

学位論文

**Cancer Class Prediction and Biomarkers
Detection Using Microarray Data with
Evolutionary Computation**

(進化論的計算を用いたマイクロアレイデータの分類と
ガン関与遺伝子群の検出)



平成 18 年 12 月 15 日 博士(科学)申請

指導教員 伊庭 齊志 教授

東京大学大学院 新領域創成科学研究科 基盤情報学専攻

ポール トポン クマル

To my late mother: Geeta Rani Paul

Abstract

Cancer diagnosis based on the morphological appearance of tumor is sometimes difficult or impossible because tumors of different cancers may have identical appearances and show unequal responses to the same initial treatment. Recently many researchers are investigating whether gene expression profiling, coupled with class prediction methodology can be used to classify different types of tumor samples more reliably. Though different machine learning approaches have been proposed in this context, their success is limited due to smaller number of available samples compared to huge number of genes and due to many redundant genes.

The aim of this work is to develop a reliable and robust computational model for gene expression based diagnosis of cancer and identification of potential biomarkers of cancer. For this purpose, we propose two methods: random probabilistic model building genetic algorithm (RPMBGA), and majority voting genetic programming classifier (MVGPC). RPMBGA, a variant of genetic algorithm, is a gene selection method and requires a classifier. Therefore, its accuracy as well the selected genes is dependent on the classifier used to calculate the goodness of a gene subset.

MVGPC is based on genetic programming (GP) and majority voting technique and improves the classification accuracy of GP. It uses an ensemble of GP rules and predicts the label of a sample by employing majority voting technique. For identification of potential biomarkers, we propose that classifier be first devised, which will obtain higher classification accuracy, and then the evolved rules be analyzed to determine the most frequently occurring genes, i.e. first classification, then gene selection. To get a more stable frequency distribution of selected genes, MVGPC should be repeated several times on the microarray data. The ways the optimum ensemble size be determined, the label of a test sample be predicted using the ensemble, and the potential biomarkers be extracted from microarray data are our main contributions in MVGPC.

By performing experiments on microarray data sets, we have found that MVGPC is more reliable than RPMBGA, and the test accuracies obtained with MVGPC are significantly better than those with other competitive methods of gene selection and classification including AdaBoost+GP, and some of the more frequently selected genes in the ensemble of MVGPC are known to be associated with the types of cancers being studied in this dissertation.

論文内容の要旨

1. 背景:

癌治療においては、正確な癌患者サンプルの分類は非常に重要であると考えられている。しかし、腫瘍の形態、発生、微視的な外見及び位置にもとづく診断は非常に困難である。なぜなら、異なる癌の腫瘍であっても同じ外見である場合や、同じ処置を施しても、異なる反応を示す場合があるためである。さらに、癌細胞の採取には外科手術を伴う場合があり、危険性を有している。遺伝子発現データに対してクラス分類手法を用いることで、従来の病理学的な手法と比較して、客観的、明白かつ一貫した癌の分類手法の研究が近年盛んに行われている。本研究は、遺伝子発現量は多くの外的要因によって影響されるという仮説にもとづいている。ここで外的要因とは、温度、光、種々の信号など、ホルモンの分泌に影響を及ぼすものや、特定の遺伝子の発現量に影響を及ぼすような種々の病気を指す。

通常、癌細胞は通常の細胞の DNA が突然変異することによって生じる。そのため、通常の細胞と癌細胞の発現量を比較することで、癌の病状を起こす遺伝子を特定することができると考えられている。本研究の目的は DNA マイクロアレイの遺伝子発現量データから、バイオマーカーを同定し、正確でロバストな癌分類モデルを構築することにある。このような研究においては、種々の機械学習的手法にもとづく方法が提案されている。しかし、データサンプルの数に比べて、冗長な部分などを含む遺伝子の数が非常に多いため、これらの手法は、限定された状況下においてのみ有効である。

2. 手法:

本論文では、二つの手法を提案します：random probabilistic model building genetic algorithm (RPMBGA), majority voting genetic programming classifier (MVGPC)。これら二つの手法は、テストデータにおいて、他の手法と比べて非常に高い精度で癌を分類することが可能である。遺伝的アルゴリズムにもとづく RPMBGA は遺伝子の同定のみを行い、クラス分類器を別に必要とするが、MVGPC はそれ自体が分類及び遺伝子同定を行う。

RPMBGA は、遺伝的アルゴリズムのような従来の手法と比較して高速である。また、RPMBGA には交叉や突然変異はなく、他の手法と比較してコンパクトな遺伝子セットを同定し、高い精度で分類することが可能である。RPMBGA の初期集団は、多くの遺伝子を選択する状態にある個体によって形成される。RPMBGA は、徐々に個体を選択する無関係な遺伝子を減らし、最終的には少数の遺伝子を選択する個体のみを残す。RPMBGA は一度にひとつ以上の遺伝子を選択するような集団を生成することで、遺伝子間の相互作用を考慮することが出来る。このような方法は、一つの遺伝子の分類精度に基づいて、一度に一つの遺伝子を選ぶランクベースの方法より優れている。なぜなら、筆者らは最も高い精度をもたらす遺伝子のセットは、個々の遺伝子ではそれ以上の分類精度をもたらさないということを見つけたためである。さらに、個々の遺伝子及び多くの遺伝子を含む遺伝子のセットでは、完全な分類を行うことは出来ない。多くの遺伝子を含む集合では、無関係な遺伝子が含まれることで分類精度を下げってしまう。RPMBGA は以上の点で他手法と比較して優れているものの、RPMBGA においては同定される遺伝子セット及び分類精度は、適合度の算出に用いる分類器に大きく依存するという問題点がある。

MVGPC は遺伝的プログラミング (GP) に多数決手法を導入することで、GP より正確に、さらに RPMBGA より高い信頼性で分類することが可能である。MVGPC は異なる GP のルールを統合することで、テストサンプルの種類の推定を確実かつロバストに行うことが出来る。MVGPC においては、独立した GP の進化において得られた複数のルールをひとつずつテストサンプルに適用し、それぞれのルールは同定した癌の種類に対して投票を行う。テストサ

ンプルの種類は、最も支持数の多かった種類に決定される。MVGPC の基本的なアイデアは、GP によって進化した個々のルールでは、サンプルの種類を正確に推定することは困難であるが、ルールが集団で推定した場合は高い信頼性で推定することができるという考えに基づいている。しかし、多数決手法が有効であるかどうかは、多数決に用いるルールに数 (Ensemble size) 及び、一つ一つのルールの誤判定率に依存する。Ensemble size が小さい場合や、それぞれのルールの誤判定率が 0.5 以上である場合、MVGPC は個々のルールを単独で適用した場合より低い性能しか示すことが出来ない。そこで、本論文では最も高い性能を示す、最適 Ensemble Size を調査する。

本論文ではさらに、バイオマーカーの同定には、まず高い精度の分類器を生成し、その後で、分類器に含まれるルールの中で、高い頻度で出現する遺伝子を選出する方法を提案する。選出される遺伝子の定常的な頻度分布を得るには、マイクロアレイデータに対して複数回 MVGPC を適用する必要がある。この手法は、ある特定の遺伝子はどのような遺伝子選択アルゴリズム及び分類器を用いた場合でも、高い頻度で出現するという点にもとづいている。高い頻度で選択される遺伝子は、癌のバイオマーカーである場合と、生物学的には無関係であるが、トレーニング及びテストサンプルと非常に相関のある遺伝子である場合がある。

本論文の主要な提案は以下の点である：

- ・最適な Ensemble Size の決定手法。
- ・多数決を用いたテストサンプルの種類と同定。
- ・マイクロアレイデータのバイオマーカーの抽出。

3. 結果：

本論文では、Affymetrix の GeneChip ソフトが生成する遺伝子発現データを用い、二分類及び多分類の分類を行った。上記のマイクロアレイデータに対して RPMBGA 及び MVGPC を適用することで、他の手法と比較して高い精度で分類することに成功した。MVGPC は RPMBGA より正確に分類することが可能である。MVGPC におけるテストデータの正確度は、AdaBoost と GP の統合手法を含む他の手法と比較して、非常に高い結果を示す。さらに、MVGPC によって選択された遺伝子のうちいくつかは、本論文で扱った癌と関係があることが知られている。さらに、MVGPC をマイクロアレイ以外のデータに適用し、MVGPC による正確度は、GP で獲得した単独のルール及び複数のルールで単純に同定を行った場合より高い結果を示した。

4. 結論：

MVGPC は遺伝子発現量に基づく癌診断、そして癌のバイオマーカーの同定を行うのに、正確かつロバストな計算手法であると考えられる。AdaBoost は、弱学習器を統合することで、推定精度を改善する手法であるが、遺伝子発現データの分類においては、MVGPC が AdaBoost と GP の統合手法を上回る性能を発揮することが分かった。このような結果となった理由は、AdaBoost によって GP で獲得されたルールは全てのテストサンプルを用いない場合があるのに対して、MVGPC によって獲得されたルールは全てのテストサンプルを用いるためである。

しかし、MVGPC が有効であるかは、個々のルールの性能に依存し、MVGPC によって扱われる遺伝子の数は非常に多いものとなる。さらに、MVGPC の実行時間は、大きな多分類のマイクロアレイデータの場合、他の手法と比較して、非常に長くなるという問題点がある。これらの点は今後の課題であると考えられる。

Acknowledgements

I would like to thank Professor Hitoshi Iba, my research supervisor who is an expert in the theories and applications of genetic and evolutionary computations, for his many suggestions and constant support during this research. Without his support, I would not have been able to continue my research.

I am indebted to Japanese Government for awarding me the Monbukagakusho scholarship for postgraduate studies in the University of Tokyo for the period April 2001–March 2007. This support was crucial for my survival in Japan and the successful completion of this work.

I owe a debt of gratitude to Ms. Juan Liu, ex-member of Iba laboratory, for her work that inspired me to conduct research on the topic. I am equally grateful to the fellow members of my laboratory who are kind, sympathetic and helpful to assist me in solving many problems occurred during my experiments; their constructive criticism about my research during laboratory meetings helped me improve the quality of my research. I would specially thank Mr. Yoshihiko Hasegawa who translated the summary of this dissertation into Japanese for official purposes.

I would like to acknowledge different reviewers of my research papers on the topic for their masterly suggestions about the writing of a better quality research paper, and their supportive criticism of those papers.

Finally, I wish to thank Mr. Budrul Ahsan for helping me understand different topics of computational biology and bioinformatics.

Kashiwanoha 5-1-5, Kashiwa-shi, Chiba 277-8561
December 15, 2006

Topon Kumar Paul

Table of Contents

Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Abbreviations	xv
1 Introduction	1
1.1 What is cancer?	1
1.2 Cancer related deaths	1
1.3 Traditional method of cancer diagnosis	2
1.4 Molecular diagnostics	2
1.4.1 Different computational methods for molecular diagnostics	4
1.5 Objective and scope of the dissertation	6
1.6 Outline of the dissertation	7
2 Gene Expression and Microarray Data Files	9
2.1 Gene expression	9
2.2 DNA microarray	11
2.3 Affymetrix's GeneChip data	12
2.4 Preprocessing of microarray data	14
2.5 Some widely used microarray data sets	16
2.5.1 Brain cancer data	16
2.5.2 Breast cancer data	17
2.5.3 Colon cancer data	17
2.5.4 Diffused large B-cell lymphoma data	18
2.5.5 Leukemia data	18

2.5.6	Lung carcinoma data	18
2.5.7	Prostate cancer data	19
2.5.8	Scleroderma data	19
2.5.9	Small round blue-cell tumors	20
2.6	Summary	20
3	Related Works on Gene Selection	21
3.1	Introduction	21
3.2	Ranking of genes	23
3.2.1	Signal-to-noise ratio	23
3.2.2	t-Test method	24
3.3	Principal component analysis	25
3.4	Sequential forward search	26
3.5	Genetic algorithm	26
3.5.1	Representation of an individual	27
3.5.2	Initial population generation	28
3.5.3	Evaluation of an individual	28
3.6	Summary	29
4	Related Works on Class Discovery and Class Prediction	31
4.1	Introduction	31
4.2	Clustering of samples for class discovery	32
4.2.1	Self-organizing map	32
4.2.2	Hierarchical clustering	33
4.3	Different classifiers for class prediction	36
4.3.1	Weighted voting classifier	36
4.3.2	Naive-Bayes classifier	37
4.3.3	Decision tree: C4.5	39
4.3.4	k-Nearest neighbor classifier	40
4.3.5	Artificial neural network	42
4.3.6	Support vector machine	45
4.4	Accuracy estimation through cross-validation	46

4.5	Summary	48
5	Gene Selection by Random Probabilistic Model Building Genetic Algorithm	49
5.1	Introduction	49
5.2	Motivation	49
5.3	Details of RPMBGA	50
5.3.1	Notations	50
5.3.2	Initial population generation	51
5.3.3	Generation of new solutions (offspring)	51
5.3.4	Evaluation of a gene subset	52
5.3.5	Overall gene selection procedure	52
5.3.6	Example of offspring generation by RPMBGA	53
5.4	Evaluations of RPMBGA on microarray datasets	54
5.4.1	Microarray datasets	54
5.4.2	Experimental setup	55
5.4.3	Results	56
5.5	Summary	66
6	Classification and Gene Selection by Genetic Programming	67
6.1	Introduction	67
6.2	Genetic programming	68
6.2.1	Components of an S-expression in GP	70
6.2.2	Generation of initial population	70
6.2.3	Evaluation of a rule	71
6.2.4	Offspring generation through crossover and mutation	74
6.2.5	Evolution of rules for multiclass classification	74
6.3	Related works with genetic programming	76
6.4	Evaluation of genetic programming classifier	77
6.4.1	Results	78
6.5	Summary	85

7	Majority Voting Genetic Programming Classifier	86
7.1	Introduction	86
7.2	Class prediction through majority voting	87
7.2.1	Majority voting technique	87
7.2.2	Dependency of MVGPC on the performance of single rules	89
7.2.3	Optimal number of rules for MVGPC	92
7.2.4	Majority voting with LOOCV rules	93
7.2.5	Difference between AdaBoost and MVGPC	93
7.3	Evaluation of MVGPC	95
7.3.1	Test accuracies on the data sets	96
7.3.2	More frequently occurring genes	100
7.3.3	Effects of scaling of the values on the classification accuracy	105
7.3.4	Speed of convergence and computational time of MVGPC	105
7.3.5	Comparative test accuracies of different methods	109
7.3.6	Overfitting on the data sets	112
7.4	Discussion	113
7.5	Summary	114
8	Evaluation of MVGPC on Non-Microarray Data	115
8.1	Non-microarray data sets	115
8.1.1	Wisconsin breast cancer data	115
8.1.2	Monk's problem	116
8.2	Results	116
8.3	Summary	118
9	Conclusions and Future Works	119
9.1	Conclusions	119
9.1.1	Random probabilistic model building genetic algorithm	119
9.1.2	Majority voting genetic programming classifier	120

9.2	Future works	122
9.2.1	Classification of unbalanced data	122
9.2.2	Use of logical functions instead of arithmetic functions	124
9.2.3	Determination of ensemble size dynamically	124
9.2.4	MVGPC on the data of more frequently selected genes	125
9.2.5	Comparison of MVGPC with other ensemble methods	125
9.2.6	Investigation of influences of different GP parameters	126
A	EGPC: A Powerful Tool for Data Classification and Important Features Identification	127
A.1	Data format	128
A.2	Training and test subsets constructions	128
A.3	Attributes/Genes	129
A.4	Functions	129
A.5	Ensemble size	130
A.6	Evolved rules (S-expressions)	130
A.7	Default settings of some GP parameters	131
A.8	Example files included with this software bundle	131
A.9	Execution of the software	132
A.9.1	Execution of EGPCpre.jar in CLI mode	133
A.9.2	Execution of EGPCcom.jar in CLI mode:	134
A.9.3	Execution of example files from command prompt	135
A.9.4	Execution of EGPCgui.jar in GUI mode:	136
	Bibliography	141
	Publication List	155

List of Figures

1.1	Systematic differences in gene expression levels of some genes across cancerous and healthy (normal) samples	3
1.2	Objectives of the research	4
2.1	Central dogma of molecular genetics	10
2.2	Central dogma in action	10
2.3	Sequence of steps required to measure the expression levels of genes using DNA microarrays	13
2.4	A snapshot of a DNA microarray data file opened with Microsoft Excel	15
3.1	Filter approach of gene selection	22
3.2	Wrapper approach of gene selection	22
3.3	Fitness calculation in genetic algorithm	29
4.1	An example of hierarchical clustering of samples A, B, C, and D	35
4.2	An example of classification by decision tree	40
4.3	Class prediction by nearest neighbor classifier (IB1)	41
4.4	Class prediction by k-nearest neighbor (kNN) classifier	41
4.5	Classification by a feed forward back propagation neural network	42
4.6	Examples of hyper-planes in support vector machine	46
4.7	An example of leave-one-out-cross-validation technique for accuracy estimation	47
5.1	Plot of training and test accuracies of different gene subsets of lung carcinoma data	63
5.2	Plot of training and test accuracies of different gene subsets of brain cancer data	63
5.3	Plot of training and test accuracies of different gene subsets of prostate cancer data	64

6.1	An example of creation of an S-expression in GP	72
6.2	Fitness evaluation of a GP rule (R6)	73
6.3	Graphical plots of two fitness functions	73
6.4	An example of crossover in genetic programming	75
6.5	One-vs-rest approach for evolution of rules for multi-class classification	76
7.1	Steps in classification of gene expression data with MVGPC	87
7.2	Probability that MVGPC will be worse than a single rule	91
7.3	Class prediction by AdaBoost+GP and MVGPC	95
7.4	Test accuracies on brain cancer data under different conditions. For each experiment, in addition to the average accuracy, the maximum and the minimum accuracies are plotted on the graphs using error bars	98
7.5	Test accuracies on prostate cancer data under different conditions. For each experiment, in addition to the average accuracy, the maximum and the minimum accuracies are plotted on the graphs using error bars	99
7.6	Test accuracies on breast cancer data under different conditions. For each experiment, in addition to the average accuracy, the maximum and the minimum accuracies are plotted on the graphs using error bars	101
7.7	Test accuracies on lung carcinoma data under different conditions. For each experiment, in addition to the average accuracy, the maximum and the minimum accuracies are plotted on the graphs using error bars	102
7.8	Test accuracies on the brain cancer data that are normalized in base-10 logarithmic scale	106
7.9	Test accuracies on the prostate cancer data that are normalized in base-10 logarithmic scale	107

7.10 Plot of gene expressions of 38032_at (X2744) and 39601_at (X2289) across different samples of lung carcinoma	109
9.1 MVGPC on the data of more frequently selected genes	126
A.1 A screen shot of GUI of EGPC (<i>Preprocess Data</i> page)	137
A.2 A screen shot of GUI of EGPC (<i>Run EGPC</i> page)	138
A.3 A screen shot of GUI of EGPC (<i>View Accuracy</i> page)	139
A.4 A screen shot of GUI of EGPC (<i>Feature Ranking</i> page)	140

List of Tables

5.1	Microarray data sets used in the experiments	55
5.2	Overall accuracies by single genes on each data set	57
5.3	Overall accuracy by all genes on each data set	58
5.4	Best results of RPMBGA on all samples	59
5.5	Average results of RPMBGA on all samples	59
5.6	Overall accuracy by the genes selected by SNR	59
5.7	Best results of RPMBGA on training and test samples	62
5.8	Average results of RPMBGA on training and test samples	62
5.9	Training and test accuracies of the genes selected by SNR. For each gene subset, first row contains training accuracy while second row contains test accuracy	62
5.10	Description of the selected genes in the best subset of brain cancer data	64
5.11	Description of the selected genes in the best subset of prostate cancer data	65
6.1	Typical GP parameter settings	78
6.2	Training accuracies on the four microarray data sets	79
6.3	Test accuracies on the four microarray data sets	79
6.4	Descriptions of the genes more frequently selected during training and test accuracy estimation on the microarray data sets	81
7.1	Votes of different rules in the example of MVGPC	90
7.2	More frequently selected genes of the data sets	103
7.3	Comparative test accuracies on the data sets	110
7.4	p-values in statistical tests of significance (MVGPC vs other method)	110
8.1	Values of different GP parameters	117

8.2	Accuracies of MVGPC and single rules on Wisconsin breast cancer data	117
8.3	Accuracies of MVGPC and single rules on Monk-3 problem	118

List of Abbreviations

ANN	Artificial neural network
CLI	Command line interface
DNA	Deoxyribonucleic acid
EGPC	Ensemble of genetic programming classifiers
GA	Genetic algorithm
GP	Genetic programming
GUI	Graphical user interface
kNN	k-Nearest neighbor
LOOCV	Leave-one-out-cross-validation
MVGPC	Majority voting genetic programming classifier
NBC	Naive-Bayes classifier
PCA	Principal component analysis
PMBGA	Probabilistic model building genetic algorithm
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
RPMBGA	Random probabilistic model building genetic algorithm
SFS	Sequential forward search
SOM	Self-organizing maps
SNR	Signal-to-noise ratio
SRBCTs	Small round blue-cell tumors
SVM	Support vector machine
WVC	Weighted voting classifier

Chapter 1

Introduction

1.1 What is cancer?

Cancer is a genetic disease or disorder that is caused by the damages in genes that regulates the cell growth and division. Proto-oncogenes promote cell growth while tumor suppressor genes discourage cell growth or temporarily halt cell division to carry out DNA repair. In general, mutations in both types of genes are required to create a malignant tumor. In normal condition, the cells proliferate to produce new cells and some old cells die in a programmed way. However, when these two types of genes are mutated, the cells grow and divide in an abnormal way and the old cells, that are supposed to die, survive and form a tumor. Sometimes, these tumors are benign and do not cause cancer; sometimes, they are malignant and cause cancer. These cancerous cells invade and destroy other cells either by direct growth into adjacent cells through invasion or by implantation into distant sites through metastasis.

1.2 Cancer related deaths

In Japan, cancer ranks the top causes of deaths since 1981 (National Cancer Center, 2006). In 2004, the cancer related deaths in Japan were 320,315 and accounted for 31.1% of total deaths. For males, the three most common causes of cancer related deaths are lung cancer (22.3%), stomach cancer (17.2%) and hepatic cancer (12.5%) while for females, the three most common causes of cancer related deaths are colorectal cancer (14.6%), stomach cancer (14.2%) and lung cancer (12.3%).

In USA, cancer is the second leading cause of deaths, exceeding by heart diseases and it accounts for 25% of overall deaths (American Cancer Society, Inc, 2005). The three most common causes of cancer deaths for males are lung cancer (31%), prostate cancer (10%) and colorectal cancer (10%) while for females, lung cancer (27%), breast cancer (15%) and colorectal cancer (10%) are the three leading causes of cancer deaths.

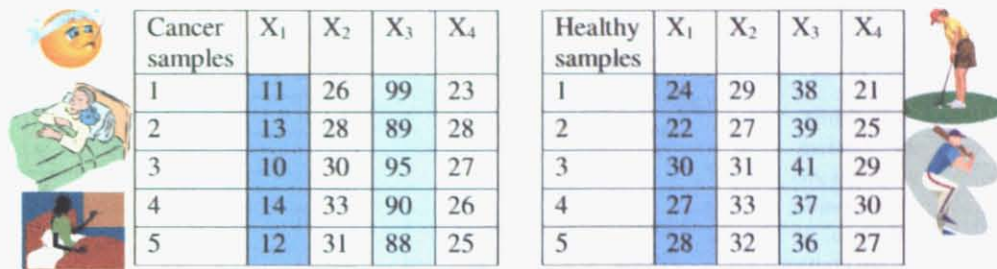
1.3 Traditional method of cancer diagnosis

Traditional method of cancer diagnosis is based on the morphological appearance of the tumor—the origin, the microscopic appearance and the location of the cancerous cells. For this purpose, tissues are collected through either biopsy or surgery. A cancerous tissue has a distinctive appearance under the microscope. Among the distinguishing traits are a large number of dividing cells, variation in nuclear size and shape, variation in cell size and shape, loss of specialized cell features, loss of normal tissue organization, and a poorly defined tumor boundary. Immunohistochemistry and other molecular methods may characterize specific markers on tumor cells, which may aid in diagnosis and prognosis.

However, sometimes the task of cancer diagnosis is difficult or impossible because of atypical clinical presentation or histopathology (Ramaswamy *et al.*, 2001). Tumors of identical appearance may progress at different speeds; some progress aggressively and require aggressive treatment while others remain inactive for a long period and may not need any treatment. These tumors can only be distinguished by observing a patient over a long period of time and discovering whether or not the initial treatment was aggressive (Golub *et al.*, 1999). Moreover, collection of tissue samples sometimes requires surgery and may be risky.

1.4 Molecular diagnostics

Recently many researchers are investigating whether gene expression profiling, coupled with class prediction methodology, can be used to classify different types of tumor samples in a manner more objective, explicit and consistent than standard



Cancer samples	X ₁	X ₂	X ₃	X ₄
1	11	26	99	23
2	13	28	89	28
3	10	30	95	27
4	14	33	90	26
5	12	31	88	25

Healthy samples	X ₁	X ₂	X ₃	X ₄
1	24	29	38	21
2	22	27	39	25
3	30	31	41	29
4	27	33	37	30
5	28	32	36	27

★ Genes X₁ and X₃ are potential bio-markers

Figure 1.1: Systematic differences in gene expression levels of some genes across cancerous and healthy (normal) samples

pathology. The hypothesis behind this research is that gene expression levels are affected by a large number of environmental factors, including temperature, stress, light, and other signals, that lead to change in the levels of hormones and other signaling substances, and many or all human diseases may be accompanied by specific changes in the expression levels of some genes (Schena, 2000). In Fig. 1.1, we have illustrated this hypothesis.

Recent advances in DNA microarray technology allow scientists to measure expression levels of thousands of genes simultaneously in a biological organism and have made it possible to create larger data sets of molecular information that represent molecular snapshots of biological systems of interest. Since the cancer cells usually evolve from normal cells due to mutations in genomic DNA, comparison of the gene expression levels of cancerous and normal tissues or different cancerous tissues may be useful to identify those genes that might anticipate the clinical behavior of cancers. The objective of the research is to extract the possible biomarkers of the cancers from the DNA microarray data of gene expression and then to design a reliable and robust predictive model that will perfectly classify different types of samples (see Fig. 1.2). In other words, there are two objectives: minimization of number of selected important genes and maximization of classification accuracy.

The main challenges for this research are:

1. availability of a smaller number of training and validation samples compared

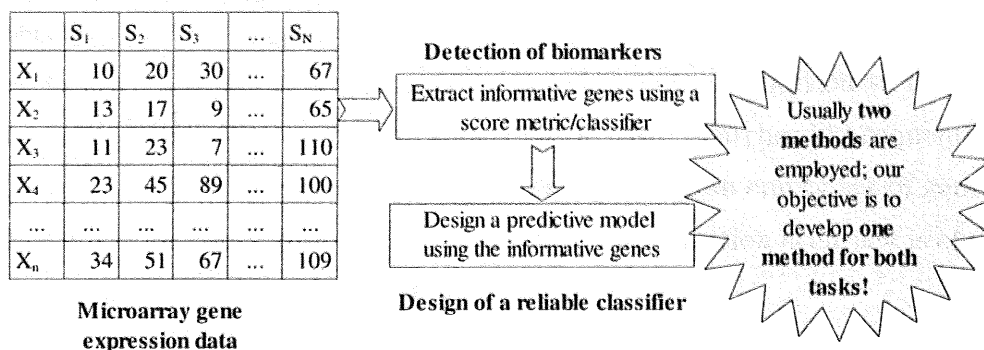


Figure 1.2: Objectives of the research

to huge number of genes;

2. class imbalance—some classes have more training samples than some other classes;
3. many irrelevant and redundant genes in gene expression profiles, which sometimes negatively affect the acquired classification accuracy of informative genes; and
4. noisy nature of microarray data.

1.4.1 Different computational methods for molecular diagnostics

Various machine learning techniques have been proposed for the identification of informative genes and for the classification of gene expression data. There are two categories of methods for identification of informative genes—deterministic and stochastic methods.

Deterministic methods use score metrics to filter out some informative genes. Then these genes are used to build a classification model. Widely used score metrics for ranking of genes are signal-to-noise ratio (SNR) (Golub *et al.*, 1999), disorder score (Park *et al.*, 2001), likelihood score (Keller *et al.*, 2000a) and threshold number of misclassification score (Ben-Dor *et al.*, 2000). The problem of this approach of gene selection is that it totally ignores the effects of the classifier on

the selected genes, whereas an optimal selection of genes may not be independent of the algorithm to be used to construct the classifier. In addition to rank-based methods, there have been proposed some deterministic searches like sequential forward search (SFS) for gene selection, which starts from an empty set of genes and sequentially adds genes until no improvement in classification accuracy is obtained (Inza *et al.*, 2002).

The most widely used stochastic methods for gene selection are evolutionary computation methods that employ wrapper approaches (Kohavi and John, 1997) of gene selection, where classifiers are used to measure the goodness of gene subsets (Liu and Iba, 2001, 2002; Ando and Iba, 2004; Deb and Reddy, 2003; Li *et al.*, 2004; Deutsch, 2003; Kim *et al.*, 2004; Liu *et al.*, 2005; Rowland, 2003; Ooi and Tan, 2003).

To build a predictive model, clustering, weighted voting classifier (Golub *et al.*, 1999), k-nearest neighbor (kNN) (Dasarathy, 1991), support vector machine (SVM) (Vapnik, 1995), Bayesian classifier, to name a few prominent ones, have been applied (Ben-Dor *et al.*, 1999; Eisen *et al.*, 1998; Slonim *et al.*, 2000; Ding, 2000; Pan *et al.*, 2004; Bhattacharjee *et al.*, 2001; Nutt *et al.*, 2003; Singh *et al.*, 2002; Boser *et al.*, 1992; Guyon *et al.*, 2002; Li *et al.*, 2006; Ramaswamy *et al.*, 2001; Shen and Tan, 2006). The main disadvantage of the above methods is that it is very difficult to find an optimal pair of a gene selection algorithm and a classifier.

Recently, genetic programming (GP) (Koza, 1992), an evolutionary computation method based on natural selection and evolution, has been applied to the classification of gene expression data (Mitra *et al.*, 2006; Moore *et al.*, 2002; Driscoll *et al.*, 2003; Hong and Cho, 2004; Langdon and Buxton, 2004). The main advantage of GP is that it acts as a classifier as well as a gene selection algorithm. In its typical implementation, a training set of gene expression data of patient-samples is presented to GP to evolve a Boolean or an arithmetic expression of genes describing whether a given sample belongs to a given class or not. Then the evolved best rule (s) is (are) applied to the test samples to get the generalized accuracy on unknown samples. However, the potential challenge for genetic programming is that it has to search two large spaces of functions and genes simultaneously to

find an optimal solution. In most cases, the evolved single rules or sets of rules produce very poor test accuracies.

1.5 Objective and scope of the dissertation

The aim of this dissertation is to develop a reliable and robust computational model for classification of microarray gene expression data and identification of potential biomarkers of cancers. In this dissertation, we concentrate on Affymetrix's GeneChip software generated gene expression data and take into consideration both binary and multi-category public cancer data sets.

To this end, we propose two methods: *random probabilistic model building genetic algorithm (RPMBGA)* and *majority voting genetic programming classifier (MVGPC)*. RPMBGA, a variant of genetic algorithm, is a gene selection method and requires a classifier. Therefore, its accuracy as well the selected genes is dependent on the classifier used to calculate the goodness of a gene subset.

MVGPC, based on genetic programming (GP) and majority voting technique, improves the classification accuracy of GP. It uses an ensemble of GP rules and predicts the label of a sample by employing majority voting technique. In its typical implementation, we evolve multiple rules in different GP runs, apply them one by one to a test sample and count their votes in favor of a particular class. Then the sample is assigned to the class that gets the highest number of votes in favor of it. However, the success of majority voting depends on the number of rules in the ensemble. Here we also investigate the optimal number of rules in the ensemble that produces the best results.

For identification of potential biomarkers, we propose that classifier be first devised, which will obtain higher classification accuracy, and then the evolved rules be analyzed to determine the most frequently occurring genes, i.e. first classification, then gene selection. To get a more stable frequency distribution of selected genes, MVGPC should be repeated on the microarray data for several times. Our proposal is based on the observation that some genes are frequently always selected whatever gene selection algorithms and classifiers are used. These

more frequently selected genes may be either potential biomarkers of cancers or junk genes that are highly correlated with distinction of different training and test samples but have no biological significance.

We evaluate RPMBGA on publicly available microarray data sets and MVGPC on both microarray and non-microarray data sets.

1.6 Outline of the dissertation

In Chap. 2, the process of gene expression, the method for measurement of gene expression levels, and the different attributes of a microarray data set generated by Affymetrix GeneChip's software are discussed. Descriptions of some widely used and publicly available microarray data sets and their preprocessing techniques are provided here.

In Chap. 3, different deterministic and stochastic computational methods for extraction of informative genes from a microarray data set containing huge number of genes compared to a smaller number of available samples are described.

In Chap. 4, different class discovery and class prediction methods are described. Class discovery refers to the process of dividing samples into reproducible classes that have similar behavior or properties while class prediction places new samples into already known classes. Class discovery is an unsupervised learning method and clustering is the widely used method for it; class prediction is a supervised learning method and different machine learning classifiers are used as a class predictor. Therefore, this chapter gives an overview of different clustering methods and classifiers.

In Chap. 5, random probabilistic model building genetic algorithm (RPMBGA), a variant of genetic algorithm, has been proposed for selection of informative genes from microarray data. In RPMBGA, a gene is selected based on its probability distribution and this probability is updated by incorporating randomness in the weighted average of probability of previous generation and the marginal distribution of current generation. The performance of RPMBGA has been evaluated by applying it to microarray data sets.

In Chap. 6, the application of genetic programming (GP) to the analysis of microarray data is discussed. The advantages of genetic programming are that it acts as a classifier as well as a gene selection algorithm and its transparent algebraic rules provide an insight into the quantitative relationships among the selected genes.

In Chap. 7, majority voting genetic programming classifier (MVGPC) has been proposed. MVGPC uses an ensemble of different genetic programming rules, improves the test accuracy of genetic programming rules and appears to be a reliable and robust method for prediction of the label of a test sample. In this chapter, we show how the optimum ensemble size be determined, the label of a test sample be predicted using the ensemble, and the potential biomarkers be extracted from microarray data. Here MVGPC has been evaluated and compared with other methods on different microarray data sets.

In Chap. 8, MVGPC has been applied to non-microarray data sets to investigate whether MVGPC will get competitive accuracies on these data sets or not. Unlike microarray data sets, these non-microarray data sets have larger number of available samples compared to number of features.

In Chap. 9, the conclusions of the dissertation and future works concerning different unresolved technical issues are provided.

Finally, the documentation of the software for classification of microarray data and identification of potential biomarkers is added in Appendix A.

Chapter 2

Gene Expression and Microarray Data Files

2.1 Gene expression

The central dogma of molecular genetics states that DNA (deoxyribonucleic acid) codes for RNA (ribonucleic acid) and RNA codes for protein. That is, the sequence of nucleotides in a DNA molecule specifies the sequence of nucleotides in a molecule of messenger RNA (mRNA); in turn, the sequence of nucleotides in mRNA specifies the sequence of amino acids in the polypeptide chain of a protein (Hartl and Jones, 2005). There are two steps: *transcription* and *translation* involved in synthesis of protein. The step DNA→RNA is called transcription while the step RNA→protein is called translation. Figs. 2.1 and 2.2 illustrate the central dogma of molecular genetics. The whole process by which RNA and eventually protein is synthesized from the DNA template of each gene is called *gene expression*. Gene expression level indicates the amount of mRNA produced in a cell during protein synthesis; and is thought to be correlated with the amount of corresponding protein made.

In protein synthesis, three types of RNA: mRNA (messenger RNA), rRNA (ribosomal RNA) and tRNA (transfer RNA) take part (Hartl and Jones, 2005). The roles of these RNA molecules are given below:

- mRNA carries information from DNA and is used as a template for protein synthesis. An mRNA molecule consist of coding and non-coding regions; however, in most mRNA molecules, a high portion of nucleotides code for amino acids.

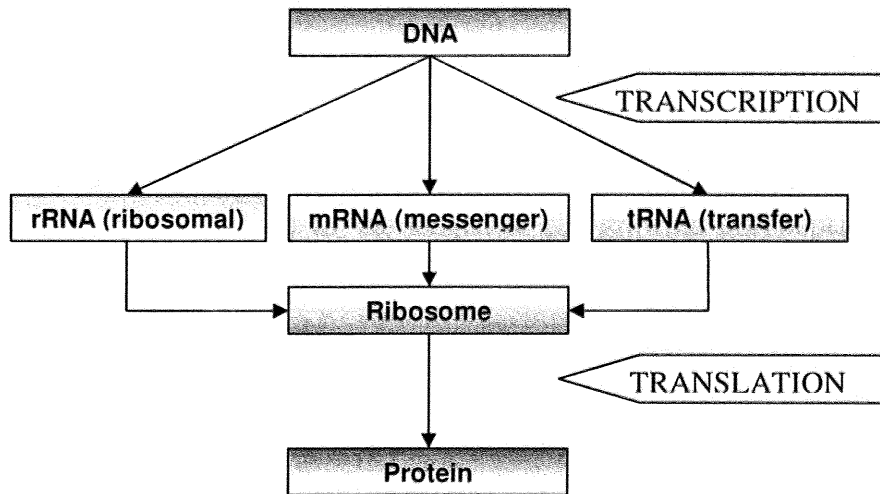


Figure 2.1: Central dogma of molecular genetics

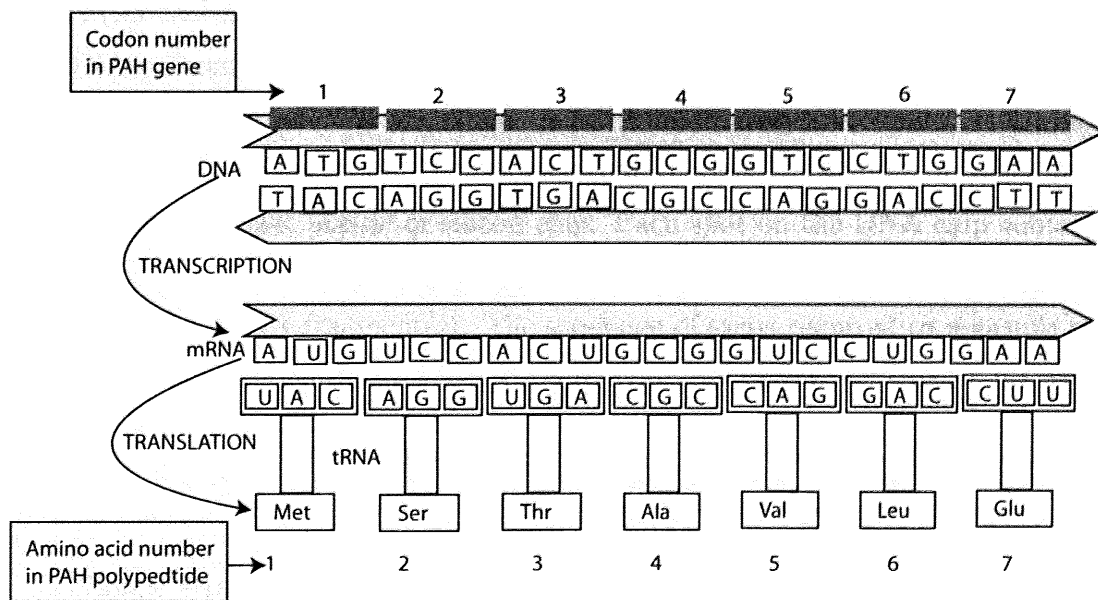


Figure 2.2: Central dogma in action

- rRNA molecules are the major constituents of ribosome on which polypeptide synthesis takes place. Four types of rRNA take part in protein synthesis.
- A tRNA molecule carries a particular amino acid and a three-base recognition regions that base-pairs with a group of three adjacent bases in the mRNA. After translation, the amino acid of a tRNA molecule becomes the terminal subunit added to the length of the growing polypeptide chain of a protein. A set of about 45 tRNA molecules take part in the translation phase of protein synthesis.

2.2 DNA microarray

The identification of those genes that are differentially expressed across normal and tumor tissues is the first step in cancer diagnosis and treatment. The expression level of a gene may be changed in response to a number of intra-cellular and extra-cellular signals. The changes in gene expression can be monitored using northern blotting (Thomas, 1980), reverse transcription -polymerase chain reaction (RT-PCR) (Mocharla *et al.*, 1990) and DNA microarray (Shalon *et al.*, 1996).

DNA microarray technology provides a rough measure of the cellular concentration of different mRNAs at a time. A DNA microarray (also called DNA chip or gene chip) is a collection of thousands of microscopic DNA spots attached on a slid surface like glass, plastic or silicon chip. Each spot on the DNA chip contains a known DNA sequence (a probe) and acts as a template for the binding of one or two labeled cDNA fragment(s). The sequence of steps required to measure the expression labels of genes using DNA microarrays are as follows (Reece, 2005):

1. RNA samples from two sets of cells grown in two different conditions (for example, a tissue grown in normal and cancerous states) are collected.
2. Single strands of cDNA are produced using reverse transcriptase from an oligo-dT primer. One of cDNA samples is labeled with a green fluorescent dye (Cy3) while the other one is labeled with a red fluorescent dye (Cy5).

3. The cDNA samples are mixed in equal proportions and hybridize to the microarray.
4. The color and intensity of each spot on the microarray is then monitored using a fluorescent scanning confocal microscope. The labels are excited using a laser, and the fluorescent at each spot detected with the microscope gives an indication of the relative amount of each mRNA species within the original sample. If a gene is expressed at equal levels in both samples, the corresponding spot will be yellow.

In Fig. 2.3, the sequence of steps required for preparation of a DNA microarray is shown. DNA microarray chips are now commercially available from companies like GE Healthcare, Affymetrix, or Agilent. These microarrays give estimations of the absolute value of gene expression, and therefore the comparison of two conditions requires the use of two separate microarrays.

2.3 Affymetrix's GeneChip data

For measurement of gene expression levels, Affymetrix's GeneChip is widely used. In this context, different GeneChip assays are prepared, scanned and the images are processed by Affymetrix software, Microarray Suite (MAS). (MAS has recently been replaced by GeneChip Operating Software (GCOS).) MAS 5.0 generates five types of files (*.EXP, *.DAT, *.CEL, *.CHP and *.RPT) during process of a GeneChip array experiment. Among these files, the chip file (*.CHP) is the output file from the MAS expression analysis of the Probe Array and contains the data that are used for statistical and data mining analysis.

MAS analysis metrics are (retrieved from http://www.ohsu.edu/gmsr/amc/amc_technology.html on Feb 5, 2007):

- Signal: a measure of the abundance of transcript.
- Detection: the call that indicates whether the transcript is detected (P, present), undetected (A, absent), or at the limit of detection (M, marginal).

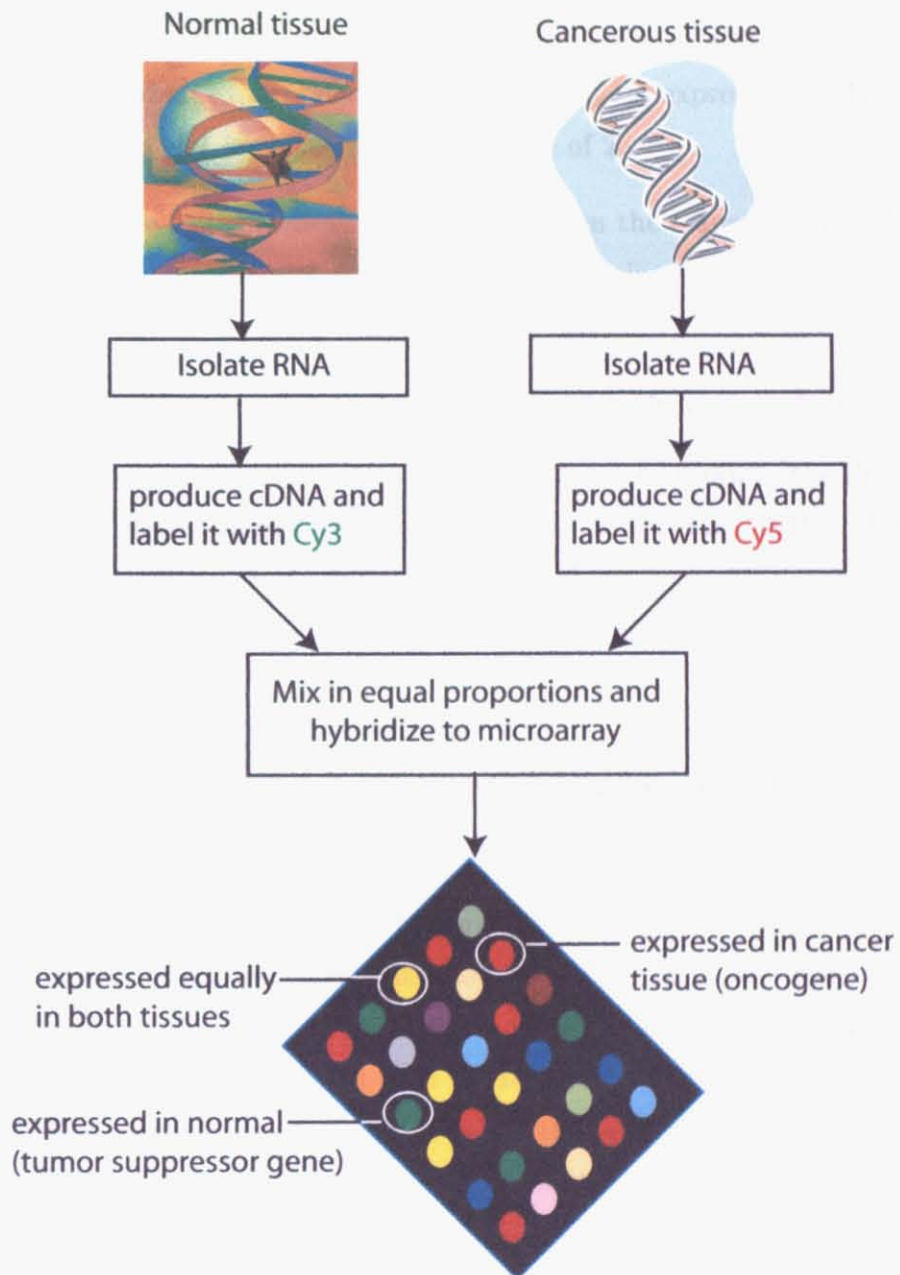


Figure 2.3: Sequence of steps required to measure the expression levels of genes using DNA microarrays

- Detection p-value: p-value that indicates the significance of the detection call.
- Signal log ratio: the change in expression level of a transcript between a baseline and an experiment array. This change is expressed as the \log_2 ratio. A \log_2 ratio of 1 is equal to a fold change of 2.
- Change: the call that indicates the change in the transcript level between a baseline and an experiment (increase (I), marginal increase (MI), no change (NC), marginal decrease (MD), decrease (D)).
- Change p-value: p-value that indicates the significance of the change call.

Each probe set on a GeneChip array has a unique name known as the Probe set ID. Probe set ID's have different extensions (examples: 1013_at, 1016_s_at, 1022_f_at, 1089_i_at) that denote important information about how the probe set was designed. Detailed information about Affymetrix GeneChip can be found at <http://www.affymetrix.com/>.

2.4 Preprocessing of microarray data

Usually, an Affymetrix's GeneChip gives estimations of the absolute values of gene expressions. The MAS output files are organized such that each column contains expression levels of different genes in a single sample, and each row contains expression levels of a single gene in different samples. In Fig. 2.4, a snapshot of a DNA microarray data file opened with Microsoft Excel is shown.

These files may have many negative values that are replaced by using a threshold of θ_l and a ceiling of θ_h . If a value is less than θ_l , it is replaced with θ_l ; similarly, if a value is greater than θ_h , it is replaced with θ_h . Missing values, if any, are determined by applying kNN method. Then variation filters are applied to exclude those genes that violate $\max(g) - \min(g) > \Delta$ and $\max(g)/\min(g) > \Omega$. Different researchers have applied different values of θ_l , θ_h , Δ and Ω for preprocessing of their microarray data. Then, these values (sometimes after taking log) are scaled. If y is the expression value of a gene g , its linearly scaled value in

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Description	Accession	Brain_NG_1	Brain_NG_10	Brain_NG_11	Brain_NG_12	Brain_NG_13	Brain_NG_14	Brain_NG_2						
2		CL2001041010AA		CL2001041705AA/sci	CL2001041128AA/sci	CL2001041705AA/sci	CL2001041006AA/sci	CL2001041004AA/sci	CL2001051650AA/sci						
3		12625													
4	X60188 /F/1000_at		70.9 P	68.12266 P	226.5166 P	110.0779 P	51.89813 P	69.19833 P	43.26284 P						
5	X60957 /F/1001_at		12.7 A	14.55612 P	27.32708 P	31.59844 P	14.01987 A	32.82863 P	16.29997 P						
6	X65962 /F/1002_f_at		-1.1 A	3.639031 A	-5.48154 A	-1.30527 A	2.377639 A	-0.88527 A	-13.9908 A						
7	X68149 /F/1003_s_at		-59.6 A	-23.1442 A	-31.5995 A	-33.23 A	-37.0584 A	-37.9189 A	-92.5703 A						
8	X68149 /F/1004_at		9.9 A	0.327513 A	9.634525 P	3.263178 A	-3.60745 A	6.418181 P	20.5108 A						
9	X68277 /F/1005_at		83.1 P	100.037 P	49.7369 P	130.2552 P	35.66459 P	44.48463 P	9.032901 A						
10	X07820 /F/1006_at		5.5 A	12.04519 A	-0.7255 A	5.221085 A	8.198756 A	9.221525 A	7.67457 A						
11	U48705 /F/1007_s_at		413.5 P	426.3125 P	414.0174 P	729.2658 P	144.2161 P	786.633 P	859.3481 P						
12	U50648 /F/1008_f_at		511.9 P	87.66425 P	129.7835 P	189.2099 P	290.9739 P	141.2738 P	408.9255 P						
13	U51004 /F/1009_at		74.3 P	217.3957 P	86.25362 P	65.97057 P	350.3328 P	41.60752 P	73.62154 P						
14	Y08200 /F/100_g_at		59.3 A	65.57534 P	74.9681 P	49.92662 P	43.89937 A	40.27962 P	91.41688 A						
15	U53442 /F/1010_at		0.4 A	8.07865 P	9.189638 A	4.242131 A	33.20496 P	3.319749 A	7.199154 A						
16	U54778 /F/1011_s_at		76.2 P	82.89713 P	81.98124 P	106.7059 P	244.6509 P	78.2723 P	45.9795 P						
17	U57317 /F/1012_at		3.7 A	3.930154 A	43.85231 P	22.6247 P	9.018632 P	57.321 P	14.12664 P						
18	U59913 /F/1013_at		-1.4 A	2.765664 A	2.176493 A	3.426337 A	4.099378 P	1.549216 A	0.950832 A						
19	U60325 /F/1014_at		7.9 P	10.44402 A	19.66905 P	15.44571 P	37.96024 P	22.60052 P	-1.69791 P						
20	U62293 /F/1015_s_at		-61.9 A	-43.1953 A	-66.1815 A	-25.4528 A	-50.0124 A	-34.3778 A	-52.092 A						
21	U70981 /F/1016_s_at		6.1 A	-1.81952 A	4.917263 A	33.17564 P	26.80993 P	1.549216 A	7.131238 A						
22	U73737 /F/1017_at		3.1 A	10.66236 P	7.658032 A	11.20358 P	16.61118 P	3.762382 A	1.76583 A						
23	U81787 /F/1018_at		-18.3 A	-7.86031 A	-9.75391 A	-2.12107 A	6.313042 A	4.131243 A	-24.7216 A						
24	U81787 /F/1019_g_at		9.9 A	18.12237 A	36.91977 M	23.65804 A	32.87701 A	16.08234 A	39.52743 A						
25	Y08305 /F/101_at		-0.3 A	15.10198 P	7.013145 A	-8.97374 A	-2.13168 A	-0.22132 A	-7.13124 A						
26	U85611 /F/1020_s_at		44.1 P	124.2001 P	57.07249 P	47.64239 P	33.53291 P	45.00104 P	68.25613 P						
27	J00219 /F/1021_at		0.3 A	0.436684 A	-1.93466 A	4.785994 A	1.475776 A	-0.73772 A	-1.96959 A						
28	V00542 /F/1022_f_at		1.8 A	1.601174 A	5.684693 A	-0.10877 A	10.5764 A	2.803344 A	6.791656 A						
29	X02158 /F/1023_at		-29.3 A	-20.4877 A	-5.56215 A	-21.6457 A	-31.3192 A	-10.033 A	-42.7195 A						
30	X02612 /F/1024_at		2.7 A	-2.83844 A	-11.7692 A	-2.50177 A	-7.70683 A	-4.27879 A	-2.64875 A						
31	X02612 /F/1025_g_at		7.2 A	-0.72781 A	5.330317 A	0 A	10.62236 A	3.02466 A	5.297491 A						
32	U41059 /F/1026_s_at		15.8 A	0.61964 A	-3.64699 A	6.096979 A	0.901993 A	0.073777 A	46.90701 A						

Figure 2.4: A snapshot of a DNA microarray data file opened with Microsoft Excel

the range $[a, b]$ will be $(b - a) \frac{y - \minval}{\maxval - \minval} + a$ where \minval and \maxval are the minimum and maximum values of gene expressions across all genes and samples. The standard normalized value of y will be $\frac{y - \mu}{\sigma}$ where μ and σ are the mean and standard deviation of genes across all genes and samples.

2.5 Some widely used microarray data sets

In this section, descriptions of different microarray data sets, publicly available and widely used as bench-mark data, are provided. Among these data sets, we have used brain cancer (Nutt *et al.*, 2003), breast cancer (Hedenfalk *et al.*, 2001), lung carcinoma (Bhattacharjee *et al.*, 2001), prostate cancer (Singh *et al.*, 2002), scleroderma (Whitfield *et al.*, 2003), and small round blue cell tumors (SRBCTs) (Khan *et al.*, 2001) data sets in different experiments for this work. The SRBCTs data set is already divided into training and test subsets; we divide the other data sets into mutually exclusive training and test subsets for our experiments. The colon cancer (Alon *et al.*, 1999), diffused large B-cell lymphoma (DLBCL) (Alizadeh *et al.*, 2000) and leukemia (Golub *et al.*, 1999) data sets have been used by others authors in related works.

2.5.1 Brain cancer data

The brain cancer data set (Nutt *et al.*, 2003) contains expression levels of 12625 genes of 50 gliomas samples: 28 glioblastomas (GBM) and 22 anaplastic oligodendrogliomas (AO) divided into two subsets of classic and non-classic gliomas. The classic subset contains 14 glioblastomas and 7 anaplastic oligodendrogliomas with classic histology while the non-classic subset contains 14 glioblastomas and 15 anaplastic oligodendrogliomas samples that are clinically common but histologically non-classic gliomas. The complete set of data is available at <http://www-genome.wi.mit.edu/cancer/pub/glioma>. After preprocessing of the data with $\theta_l = 20$, $\theta_h = 16000$, $\Delta = 100$ and $\Omega = 3$, only 4434 genes were left. For experiments with random probabilistic model building genetic algorithm (RPM-BGA) (Chap. 5) and genetic programming classifier (Chap. 6), we use the classic

subset as the training data and the non-classic subset as test data. However, for experiments with majority voting genetic programming classifier (MVGPC) (Chap. 7), we divide the data set into mutually exclusive training and test subsets containing 28 and 22 samples, respectively. The training subset consists of 14 glioblastomas and 14 anaplastic oligodendrogliomas samples; the test subset consists of 14 glioblastomas and 8 anaplastic oligodendrogliomas samples.

2.5.2 Breast cancer data

The breast cancer data set contains 22 cDNA microarrays, each representing 5361 genes based on biopsy specimens of primary breast tumors of patients with germline mutations of BRCA1 and BRAC2 and with sporadic cases. After preprocessing of this data set, only 3226 genes were left. The preprocessed data set is available at http://research.nhgri.nih.gov/microarray/NEJM_Supplement/. One sample in this data set labeled as 'Sporadic/Meth.BRCA1' was treated as 'BRAC1' in our experiments because we have found by performing many experiments with different methods described later that it has much similarity with samples of 'BRAC1' class rather than with samples of 'sporadic' class. Therefore, the numbers of BRAC1, BRAC2 and sporadic samples in the data set were 8, 8 and 6, respectively.

For experiments with genetic programming classifier (Chap. 6), we construct a training subset containing 6 BRAC1, 6 BRAC2 and 5 sporadic samples and the test subset containing the remaining 5 samples. However, for experiments with majority voting genetic programming classifier (MVGPC) (Chap. 7), 22 samples were randomly divided into mutually exclusive training and test subsets into approximately 2:1 ratio. The numbers of BRAC1, BRAC2 and sporadic samples in the training subset were $6(=\lceil 8 * 2/3 \rceil)$, $6(=\lceil 8 * 2/3 \rceil)$, and $4(=\lceil 6 * 2/3 \rceil)$, respectively; the remaining 6 samples went to the test subset.

2.5.3 Colon cancer data

This data set, a collection of expression values of 62 colon biopsy samples measured using high density oligonucleotide microarrays containing 2000 genes, is reported

by Alon *et al.* (1999). It contains 22 normal and 40 colon cancer samples. It is available at <http://microarray.princeton.edu/oncology>.

2.5.4 Diffused large B-cell lymphoma data

The diffused large B-cell lymphoma (DLBCL) data set (Alizadeh *et al.*, 2000) contains gene expression measurements of 96 normal and malignant lymphocyte samples, each measured using a specialized cDNA microarray, containing 4026 genes that are either preferentially expressed in lymphoid cells or of known immunological or oncological importance. The expression data in raw format are available at <http://llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt>. It contains 42 samples of DLBCL and 54 samples of other types. There are some missing gene expression values in this data set. In (Deb and Reddy, 2003), the authors have applied k -nearest neighbor algorithm to determine the missing values.

2.5.5 Leukemia data

This data set is a collection of gene expressions of 7129 genes of 72 leukemia samples reported by Golub *et al.* (1999). The data set is divided into an initial training set of 27 samples of acute lymphoblastic leukemia (ALL) and 11 samples of acute myeloblastic leukemia (AML), and an independent test set of 20 ALL and 14 AML samples. The data sets can be downloaded from <http://www.genome.wi.mit.edu/MPR>. These data sets contain many negative values which are meaningless for gene expressions, and need to be preprocessed. After preprocessing of this data set with $\theta_l = 20$, $\theta_h = 16000$, $\Delta = 500$ and $\Omega = 5$, only 3859 genes are left (Deb and Reddy, 2003).

2.5.6 Lung carcinoma data

The lung carcinoma data set (Bhattacharjee *et al.*, 2001) contains mRNA expression levels corresponding to 12,600 transcript sequences in 203 lung tumor and normal samples. The 203 samples consist of 139 lung adenocarcinomas (AD), 21 squamous (SQ) cell carcinoma cases, 20 pulmonary carcinoid (COID) tumors

and 6 small cell lung cancers (SCLC), as well as 17 normal lung (NL) samples. Negative gene expression values were replaced by setting a lower threshold of 0. Using a standard deviation threshold of 50 expression units, only 3312 genes were selected out of 12600. The original data sets are available at <http://research.dfci.harvard.edu/meyersonlab/lungca.html>. This data set is a five-class (AD, SQ, COID, SCLC and NL) classification problem. Since this data set is not divided into training and test sets, we divide it into mutually exclusive training subset containing 103 samples (AD:70, SQ:11, COID:10, SCLC:3 and NL:9), and test subset containing 100 samples (AD:69, SQ:10, COID:10, SCLC:3 and NL:8) for our experiments.

2.5.7 Prostate cancer data

The initial data set of prostate cancer (Singh *et al.*, 2002) contains gene expression profiles that were derived from 52 prostate tumors (PT) and 50 non-tumor prostate (normal)(NL) samples using oligonucleotide microarrays containing probes for approximately 12,600 genes and ESTs. The independent data set contains 8 normal and 27 tumor prostate samples. Raw data of initial set are available at <http://www-genome.wi.mit.edu/MPR/prostate>. Raw expression values were preprocessed with $\theta_l = 10$, $\theta_h = 16000$, $\Delta = 50$ and $\Omega = 5$. After preprocessing of raw expression values, only 5966 genes were left. Due to unavailability of the independent data set, we divide the initial set into mutually exclusive training and test sets, each set containing 51 samples (prostate tumors: 26 and normal: 25).

2.5.8 Scleroderma data

The scleroderma data set (Whitfield *et al.*, 2003) contains the expression levels of more than 12000 genes across 27 oligonucleotide microarrays of systemic sclerosis and normal biopsies. Out of these 27 oligonucleotide microarrays, 12 are signal amplification replicates. These data were preprocessed to get rid of negative values. After preprocessing, 7777 genes were left. The full set of data is available at <http://genome-www.stanford.edu/scleroderma>.

2.5.9 Small round blue-cell tumors

The small round blue-cell tumors (SRBCTs) data set (Khan *et al.*, 2001) contains expression levels of 6567 genes of 88 cells of four sub-types of SRBCTs: the Ewing family of tumors (EWS), Neuroblastoma (NB), Rhabdomyosarcoma (RMS), and Burkitt Lymphomas (BL), a type of non-Hodgkin lymphoma. These data were filtered to remove those genes whose expression levels were below a minimum level of expression leaving a total of 2308 genes. The data of 88 cells are divided into mutually exclusive training set of 63 samples, and a test set of 25 samples. The complete set of data including supplementary information is available at <http://research.nhgri.nih.gov/microarray/Supplement/>.

2.6 Summary

In this chapter, we describe the process of gene expression and measurement of gene expression levels, and some widely used microarray data sets. The Affymetrix's GeneChip generated microarray data files are widely used for investigation of genetic diagnosis of cancer and identification of biomarkers of the diseases. These data sets usually contain gene expression levels of cancerous and normal tissues or different types of tumorous tissues. The main characteristic of these microarray data files is that it contain a huge number of genes compared to a smaller number of samples. The identification of those genes that might anticipate the clinical behavior of cancers and the development of a reliable classifier using those genes are very important steps in gene expression based cancer diagnosis and treatment. In the next two chapters, we describe different gene selection and classification methods.

Chapter 3

Related Works on Gene Selection

3.1 Introduction

Since the cancerous cells usually evolve from normal cells due to mutations in genomic DNA, comparison of the gene expression levels of cancerous and normal tissues or different cancerous tissues may be useful to identify those genes that might anticipate the clinical behavior of cancers. However, this gene identification task faces many challenges due to availability of a smaller number of patient samples compared to a huge number of genes, class imbalance and the noisy nature of microarray data.

The main target of gene identification task is to maximize the classification accuracy (sensitivity and specificity as well) and minimize the number of selected genes. For a given classifier and a training set, the optimality of a gene identification algorithm can be ensured by an exhaustive search over all possible gene subsets. For a data set with n genes, there are 2^n gene subsets. So, it is impractical to search whole space exhaustively, unless n is small. There are two approaches, filter and wrapper approaches (Kohavi and John, 1997), for selection of gene subsets. The two approaches are illustrated in Figs. 3.1 and 3.2.

In filter approach, the data are preprocessed and some top rank genes are selected using a quality metric, independently of the classifier. Class prediction based on gene expression using filter approach has been proposed in the works (Golub *et al.*, 1999; Slonim *et al.*, 2000; Keller *et al.*, 2000b). Though the filter approach is computationally more efficient than wrapper approach, it ignores the effects of the selected genes on the performance of the classifier but the selection

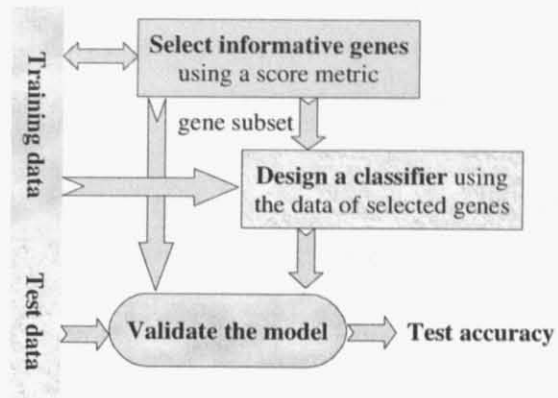


Figure 3.1: Filter approach of gene selection

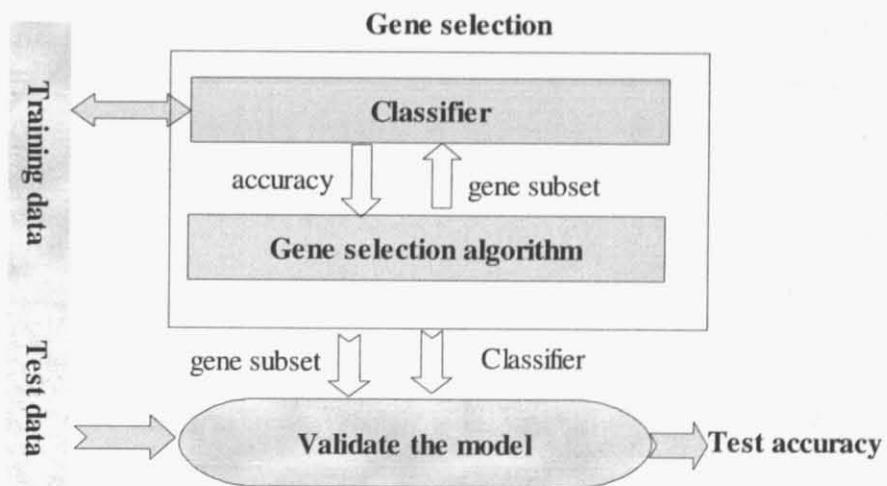


Figure 3.2: Wrapper approach of gene selection

of optimal gene subset is always dependent on the classifier.

In wrapper approach, the gene subset selection algorithm conducts the search for a good subset by using the classifier itself as a part of evaluation function. The classification algorithm is run on the training set, partitioned into internal training and validation sets, with different gene subsets. The internal training set is used to estimate the parameters of a classifier, and the internal validation set is used to estimate the fitness of a gene subset with that classifier. The gene subset with the highest estimated fitness is chosen as the final set on which the classifier is run. Usually in the final step, the classifier is built using the whole training set and the final gene subset, and then accuracy is estimated on the test set. A major disadvantage of the wrapper approach is that it requires much computation time.

Ranking of genes, principal component analysis, sequential forward search, and genetic algorithm are some widely used gene selection methods; in this chapter, we describe these methods.

3.2 Ranking of genes

Widely used score metrics for ranking of genes are signal-to-noise ratio (SNR) (Golub *et al.*, 1999), disorder score (Park *et al.*, 2001), likelihood score (Keller *et al.*, 2000a) and threshold number of misclassification score (Ben-Dor *et al.*, 2000).

3.2.1 Signal-to-noise ratio

The traditional gene selection method in molecular classification selects those genes that individually classify best the training samples. Widely used method for evaluation of how well a gene separates training samples is signal-to-noise ratio (Golub *et al.*, 1999). The signal-to-noise ratio of a gene X_i in binary classification is defined as

$$SNR(X_i) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \quad (3.2.1)$$

where μ_1 and σ_1 , and μ_2 and σ_2 are means and standard deviations of expression levels of gene X_i in class 1 and 2, respectively. Genes with the most positive and the most negative $SNR(X_i)$ values are selected in parallel and are grouped together in equal number in the final classifier. For a multi-class problem having k classes, $\frac{k(k-1)}{2}$ pairwise classifiers are considered. If we need to select n genes in total, we select $\frac{2n}{k(k-1)}$ distinct genes from each pair i and j . For each pair, we apply the above rule of binary classification to select top rank genes. If some of the selected genes of each pair are already included in the final subset, we exclude those genes and select the genes next to them in the ranking list. For binary classification, if n is odd, we select $\lceil \frac{n}{2} \rceil$ genes having the most positive values, and $\lfloor \frac{n}{2} \rfloor$ genes having the most negative values.

Variant forms of signal-to-noise ratio have also been used for gene ranking. For example, Furey *et al.* (2000) have used the absolute value of $SNR(X_i)$ as the ranking criterion while Pavlidis *et al.* (2001) have used $\frac{(\mu_1)^2 - (\mu_2)^2}{(\sigma_1)^2 + (\sigma_2)^2}$ as the score metric for gene ranking.

The serious limitation of this method of gene selection is that it may include some redundant genes and exclude those complementary genes that individually do not separate data well.

3.2.2 t-Test method

Ding (2003) used t-test to filter out important genes from microarray data.

The t-test score of a gene X_i is defined as follows:

$$t = \frac{\mu_1 - \mu_2}{\sigma}; \sigma^2 = \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n - 2} \quad (3.2.2)$$

where μ_1 and σ_1 , and μ_2 and σ_2 are means and standard deviations of expression levels of gene X_i in class 1 and 2, respectively; n_1 and n_2 are the numbers of samples in class 1 and 2, respectively. Then the genes having higher absolute t values are selected as the discriminative genes.

3.3 Principal component analysis

Principal component analysis (PCA)(Jolliffe, 2002), a technique for reduction of dimensionality of a problem, has been applied for selection of biomarkers from microarray data in (Khan *et al.*, 2001). It is a way of identifying patterns in data, and expressing the data in such a way that represent the similarities and the differences.

To use PCA, the gene expression data should be represented in the following $N \times n$ matrix form:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix}$$

where x_{ik} is the gene expression value of gene k in sample i (\mathbf{x}_i), n and N are the number of genes and the number of samples in the data set. To identify the principal components (informative genes), the following steps are done:

1. Calculate the covariance matrix using the data. Since the gene expression data of a sample is n -dimensional, the covariance will be $n \times n$.
2. Calculate the eigenvectors and eigenvalues of the covariance matrix. The eigenvector corresponding to the highest eigenvalue is the principal component of the data.
3. Take the genes corresponding to the higher eigenvalues as the informative genes. This way the dimensionality of the original data is reduced.
4. Apply a classifier to the reduced dimension data and get the classification accuracy.

PCA is also a filter approach since it selects genes independently of the classifier and suffers the same problems of overfitting like other filter approaches.

3.4 Sequential forward search

Sequential forward search (SFS) is a hill-climbing, deterministic heuristic search algorithm that starts from an empty set of genes. It selects genes until no improvement is achieved in evaluation value (Inza *et al.*, 2002). The goodness of a gene subset is estimated using LOOCV procedure with a machine learning classifier. SFS performs major part of its search near the empty feature set. SFS is used based on the idea of microarray data analysis that few genes are needed to classify patient samples among different classes.

3.5 Genetic algorithm

Genetic algorithm (GA) (Holland, 1975), based on natural selection and adaptation, is developed to solve complex real-world problems. GAs are widely used to solve those problems which are highly non-linear, contain inaccurate and noisy data, and whose objective function can not be expressed mathematically.

A typical genetic algorithm might consists of the following:

1. a population, guesses of the solution to the problem;
2. a way of calculating the goodness of each candidate solution in the population;
3. a method for selection of good solutions for mixing;
4. a method of mixing fragments of better solutions to form new solutions;
5. a mutation operator to maintain diversity in the population; and
6. a strategy to create the new population from old population and new solutions.

In a GA, each candidate solution is evaluated using some score metrics, and some of the better solutions are selected for reproduction. There are two main genetic operators, crossover and mutation, for reproduction of offspring. Crossover

operator creates offspring by combining parts from two or more parents (selected solutions); whereas, mutation generates an offspring by making changes in a single solution. The offspring are evaluated, and some of them are combined with the old population to generate the new population. This completes one cycle of generation. After several generations, the algorithm terminates converging to an optimal or a sub-optimal solution.

Recently, genetic algorithm and its variants (parallel genetic algorithm, multi-objective evolutionary algorithm, probabilistic model building genetic algorithm) have been applied to selection of informative genes from the microarray data (Liu and Iba, 2001, 2002; Deb and Reddy, 2003; Ooi and Tan, 2003; Deutsch, 2003; Ando and Iba, 2004; Li *et al.*, 2004; Kim *et al.*, 2004; Liu *et al.*, 2005). These methods usually employ a wrapper approach (Kohavi and John, 1997) of gene selection, where a classifier is used to measure the goodness of a gene subset. These GA-based methods obtain better classification accuracies than ranking based gene selection methods because different combinations of genes are evaluated in evolutionary computations through generation of different individuals of a population. However, the success of identification of a smaller size predictive gene subset depends on the choice of the appropriate recombination operators of an evolutionary computation method as well as on the choice of the appropriate classifier.

3.5.1 Representation of an individual

In selection of informative genes from microarray gene expression data, an individual (gene subset) of a genetic algorithm population is encoded as a binary string with each bit for each gene. If a bit is '1', it means that the gene is selected in the gene subset; '0' means its absence. For example, in the binary string "1000100111", genes 1, 5, 8, 9, 10 of a microarray data set are selected as possible informative genes. The main advantage of these binary-coded individuals is that the design of a crossover and a mutation operator is easy.

3.5.2 Initial population generation

Individuals of initial population for the problem are usually generated by random compositions of 1 and 0's. Half of the genes of the microarray data are expected to be selected in each individual. However, due to this huge number of selected genes in each individual, sometimes it becomes difficult to obtain compact gene subsets through the applications of recombination operators, and the initial best fitness does not improve over generations. To get compact gene subsets, the number of selected genes in each individual of initial population can be restricted to be an integer in $[1, m]$ where $m \ll n$ and n is the number of genes in the microarray data.

3.5.3 Evaluation of an individual

Different methods have been proposed for evaluation of an individual. Some have used the weighted average of the objectives (selection of minimum number of genes and maximization of classification accuracy) as the fitness measure of a gene subset; some have evaluated a gene subset using the pareto-optimal idea of multi-objective optimizations. In Fig. 3.3, we have shown the steps in calculation of fitness of a gene subset. Commonly, the fitness of an individual is a function of the classification accuracy obtained by that individual and the number of genes selected in that individual.

In the works (Liu and Iba, 2001; Paul and Iba, 2004b,a), average of the two objectives of the gene identification task is used as a fitness function:

$$fitness(X) = w_1 * Accuracy(X) + w_2 * (1 - d(X)/n) \quad (3.5.1)$$

where w_1 and w_2 are weights from $[0, 1]$, $a(X)$ is the accuracy obtained by the individual X , $d(X)$ is the number of genes selected in X , and n is the total number of genes in the microarray data set.

To select informative genes, Liu and Iba (2002) have used a multi-objective evolutionary algorithm (MOEA), in which the authors have identified three objectives: minimization of the size of the gene subset, minimization of mismatches in

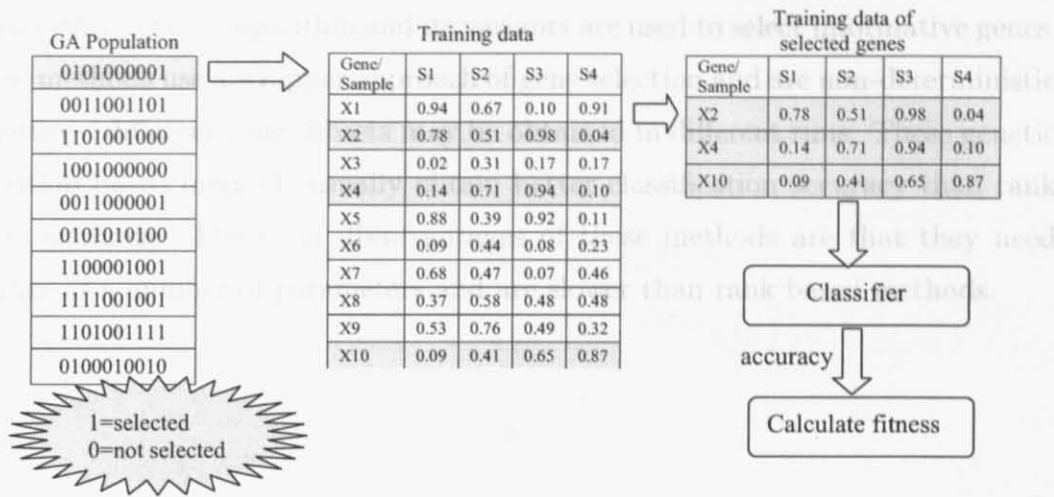


Figure 3.3: Fitness calculation in genetic algorithm

the training data, and minimization of the difference in error rate among classes, which is used to avoid bias due to unbalanced test patterns in different classes. For fitness calculation, these three objectives have been aggregated.

In optimization using non-dominated sorting genetic algorithm-II (NSGA-II) (Deb and Reddy, 2003), three objectives have been used. The first two objectives are the same as above; the third objective is minimization of mismatches in the test set. The number of mismatches in the training set is calculated using the leave-one-out-cross-validation (LOOCV) technique, and that in the test set is calculated by first building a classifier with the training data and the gene subset and then predicting the class of the test samples by using that classifier.

3.6 Summary

In this chapter, we discuss different gene selection methods. Rank based methods like signal-to-noise ratio and its variants are widely used to select a set of genes that are differentially expressed across normal and cancerous tissues or different tumorous tissues. These methods selects those genes that individually classify best the training data. However, this method may miss those complimentary genes that do not separate the data well but as a set of genes, they separate data very well.

Recently genetic algorithm and its variants are used to select informative genes. These methods use a wrapper approach of gene selection and are non-deterministic in nature—different gene subsets may be obtained in different runs. These genetic algorithm based methods usually obtain better classification accuracy than rank based methods. The main disadvantages of these methods are that they need settings of a number of parameters and are slower than rank based methods.