectin-3 modulates MUC2 mucin expression in human colon cancer cells at the level of transcription via AP-1 activation. Gastroenterology, 129*:* 1581-91, 2005.
308. Mizoshita T, Tsukamoto T, Inada KI, et al. Loss of MUC2 expression correlates with progression along the adenoma-carcinoma sequence pathway as well as de novo carcinogenesis in the colon. Histol Histopathol, 22*:* 251-60, 2007.
309. Rust R, Visser L, van der Leij J, et al. High expression of calcium-binding proteins, S100A10, S100A11 and CALM2 in anaplastic large cell lymphoma. Br J Haematol, 131*:* 596-608, 2005.
310. Rushmere NK, Knowlden JM, Gee JM, et al. Analysis of the level of mRNA expression of the membrane regulators of complement, CD59, CD55 and CD46, in breast cancer. Int J Cancer, 108*:* 930-6, 2004.

311. Collart MA. Global control of gene expression in yeast by the Ccr4-Not complex. Gene, 313*:* 1-16, 2003.
312. Le Sourd F, Boulben S, Le Bouffant R, et al. eEF1B: At the dawn of the 21st century. Biochim Biophys Acta, 1759*:* 13-31, 2006.
313. Tanaka H, Shirkoohi R, Nakagawa K, et al. siRNA gelsolin knockdown induces epithelial-mesenchymal transition with a cadherin switch in human mammary epithelial cells. Int J Cancer, 118*:* 1680-91, 2006.
314. Liloglou T, Walters M, Maloney P, Youngson J, and Field JK. A T2517C polymorphism in the GSTM4 gene is associated with risk of developing lung cancer. Lung Cancer, 37*:* 143-6, 2002.
315. DeRubertis FR, Chayoth R, and Field JB. The content and metabolism of cyclic adenosine 3', 5'-monophosphate and cyclic guanosine 3', 5'-monophosphate in adenocarcinoma of the human colon. J Clin Invest, 57*:* 641-9, 1976.
316. Ushigome M, Ubagai T, Fukuda H, et al. Up-regulation of hnRNP A1 gene in sporadic human colorectal cancers. Int J Oncol, 26*:* 635-40, 2005.
317. Shang L and Tomasi TB. The heat shock protein 90-CDC37 chaperone complex is required for signaling by types I and II interferons. J Biol Chem, 281*:* 1876-84, 2006.
318. Tahara E, Jr., Tahara H, Kanno M, et al. G1P3, an interferon inducible gene 6-16, is expressed in gastric cancers and inhibits mitochondrial-mediated apoptosis in gastric cancer cell line TMK-1 cell. Cancer Immunol Immunother, 54*:* 729-40, 2005.
319. Cherbonnel-Lasserre CL, Linares-Cruz G, Rigaut JP, Sabatier L, and Dutrillaux B. Strong decrease in biotin content may correlate with metabolic alterations in colorectal adenocarcinoma. Int J Cancer, 72*:* 768-75, 1997.
320. Lauer C, Volkl A, Riedl S, Fahimi HD, and Beier K. Impairment of peroxisomal biogenesis in human colon carcinoma. Carcinogenesis, 20*:* 985-9, 1999.
321. Bao S, Ouyang G, Bai X, et al. Periostin potently promotes metastatic growth of colon cancer by augmenting cell survival via the Akt/PKB pathway. Cancer Cell, 5*:* 329-39, 2004.
322. Takada T, Noguchi T, Inagaki K, et al. Induction of apoptosis by stomach cancer-associated protein-tyrosine phosphatase-1. J Biol Chem, 277*:* 34359-66, 2002.
323. Seo Y, Matozaki T, Tsuda M, et al. Overexpression of SAP-1, a transmembrane-type protein tyrosine phosphatase, in human colorectal cancers. Biochem Biophys Res Commun, 231*:* 705-11, 1997.
324. Buchholz M, Biebl A, Neesse A, et al. SERPINE2 (protease nexin I) promotes extracellular matrix production and local invasion of pancreatic tumors in vivo. Cancer Res, 63*:* 4945-51, 2003.
325. Valineva T, Yang J, Palovuori R, and Silvennoinen O. The transcriptional co-activator protein p100 recruits histone acetyltransferase activity to STAT6 and mediates interaction between the CREB-binding protein and STAT6. J Biol Chem, 280*:* 14989-96, 2005.
326. Pena C, Garcia JM, Silva J, et al. E-cadherin and vitamin D receptor regulation by SNAIL and ZEB1 in colon cancer: clinicopathological correlations. Hum Mol Genet, 14*:* 3361-70, 2005.
327. Guaita S, Puig I, Franci C, et al. Snail induction of epithelial to mesenchymal transition in tumor cells is accompanied by MUC1

repression and ZEB1 expression. J Biol Chem, 277*:* 39209-16, 2002.
328. Kulawiec M, Arnouk H, Desouki MM, et al. Proteomic analysis of mitochondria-to-nucleus retrograde response in human cancer. Cancer Biol Ther, 5*:* 967-75, 2006.
329. Hayashi M, Fearns C, Eliceiri B, Yang Y, and Lee JD. Big mitogen-activated protein kinase 1/extracellular signal-regulated kinase 5 signaling pathway is essential for tumor-associated angiogenesis. Cancer Res, 65*:* 7699-706, 2005.
330. Lui HM, Chen J, Wang L, and Naumovski L. ARMER, apoptotic regulator in the membrane of the endoplasmic reticulum, a novel inhibitor of apoptosis. Mol Cancer Res, 1*:* 508-18, 2003.
331. Kanehisa M, Goto S, Kawashima S, Okuno Y, and Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res, 32*:* D277-80, 2004.
332. Vastrik I, D'Eustachio P, Schmidt E, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol, 8*:* R39, 2007.
333. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer, 4*:* 177-83, 2004.
334. Vogelstein B and Kinzler KW. Cancer genes and the pathways they control. Nat Med, 10*:* 789-99, 2004.
335. de Castro Junior G, Puglisi F, de Azambuja E, El Saghir NS, and Awada A. Angiogenesis and cancer: A cross-talk between basic science and clinical trials (the "do ut des" paradigm). Crit Rev Oncol Hematol, 59*:* 40-50, 2006.
336. Mancuso A and Sternberg CN. Colorectal cancer and antiangiogenic therapy: what can be expected in clinical practice? Crit Rev Oncol Hematol, 55*:* 67-81, 2005.
337. Irish JM, Kotecha N, and Nolan GP. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. Nat Rev Cancer, 6*:* 146-55, 2006.
338. Muller AJ and Scherle PA. Targeting the mechanisms of tumoral immune tolerance with small-molecule inhibitors. Nat Rev Cancer, 6*:* 613-25, 2006.
339. Maglott D, Ostell J, Pruitt KD, and Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res, 35*:* D26-31, 2007.
340. http://bidd.nus.edu.sg/group/cjttd/ttd.asp.
341. Cruz JA and Wishart DS. Applications of Machine Learning in Cancer Predictionand Prognosis. Cancer Informatics, 2*:* 59-78, 2006.
342. http://www.merck.com/mmpe/sec05/ch062/ch062b.html#sec05-ch062-ch062b-1405. Lung Carcinoma: Tumors of the Lungs, Online edition. Merck Manual Professional Edition.
343. Huber RM and Stratakis DF. Molecular oncology--perspectives in lung cancer. Lung Cancer, 45 Suppl 2*:* S209-13, 2004.
344. Sorensen JB, Hirsch FR, Gazdar A, and Olsen JE. Interobserver variability in histopathologic subtyping and grading of pulmonary adenocarcinoma. Cancer, 71*:* 2971-6, 1993.
345. Gail MH, Eagan RT, Feld R, et al. Prognostic factors in patients with resected stage I non-small cell lung cancer. A report from the Lung Cancer Study Group. Cancer, 54*:* 1802-13, 1984.
346. Takise A, Kodama T, Shimosato Y, Watanabe S, and Suemasu K. Histopathologic prognostic factors in adenocarcinomas of the peripheral

lung less than 2 cm in diameter. Cancer, 61*:* 2083-8, 1988.
347. Okada M, Tsubota N, Yoshimura M, Miyamoto Y, and Nakai R. Evaluation of TMN classification for lung carcinoma with ipsilateral intrapulmonary metastasis. Ann Thorac Surg, 68*:* 326-30; discussion 31, 1999.
348. Harpole DH, Jr., Herndon JE, 2nd, Wolfe WG, Iglehart JD, and Marks JR. A prognostic model of recurrence and death in stage I non-small cell lung cancer utilizing presentation, histopathology, and oncoprotein expression. Cancer Res, 55*:* 51-6, 1995.
349. Guo L, Ma Y, Ward R, et al. Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. Clin Cancer Res, 12*:* 3344-54, 2006.
350. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med, 8*:* 816-24, 2002.
351. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A, 98*:* 13790-5, 2001.
352. Edgerton E, H. Fisher, Lianhong Tang, Lewis J. Frey, and Chen Z. Data Mining for Gene Networks Relevant to Poor Prognosis in Lung Cancer Via Backward-Chaining Rule Induction. Cancer Informatics, 2*:* 93-114, 2007.
353. Lu Y, Lemon W, Liu PY, et al. A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLoS Med, 3*:* e467, 2006.
354. Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med, 356*:* 11-20, 2007.
355. Xu J, Yang Y, and Ott J. Survival analysis of microarray expression data by transformation models. Comput Biol Chem, 29*:* 91-4, 2005.
356. http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html.
357. http://www.genome.wi.mit.edu/MPR/lung.
358. http://www.xlstat.com/en/support/tutorials/km.htm.
359. Joo YE, Sohn YH, Lee WS, et al. Expression of vascular endothelial growth factor and p53 in pancreatic carcinomas. Korean J Intern Med, 17*:* 153-9, 2002.
360. Strohmeyer D, Rossing C, Bauerfeind A, et al. Vascular endothelial growth factor and its correlation with angiogenesis and p53 expression in prostate cancer. Prostate, 45*:* 216-24, 2000.
361. Maeda K, Kang SM, Onoda N, et al. Expression of p53 and vascular endothelial growth factor associated with tumor angiogenesis and prognosis in gastric cancer. Oncology, 55*:* 594-9, 1998.
362. Liu DH, Zhang XY, Fan DM, et al. Expression of vascular endothelial growth factor and its role in oncogenesis of human gastric carcinoma. World J Gastroenterol, 7*:* 500-5, 2001.
363. Lee JS, Kim HS, Jung JJ, Lee MC, and Park CS. Expression of vascular endothelial growth factor in adenocarcinomas of the uterine cervix and its relation to angiogenesis and p53 and c-erbB-2 protein expression. Gynecol Oncol, 85*:* 469-75, 2002.
364. Gills JJ, Granville CA, and Dennis PA. Targeting aberrant signal transduction pathways in lung cancer. Cancer Biol Ther, 3*:* 147-55, 2004.
365. Muraoka K, Nabeshima K, Murayama T, Biswas C, and Koono M. Enhanced expression of a tumor-cell-derived collagenase-stimulatory factor in urothelial carcinoma: its usefulness as a tumor marker for bladder

cancers. Int J Cancer, 55*:* 19-26, 1993.
366. Caudroy S, Polette M, Nawrocki-Raby B, et al. EMMPRIN-mediated MMP regulation in tumor and endothelial cells. Clin Exp Metastasis, 19*:* 697-702, 2002.
367. Tang Y, Nakada MT, Kesavan P, et al. Extracellular matrix metalloproteinase inducer stimulates tumor angiogenesis by elevating vascular endothelial cell growth factor and matrix metalloproteinases. Cancer Res, 65*:* 3193-9, 2005.
368. Yan L, Zucker S, and Toole BP. Roles of the multifunctional glycoprotein, emmprin (basigin; CD147), in tumour progression. Thromb Haemost, 93*:* 199-204, 2005.
369. Klein CA, Seidl S, Petat-Dutter K, et al. Combined transcriptome and genome analysis of single micrometastatic cells. Nat Biotechnol, 20*:* 387-92, 2002.
370. Strieter RM, Belperio JA, Burdick MD, et al. CXC chemokines: angiogenesis, immunoangiostasis, and metastases in lung cancer. Ann N Y Acad Sci, 1028*:* 351-60, 2004.
371. Strieter RM, Belperio JA, Burdick MD, and Keane MP. CXC chemokines in angiogenesis relevant to chronic fibroproliferation. Curr Drug Targets Inflamm Allergy, 4*:* 23-6, 2005.
372. Tzouvelekis A, Anevlavis S, and Bouros D. Angiogenesis in interstitial lung diseases: a pathogenetic hallmark or a bystander? Respir Res, 7*:* 82, 2006.
373. Strieter RM, Belperio JA, Phillips RJ, and Keane MP. Chemokines: Angiogenesis and Metastases in Lung Cancer, p. 173-88: John Wiley & Sons, 2004.
374. Schuller HM. Neurotransmitter receptor-mediated signaling pathways as modulators of carcinogenesis. Prog Exp Tumor Res, 39*:* 45-63, 2007.
375. Ho YS, Chen CH, Wang YJ, et al. Tobacco-specific carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) induces cell proliferation in normal human bronchial epithelial cells through NFkappaB activation and cyclin D1 up-regulation. Toxicol Appl Pharmacol, 205*:* 133-48, 2005.
376. Hakomori S. Aberrant glycosylation in tumors and tumor-associated carbohydrate antigens. Adv Cancer Res, 52*:* 257-331, 1989.
377. Friederichs J, Zeller Y, Hafezi-Moghadam A, et al. The CD24/P-selectin binding pathway initiates lung arrest of human A125 adenocarcinoma cells. Cancer Res, 60*:* 6714-22, 2000.
378. Martin-Satue M, Marrugat R, Cancelas JA, and Blanco J. Enhanced expression of alpha(1,3)-fucosyltransferase genes correlates with E-selectin-mediated adhesion and metastatic potential of human lung adenocarcinoma cells. Cancer Res, 58*:* 1544-50, 1998.
379. Ohyama C, Tsuboi S, and Fukuda M. Dual roles of sialyl Lewis X oligosaccharides in tumor metastasis and rejection by natural killer cells. Embo J, 18*:* 1516-25, 1999.
380. Grzmil M, Voigt S, Thelen P, et al. Up-regulated expression of the MAT-8 gene in prostate cancer and its siRNA-mediated inhibition of expression induces a decrease in proliferation of human prostate carcinoma cells. Int J Oncol, 24*:* 97-105, 2004.
381. Kayed H, Kleeff J, Kolb A, et al. FXYD3 is overexpressed in pancreatic ductal adenocarcinoma and influences pancreatic cancer cell growth. Int J

Cancer, 118*:* 43-54, 2006.
382. Kim JH, Kim HW, Jeon H, Suh PG, and Ryu SH. Phospholipase D1 regulates cell migration in a lipase activity-independent manner. J Biol Chem, 281*:* 15747-56, 2006.
383. Zhong M, Shen Y, Zheng Y, et al. Phospholipase D prevents apoptosis in v-Src-transformed rat fibroblasts and MDA-MB-231 breast cancer cells. Biochem Biophys Res Commun, 302*:* 615-9, 2003.
384. Kahlina K, Goren I, Pfeilschifter J, and Frank S. p68 DEAD box RNA helicase expression in keratinocytes. Regulation, nucleolar localization, and functional connection to proliferation and vascular endothelial growth factor gene expression. J Biol Chem, 279*:* 44872-82, 2004.
385. Kvissel AK, Ramberg H, Eide T, et al. Androgen dependent regulation of protein kinase A subunits in prostate cancer cells. Cell Signal, 19*:* 401-9, 2007.
386. Burns JM, Summers BC, Wang Y, et al. A novel chemokine receptor for SDF-1 and I-TAC involved in cell survival, cell adhesion, and tumor development. J Exp Med, 203*:* 2201-13, 2006.
387. Strieter RM, Belperio JA, Burdick MD, and Keane MP. CXC Chemokines in Angiogenesis Relevant to Chronic Fibroproliferation. Current Drug Targets - Inflammation & Allergy, 4*:* 23-6, 2005.
388. Oue N, Mitani Y, Aung PP, et al. Expression and localization of Reg IV in human neoplastic and non-neoplastic tissues: Reg IV expression is associated with intestinal and neuroendocrine differentiation in gastric adenocarcinoma. J Pathol, 207*:* 185-98, 2005.
389. Sekikawa A, Fukui H, Fujii S, et al. Possible role of REG Ialpha protein in ulcerative colitis and colitic cancer. Gut, 54*:* 1437-44, 2005.
390. Jang CY, Lee JY, and Kim J. RpS3, a DNA repair endonuclease and ribosomal protein, is involved in apoptosis. FEBS Lett, 560*:* 81-5, 2004.
391. Kim SH and Kim J. Reduction of invasion in human fibrosarcoma cells by ribosomal protein S3 in conjunction with Nm23-H1 and ERK. Biochim Biophys Acta, 1763*:* 823-32, 2006.
392. Robert C, Bolon I, Gazzeri S, et al. Expression of plasminogen activator inhibitors 1 and 2 in lung cancer and their role in tumor progression. Clin Cancer Res, 5*:* 2094-102, 1999.
393. Almholt K, Nielsen BS, Frandsen TL, et al. Metastasis of transgenic breast cancer in plasminogen activator inhibitor-1 gene-deficient mice. Oncogene, 22*:* 4389-97, 2003.
394. Speleman L, Kerrebijn JD, Look MP, et al. Prognostic value of plasminogen activator inhibitor-1 in head and neck squamous cell carcinoma. Head Neck, 29*:* 341-50, 2007.
395. Gil-Bazo I, Paramo JA, and Garcia-Foncillas J. [New prognostic and predictive factors in advanced colorectal cancer]. Med Clin (Barc), 126*:* 541-8, 2006.
396. Goldman NA, Katz EB, Glenn AS, et al. GLUT1 and GLUT8 in endometrium and endometrial adenocarcinoma. Mod Pathol, 19*:* 1429-36, 2006.
397. Tesfaigzi Y, Wright PS, and Belinsky SA. SPRR1B overexpression enhances entry of cells into the G0 phase of the cell cycle. Am J Physiol Lung Cell Mol Physiol, 285*:* L889-98, 2003.
398. Patterson T, Vuong H, Liaw YS, et al. Mechanism of repression of squamous differentiation marker, SPRR1B, in malignant bronchial

epithelial cells: role of critical TRE-sites and its transacting factors. Oncogene, 20*:* 634-44, 2001.
399. Zhong H and Bowen JP. Antiangiogenesis drug design: multiple pathways targeting tumor vasculature. Curr Med Chem, 13*:* 849-62, 2006.
400. Kyu-Ho Han E, Gehrke L, Tahir SK, et al. Modulation of drug resistance by alpha-tubulin in paclitaxel-resistant human lung cancer cell lines. Eur J Cancer, 36*:* 1565-71, 2000.
401. Dumontet C and Sikic BI. Mechanisms of action of and resistance to antitubulin agents: microtubule dynamics, drug transport, and cell death. J Clin Oncol, 17*:* 1061-70, 1999.
402. Kommagani R, Caserta TM, and Kadakia MP. Identification of vitamin D receptor as a target of p63. Oncogene, 25*:* 3745-51, 2006.
403. Ilhan N, Ilhan N, and Deveci F. Functional significance of vascular endothelial growth factor and its receptor (receptor-1) in various lung cancer types. Clin Biochem, 37*:* 840-5, 2004.
404. Dudek AZ and Mahaseth H. Circulating angiogenic cytokines in patients with advanced non-small cell lung cancer: correlation with treatment response and survival. Cancer Invest, 23*:* 193-200, 2005.
405. Kaiser U, Schilli M, Wegmann B, et al. Expression of vitamin D receptor in lung cancer. J Cancer Res Clin Oncol, 122*:* 356-9, 1996.
406. Cooper R, Sarioglu S, Sokmen S, et al. Glucose transporter-1 (GLUT-1): a potential marker of prognosis in rectal carcinoma? Br J Cancer, 89*:* 870-6, 2003.
407. Noh DY, Ahn SJ, Lee RA, et al. Overexpression of phospholipase D1 in human breast cancer tissues. Cancer Lett, 161*:* 207-14, 2000.
408. Zhao Y, Ehara H, Akao Y, et al. Increased activity and intranuclear expression of phospholipase D2 in human renal cancer. Biochem Biophys Res Commun, 278*:* 140-3, 2000.
409. Oka M, Kageshita T, Ono T, et al. Protein kinase C alpha associates with phospholipase D1 and enhances basal phospholipase D activity in a protein phosphorylation-independent manner in human melanoma cells. J Invest Dermatol, 121*:* 69-76, 2003.
410. Madjd Z, Parsons T, Watson NF, et al. High expression of Lewis y/b antigens is associated with decreased survival in lymph node negative breast carcinomas. Breast Cancer Res, 7*:* R780-7, 2005.
411. Castello R, Espana F, Vazquez C, et al. Plasminogen activator inhibitor-1 4G/5G polymorphism in breast cancer patients and its association with tissue PAI-1 levels and tumor severity. Thromb Res, 117*:* 487-92, 2006.
412. Bhuvarahamurthy V, Schroeder J, Denkert C, et al. In situ gene expression of urokinase-type plasminogen activator and its receptor in transitional cell carcinoma of the human bladder. Oncol Rep, 12*:* 909-13, 2004.
413. Shetty S and Idell S. Posttranscriptional regulation of urokinase receptor gene expression in human lung carcinoma and mesothelioma cells in vitro. Mol Cell Biochem, 199*:* 189-200, 1999.
414. Pedersen H, Grondahl-Hansen J, Francis D, et al. Urokinase and plasminogen activator inhibitor type 1 in pulmonary adenocarcinoma. Cancer Res, 54*:* 120-3, 1994.
415. Yonemura Y, Sakurai S, Yamamoto H, et al. REG gene expression is associated with the infiltrating growth of gastric carcinoma. Cancer, 98*:* 1394-400, 2003.
416. Dhar DK, Udagawa J, Ishihara S, et al. Expression of regenerating gene I

in gastric adenocarcinomas: correlation with tumor differentiation status and patient survival. Cancer, 100*:* 1130-6, 2004.
417. Kojima H, Shijubo N, and Abe S. Thymidine phosphorylase and vascular endothelial growth factor in patients with Stage I lung adenocarcinoma. Cancer, 94*:* 1083-93, 2002.
418. Sfar S, Hassen E, Saad H, Mosbah F, and Chouchane L. Association of VEGF genetic polymorphisms with prostate carcinoma risk and clinical outcome. Cytokine, 35*:* 21-8, 2006.
419. Jordan A, Hadfield JA, Lawrence NJ, and McGown AT. Tubulin as a target for anticancer drugs: agents which interact with the mitotic spindle. Med Res Rev, 18*:* 259-96, 1998.
420. Zhou W, Heist RS, Liu G, et al. Polymorphisms of vitamin D receptor and survival in early-stage non-small cell lung cancer patients. Cancer Epidemiol Biomarkers Prev, 15*:* 2239-45, 2006.
421. Shetty S, Bdeir K, Cines DB, and Idell S. Induction of plasminogen activator inhibitor-1 by urokinase in lung epithelial cells. J Biol Chem, 278*:* 18124-31, 2003.
422. Shetty S, Gyetko MR, and Mazar AP. Induction of p53 by urokinase in lung epithelial cells. J Biol Chem, 280*:* 28133-41, 2005.
423. Offersen BV, Pfeiffer P, Andreasen P, and Overgaard J. Urokinase plasminogen activator and plasminogen activator inhibitor type-1 in nonsmall-cell lung cancer: relation to prognosis and angiogenesis. Lung Cancer, 56*:* 43-50, 2007.
424. Cummings R, Parinandi N, Wang L, Usatyuk P, and Natarajan V. Phospholipase D/phosphatidic acid signal transduction: role and physiological significance in lung. Mol Cell Biochem, 234-235*:* 99-109, 2002.
425. Buchanan FG, McReynolds M, Couvillon A, et al. Requirement of phospholipase D1 activity in H-RasV12-induced transformation. Proc Natl Acad Sci U S A, 102*:* 1638-42, 2005.
426. Li XW, Ding YQ, Cai JJ, et al. Studies on mechanism of Sialy Lewis-X antigen in liver metastases of human colorectal carcinoma. World J Gastroenterol, 7*:* 425-30, 2001.
427. Varambally S, Dhanasekaran SM, Zhou M, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature, 419*:* 624-9, 2002.
428. Petrovsky N, Schonbach C, and Brusic V. Bioinformatic strategies for better understanding of immune function. In Silico Biol, 3*:* 411-6, 2003.
429. Rammensee HG. Immunoinformatics: bioinformatic strategies for better understanding of immune function. Introduction. Novartis Found Symp, 254*:* 1-2, 2003.
430. Geneva. WORLD HEALTH ORGANIZATION: International statistical classification of diseases and related health problems, 1992.
431. Senis YA, Tomlinson MG, Garcia A, et al. A comprehensive proteomics and genomics analysis reveals novel transmembrane proteins in human platelets and mouse megakaryocytes including G6b-B, a novel ITIM protein. Mol Cell Proteomics, 2006.
432. Le Naour F. Identification of tumor antigens by using proteomics. Methods Mol Biol, 360*:* 327-34, 2007.
433. Rolla G, Ferrero N, Bergia R, and Guida G. [Perspectives in clinical immunology]. Recenti Prog Med, 97*:* 787-96, 2006.

434. Adair J. Antibody engineering and expression, second edition edition, Vol. 5a, p. 219-44. Weinheim, New York, Chichester,Brisbane, Singapore, Toronto: WILEY-vch, 1999.
435. Roskos LK, Davis CG, and Schwab GM. The clinical pharmacology of therapeutic monoclonal antibodies. Drug Development Research, 61*:* 108-20, 2004.
436. Pantophlet R and Burton DR. GP120: target for neutralizing HIV-1 antibodies. Annu Rev Immunol, 24*:* 739-69, 2006.
437. Presta LG. Selection, design, and engineering of therapeutic antibodies. J Allergy Clin Immunol, 116*:* 731-6; quiz 7, 2005.
438. Imai K and Takaoka A. Comparing antibody and small-molecule therapies for cancer. Nat Rev Cancer, 6*:* 714-27, 2006.
439. Bae K, Mallick BK, and Elsik CG. Prediction of protein interdomain linker regions by a hidden Markov model. Bioinformatics, 21*:* 2264-70, 2005.
440. Bateman A, Birney E, Cerruti L, et al. The Pfam protein families database. Nucleic Acids Res, 30*:* 276-80, 2002.

APPENDICES

Table S1 The information of colon microarray data collected from Standford Microarray Database (SMD) (284)

ExptID in SMD	Tissue Description in SMD	Tissue type
33061	CACO-2	Colon cancer cell line
48604	COLO205	Colon cancer cell line
48603	COLO320	Colon cancer cell line
48602	COLO741	Colon cancer cell line
33055	DLD-1	Colon cancer cell line
48599	GP2D	Colon cancer cell line
48607	GP5D	Colon cancer cell line
48608	HCA7	Colon cancer cell line
60328	HCT 116	Colon cancer cell line
33052	HCT116	Colon cancer cell line
60329	HCT116+Ch2	Colon cancer cell line
60327	HCT116+Ch3	Colon cancer cell line
33058	HCT15	Colon cancer cell line
30367	HT-29	Colon cancer cell line
48606	LIM1215	Colon cancer cell line
48605	LIM2412	Colon cancer cell line
30370	LOVO	Colon cancer cell line
33050	LS-174T	Colon cancer cell line
33048	LS-180	Colon cancer cell line
48600	LS411	Colon cancer cell line
33062	NCI-H508	Colon cancer cell line
33060	NCI-H747	Colon cancer cell line
30369	RKO	Colon cancer cell line
33059	SK-CO-1	Colon cancer cell line
48609	SNUC2B	Colon cancer cell line
33057	SW-1116	Colon cancer cell line
33051	SW-403	Colon cancer cell line
33053	SW-480	Colon cancer cell line
30371	SW-837	Colon cancer cell line
48597	SW1417	Colon cancer cell line
30297	SW48	Colon cancer cell line
33056	SW620	Colon cancer cell line
48598	SW948	Colon cancer cell line
48601	T84	Colon cancer cell line
32614	Colon 2075	Normal colon tissue
20950	Colon, ascending 0222	Normal colon tissue
32615	Colon, sigmoid 0361	Normal colon tissue
16373	colon,autopsy	Normal colon tissue
56125	6.01.04 colonic omentum	Normal colon tissue
56109	7.16.04 colonic omentum	Normal colon tissue

56123	8.30.04 colonic omentum	Normal colon tissue
56121	6.01.04 pericolonic/mesenteric	Normal colon tissue

Table S2 List of 20 derived colon cancer gene signatures selected by SVM class-differentiation systems

Gene Name (EST number)	Number of signatures which included this gene	Gene rank in each signature (Number of selected gene in each signaure)																			
		1 (155)	2 (135)	3 (156)	4 (146)	5 (116)	6 (112)	7 (119)	8 (127)	9 (133)	10 (156)	11 (139)	12 (115)	13 (144)	14 (157)	15 (149)	16 (136)	17 (136)	18 (127)	19 (146)	20 (122)
DES (M63391)	20	1	1	3	3	1	3	1	3	1	1	3	3	1	1	3	1	3	3	1	1
GUCA2B (Z50753)	20	2	3	1	1	2	1	3	1	2	2	1	1	3	2	1	2	1	1	3	2
MYH9 (R87126)	20	3	2	2	2	3	2	2	2	3	3	2	2	2	3	2	3	2	2	2	3
WDR77 (H08393)	20	4	4	5	4	4	5	5	4	4	5	5	5	4	5	5	5	4	5	5	5
CSRP1 (M76378)	20	5	5	4	5	5	4	4	5	5	4	4	4	5	4	4	4	5	4	4	4
VIP (M36634)	20	6	6	6	6	6	6	6	6	7	7	6	6	6	6	6	6	6	6	6	6
MYL9 (J02854)	20	7	7	7	7	7	7	8	7	8	8	7	7	7	7	7	7	7	7	7	8
NR3C2 (R44301)	20	8	11	9	8	8	8	7	8	9	9	8	11	11	8	9	8	10	9	9	9
MYL6 (H20709)	20	9	8	12	9	10	9	14	10	10	10	9	8	9	9	11	9	8	16	10	13
GTF3A (R36977)	20	10	9	11	10	11	10	10	11	11	11	10	9	10	10	10	10	9	11	11	11
HNRPA1 (X12671)	20	11	13	10	12	12	12	12	9	12	12	12	12	12	11	12	11	11	12	12	10
CDH3 (X63629)	20	12	10	8	11	9	11	11	12	6	6	11	10	8	12	8	14	12	8	8	12
CFD (H43887)	20	13	21	14	17	16	21	21	22	19	15	20	21	21	13	20	21	19	22	20	18
MT1G (T71025)	20	14	17	16	19	19	14	9	14	22	17	22	17	14	18	25	12	21	10	13	7
GSN (H06524)	20	15	14	13	14	14	13	15	15	14	13	14	16	15	23	14	15	14	13	14	14
H40095	20	16	20	17	18	21	20	19	20	18	18	21	20	20	16	19	19	18	19	19	19
H64807	20	17	16	18	16	17	17	18	16	16	16	17	15	16	17	18	16	15	14	17	15
TPM2 (T92451)	20	18	15	19	15	23	18	17	18	15	21	16	14	17	14	16	17	16	18	16	16
SNRPB (R84411)	20	19	43	21	20	37	19	16	19	17	42	19	45	19	15	28	18	17	64	44	17
MXI1 (L07648)	20	20	12	20	13	13	15	13	13	13	19	13	13	13	19	13	13	13	15	15	20
DARS (J05032)	20	21	22	15	21	20	16	20	17	21	20	15	18	18	22	15	22	20	17	18	21
SERPINE2 (T51261)	20	22	19	24	23	18	23	22	21	23	24	25	23	23	20	22	20	22	21	25	25
MXI1 (H49870)	20	23	18	28	25	22	28	24	27	25	22	26	19	26	24	21	26	23	20	26	26

HMGA1 (X14958)	20	24	26	22	29	28	31	28	31	24	23	32	26	32	27	34	29	24	28	29	28
S100P (T47377)	20	25	24	23	22	15	22	23	23	20	14	18	22	22	21	17	23	25	23	21	22
FUCA1 (M80815)	20	26	31	25	30	38	29	27	33	29	34	43	32	34	29	32	28	34	32	36	27
HSD11B2 (U14631)	20	27	33	27	26	33	30	25	34	30	36	30	35	27	25	29	32	37	33	38	24
CKS2 (X54942)	20	28	23	26	24	24	26	26	24	28	26	23	24	25	26	26	25	27	24	22	23
CDKN1A (R46753)	20	29	27	30	33	27	32	30	38	26	25	28	28	33	28	36	37	26	29	30	32
C15orf15 (T62947)	20	30	28	29	32	26	27	31	26	27	28	24	30	30	30	24	27	28	27	32	33
TPM1 (X12369)	20	31	25	31	28	29	25	29	35	31	29	27	29	28	39	27	35	29	30	31	29
SPARCL1 (X86693)	20	33	36	33	31	36	36	36	28	36	32	34	37	29	33	31	30	32	37	27	35
EIF2S2 (R54097)	20	34	34	45	44	31	37	32	36	38	30	33	33	31	38	46	31	38	38	34	30
TSPAN1 (H64489)	20	35	29	40	34	30	33	35	25	37	27	35	25	36	31	35	24	30	26	28	34
SND1 (U22055)	20	36	35	34	38	34	38	34	29	35	33	31	36	48	34	30	33	33	36	33	36
PCNP (T95018)	20	37	38	35	37	44	39	33	42	34	39	87	31	35	37	42	34	35	39	35	31
SRPK1 (U09564)	20	38	32	36	40	32	35	52	61	39	35	39	34	38	46	38	40	39	34	37	49
UQCRC1 (R55310)	20	39	41	37	41	41	43	46	43	41	40	40	41	45	41	43	41	41	40	42	46
CST3 (T51534)	20	40	44	41	42	43	40	40	40	43	44	46	43	40	44	40	45	44	43	46	40
S100A11 (T51571)	20	41	39	32	35	25	41	38	30	32	37	41	39	42	35	23	38	42	25	23	37
EEF1B2 (X60489)	20	42	37	38	36	39	34	37	32	33	38	36	38	37	32	33	36	36	35	39	38
ATP5J (R88740)	20	43	45	43	47	45	42	42	46	44	43	48	47	43	42	45	43	43	45	45	42
CLNS1A (U17899)	20	44	30	39	46	40	24	50	41	42	31	37	27	39	43	39	44	31	31	47	39
GYG1 (U31525)	20	45	59	52	51	53	60	60	53	52	55	55	61	51	51	57	58	55	56	52	56
SRPK1 (R62549)	20	47	42	53	54	42	50	45	39	48	50	49	46	57	45	48	49	49	48	53	45
PPP2R5C (T51493)	20	48	57	47	49	49	46	44	49	51	52	51	51	49	60	47	54	56	41	62	53
H11084	20	49	58	51	61	54	48	39	51	50	48	53	44	52	50	37	50	57	42	64	55
MARCKSL1 (X70326)	20	50	46	44	52	48	47	41	47	45	45	47	42	53	40	44	42	45	44	57	41
PRKACA (X07767)	20	51	52	55	50	56	57	47	60	59	56	57	54	46	57	56	57	47	57	56	52
TCF8 (U19969)	20	52	51	57	60	59	62	54	58	58	51	61	49	56	52	54	55	58	62	40	48
SPARC (J03040)	20	53	49	56	53	50	54	53	57	53	53	56	50	55	53	55	56	54	51	58	50

HSPD1 (M22382)	20	54	50	49	43	63	52	49	44	65	61	42	59	44	56	51	62	62	61	59	44
GABRB3 (M82919)	20	55	48	50	39	46	44	48	45	47	47	29	48	47	47	50	53	48	46	43	47
PPP1R9B (T86444)	20	57	56	60	55	55	49	51	52	49	49	52	53	54	54	49	51	50	50	60	54
IFI6 (X02492)	20	58	40	61	58	35	59	55	37	40	41	38	40	62	36	58	39	40	55	55	66
CREB5 (R33481)	20	59	54	67	70	60	56	56	59	75	85	58	55	60	63	65	48	84	58	68	60
ZNF358 (H81558)	20	60	55	64	57	51	55	57	56	54	54	59	52	61	61	63	52	53	54	50	61
MLC1 (D25217)	20	61	60	42	72	52	61	64	62	61	62	44	94	70	49	62	64	52	59	61	62
HSP90AB1 (T51023)	20	62	72	59	71	77	76	66	76	73	73	72	66	73	67	76	79	74	67	41	75
CNNM4 (R67343)	20	63	53	63	56	62	53	59	54	56	58	65	58	59	62	41	47	60	52	51	57
ARL6IP (D31885)	20	64	91	62	59	82	58	71	71	67	66	69	100	69	83	61	91	51	74	54	80
MUC2 (M94132)	20	66	68	74	75	69	73	70	87	85	80	70	68	75	69	64	67	72	71	77	78
NCOA4 (X77548)	20	67	67	75	76	67	75	69	75	72	67	73	69	74	68	77	77	73	73	71	76
RPS6 (T57619)	20	68	62	65	62	81	63	58	66	63	68	64	57	65	72	53	59	63	87	69	59
KRT8 (T54303)	20	69	61	66	68	71	64	61	64	62	60	63	60	71	58	52	65	61	63	65	63
H73908	20	70	64	69	69	61	65	62	63	60	63	62	62	66	59	60	61	66	60	70	70
MT1M (R06601)	20	71	65	79	79	58	69	90	68	57	57	67	85	64	55	87	68	96	53	87	65
MMP9 (T41204)	20	72	71	81	91	74	79	87	82	80	83	99	72	80	74	82	73	67	68	73	73
POLD2 (U21090)	20	73	69	90	87	72	82	80	83	86	79	81	73	88	76	71	87	77	77	80	81
GPD1L (D42047)	20	74	80	80	84	102	112	88	125	93	84	89	89	102	99	78	72	83	99	72	122
ETS2 (J04102)	20	75	78	73	65	110	67	72	67	66	77	68	76	68	119	70	69	65	92	86	69
SCNN1B (X87159)	20	76	63	71	66	57	68	65	65	64	64	78	56	63	65	68	63	75	47	67	58
WDR7 (T64012)	20	77	66	83	74	75	74	68	74	69	65	71	70	76	75	73	75	68	72	81	74
COX6A1 (H48072)	20	78	92	86	80	78	81	78	77	70	69	76	88	84	80	89	93	85	85	91	86
CD46 (R33367)	20	79	93	92	89	100	78	77	86	81	94	75	84	79	82	86	95	87	84	90	90
PECAM1 (L34657)	20	80	74	82	77	70	71	75	73	76	72	77	65	83	70	72	71	71	69	79	79
RPL30 (T58861)	20	81	85	72	67	66	66	63	55	55	59	54	64	58	64	66	70	59	65	66	64
UQCRFS1 (L32977)	20	82	87	70	64	65	70	73	72	68	71	79	92	67	87	69	81	70	78	82	82
ACAA2 (D16294)	20	84	89	85	98	84	87	79	91	83	78	84	83	93	86	80	80	82	79	88	87
WASF2 (H81068)	20	85	77	84	90	86	77	74	69	78	76	74	75	92	84	79	76	90	80	89	93

MGC22793 (H87135)	20	86	84	77	82	85	86	82	100	90	89	85	90	86	85	84	84	79	75	75	88
PMP22 (T94350)	20	87	90	93	86	87	105	89	107	99	98	90	93	94	92	85	86	103	106	99	83
TPM1 (Z24727)	20	88	75	95	78	89	90	76	92	77	93	100	80	81	91	83	96	91	95	94	68
POSTN (D13665)	20	89	82	106	108	95	94	94	99	89	92	97	82	107	90	103	113	88	98	110	109
PTPRH (D15049)	20	90	76	88	81	79	85	81	80	74	88	83	78	85	78	75	92	78	81	84	71
PRPSAP1(T65380)	20	91	95	104	120	90	98	110	98	97	86	109	79	99	93	110	100	92	91	85	118
CDX1 (U15212)	20	92	81	91	93	91	84	83	85	84	82	86	81	89	88	81	82	81	82	93	98
CCL14 (Z49269)	20	93	123	78	114	76	80	95	105	122	87	104	101	77	104	97	107	117	89	74	85
COX8A (T51250)	20	94	104	103	88	107	93	104	110	109	75	103	74	117	114	121	99	101	94	95	101
KIF5A (U06698)	20	95	73	121	96	80	83	111	111	71	70	80	71	91	73	74	101	76	76	109	114
DTWD2 (R98842)	20	96	96	87	103	114	99	85	81	103	108	95	108	106	98	98	85	95	90	104	94
CBX3 (U26312)	20	97	79	99	92	73	88	93	84	82	81	82	102	82	77	99	102	80	83	105	96
FXYD1 (T67077)	20	99	135	76	105	103	92	112	90	87	97	105	77	87	100	100	109	93	105	98	103
IFITM2 (X57351)	20	100	110	94	85	93	95	107	114	101	119	94	104	78	96	92	110	110	108	112	67
CPSF1 (U37012)	20	101	112	101	106	111	107	108	115	114	112	107	106	110	107	95	105	104	109	106	106
CNOT1 (T64885)	20	103	88	98	97	83	89	92	88	92	90	96	87	96	81	93	90	86	93	96	95
MORF4L2(D14812)	20	108	99	89	95	99	97	86	79	88	102	93	95	98	79	88	94	97	86	107	99
GSTM4 (M96233)	20	109	70	108	102	101	108	96	106	110	74	92	67	100	71	91	89	102	70	103	77
ATP2A2 (M23115)	20	110	106	109	116	115	111	97	113	118	109	123	112	143	126	125	121	114	124	145	121
FBL (X56597)	20	111	101	112	99	104	100	99	89	94	118	117	86	103	106	105	74	89	107	92	91
CALM2 (D45887)	20	112	109	107	118	105	101	101	95	113	95	106	96	97	95	108	97	99	96	100	105
PCCB (X73424)	20	116	100	116	110	108	106	98	94	105	99	98	97	109	111	104	103	108	100	117	110
ITPR3 (H25136)	20	119	113	114	113	113	104	109	109	108	107	102	107	111	103	101	106	105	103	121	112
IL1R2 (H78386)	20	130	94	110	117	96	91	84	78	98	115	88	91	95	94	90	83	136	88	97	89
HSP90AA1(X15183)	19	83	86	96	100	94	103		102	91	91	101	99	116	89	96	78	69	115	83	72
SFRS9 (U30825)	19	107	116	113	101	112	96	116	103	115	106	91		104	108	109	111	111	110	101	107
THBS2 (L12350)	19	125	102	119	111	116		119	101	119	105	131	109	115	133	106	88	109	101	111	104
CALM2 (M19311)	18	113	111	120	125	92		102	108	95	100	125	105	130	101	135	104	116	114	119	

PCK1 (L05144)	18	115	105	123	104			105	112	131	123	121	111	113	117	117	112	127	112	115	108
ALDH1A1 (M31994)	17	56	47	46	45	47	45		48	46	46	45		50	48		46	46	49	49	51
H05803	17	65		68	73	64	72	113	70	107		66	98	72	66	67	66	98	66	78	
HDGF (D16431)	17	102	103	97	83	88		103	97	102	103	113		105	113	114	98	106		108	100
MAPK3 (M84490)	17	117	120	126	139			118	126	112	117	116	115	136	132	119	122	122	120	139	
NXPH3 (H40699)	17	123	114	125	138	97			119	106	116	110	114	129	130	133	129	121	127	138	
ZNRF1 (H11460)	16	131	126	124	145				118	125	131	130	113	128	118	118	128	120	126	137	
CANX (L10284)	16	134	98	128	94				104	96	120	108		112	121	94	117	100	117	114	117
MYH9 (T57882)	16	141	122	134	126	106		115	116		142	112		132	127	113	123	130	116		111
MSTP9 (T51539)	15	98	83	140	63			67	93		96		63	121	110	59	60	64		76	102
CD46 (T83368)	15	105	107	111	109			106		124	104	115		108	105	140	114	107		136	97
FGFR2 (T94993)	15	122	119	133	123			117			110	137	103	144	140	112	108	129	125	116	
CCDC106 (T47424)	15	126	97	150	112		102		122	120	101	126		131	102	136	116		102	120	
CD44 (M59040)	14	137		132	122		110			132	124	135		127	124	131	127		113	129	120
PRPS1 (D00860)	14	138	118	145	129					117	130	122		142	125	132	120	128	123	131	
AVPI1 (R60883)	14	145	125	138	143				124		122	128		135	122	123	124	118	122	135	
ATP5A1 (T74556)	13	106	115	127	141			114			149	127		138	151	120		123		113	116
FAS (M67454)	13	118	108	148		98		100	120	100	111	124			112		115	115		118	
IARS (U04953)	13	120	132	139	128					129	128	129		119	129	124	133	119		122	
PFN1 (T61661)	13	149	129		140			91	121		133		110	137	120	107	130		121	140	
PSMC2 (H72965)	12	121	133	131	121				117		129	120			138	129	134			127	119
ITGA7 (X74295)	12	128		122	135					123	137	134		125	123	130	126	126		128	
ZNF3 (X07290)	12	132	127	146	130					126	140	138		122	131	148				124	115
PTPRO (Z48541)	11	127		129	127				123		121	118		118	154	122		132			113
DPT (R48303)	11	129		100	115					111	125			120	139	111		113	104	123	
RIMS2 (R75843)	11	151	124	130						127	135	133		140	116	145		133		134	
T47383	10	46		48	48		51	43	50			50		41						48	43
AMPD3 (M84721)	9	124		117	146	109			127	133		111		114	149						
SULT1E1 (H67764)	9	135	117								151	114		133		115	118	112		132	

MXD1 (L06895)	9	142		135	133						127				128	137	132	131	118		
T47342	9	144		136						121	114				115	138	119		119	133	
IGFBP3 (M35878)	8	104		105	107					104				90		102		94		130	
CYP2A7 (K03192)	8	155	128	155							147	139			141	141	136				
IMPDH2 (R42501)	8		130	142	132						126			139	136		131			142	
DNAJA1 (L08069)	7	136								128	153	119			156		125	125			
CEACAM1(X16354)	7		134		136					130				101	109					144	92
RPS19 (T52185)	6	32		54	27							60		24						24	
PLP2 (L09604)	6	133		141								132		123	134	127					
PTPRD (L38929)	6	139			144						145	136			157	126					
TNIP1 (D30755)	6	143		143						116	113			134		116					
RNASE3 (M28128)	6			102		68			96						97				97		84
POLD4 (R44418)	5	114		115	119														111	102	
COL1A1 (T51558)	5	146		152							144				137			134			
EIF1 (T61599)	5		131	137	134						143									126	
GCN5L2 (R52081)	5			154							156				146	146				146	
MUC1 (X52228)	4	147					109				138			126							
RBMY2BP(U36621)	4	148		156							148					149					
ACTA2 (T60155)	4	150													135	144				141	
SLC2A4 (M91463)	4		121	147	131						132										
KCNH2 (X86779)	4			149							134				142	134					
HIVEP2 (R39209)	4			153							155				145			135			
CDK4 (T86749)	3	140			137										147						
GYPC (X12496)	3	153									154			141							
COX5B (T71049)	3	154														147	135				
NPM1 (M26697)	3										141				150	143					
TUBB (T56604)	3				142									124						125	
C1R (T53889)	3										136				144			124			
NME1 (T86473)	2	152													155						

GTF2A2 (R01221)	2										150				153						
ITGB1 (H65425)	2										146				148						
CD55 (M31516)	2			58																63	
ATP6V1E1(X76228)	1															139					
FTL (H87344)	1																			143	
PI3 (Z18538)	1										152										
WEE1 (X62048)	1			118																	
H61410	1														152						
FLNA (R78934)	1															142					
SELENBP1(T59162)	1			151																	
AURKB (R97912)	1			144																	
AP3B2 (U37673)	1				124																
ANXA13 (Z11502)	1									79											
PPIB (T59878)	1															128					
NPTN (R61359)	1														143						
SEMG2 (M81651)	1										139										

Table S3 The clinical information of 86 lung adenocarcinoma samples from Beer et al (350)

Sample ID	cluster ID[1]	Age	Sex	Tumor stage. either 1 or 3	T (tumor size)	N (nodal status)	Survival times (month)[2]	Patient's survival status	classification (tumor histological type)[3]	Tumor differentiation	p53 nuclear accumulation status	12/13th codon K-ras mutation status	Smoking [4]
AD2	Cluster 1	65.6	F	1	1	0	91.8	alive	BD	Poor	+	-	48
AD5	Cluster 1	62	F	1	2	0	108.2	alive	BA	Well	-	+	positive
L01	Cluster 1	76.7	M	1	2	0	47	alive	BD/CC	Poor	-	-	100
L06	Cluster 1	57.9	F	1	1	0	91.9	alive	BD	Poor	-	+	NA
L26	Cluster 1	61.4	M	1	2	0	17.7	alive	BD	Poor	-	+	90
L33	Cluster 1	53.5	F	3	4	0	29.4	alive	BD	Moderate	-	-	23
L43	Cluster 1	50.6	F	1	2	0	78.5	alive	BD	Moderate	-	-	57
L56	Cluster 1	60.2	M	1	1	0	61.8	alive	BD/CC	Moderate	+	-	90
L62	Cluster 1	52.3	F	3	3	2	52.4	alive	BD	Moderate	-	-	none
L83	Cluster 1	62	F	1	2	0	30.6	alive	BA/mucinous	Well	-	-	none
L91	Cluster 1	63.7	M	3	2	2	6.1	alive	BD/mucinous	Poor	-	-	30
L92	Cluster 1	55.4	M	3	4	0	8.5	alive	BD	Poor	-	-	50
AD10	Cluster 1	65	M	1	1	0	84.1	death	BD	Moderate	-	NA	60
L04	Cluster 1	51.7	M	1	2	0	45.8	death	BD	Poor	-	-	50
L13	Cluster 1	67.1	M	1	1	0	79.5	death	BD	Moderate	+	+	25
L19	Cluster 1	56.5	M	3	3	2	9.6	death	BD	Moderate	-	+	40
L34	Cluster 1	77.2	M	3	1	2	14.9	death	BD	Moderate	+	-	45
L36	Cluster 1	69.7	M	3	1	2	7.2	death	BD/PA	Moderate	-	+	25
L37	Cluster 1	64.4	M	3	1	2	2.6	death	BD	Poor	-	+	84
L41	Cluster 1	73.1	F	1	2	0	8.4	death	BD/CC	Poor	-	+	26
L54	Cluster 1	45.8	F	3	3	1	4	death	BD	Poor	+	+	75

[1]These clusters are obtained from hierarchical cluster analysis of the 86 samples and 21 survival marker genes share by 10 signatures
[2]This is patient's survival time from operation date to death or last follow up as of May 2001
[3]BD: bronchial derived; BA: bronchial alveolar; CC: clear cell; PA: papillary; Note that some tumors contained a mixture of two histological types
[4]Patient smoking history in packs per year

L40	Cluster 1	54.9	F	3	1	2	20.1	death	BD	Moderate	-	-	7.5
L80	Cluster 1	68.2	F	1	2	0	10.1	death	BD/mucinous	Moderate	+	+	50
L61	Cluster 1	63.1	F	1	2	0	20.6	death	BD	Moderate	-	-	30
L95	Cluster 1	72	F	3	2	2	5.4	death	BD	Poor	-	+	50
L96	Cluster 1	64	F	3	3	1	21.2	death	BD	Moderate	-	+	50
AD7	Cluster 2	56	M	1	1	0	68.1	alive	BD	Moderate	+	-	80
L02	Cluster 2	63.2	M	1	1	0	39.1	alive	BD	Poor	-	-	27
L09	Cluster 2	48.2	F	1	1	0	98.7	alive	BD	Moderate	-	+	none
L101	Cluster 2	46.3	F	1	1	0	40	alive	B/A/mucinous	Well	-	-	NA
L103	Cluster 2	84.6	F	1	1	0	30.8	alive	B/A	Well	-	-	none
L104	Cluster 2	68.5	F	1	1	0	24.4	alive	B/A	Well	-	-	5
L105	Cluster 2	74.2	F	1	1	0	28.3	alive	B/A with PA	Well	-	+	75
L108	Cluster 2	61	F	1	1	0	19.5	alive	B/A	Well	-	+	100
L111	Cluster 2	54.9	F	1	1	0	1.5	alive	B/A	Well	-	+	40
L12	Cluster 2	44.6	F	1	1	0	85.2	alive	BD	Moderate	-	-	15
L18	Cluster 2	82.5	F	1	1	0	48.2	alive	BD	Well	-	-	none
L23	Cluster 2	62.2	M	3	2	2	15.1	alive	BD/PA	Moderate	-	+	20
L25	Cluster 2	62.6	F	1	2	0	14.5	alive	BD	Well	-	+	50
L27	Cluster 2	70	M	1	1	0	21.1	alive	BD	Poor	+	-	60
L38	Cluster 2	78.5	F	3	4	2	10	alive	BD	Poor	+	+	2
L42	Cluster 2	76	F	1	1	0	63.4	alive	BD	Well	-	-	40
L46	Cluster 2	60.4	M	1	2	0	82.4	alive	BD	Poor	-	+	160
L47	Cluster 2	60	M	1	2	0	60.5	alive	BD	Moderate	-	+	27
L48	Cluster 2	42.8	M	1	1	0	77.8	alive	BD	Moderate	-	-	60
L52	Cluster 2	67.3	M	1	1	0	65.4	alive	BA	Well	-	-	30
L57	Cluster 2	73.6	F	1	2	0	54.8	alive	BD/PA	Moderate	-	+	50
L65	Cluster 2	59.6	M	1	1	0	52.9	alive	BD	Moderate	-	-	60
L78	Cluster 2	75.6	F	1	1	0	36.5	alive	BD	Moderate	-	+	108

L82	Cluster 2	69.2	F	1	1	0	34.1	alive	BA/BD	Well	-	-	40
L85	Cluster 2	60.2	M	1	1	0	26.8	alive	BD/mucinous	Moderate	-	+	60
L97	Cluster 2	63.6	F	1	1	0	4.9	alive	B/A	Well	-	+	34
L50	Cluster 2	72.1	M	1	1	0	19	death	BD/PA	Moderate	+	+	100
AD3	Cluster 3	59.5	F	1	2	0	93.7	alive	BD	Moderate	-	-	positive
AD8	Cluster 3	75	M	1	2	0	34.2	alive	BD	Moderate	-	-	14
L05	Cluster 3	54.6	F	1	1	0	110.6	alive	BD/CC	Moderate	-	-	29
L08	Cluster 3	59.9	F	1	1	0	107.9	alive	BD	Moderate	-	+	80
L102	Cluster 3	74.6	F	1	1	0	40	alive	BD	Moderate	-	-	50
L106	Cluster 3	82.8	F	1	1	0	25.3	alive	B/A	Well	-	-	none
L107	Cluster 3	59.4	F	1	1	0	13	alive	BD	well/mod.	-	+	none
L17	Cluster 3	40.9	F	1	2	0	83.7	alive	BD/PA	Moderate	-	+	15
L22	Cluster 3	65.6	M	1	1	0	12.5	alive	BD	Moderate	+	-	90
L30	Cluster 3	51.8	F	1	1	0	20.2	alive	BD	Moderate	+	-	20
L31	Cluster 3	62.1	F	1	1	0	25.2	alive	BA/mucinous	Well	-	-	20
L49	Cluster 3	65.8	F	1	1	0	70.7	alive	BD	Moderate	-	+	20
L59	Cluster 3	71.5	F	3	2	2	54.6	alive	BD/PA	Moderate	-	+	25
L64	Cluster 3	65.4	M	1	2	0	48.1	alive	BD	Moderate	+	-	12
L76	Cluster 3	46.2	M	1	1	0	87.7	alive	BD	Poor	-	+	50
L81	Cluster 3	58.4	M	1	1	0	36	alive	BA	Well	-	-	90
L84	Cluster 3	66.8	F	1	2	0	32.2	alive	BD	Poor	-	-	15
L86	Cluster 3	62.7	F	1	1	0	10.1	alive	B/A	Well	-	-	45
L87	Cluster 3	66.3	M	1	1	0	10.4	alive	BD	Moderate	+	-	18
L88	Cluster 3	52.9	F	1	1	0	8.3	alive	BD	Poor	+	+	60
L89	Cluster 3	58.8	M	3	2	2	12.2	alive	BD	Moderate	NA	+	48
L99	Cluster 3	73.8	M	1	2	0	4.5	alive	B/A/mucinous	Well	-	+	55
L100	Cluster 3	72.9	F	1	1	0	43.8	censored	B/A	Well	-	-	2.5
L24	Cluster 3	84.5	F	1	1	0	1.6	censored	BD	Poor	-	-	75
AD6	Cluster 3	66.2	M	1	2	0	34.6	death	BA	Well	-	+	NA

L11	Cluster 3	68.2	F	1	2	0	34.7	death	BA	Well	-	+	none
L20	Cluster 3	79.8	M	1	2	0	19.9	death	BA	Well	-	-	30
L35	Cluster 3	64.4	M	3	2	2	28.2	death	BD	Moderate	+	+	4
L45	Cluster 3	74.9	F	1	1	0	29.6	death	BD	Poor	-	+	30
L53	Cluster 3	58.5	F	3	2	2	16.6	death	BD/PA	Moderate	-	-	none
L79	Cluster 3	49	F	1	2	0	8.7	death	BD	Poor	-	-	60
L90	Cluster 3	63.8	F	1	1	0	5.8	death	BD/PA	Moderate	-	-	100
L94	Cluster 3	72	M	3	3	2	2.4	death	BD/mucinous	Moderate	-	-	50

Table S4 The clinical information of 84 lung adenocarcinoma samples from Bhattacharjee et al (351)

Sample ID	Cluster ID[1]	Age	Sex	Stage:AJCC TNM	Stage Summary	Survival time (month)[2]	Patient's status*	Clinical Path (type diameter features) [3]	Path II[4]	Site of elapse/ metastasis	Smoking[5]
AD111	Cluster 3	76	F	T1NxMx	IA	72.4	1	ad 2.0 m-p			40
AD115	Cluster 2	70	F	T2N1M0	IIB	21.9	3	ad 6.5 m	adm/adw	lung, LN	75
AD118	Cluster 3	69	M	T1N0Mx	IA	49.6	3	ad 2.5 m	adm	lung, LN	25
AD120	Cluster 3	68	M	T2N0Mx	IB	38.9	3	ad 8.0 m	adm	bone	54
AD122	Cluster 1	73	F	T2N1Mx	IIB	33.9	3	ad 5.0 m	adm	lung	0
AD123	Cluster 1	60	M	T3N0Mx	IIB	74	1	ad 5.0 m	adm,pap		126
AD127	Cluster 3	65	F	T1N2Mx	IIIA	8.2	3	ad 1.8 p	adp	LN	69
AD130	Cluster 1	75	M	T2N1Mx	IIB	7.1	d	ad 15.0 BAC	BAC		100
AD136	Cluster 1	66	F	T2N0Mx	IB	31.4	1	ad 4.0 m	adm		100
AD159	Cluster 1	71	M	T2N1Mx	IIB	19.7	d	ad 5.5 m-p	adw,acinar		80
AD162	Cluster 3	75	F	T2N0Mx	IB	41.7	1	ad 3.5 m	admod,acinar		60
AD164	Cluster 1	68	M	T3N0Mx	IIB	15	3	ad 4.5 p	adpoor, acinar	LN	80
AD167	Cluster 2	77	M	T2N0Mx	IB	41.7	1	ad 2.5 w w/BAC	adw,acinar/adm bac		0
AD169	Cluster 1	47	F	T2N0Mx	IB	20	3	ad 2.5 m	adw/pap or BAC,mucinous w/pap	bone, myocardium	21.6
AD170	Cluster 3	61	F	T1N0M0	IA	78.4	1	ad 2.5 w w/pap	BAC & pap,well		60
AD173	Cluster 1	57	F	T2N1Mx	IIB	22.3	d	ad 5.0 m-p	admod,acinar		27
AD179	Cluster 2	85	M	T2N0Mx	IB	24.3	3	ad 5.6 m w/BAC	adw//adw,acinar	lung, bone	24.75
AD187	Cluster 1	69	M	T1N0Mx	IA	86.3	3	ad 1.8 p	adp	lung	120

[1]These clusters are obtained from hierarchical cluster analysis of the 84 samples and 21 survival marker genes we selected.
[2]Patient status at last followup or death (1= alive; 2=alive with recurrence; 3= dead with recurrence; 4= dead without evidence of recurrence; d= dead, disease status unknown)
[3,4]diameter (cm) subtype (BAC = bronchioloalveolar carcinoma). type (ad = adenocarcinoma) differentiation (p, m-p, m, m-w, w) /w= with
[5]Smoking: patient smoking history (self-reported) in pack/year

AD183	Cluster 2	75	F	T1N0Mx	IA	42.2	2	ad 2.0 m BAC	adw//adw,acinar		22.5
AD188	Cluster 1	74	F	T2NxMx	IB	21.6	d	ad 2.7 BAC	adw,acinar		116
AD201	Cluster 1	46	M	T1N2	IIIA	12.3	3	ad 1.5 m		lung, bone	90
AD203	Cluster 3	60	F	T1N0Mx	IA	106.1	1	ad 2.2 m-p			0
AD207	Cluster 3	64	F	T2	IB	66.8	4	ad 3.5 w BAC	ad m		0
AD212	Cluster 3	55	F	T2N0M0	IB	59	1	ad 3.0 m-p			54
AD213	Cluster 2	69	M	T1Nx	IA	48.8	d	ad 2.5 m			111
AD225	Cluster 1	88	M	T2NxMx	IB	2.6	4	ad 3.5 m			72
AD226	Cluster 1	56	F	T1N0Mx	IA	60.5	1	ad 2.0 m			18
AD228	Cluster 1	60	F	T2N0	IB	41.2	3	ad 3.0 m		brain	75
AD230	Cluster 3	56	M	T1N0	IA	56.7	1	ad 2.5 p	adp		60
AD232	Cluster 3	73	M	T1Nx	IA	56.3	a	ad 2.4 w BAC	adm (BAC cluster)		25
AD236	Cluster 1	53	F	T2N0Mx	IB	14.2	3	ad 5.5 m-p		lung, brain	40
AD239	Cluster 3	60	M	T2N0M0	IB	58.5	1	ad 2.9 m w/BAC	BAC		40
AD240	Cluster 1	77	F	T1N0M0	IA	43.5	1	ad 2.0 m-w			5
AD243	Cluster 2	64	F	T1N0M0	IA	50.1	1	ad 1.5 w w/BAC	adw resemblance to BAC		30
AD247	Cluster 3	49	M	T1N0	IA	71.1	1	ad 2.0 m			32
AD249	Cluster 1	67	M	T1Nx	IA	31	4	ad 1.2 m			45
AD250	Cluster 3	61	F	T1Nx	IA	91	2	ad 2.0 w w/BAC	adm	lung	10
AD252	Cluster 2	66	F	T1N0	IA	16.5	3	ad 1.4		LN, CSF, brain	50
AD255	Cluster 3	79	M	T2N0	IB	44.8	1	ad 3.5 m			50
AD258	Cluster 1	67	M	T2Nx	IB	12.3	3	ad 4.5 p		bone	54
AD259	Cluster 2	58	M	T3N0	IIB	20.5	d	ad 5.0			45
AD260	Cluster 1	61	M	T2Nx	IB	21	d	ad 3.0 m	adm some BACpattern		50
AD261	Cluster 2	66	F	T1N0	IA	57.6	1	ad 2.7 w w/BAC			75

AD262	Cluster 2	63	F	T4N1Mx	IIIB	16.6	4	ad 2.0 m-p			10
AD266	Cluster 2	65	F	T1N0	IA	41.9	3	ad 2.5 w w/BAC	adm	lung, bone, liver	0
AD267	Cluster 3	61	M	T2N0M0	IB	56	1	ad 2.8 m-p			120
AD268	Cluster 3	50	F	T2N0M0	IB	50.1	1	ad 3.5 p			10
AD276	Cluster 2	68	M	T2N2	IIIA	4.5	3	ad 2.1 m-p		pleura, brain	140
AD277	Cluster 3	72	F	T1Nx	IA	8.2	3	ad 3.0 m		liver, ?bone	27
AD283	Cluster 3	78	M	T1N0	IA	47.2	3	ad 2.5 m w/pap		lung, LN, bone, groin	20
AD287	Cluster 3	36	F	T4Nx	IIIB	7.4	d	ad 4.0 p	adp		10
AD296	Cluster 1	63	M	T1N1	IIA	9.3	3	ad 2.4 m-p w/pap		liver	88
AD299	Cluster 1	78	F	T1N0M0	IA	37.9	3	ad 2.2 m-p		lung	50
AD301	Cluster 1	59	F	T2N0M0	IB	7.8	3	ad 4.0 p		brain	40
AD302	Cluster 3	65	F	T2N3Mx	IIIB	57.8	3	ad 3.7 w BAC	adm w/BAC	lung	0
AD304	Cluster 2	71	F	T2N0	IB	8.2	3	ad 5.0 p		lung, liver, spleen	35
AD308	Cluster 3	62	M	T2N0	IB	79	2	ad 4.0 m		brain	66
AD309	Cluster 1	77	F	T2N0	IB	37.6	3	ad 3.4 w	adw	lung	0
AD311	Cluster 1	63	F	T2N0	IB	50.5	1	ad 5.0 m	ok 50%		13
AD313	Cluster 1	74	F	T1N0	IA	25.3	3	ad 1.5 m-p	adp	LN	90
AD317	Cluster 3	41	F	T2Nx	IB	99.1	1	ad 3.5 m pap			7
AD318	Cluster 1	54	M	T2N0M0	IB	83	1	ad 4.0 muc	adm		100
AD323	Cluster 1	56	F	T2N1	IIB	6.8	d	ad 4.0 p			39
AD327	Cluster 1	50	F	T2N0	IB	81.9	1	ad 6.5 m			27
AD330	Cluster 3	50	F	T1N1	IIA	7.3	3	ad 2.4 m		brain	40
AD331	Cluster 3	59	M	T1N0M0	IA	52.9	1	ad 2.0 m			45
AD332	Cluster 1	52	M	TxN0	I	6	3	ad m		pleura, liver, colon, ?adrenal, ?pancreas	75
AD335	Cluster 2	40	F	T3N0	IIB	46.9	1	ad 4.5 m			20
AD336	Cluster 3	71	M	T2N0Mx	IB	21.1	4	ad 1.7 m			0
AD338	Cluster 3	55	F	T2NxMx	IB	75.4	1	ad 5.0 w BAC	(1) ad w/BAC or (2)BAC		15
AD346	Cluster 2	65	F	T1N0	IA	17.3	1	ad 2.5 m			50

AD347	Cluster 1	65	F	T2N0Mx	IB	0.5	1	ad 3.5 m BAC	adm		20
AD351	Cluster 2	43	F	T2N1	IIA	24.3	3	ad 5.5 m		lung, LN	0
AD353	Cluster 3	69	M	T2N0Mx	IB	13.7	1	ad 3.5 m BAC	adw w/bac		30
AD356	Cluster 2	72	M	T2N0	IB	49.2	1	ad 4.0 w BAC			50
AD361	Cluster 1	54	F	T2N	IB	6.4	4	ad 4.5 p			0
AD362	Cluster 3	56	M	T2N0	IB	71.5	d	ad 6.5 BAC	BAC muc		40
AD366	Cluster 2	71	M	T2N2	IIIA	9.4	3	ad 6.2 m-p w/pap		lung	23
AD367	Cluster 1	55	F	T2N0	IB	76.1	2	ad 6.5 m-p		brain	25
AD368	Cluster 3	33	F	T2N0	IB	62.6	1	ad 6.0 m-p w/muc			32
AD374	Cluster 1	51	M	T2N0	IB	8.8	3	ad 11.0 p		lung, pleura, pericardium, diaphragm	100
AD375	Cluster 2	47	F	T2N0	IB	23.4	d	ad 7.2 p	adm		13
AD379	Cluster 2	65	M	T2N1	IIB	35.4	2	ad 5.5 w/clear		lung, adrenal, brain	80
AD382	Cluster 2	51	F	T2N2Mx	IIIA	30.1	3	ad 5.0 p		brain	31

Table S5 List of 10 derived lung adenocarcinoma prognosis marker gene signatures selected by SVM class-differentiation systems

Gene Name (EST number)	Number of signatures which included this gene	Gene rank in each signature (Number of selected gene in each signaure)									
		1 (51)	2 (54)	3 (42)	4 (34)	5 (46)	6 (54)	7 (57)	8 (50)	9 (53)	10 (47)
ADFP(X97324)	10	1	46	35	28	19	22	15	18	3	13
CXCL3(X53800)	10	2	37	24	7	23	3	14	4	6	19
PLD1(U38545)	10	5	7	2	31	41	17	8	9	11	3
SLC2A1(K03195)	10	6	3	12	3	3	8	13	12	2	11
SPRR1B (M19888)	10	7	10	29	11	10	7	9	10	5	12
GALNT4 (Y08564)	10	8	23	25	27	11	32	25	14	1	28
LDHB(X13794)	10	10	11	1	1	15	16	11	8	15	1
FXYD3(U28249)	10	11	6	7	29	14	52	18	42	22	5
REG1A(J05412)	10	13	8	9	23	9	15	16	45	14	6
CHRNA2 (U62431)	10	14	24	26	30	8	46	28	40	27	27
SERPINE1 (J03764)	10	18	30	16	22	12	2	1	31	4	15
FUT3(U27326)	10	19	14	19	21	2	28	10	15	30	21
PRKACB (M34181)	10	20	5	5	15	6	1	3	1	33	4
TUBA4A(X06956)	10	21	1	14	25	13	53	49	29	26	14
VEGF(M27281)	10	22	33	8	26	30	14	26	19	23	32
RPS3(X55715)	10	25	2	10	2	5	39	55	13	17	36
ANXA8(X16662)	10	28	32	18	12	21	20	4	22	18	26
VDR(J03258)	10	32	39	33	6	4	30	2	11	16	37
CXCR7(U67784)	10	33	47	30	24	43	41	37	27	39	29
POLD3(D26018)	10	35	25	15	18	1	11	50	2	31	8
BSG(X64364)	10	36	38	39	17	33	48	27	3	20	33
CYP24(L13286)	9	23	13	34	20	22	23		41	19	25
HLA-G (HG273-HT273)	9	30	27	11		25	19	34	32	24	31
WNT10B (U81787)	9	39	35	28		39	36	29	25	38	41
GARS(U09510)	9	41	26	31		31	26	19	46	44	20
SPRR2A(M21302)	9		21	13	34	40	21	21	34	47	18
NULL (HG2175-HT2245)	9		49	37	5	44	34	56	35	53	16
CD58(Y00636)	8	16	12	3	14	17	6		44		34
KRT14(J00124)	8			20	9	34	25	12	23	12	22
E48(X82693)	7	9	15			20	33	22	5	48	
FADD(X84709)	7	12		6		35	51	17		8	9
STX1A(L37792)	7	15	18	22		46	5		6		24
ENO2(X51956)	7	24	4			32	38	32		45	47
SPRR2A(L05188)	7	29	41			7	45	44	48	28	
FEZ2(U69140)	7	38		23			42	30	26	9	17
KRT18(X12876)	7	43	42	41		26	44	6	43		
ALDH2(X05409)	7		19		10	45	4	20	21		23
UCN(U43177)	6	4		36	13	18	9				10
SCYB5(L37036)	6	31	16		33	42	31			29	

AIP-1(U23435)	6	37		42		28	18			32	7
NULL(U92014)	6	42	17			27		39		36	42
NULL(L43579)	6	47	54				35	5	24	37	
CEBPA(U34070)	6			17		24	12		47	25	30
KIAA0138 (D50928)	5	34	29		19		37				2
TFF1(X52003)	5	40	34			37	24			43	
KRT19(Y00503)	5	49			4		40	54	20		
RPS26(X69654)	4	17	28				49			21	
S100A2(Y07755)	4	26	51					40		34	
GS3686 (AB000115)	4	46	36						49	41	
EMP1(Y07909)	4		9			38	27		38		
HPCAL1(D16227)	4		43		8			33	36		
LCN2(S75256)	4			38				41	37		44
PEX7(U88871)	4			4		29				40	43
EFNB2(U81262)	3	44							30		40
ALDH8(U37519)	3	45	52						17		
EPS8(U12535)	3		20				50			51	
NDRG1(D87953)	3		22					48			46
CSTB(U46692)	3		40					45		10	
PSPH(Y10275)	3		44	27				23			
CYBA(M21186)	3						29	7	7		
CNN3(S80562)	3							57	39	49	
VIPR1(X77777)	3			40					50		35
NULL(U49020)	2	51							16		
ALDH7(U10868)	2		45				10				
AXL (HG162-HT3165)	2		53							35	
TYRO3(U02566)	2			32		36					
P2RX5(U49395)	2				32	16					
GRO1(X54489)	2								28	42	
ERBB3(M34309)	2							51		7	
BM-002(Z70222)	2				16		13				
LAMB3(U17760)	2			21							39
INHA(X04445)	2							38		46	
TAX1BP2 (U25801)	1	3									
IGHM(V00563)	1	27									
SPRR2A(X53065)	1	48									
NP(K02574)	1	50									
P63(X69910)	1		31								
AP3B1(U91931)	1		48								
C6(X72177)	1		50								
HFL1(M65292)	1										38
PRKCN (HG2707-HT2803)	1							24			
SHB(X75342)	1									13	
EIF5A(S72024)	1								33		
FCGR3B(J04162)	1							47			

GRIN1 (HG4188-HT4458)	1						47				
SLC2A3(M20681)	1										45
CA9(X66839)	1							42			
FLJ20746 (U61836)	1						43				
PPBP(M54995)	1							52			
TUBA4A (HG2259-HT2348)	1						54				
EMS1(M98343)	1							53			
IGF2(M17863)	1							36			
CHAT (HG4051-HT4321)	1							31			
LAMC2(U31201)	1									50	
BMP2(M22489)	1							43			
KIAA0111 (D21853)	1									52	
TNFAIP6 (M31165)	1							35			
NULL (HG415-HT415)	1							46			

LIST OF PUBLICATIONS

1. Tang Zhiqun, Han Lianyi, Xie Bin, Cui Juan, Ung Choong Yong, Jiang Li, Wang Rong, Cao Zhiwei, Chen Yuzong, *"Antibody Antigen Information Resource Database and Its Potential Application to Antibody Discovery and Studies of Antigen Recognition",* (Under review)

2. Tang Zhiqun, Han Lianyi, Lin Honghuang, Cui Juan, Jia Jia, Low Boon Chuan, Li Baowen, Chen Yuzong, *"Derivation of Stable Microarray Cancer-differentiating Signatures by a Feature-selection Method Incorporating Consensus Scoring of Multiple Random Sampling and Gene-Ranking Consistency Evaluation",* ***Cancer Research 67: 9996-100003, 2007***

3. Tang Zhiqun, Han Lianyi, Xie Bin, Ung Choong Yong, Jiang Li, Chen Yuzong, "*AAIR: Antibody Antigen Information Resource*", ***J Immunol*** 178(8): 4705, 2007

4. Tang Zhiqun, Lin Honghuang, Zhang Hailei, Han Lianyi, Chen Xin, Chen Yuzong, *"Prediction of Functional Class of Proteins and Peptides Irrespective of Sequence Homology by Support Vector Machines"*, *Bioinformatics and Biology Insights*. 1: 19-47, 2007

5. Cui Juan, Han Lianyi, Lin Honghuang, Tang Zhiqun, Ji Zhiliang, Cao Zhiwei, Li Yixue and Chen Yuzong, *"Advances in exploration of machine learning methods for predicting functional class and interaction profiles of proteins and peptides irrespective of sequence homology"*, ***Curr. Bioinformatics*** 2(2): 95-112, 2007

6. Cui Juan, Han Lianyi, Lin Honghuang, Tang Zhiqun, Zheng Chanjuan, Cao Zhiwei, Chen Yuzong, *"Prediction of MHC-Binding Peptides of Flexible Lengths from Sequence-Derived Structural and Physicochemical Properties",* ***Mol. Immunol.*** 44: 866-877, 2007

7. Cui Juan, Han Lianyi, Lin Honghuang, Tang Zhiqun, Zheng Chanjuan, Cao Zhiwei, Chen Yuzong, *"Computer Prediction of Allergen Proteins from Sequence-Derived Protein Structural and Physicochemical Properties",* ***Mol. Immunol.*** 44(4): 514-520, 2007

8. Zheng Chanjuan, Han Lianyi, Xie Bin, Liew CY, Ong Serene, Cui Juan, Zhang Hailei, Tang Zhiqun, Gan Shoo Hui, Jiang Li, Chen Yuzong, *"PharmGED: Pharmacogenetic Effect Database",* ***Nucleic Acids Res.*** 35:D794-D799, 2007

9. Cui Juan, Han Lianyi, Lin Honghuang, Tang Zhiqun, Zheng Chanjuan, Cao Zhiwei, Chen Yuzong, *"MHC-BPS: MHC-Binder Prediction Server for Identifying Peptides of Flexible Lengths from Sequence-Derived Physicochemical Properties"*, ***Immunogenetics*** 58(8):607-13, 2006

10. Han Lianyi, Zheng Chanjuan, Lin Honghuang, Cui Juan, Li Hu, Zhang Hailei, Tang Zhiqun, Chen Yuzong, *"Prediction of Functional Class of Novel Plant Proteins by a Statistical Learning Method",* ***New Phytologist.*** 168:109-121, 2005