Hyper-Box Enclosure Method and its Application to Microarray Analysis

by

Onur Dağlıyan

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Chemical and Biological Engineering

Koç University

July, 2010

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Onur Dağlıyan

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

İbrahim Halil Kavaklı, Ph. D. (Advisor)

_____

Metin Türkay, Ph. D.

_____

Alkan Kabakçıoğlu, Ph. D.

Date: _____

to my family…

ABSTRACT

Data mining is an important tool employed in many bioinformatics domains including genomics, proteomics, evolution, systems biology, and microarray analysis. A recently developed classifier, hyper-box enclosure (HBE) algorithm is an efficient method for classification problems, and it does not require parameter optimization depending on data type for higher prediction accuracy. The goal of this thesis is to understand, improve HBE algorithm, and apply it for microarray analysis which is an important bioinformatics problem.

The most important use of data obtained from microarray measurements is the classification of tumor types with respect to specific genes that are either up or down regulated in specific cancer types. However, almost all classification algorithms used in microarray analysis usually require optimization to obtain accurate results depending on the data type. Additionally, it is highly critical to find an optimal set of markers among those up or down regulated genes that can be clinically utilized to build assays for the diagnosis or to follow progression of specific cancer types. On the base of these necessities, we employ a mixed integer programming based classification algorithm named hyper-box enclosure method (HBE) for the classification of some cancer types with a minimal set of predictor genes. This method, a user friendly and efficient classifier, may also allow the clinicians to diagnose and follow progression of certain cancer types.

ÖZETÇE

Veri madenciliği genetik, protein bilimi, evrim, sistem biyolojisi ve mikrodizi gibi birçok biyobilişim alanında kullanılmaktadır. Son zamanlarda geliştirilen çok boyutlu kutu kapsama metodu sınıflandırma problemleri için oldukça etkili bir yöntemdir ve veri tipine göre parametre eniyileştirmesi gerektirmez. Bu tezin amacı çok boyutlu kutu kapsama yönteminin anlaşılması, geliştirilmesi ve önemli bir biyobilişim problemi olan mikrodizi analizi için kullanılmasıdır.

Mikrodizi ölçümlerinden elde edilen verinin en önemli uygulaması tümör tiplerinin belirli genlere (bu genler farklı kanser tipleri için az ya da çok ifade edilebilir) göre sınıflandırılmasıdır. Ancak mikrodizi analizi için kullanılan algoritmaların birçoğu veri tipine göre daha yüksek doğruluk için eniyileştirme gerektirir. Ayrıca bu az ya da çok ifade edilmiş genler arasından optimum seti bulmak da kanser tiplerinin tespiti ve gelişimini takip etmek için kritiktir. Bu gereklilikler bağlamında, asgari sayıda gen kullanılarak bazı kanser tiplerinin sınıflandırılması tamsayı karışık programlamaya dayalı çok boyutlu kutu kapsama yöntemi ile yapılmıştır. Kullanıcı dostu ve etkili bu yöntem klinik araştırmacılara bazı kanser tiplerinin gidişatını izlemesi için de yardımcı olabilir.

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my advisor Dr. İ. Halil Kavaklı and Dr. Metin Türkay for their great guidance, patience and continuous support during my graduate study. Also, I would like to thank my thesis committee member Dr. Alkan Kabakçıoğlu for his critical reading and useful comments.

I would like to thank also the Scientific and Technological Research Council of Turkey (TUBITAK) for their financial support during my M.Sc. study.

I would like to thank my research group friends, Onur Öztaş, İbrahim Gür, Natali Özber, Hande Asımgil, Bengisu Seferoğlu, Işıl Tulum, İbrahim Barış, Besray Ünal, Emre Özdemir, and Pelin Armutlu. It would not be possible to finish this thesis without their help especially in the molecular biology and biochemistry laboratory.

I wish to thank my friends Bilal Çakır, Enis Demir, Zeynep Ülker, Ezgi Dağyıldız, Selimcan Azizoğlu, Gökhan Hatipoğlu, Engin Çukuroğlu, Sibel Kalyoncu, Erdal Uzunlar, Meriç Ataman, and S. Kutal Gökçe for their valuable friendships and for sharing me great moments.

I owe my loving thanks to İrem for always believing in me and for her encouragement. Her endless love and support have been a main motivation during my study.

Finally, I would like to thank my wonderful family for their continuous support and patience during every step of my education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

HBE	Hyper-box Enclosure
kNN	k-Nearest Neighbor
NN	Neural Network
K	Number of Classes
σ	Activation Function
SVM	Support Vector Machines
RBF	Radial Basis Function
WGS	Whole Genome Shotgun
QSAR	Quantitative Structure Activity Relation
I	Training samples (*i=Sample1, Sample2, …, SampleI*)
j	Test samples (*j=Sample1, Sample2, …, SampleI*)
k	Class types (*k=Class1, Class2, …, ClassK*)
l	Hyper-boxes that encloses a number of data points belonging to a class
M	Total number of attributes
L	Total number of hyper-boxes
a_{im}	Value of the attribute m for the sample *I*
D_{ik}	Class k that the data point *i* belong to
yb_l	Binary variable to indicate whether the box *l* is used or not
ypb_{il}	Binary variable to indicate whether the data point *i* is in box *l* or not
ypb_{ik}	Binary variable to indicate whether the data point *i* is assigned to class *k* or not
CFS	Correlation-based Feature Selection
*DLBC*L	Diffuse Large B-cell Lymphoma
SRBCT	Small Round Blue Cell Tumor
H(Y)	Entropy of Feature *Y*

p(y|x)	Conditional Probability of *y* given *x*
InfoGain	Information Gain
FCDF	*F* Cumulative Distribution Function
ALL	Acute Lymphoblastic Leukemia
ALL	Acute Myeloid Leukemia
10-CV	Ten-fold Cross Validation
LOOCV	Leave-one-out Cross Validation
T3F3	Transcription Factor 3
TP	True Positives
FP	False Positives
FN	False Negatives
TN	True Negatives
TPR	True Positive Rate
FPR	False Positive Rate
ACC	Accuracy
LR	Logistic Regression
RF	Random Forest
EWS	Ewing Family Tumor
BL	Burkitt Lymphoma
NB	Neuroblastoma
RMS	Rhabdomysarcoma

Chapter 1

INTRODUCTION

Machine learning, which is the investigation and extraction of patterns from data to make intelligent decisions, is highly used in computer vision, object recognition, speech recognition, natural language procession, search engines, machine perception, robot locomotion, software engineering, medical diagnosis, credit card fraud detection, stock market analysis, bioinformatics, and cheminformatics. Especially the latter ones, bioinformatics and cheminformatics are emerging disciplines and generate large amount of data that require more sophisticated algorithms.

There are different algorithm classes in machine learning based on the type of input and desired outcome such as supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and transduction. In this thesis, a supervised learning class of algorithm, data classification is investigated in detail. In Chapter 2, an overview of data classification methods including k-Nearest Neighbor classifier, decision trees, neural networks, Bayesian classifiers, logistic regression, support vector machines and their applications to bioinformatics in literature are provided. Genomics is one of the fields in which classification algorithms are highly employed. Various problems in genomics including coding region identification, alternative splicing, splice site prediction, gene function prediction, RNA structure prediction, motif identification tremendously benefit from various classification algorithms. Another application domain, proteomics include protein structure and function prediction, protein location prediction and protein-protein interaction problems. Systems biology areas such as signaling networks, metabolic pathways and genetic networks are the topics getting more popular in terms of usage of classification algorithms. Additionally image analysis, microarray analysis, text mining

problems and mass spectrometry data analysis are the topics in which data classification algorithms are being used.

Chapter 3 provides the theory and details of mixed integer programming based hyper-box enclosure classifier. This algorithm builds hyper-boxes for defining class boundaries that enclose all or some of the points in the training set. The well construction of boundaries of each class using hyper-boxes makes the prediction performance of this algorithm reasonably high. This algorithm can be used in both binary and multi-class problems and does not require any modifications or parameter optimization. Also the model built by hyper-boxes is simple and understandable.

The first application of hyper-box enclosure algorithm is given in Chapter 4. In this part, the performance of hyper-box enclosure method is tested on microarray cancer datasets. Specifically, the phenotype of a cell, cancer or normal, is predicted using gene profiles from microarray data. Furthermore, this prediction or classification of cells is performed using minimum number of genes that can be considered as best feature subset. Moreover, the biological roles of selected significant genes on specific cancer types are discussed in this chapter.

This thesis ends with Chapter 5 which includes conclusions, recommendations, and future directions of the study.

Chapter 2

LITERATURE REVIEW

2.1 Data Classification Methods

2.1.1 k-Nearest Neighbor Classifier

The kNN classifier [1-2] which is a simple method that uses the only continuity assumption of the feature variables. kNN algorithm does not use model fitting but storing the training dataset with all vector prototypes of each class instead. An instance is classified by majority voting of its neighbors and $k=b$ means that the instance is assigned to the class of its b nearest neighbor(s). When the instance x^* is to be classified, the first step is the calculation of the distance between x^* and all other instances in the training set. Euclidian distance is the most popular among the distance metrics:

$$d(x, x^*) = \sqrt{\sum_{i=1}^{n}(x_i - x^*)} \tag{2.1}$$

Though kNN is a simple, understandable, and fast algorithm, it requires large amount of memory and it is sensitive to irrelevant or redundant features. Nearest neighbor classification requires Np operations to find the neighbors per query point with N observations and p predictors. Also, the choice of k is a major drawback in kNN algorithm. Generally, trial and error approach is used to define the optimum k value to have the maximum accuracy on test set.

2.1.2 Neural Networks

The neural network term was firstly used for biological neuron circuits which are the collection of interconnected neurons. Messages electrochemically pass along these connections and synaptic connections between neurons involve in learning. In pattern recognition, neural networks are usually referred as artificial neural networks [3] in which artificial neurons are represented as nodes organized in layers. When unit connections do not constitute a directed cycle then it is called feed-forward neural network in which information moves from the input nodes through the hidden nodes to the output nodes. There are K units at the top with the kth unit modeling the probability of class k in K-class classification (Figure 2.1).

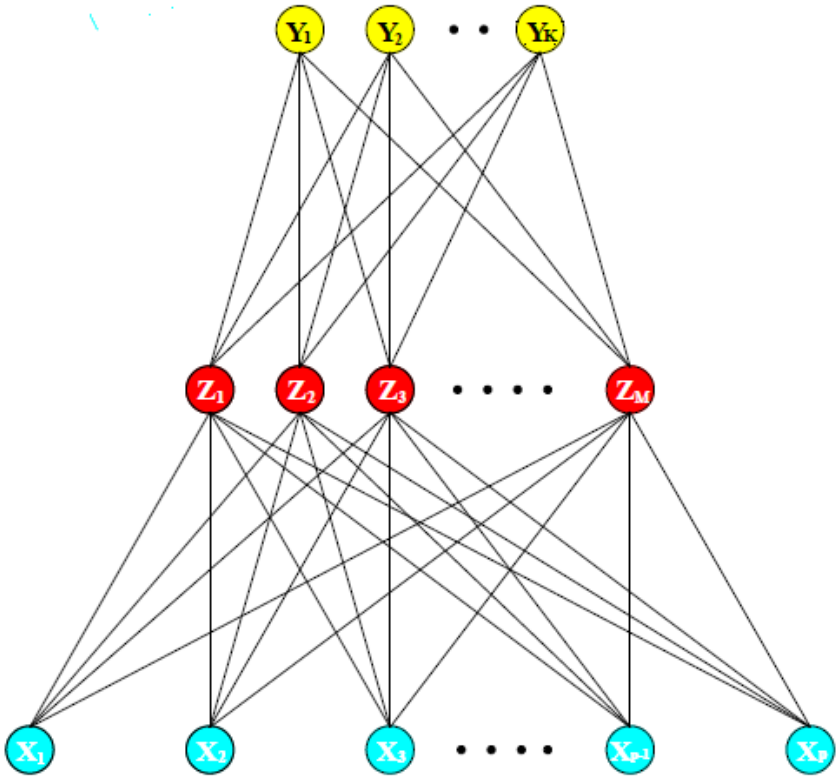

Figure 2.1 Single hidden layer, feed forward neural network

Z_m are the derived features (hidden units) obtained from the linear combinations of the inputs and then target Y_k is modeled from the linear combinations of Z_m [4].

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, \dots, M \tag{2.2}$$

$$T_k = \beta_{0k} + \beta_k^T Z, k = 1, \dots, K \tag{2.3}$$

$$f_k(X) = g_k(T), k = 1, \dots, K \tag{2.4}$$

where σ is activation function, $Z = (Z_1, Z_2, \dots, Z_M)$, and $T = (T_1, T_2, \dots, T_K)$.

σ can be sigmoid $\sigma(v) = 1/(1 + e^{-v})$, as well as it can be a radial basis function. The simplest feedforward neural network is called perceptron [5] which is a binary linear classifier. It maps the input, a real valued vector *x*, to an output value *f(x)* which is a binary value. A more complex feed forward artificial neural network is the multilayer perceptron [6], designed as by connecting perceptrons, which uses three or more layers of nodes with nonlinear activation functions. Each node in one layer is connected to every node in the following layer with a weight w_{ij}. The weight values of all nodes is determined by backpropagation algorithm [7] which is the abbreviation of backwards propagation of errors. There are some constraints that limit the performance of backpropagation algorithm: the convergence from the backpropagation is slow and not guaranteed and can result with local minimum [8-9].

2.1.3 Decision Trees

Decision tree [10] is a popular classification method which is trained by the selection of individual features iteratively. The input *X* is split into descendant subsets starting with *X* itself. The most significant feature is selected as the top split node. In decision trees, the nodes represent features, whereas branches represent values in decision trees (Figure 2.2). Top-down induction at decision trees is as follows: 1) Let *A* is best decision attribute for the next

node. 2) Assign *A* as decision attribute to node. 3) Create a new descendant of node for each value of *A*. 4) Sort training examples to leaf nodes. 5) Stop if training examples are perfectly classified, otherwise iterate over leaf nodes [11].

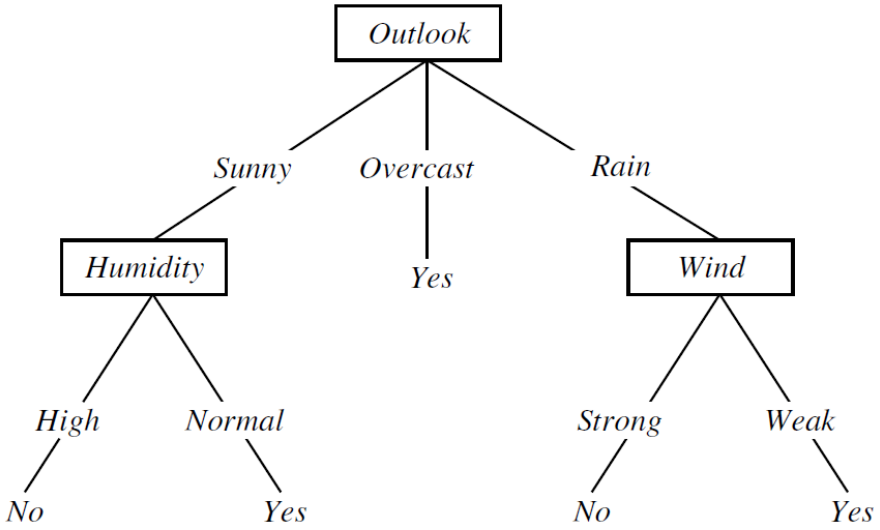

Figure 2.2 A decision tree representation

Random forest is a decision tree based classifier which was firstly developed by Leo Breiman [12]. It is a fast and effective algorithm since it can handle thousands of input variables, it can predict the significance of variables, and its accuracy is not affected by missing data. However this classifier has also some disadvantages that it cannot handle large number of irrelevant features and it has a tendency for overfitting.

Overall, decision tree classification is a simple and robust method to understand, interpret and it requires little data preparation. This method can also handle both numerical and categorical data. However there are some limitations for the usage of decision tree classification. Decision tree learning algorithms uses heuristics such as greedy algorithm in which locally optimal decisions are made at each node and cannot guarantee the global optimal decision tree [13]. Also decision trees have a tendency to create overly complex trees

which cause an overfitting problem. Another disadvantage of decision trees is some information can be lost when discretizing continuous variables.

2.1.4 Bayesian Classifiers

Bayesian classifier is a simple probabilistic algorithm based on Bayes' theorem with independent feature model [11]. Bayesian classifiers minimize total misclassification using $\gamma(x) = arg\ min_k \sum_{c=1}^{r0} cost(k,c)p(c|x_1.x_2 \dots \dots x_n)$, where *cost(k,c)* is the cost for misclassification [14]. This classifier assigns the most probable a posteriori class to a given object using $\gamma(x) = arg\ max_c p(c|x_1.x_2 \dots x_n) = arg\ max_c p(c)p(x_1.x_2 \dots x_n|c)$. In Naive Bayes classification, which is the simplest Bayesian classifier, existence of an attribute to a particular feature of a class is unrelated to existence of any other feature.

$$f_j(X) = \prod_{k=1}^{p} f_{jk}(X_k) \tag{2.5}$$

where features are represented as X_k with a given class $G=j$.

The Naive Bayes uses the most probable, maximum a posteriori decision rule with the Bayes independent rule and calculate the most probable a posteriori assignment as:

$$c^* = arg\ max_c p(c|x_1.x_2 \dots x_n) = arg\ max_c p(c) \prod_{i=1}^{n} p(x_{1i}|c) \tag{2.6}$$

Bayesian Network, directed acyclic graphical model, is a probabilistic graph that models a set of random variables and their conditional independencies. Nodes represent random variables and edges capture the direct dependencies between variables.

2.1.5 Logistic Regression

Logistic regression [15] aims to model the posterior probabilities of the K classes with linear functions in x. The model also guarantees that probabilities sum to one (Equation 2.8) and remain in [0,1].

$$p(G = k|X = x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)},\ k=1\dots,K-1 \tag{2.7}$$

$$p(G = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} \exp\,(\beta_{l0}+\beta_l^T x)} \tag{2.8}$$

Logistic regression models are generally fit by maximum likelihood. The log-likelihood for *N* observations can be shown as:

$$l(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i;\theta) \tag{2.9}$$

where $\theta = \{\beta_{10}, \beta_1^T, \dots, \beta_{(K-1)0}, \beta_{K-1}^T$ and $p_k(x_i;\theta) = p(G = k|X = x;\theta)$.
In the case of binary class problem, the log-likelihood can be written as:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N}\{y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))\} \\ &= \sum_{i=1}^{N}\{y_i\beta^T x_i - \log(1+e^{\beta^T x_i}) \end{aligned} \tag{2.10}$$

Taking the derivative and setting to zero maximizes the log-likelihood. Then this equation can be solved using Newton-Raphson algorithm and can be seen as solving weighted least squares problem [4].

Logistic regression is useful when the dependent variable is either binomial or multinomial values. Binomial logistic regression is used when the dependent variable is a dichotomy and the independents are continuous and/or categorical. Multinomial logistic regression exists to handle the case of dependent variables with more than two classes.

2.1.6 Support Vector Machines

Support vector machine (SVM) [16] is one of the popular and useful techniques for data classification, especially for two-class problems. Considering a two class problem in Figure 2.3, the task is to find the best decision boundary separating samples of different classes. SVM achieves this boundary problem by finding the decision boundary that achieves maximum margin between the two classes.

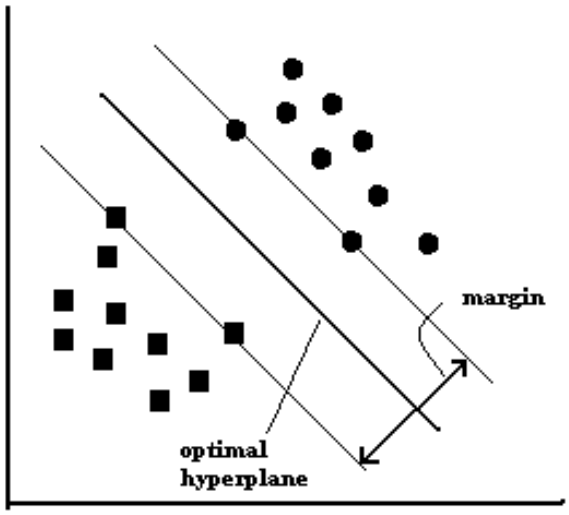

Figure 2.3 Support vector machines class boundaries

The distance between a planar decision surface that separates two classes and the closest training samples to decision surface is defined as the *margin*. SVM finds an optimal hyperplane $wx^T + b = 0$, where w is the p-dimensional vector perpendicular to the hyperplane and b is the bias. The labeled training dataset is denoted as $(\mathrm{x}_1, \mathrm{xy}_1), \ldots, (\mathrm{x}_{\mathrm{N_T}}, \mathrm{xy}_{\mathrm{N_T}})$, where $x_i \in R^p \; and \; y_i \in \{-1, +1\}$.

$$\min_z \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{N_T}\xi_i \tag{2.11}$$

subject to

$$y_i\left(wx_i^T + b\right) - 1 + \xi_i \geq 0, \forall_i \tag{2.12}$$

The objective function is given Equation 2.11 that is to find w and b such that hyperplane maximizes the margin $1/||w||^2$. $C > 0$ is the penalty parameter of the error and ξ_i represents the slack variables.

SVM can also be used for nonlinear separation by using kernel transformation in which every matrix product ($x_i x^T$) is replaced by nonlinear kernel function K($x_i x$). In other words, original input space *X* is transformed into high dimensional feature space; hence transformed feature space can be linearly separable. Also, new decision boundaries are nonlinear in the original input space and linear in new high dimensional space. The most well known kernels are polynomial $K(x,z) = (xz^T + 1)^d$ and radial basis function (RBF) $K(x,z) = exp(-\gamma\|x-z\|^2)$. However the choice of kernel function and the parameter values directly affect the performance of SVM.

2.2 Applications of data classification in bioinformatics

The field of bioinformatics has been enormously evolving, due to the rapid growth of biological information. This evolution continues with the two ways: 1) The development of the efficient management and storage of data. 2) The extraction of significant information from these data. Therefore, many machine learning techniques have been applied to computational biology and bioinformatics for the analysis, prediction and interpretation of biological phenomena.

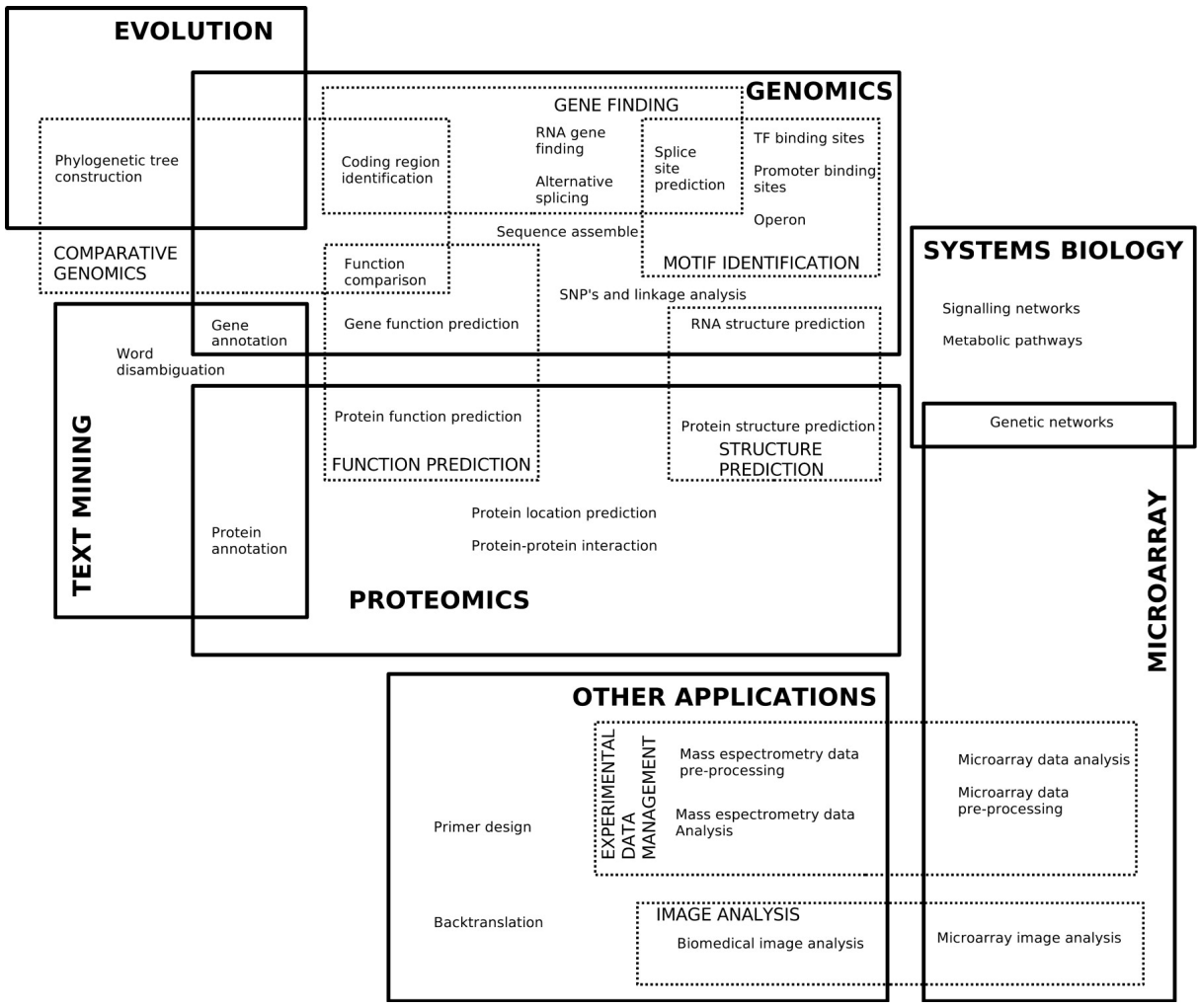

Figure 2.4 Machine learning applications in bioinformatics [14].

Figure 2.4 shows the bioinformatics disciplines such as genomics, proteomics, systems biology, evolution, microarray analysis and text mining that benefited from machine learning algorithms. Additionally, cheminformatics and drug design are the other topics taking the advantage of machine learning techniques.

2.2.1 Genomics

Genomics, the study of genomes of organisms, is an important application field of bioinformatics. The number of sequences submitted to GenBank has been exponentially increasing. As of 2007, the traditional GenBank contain over 20 billion nucleotide bases from more than 76 million individual sequences and 15 million new sequences were added in 2006. The number of nucleotide bases reach 190 billion bases with the contribution of Whole genome Shotgun (WGS) project [17]. There are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in the WGS division as of August 2009 [18]. As a result of this huge incoming, more effort is required to process these data.

As an important application of genomics, gene finding is the study to identify the biologically functional sequences. This includes coding gene identification, RNA gene finding, and splice site prediction. There are many gene prediction programs available; however these programs cannot answer all questions due to the complexity of transcription and translational processes [19].

In the study of Salzberg [20], protein coding regions in human DNA is searched using classification trees. Castelo and Guigo [21] uses Bayesian classifier for the splice site prediction problem. There are also optimization based approaches for the feature subset selection in the splice site prediction problem [22-23]. In the study of Carter et al. [24] support vector machines and neural networks are used for the identification of functional RNA genes. Lopez-Bigas and Ouzounias [25], classification trees are used in the genome-wide identification of genes responsible for genetic diseases. In the study Bao and Cui [26], support vector machines and random forests are compared in the prediction of the phenotypic effects of nonsynonmous single nucleotide polymorphisms. In another paper tree-based techniques were used for recovering the relationship between motifs and gene expression levels [27].

2.2.2 Proteomics

Protein secondary structure prediction is one of the applications of proteomics. Proteins consist of thousands of atoms and bounds making them very complex structures; therefore protein structure prediction problem is a very complicated combinatorial problem. Selbig et al. [28] uses classification tree in a consensus method for the protein secondary structure. In another work, support vector machines is employed for the classifications of protein functional families [29].

PHD is the first method giving higher than 70% accuracy for secondary structure prediction [30]. Two-level neural network was used in PHD. Another method, GORV is a secondary structure prediction method based on Bayesian statistics [31]. Yang et al. [32] uses the combination of support vector machines and Bayesian classifier to predict the hot spot residues that play an important role in protein-protein interactions. An optimization based method, functional correlation optimization method (FCOM) is used for the identification of protein function using protein-protein interaction data [33]. In another study, different classification methods were evaluated for protein interaction prediction [34]. Mohammed et al. uses random forest for human protein-protein interaction prediction [35].

Another application of proteomics, subcellular location of a protein from its amino acid sequence is predicted using k-nearest neighbor [36].

2.2.3 Systems Biology

Systems biology is a field of examination of the structure and dynamics of cellular and organism functions with an integration approach rather than an isolated part of a cell or an organism (reductionist approach). This approach can be helpful to study of the interactions between components of biological systems and the effect of these interactions on the

behavior of the system. Therefore the applications explained in this thesis are also disciplines considered the domains of systems biology.

Metabolic pathway prediction is one of the main focus in computational systems biology. Dale et al. [37] evaluated naïve Bayes, *k*-nearest neighbors, decision trees, and logistic regression methods for metabolic pathway prediction. A semi-supervised learning approach is proposed to predict synthetic genetic interactions using functional and topological features of functional gene network [38]. In another study, new drug targets are identified using support vector machines by distinguishing essential and non-essential enzymes [39].

2.2.4 Evolution

A phylogenetic tree or evolutionary tree is a schematic representation of evolutionary relationships among organisms. Construction of phylogenetic trees notably makes use of machine learning algorithms. In a study, phylogenetic tree is used to extract correlations between orthologous proteins and support vector machines is used for the prediction of protein-protein interaction [40].

2.2.5 Microarrays

DNA microarray is multiplex technology and utilized to measure changes in gene expressions, to detect single nucleotide polymorphisms and to resequence mutant genomes. Microarray data is pretty important to understand the fundamental questions of diseases. The most important application of microarray data is the classification of disease types using gene expression data. Also the selection of significant genes that play role in specific diseases is another important consideration in microarray analysis. There are many machine learning methods performed to analyze microarray data such as k-nearest-

neighbors [28], artificial neural networks [30], support vector machines [32, 33], maximal margin linear programming [34], random forest [35], bagging and adaboost [41].

2.2.6 Cheminformatics and Drug Design

The early prediction of activities of drug candidates is one of the main efforts in drug design world. Quantitative structure activity relation (QSAR) is very useful and accurate in lead optimization studies when the molecules are structurally very similar. QSAR correlates structure and function within a series of molecules in terms of physicochemical parameters and steric properties. 3D QSAR methods consider three-dimensional structures and a binding form of the ligands on the target protein [42]. This draws the attention that drug activities on specific targets (outputs) can be modeled by using a wide range of molecular descriptors (inputs). Yap et al. [43] predicted the cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates with a high accuracy by using Support Vector Machines (SVM) with 6 common molecular descriptors. Classification of 1,4-dihydropyridine calcium antagonists were performed by using the Least-Square Support Vector Machines (LSSVM) method to obtain a seven descriptor model [44]. Neugebauer uses decision trees for the prediction of protein-protein interaction inhibitors [45].

Chapter 3

HYPER-BOX ENCLOSURE CLASSIFIER

The data classification problem consists of two parts: training and testing. In the training part, characteristics of objects belonging to a defined label are determined and instances belonging to different classes are differentiated using a model. These model can be an instance based model (nearest-neighbor), a probabilistic model (bayesian, logistic regression), a linear model (perceptron, support vector machine), a decision model (decision tree, random forest). Generally, training part is also divided into two parts for the optimization of parameters. In the first part parameter values are estimated and in the second part (validation) hyperparameters are tuned and final model is built. In the testing part, the classes of test objects are predicted and the prediction accuracy of the classifier is evaluated. Hyper-box enclosure (HBE) classifier is a mixed integer linear programming based model and it does not require a parameter optimization part. This algorithm uses hyper-boxes for the definition of set boundaries that include all or some of the training objects. Depending on the complexity of data, HBE can use more than one hyper-box to cover all instances having the same label (class). A very important consideration in using hyper-boxes is the number of boxes used to define a class. If the total number of hyper-boxes is equal to the number of classes, then the data classification is very efficient. On the other hand; if there are as many hyper-boxes of a class as the number of instances in a class, then the algorithm is considered that it overfits the training data.

Figure 3.1 summarizes the steps of hyper-box enclosure algorithm. First the ‘problematic’ distances are determined as a preprocessing step and representative samples from these problematic instances are determined using integer programming (IP). The main

model, the initial boundaries of hyper- boxes are built using mixed integer linear programming in the next step. If boxes intersect, the intersections are eliminated by redefining the box boundaries iteratively.

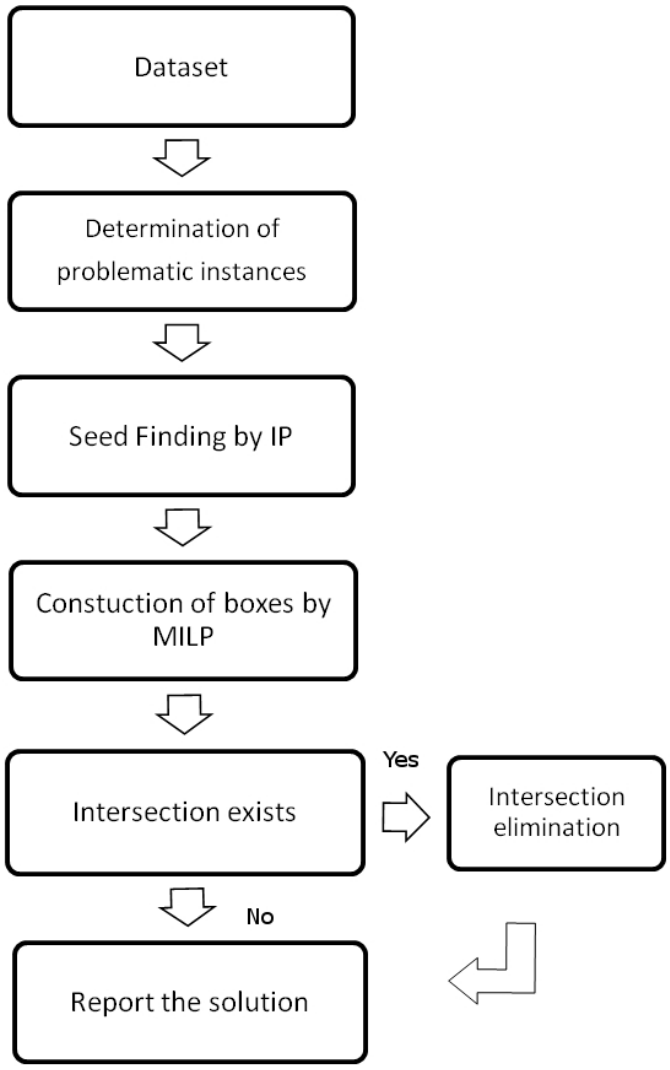

Figure 3.1 Steps of hyper-box enclosure algorithm

3.1 Preprocessing

First, rescaling (standardization) is performed to all datasets to make the attribute values between 0 and 1. Samples with corresponding attribute values were randomized for the prevention of unbiased data. Also, attributes having missing values are removed from the datasets. The maximum and the minimum attribute values are calculated for each class. Then, the boundaries of the classes are compared to identify overlapping ones. The instances enclosed by boxes belonging to other classes are identified, if the boundaries of the classes overlap. These instances, nonseperable from the instances of the other classes with a single hyper-box, are called as 'problematic' instances. When the number of 'problematic' instances is large, the same procedure is repeated to reduce the total number of problematic instances. In some cases, applying one or two times the same procedure do not reduce the number of problematic instances as desired; hence we use integer programming based seed finding algorithm to reduce this computational complexity.

3.2 Seed Finding

This step is to improve the computational efficiency by determining representative instances (seeds) for each class without changing the optimal solution. In seed finding part, an instance for each class is selected and assigned to corresponding class. The following criteria is considered for the finding of seeds: the seeds for each class must be chosen to ensure that seeds are separated well from each other as well as being a good example of the group of instances in the same class. For this purpose, a pure integer programming formulation is developed. Instances are represented by the parameter a_{im} that denotes the value of attribute m for the instances i. The class k of instance i belongs to is given by the set D_{ik}. Moreover, $PP_{ii'}$ represents the distance between two instances i and i'. This distance is calculated using Euclidean distance in m dimensional space as given in Equation 3.1.

$$PP_{ii'} = \sqrt{\sum_m (a_{im} - a_{i'm})^2} \tag{3.1}$$

$$\min z = \sum_k \sum_{i \in k} \sum_{i' \in k} PP_{ii'} YP_i - \left(\frac{1}{card(i \in k)}\right) * \sum_k \sum_{i \in k} \sum_{i' \notin k} PP_{ii'} YP_i \tag{3.2}$$

$$\sum_{i \in k} YP_i = 1 \tag{3.3}$$

Equation 3.2 gives the objective of function which is for the minimization of the distances from each seed to instance of its group (in-class distances) and for the maximization of the average distances from each seed to the instances that belong to other classes (out-class distances). Equation 3.3 states that every class must have exactly one seed.

3.3 Construction of boxes with seeds

The objective of the mixed integer linear programming model is the minimization of the misclassified instances in the data set with the minimum number of hyper-boxes. Equation 3.4 is the objective function:

$$\min z = \sum_i \sum_k yp_{ik} + \sum_k yb_l \tag{3.4}$$

where yp_{ik} indicates the misclassification of instance *i* to class *k* and the existence of hyper-box *l* is represented by binary variable yb_l.

The lower and upper bounds of the hyper-boxes are determined by the instances that are enclosed within the hyper-boxes. Hence, lower and upper bounds of hyper-boxes are

calculated. The bounds of hyper-boxes exist if and only if this hyper-box is assigned to a class. Every data point must be assigned to a single box and single class.

3.4 Intersection elimination

After the building of the model on problematic instances, the 'non-problematic instances' are assigned to hyper-boxes in a straight forward way. For each class, k hyper-boxes are defined and we assign a 'non-problematic instance' to corresponding newly defined hyper-box. Each 'non-problematic instance' is assigned to a hyper-box until no instance left outside of a hyper-box. Finally, the bounds of these new hyper-boxes are determined by considering the maximum and minimum attribute values of all instances in these hyper-boxes. It is also possible that these hyper-boxes can overlap, and then instances are separated from the original hyper-box until all intersections are eliminated.

3.5 Testing

In the testing part, the values of test instances are considered and determined which boxes include test instances to predict their classes. There are three possible cases for a test instance when we determine its class. It can be

1. within the boundaries of a single hyper-box
2. within the boundaries of more than one hyper-boxes
3. not enclosed any of the hyper-box

In the first case, test instance is directly assigned to a class which represented by hyper-box whose boundaries enclose the test instance. If the test instance is within the boundaries of more than one hyper-box, in the second case, the test instance is assigned to the classes of these hyper-boxes. As an example, if a data point belonging to class is enclosed by two different hyper-boxes whose have different classes, then this data point is assumed to be

classified with 50% accuracy and the number 0.5 is added to the number of correctly classified test samples. In the third case, the shortest distance between the test point and the hyperplanes defining the hyper-box. The number of hyper-planes is *2(M-1)* where M is the total number of attributes. If the test instance is between the bounds of smaller or equal to *M-2* attributes, then shortest distance between the test instance and edges of the hyper-box. The number of edges in a hyper-box is equal to $M2^{M-1}$; therefore $ML2^{M-1}$ distances are calculated for each test instance where *L* is the number of hyper-boxes. Then minimum of calculated distances is selected. In the case that the text instance is not within the lower and upper bounds of any attributes, the shortest distance between the test instance and extreme points are calculated. The number of extreme points in a hyper-box is 2^M; thus 2^ML distances are calculated for each test point. Then the minimum of calculated distances is selected.

For the performance evaluation of the classifier, we use some terms derived from the confusion matrix. There are four possible outcomes in binary class problem and the outcomes are labeled as *positive* (p) or *negative* (n) class. *True positive* (TP) is the case if the outcome (p) from the prediction is equal to the actual value that is also p. If the actual value is n, then it is called *false positive* (FP). If both the predicted value and the actual value are n, then, it is called *true negative* (TN). *False negative* (FN) is the case where the actual value is p and the predicted value is n. *True positive rate* (TPR) is obtained by the division of TP to P. *False positive rate* (FPR) is the rate of FPs to all negatives. Finally, accuracy (ACC) is the rate of all correctly predicted outcomes (TP+TN) to all samples in the test set (TP+TN+FP+FN).

Chapter 4

CANCER CLASSIFICATION and GENE SELECTION

4.1 Introduction

Cancer is a class of a disease where cells start to divide in an uncontrollable way. Cancerous cells, depending on their types, display different gene expression patterns which may be used for early diagnostics. Different approaches have been investigated to detect global gene expression in cancerous cells including the microarray technology. Classification of tumor types with respect to gene expression levels in specific cancer types is an important use of microarray data. Also, it is quite critical to find an optimal set of markers among those up or down regulated genes that can be clinically utilized to build assays for the diagnosis or to follow progression of specific cancer types. In this investigation, we use our algorithm, hyper-box enclosure method, for the classification of some cancer types with a minimal set of predictor genes. This method, a user friendly and efficient classifier may also allow the clinicians to diagnose and follow progression of certain cancer types.

Microarray technology provides wealth information of expression level of thousand genes simultaneously and it has been used for diagnostic and prognostic purposes for various types of diseases. The data obtained from microarray measurements will significantly lead to understanding of the fundamental questions of diseases both in biology and clinical medicine. Cancer is the most deadly genetic disease, and it is caused by either acquired mutations or epigenetic changes lead to altered gene expressions profile of cancerous cell. Consequently, microarray can be employed to identify up or down regulated genes that play a role on the specific cancers, activation of oncogenic pathways,

and to discover novel biomarkers for the clinical diagnosis [46]. However, such a technique would be an expensive and time-consuming process in terms of clinical application. Building a small set of marker genes can provide us to make antibody assays for the diagnosis of specific types of cancer or to find specific marker genes enable to assess the cancer.

The number of genes (features) considered in the analysis of microarray data is very critical. A very small number of genes usually cannot be optimal, whereas huge number of genes decreases the information criterion by adding noise [47]. Therefore, it is necessary to find an optimal set of genes as predictors that help to classify different labeled cells with high prediction accuracy. The number of genes per sample is relatively high compared to number of samples in microarray data. This high dimensionality increases the computational complexity while usually decreasing the accuracy of the classification. This fact brings the necessity of feature selection by ranking or feature reduction for the high dimensional gene space. The relevance of genes in cancer occurrence can be categorized into three classes: Strongly relevant, weakly relevant and irrelevant genes [48]. Strongly relevant genes are the ones that have been shown in cancer cell formation and always needed in the optimal set, whereas the weak relevant genes are necessary for the optimal set at some conditions. To summarize, the purpose of gene selection is 1) making the classification easier by revealing only the relevant genes 2) improving the classification accuracy 3) reducing the dimensionality of the data set [49]. In the effort to choose the optimal subset of predictor genes, different methods such as neighborhood analysis[50], gene shaving [51], principle component analysis [52], genetic evolution of subsets of expressed sequences named GESSES [53] are employed.

The effectiveness of the selected gene subset is measured by their prediction accuracy or error rate in classification. Classification is a crucial question in microarray experiments for the prediction of outcome or phenotype of cells. Different machine learning approaches

have been performed to analyze microarray data under many conditions including k-nearest-neighbors [50], artificial neural networks [52], support vector machines [54-55], maximal margin linear programming [56], random forest [57]. As an alternative classification approach, it has been proved that mixed integer programming based classification is highly effective in different applications such as protein fold type prediction [58] and drug classification [59-61].

In this application, mixed integer linear programming based hyper-box enclosure approach is employed for the classification of cancer types. We first introduce and establish a consistent classification method for different types of microarray data. Second we provide an optimal set of genes as cancer best diagnostic indicators that gives the highest accuracy in classification. For the feature selection, information gain attribute evaluator, relief attribute evaluator and correlation-based feature selection (CFS) methods are used. We conduct experiments using six well known cancer data sets including leukemia data set [50], two prostate cancer data sets [62], lymphoma [63], diffuse large B-cell lymphoma (DLBCL) [64], small round blue cell tumors (SRBCT) [52]. Moreover, biological interpretation of selected genes is presented with the explanation of their relationship to the related cancer types.

4.2 Data sets

Microarray data sets were obtained from Keng Ridge Bio-Medical (http://datam.i2r.a star.edu.sg/datasets/krbd/) and Artificial Intelligence Laboratory (http://www.ailab.si/supp/ bi-cancer/projections/index.htm) databases. The data set structure is given in Table 4.1. Information gain attribute evaluator, relief attribute evaluator and correlation-based feature selection methods are employed for the feature selection. We have selected the most popular data sets in the literature for the evaluation of our algorithm.

Table 4.1 Details of cancer microarray data sets

Data Set	Instances	Genes	Classes	Reference
Leukemia	72	7129	2	[50]
Prostate cancer	102	12600	2	[62]
Prostate outcome	21	12600	2	[62]
DLBCL	77	7129	2	[64]
Lymphoma	47	4026	2	[63]
SRBCT	83	2308	4	[52]

4.3 Preprocessing and Gene Selection

We performed rescaling to all data sets by using the expression of *(A(i) - minA) / (maxA-minA)* where *A(i)* is the gene expression value of *i*th sample. Additionally, in cross validation runs, the samples with their gene expression values and classes are randomly reordered before dividing the data into *k*-fold.

There are generally three types of approaches in feature selection: Filters, wrappers and feature weighting. Filter methods eliminate irrelevant features according to some prior knowledge. Wrapper approaches use machine learning algorithms to evaluate the feature subsets; however they have high computational complexity when they combined with classification algorithms. Feature weighting methods simply weight features instead of selecting a subset of features that is a combinatorial problem. We employed information gain attribute evaluator, relief attribute evaluator and correlation-based feature selection (CFS) from Weka machine learning package [65] for the gene selection. Information gain evaluates a feature by measuring the information gain with respect to the class:

$$H(Y) = -\sum p(y) \log_2 p(y) \tag{4.1}$$

where *Y* and *X* are the features, *p(y)* is the marginal probability density function for random variable *Y*.

Equation 4.1 provides the entropy of *Y*. Entropy is a measure of uncertainty in information theory. There is a relationship between feature *X* and *Y* when following cases are ensured: i) expression values of feature *Y* in the training set are partitioned in due to the expression values of second feature *X* ii) the entropy of *Y* prior to partitioning is higher than the entropy of *Y* with respect to the partitions induced by *X*. The entropy of *Y* after observing *X* is given in Equation 4.2.

$$\mathrm{H(Y|X)} = -\sum \mathrm{p(x)} \sum \mathrm{p(y|x)} \log_2 \mathrm{p(y|x)} \tag{4.2}$$

$$\mathrm{InfoGain} = \mathrm{H(Y)} - \mathrm{H(Y|X)} \tag{4.3}$$

where *p(y|x)* is the conditional probability of *y* given *x*. Information Gain (Equation 4.3) is a measure of additional information about *Y* provided by *X* representing the amount by which the entropy of *Y* decreases.

Relief attribute evaluator is an evaluating algorithm that rates features due to these facts: (1) how well their values distinguish among instances of different classes (2) how well they cluster instances of the same class [66].

Correlation-based feature selection (CFS) is a fast algorithm that reveals a good feature subset that contains features highly correlated with the class, yet uncorrelated with each other [67].

$$CFS_S = \frac{k\bar{r}_{cf}}{\sqrt{k+k(k-1)\bar{r}_{ff}}} \tag{4.4}$$

where CFS_S is the score of a feature subset *S* containing *k* features, $\bar{r}_{cf}$ is the average feature to class correlation ($f \in S$), and $\bar{r}_{ff}$ is the average feature to feature correlation. The numerator of Equation 4.4 indicates how predictive of the class a group of features are and the denominator is a measure of redundancy among that group of features.

4.4 Optimal Gene Set Finding

After the initial gene ranking, the optimal gene set is searched by using F cumulative distribution function (FCDF) and the classification iteratively. FCDF is computed at each values in X using the corresponding parameters in v1 and v1. FCDF is:

$$p = F(x|v_1, v_2) = \int \frac{\Gamma\left[\frac{(v_1+v_2)}{2}\right]}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)} \left(\frac{v_1}{v_2}\right)^{v_1/2} \frac{t^{(v_1-2)/2}}{[1+\left(\frac{v_1}{v_2}\right)t]^{(v_1+v_2)/2}} dt \qquad (4.5)$$

where p is the probability that single observation from an F distribution with parameters v_1 and v_2. In our case, v_1 is the number of samples and v_2 is number of samples at each class. X is the division of the variances of each gene of all samples to the variances of each gene of each sample. As a result p value for one class is calculated 1- FCDF. While defining the relatively irrelevant genes (the weakest genes) to leave the model within the optimal gene subset the gene with the maximum p value for one of the classes is selected. In this way, the weakest gene is replaced by the strongest one. The strongest gene is described as the attribute whose maximum p value for high or low classes is the minimum among other genes. As the weakest and the strongest genes were calculated by FCDF, the weakest genes are replaced by the strongest ones, and hyper-box enclosure method is used at each iteration. The set giving the highest classification accuracy is reported as the optimal gene set. Also, the genes whose ranking scores are the highest are checked whether there is a redundancy among them by considering the pair correlation-coefficients. As a result, this approach selects the most relevant genes to the target classes and minimizes the redundancy among the selected genes to define an optimal gene set which provides the highest classification accuracy.

4.5 Classification results and discussions

4.5.1 Leukemia

This data [50] consist of two types of leukemia, acute lymphoblastic leukemia (ALL), and acute myeloid leukemia (AML). Each sample obtained from bone marrow samples was analyzed using Affymetrix microarrays with 7129 genes. The training data consists of 38 samples (27 ALL and 11 AML), and the test data consists of 34 samples (20 ALL and 14 AML).

Table 4.2 reports the classification accuracies of different classification algorithms on leukemia data set. The hyper-box enclosure method classifies all test samples perfectly, and gives also the best leave-one-out (LOOCV) result (98.61%) together with the logistic regression classifier. RBF Network and hyper-box method gives 97.22% accuracy with ten-fold cross validation (10-CV). As an additional evaluation, hyper-box enclosure algorithm also perfectly classifies all test samples using 4 genes proposed by Golub et al. [50] and gives a classification accuracy of 98.57% using ten-fold cross validation. For the comparison of our results with different algorithms in the literature, we selected the papers including the highest classification accuracies. Tan and Gilbert [41] report 91.18% (10-CV) with 1038 genes using Bagging and AdaBoost methods. Dettling and Buhlmann [68] obtained the accuracy of 98.61% (LOOCV) with 3 gene clusters using aggregated trees method, where gene clusters are reported as a minimum number of 1 and maximum number of 23. Nguyen and Rocke [69] correctly classifies 33 out of 34 with 50 genes using partial least squares (PLS) and Logistic Discrimination (LD) classification. Lee et al. [70] misclassifies one sample in test using 5 genes. Deutsch [53] uses an iteration algorithm to obtain high classification accuracy with a minimum number of genes. In this work, test samples are classified after an average dimension of about 9.

Table 4.2 Classification results of leukemia data set

Classifier	Test Set	10-CV	LOOCV
Hyper-box enclosure	**100**	**97.22**	**98.61**
Bayes Net	94.12	95.83	95.83
LibSVM	58.82	91.67	91.67
SMO	97.06	93.06	94.44
Logistic Regression	91.18	94.44	**98.61**
RBF Network	97.06	**97.22**	97.22
IBk	97.06	95.83	95.83
J48	94.12	91.67	90.28
Random Forest	94.12	91.67	90.2

Antonov et al. [56] could achieve to predict all samples in test set and obtained the accuracy of 98% (LOOCV) using 132 genes. Chen et al. [71], perfect test set accuracy with minimum 7 genes is obtained. Consideration of all these results shows that hyper-box enclosure method is the most accurate classifier on leukemia data set considering all validation methods including test set validation, ten-fold cross validation and leave-one-out validation. The prediction accuracies at each fold is given in Table 4.3.

All the genes that have been selected in our studies are also selected by Lee et al. in their significant gene pool which consists of 27 genes. However there are redundant genes in this significant gene set. Table 4.4 gives the genes overlapping with the selected genes by other groups. Although 4 genes proposed as optimal gene set is reported by other groups that usually report a large number of gene set.

Table 4.3 Performance evaluation of HBE with 10-CV for leukemia data set

Fold	TP	FP	FN	TN	TPR	FPR	ACC
1	4	0	0	3	1	0	1
2	5	0	0	2	1	0	1
3	6	1	0	0	1	1	0.86
4	4	0	0	3	1	0	1
5	3	1	0	3	1	0.25	0.86
6	4	0	0	3	1	0	1
7	5	0	0	2	1	0	1
8	6	0	0	1	1	0	1
9	4	0	0	3	1	0	1
10	4	0	0	3	1	0	1

The highest accuracy is obtained with an optimal gene set including 4 genes: Myeloperoxidase (M19507_at), DF D component of complement factor, adipsin (M84526_at), CD33 antigen, differentiation antigen (M23197_at), TCF3 transcription factor 3, E2A immunoglobulin enhancer binding factors E12/E47 (M31523_at). Myeloperoxidase is a peroxidase enzyme that produces hypochlorous acid from hydrogen peroxide and chloride anion. In recent years, myeloperoxidase staining is used in the diagnosis of acute myeloid leukemia to show that the leukemic cells were obtained from the myeloid lineage [69, 72]. Our result is agreement with Chen et al. [71] where they also selected myeloperoxidase in their classification method. The membrane antigen CD33 is a sialic acid-dependent cell adhesion molecule is a membrane protein. CD33 is highly expressed on the surface of leukemic blasts. About 85-90 of acute AML cases are considered to be CD33 positive [73]. CD33 is constitutively expressed in haematopoietic progenitors, but at significantly lower membrane density than in leukemia cells [74]. Therefore, CD33 represents an interesting target for antibody-based anti-leukemic

therapies. This gene is also selected as an important gene in significant gene subset studies conducted by other researchers [75-76].

Table 4.4 Selected leukemia genes overlapping with genes selected by other groups

Gene	Reference
Myeloperoxide	[56, 70-71, 77]
Cd33	[49-50, 70, 75, 77-78]
T3F3	[70, 77-79]
Adipsin	[70, 77-79]

Transcription factor 3 (TCF3) plays an important role with tissue-specific basic helix-loop-helix (bHLH) in embryogenesis [80]. The TCF3-HLF fusion transcription factor generated by t(17;19)(q22;p13) translocation is found in a small subset of pro-B cell acute ALLs and promotes leukemogenesis by substituting for the antiapoptotic function of cytokines [81]. Also it has been shown in ALLs patients TCF3 level is up-regulated due to this translocation. Additionally, this protein is the cause of forms of pre-B-cell acute lymphoblastic leukemia [82]. Although adipsine is a serine protease homolog which is synthesized and secreted by adipose cells and is found in the bloodstream it has been shown to plays a role in myeloid cell differentiation [83]. Study carried out by Sakhinia et al. [84] indicated that gene expression is up-regulated in acute AML patients by real time PCR. All the genes that have been selected in our study are also selected by Lee et al. [70] in their significant gene pool.

4.5.2 Prostate Cancer

The prostate cancer data consists of 102 tissue samples (52 prostate tumor and 50 normal tissues) with 12600 genes. Considering the results given in Table 4.5, hyper-box enclosure method is again the most accurate classifier among others with LOOCV and it is the second most accurate classifier with 95.10% accuracy using ten-fold cross validation. Both support vector machines LibSVM and SMO report 96.08% accuracy.

Table 4.5 Classification results of prostate cancer data set

Classifier	10-CV	LOOCV
Hyper-box enclosure	95.10	**96.08**
Bayes Net	94.12	95.10
LibSVM	**96.08**	95.10
SMO	**96.08**	95.10
Logistic Regression	91.18	92.16
RBF Network	94.12	93.14
IBk	92.16	93.14
J48	85.29	90.20
Random Forest	93.14	94.12

Tan and Gilbert [41] have 75.53% (10-CV) with 2071 genes using Bagging method. Hewett and Kijsanayothin [76] obtain the accuracy of 91.18% (10-CV) with 6 genes using SVM on prostate cancer data set. Statnikov et al. [54] obtain 92% (10-CV) accuracy without any gene selection. Dettling and Buhlmann [85] report 95.10% (LOOCV) with 3 gene clusters (clusters consist of minimum 1 gene and maximum 17 genes) using nearest neighbor method. Similarly, Fort and Lambert-Lacroix [86] get 95.1% (LOOCV) with 1000 genes using Ridge PLS method. Xiong and Chen [87] choose N_f most discriminatory genes where N_f takes values between 10 and 2000, repeated the experiment 100 times for

each N_f value and obtained an average value of 94.78% using uncorrelated linear discriminant analysis. Finally, Zhang and Deng [88] reach the accuracy of 96.08% (LOOCV) using SVM with 13 genes.

Table 4.6 Performance evaluation of HBE with 10-CV for prostate cancer data set

Fold	TP	FP	FN	TN	TPR	FPR	ACC
1	5	0	1	4	0.83	0	0.9
2	5	0	1	4	0.83	0	0.9
3	5	0	1	4	0.83	0	0.9
4	4	1	0	5	1	0.16	0.9
5	6	0	0	4	1	0	1
6	4	0	0	6	1	0	1
7	3	1	1	5	0.75	0.16	0.8
8	6	0	0	4	1	0	1
9	7	0	0	3	1	0	1
10	6	0	0	4	1	0	1

The selected genes are serine protease hepsin (X07732), nel-related protein 2 (D83018), ao89h09.x1 (AI207842), Cdk-inhibitor p57KIP2 (U22398), DKFZp564I1663-r1 (AL036744), adipsin/complement factor D (M84526), glutathione transferase 4 (GSTM4) (M96233), DKFZp586K1220 (AL050152), aldose reductase (X15414), ADP/ATP translocase (J03592). In fact many of the genes that we have selected in this study have shown that their expression patterns are changed in the prostate cancer tissues. For example hepsin, a cell surface serine protease, is significantly up-regulated in human prostate cancer and it promotes prostate cancer progression and metastasis [89]. Also, the expression of p57Kip2 is dramatically decreased in human prostate cancer and the overexpression of p57Kip highly suppresses the cell proliferation [90]. Furthermore, another selected gene

glutathione transferase mediates the proliferation of androgen-independent prostate cancer cells [91]. Aldose reductase gene is responsible in carbohydrate metabolism that converts glucose to sorbitol [92]. The genes that were selected in this study have also been reported by others in prostate cancer as markers [76, 93].

4.5.3 Prostate Cancer Outcome

This data contains 8 patients having relapsed and 13 patients having non-relapse and there are 12600 genes in the data set. Table 4.7 is the summary of classification results of prostate cancer outcome data set. Since there are only 21 samples in the data, we have only performed leave-one-out-validation. As it is seen in Table 4.7, hyper-box enclosure method is one of the top classifiers (BayesNet and RBF Network) with the accuracy of 95.24% compared with other methods. There are not many studies that employed this data set in the literature. Tan and Gilbert [41] get 85.71% with ten-fold-cross validation using Bagging method with 208 genes. Comparing to the results of Tan and Gilbert [41], hyper-box enclosure approach gives an the accuracy of 90% with ten-fold cross validation using only three genes.

Table 4.7 Classification results of prostate outcome data set

Classifier	LOOCV
Hyper-box enclosure	**95.24**
Bayes Net	**95.24**
LibSVM	61.90
SMO	57.14
Logistic Regression	47.62
RBF Network	**95.24**
IBk	80.95
J48	85.71
Random Forest	90.48

Three genes were selected to be important as prostate cancer outcome. The genes are human cofactor A protein (AF038952), farnesyl-protein transferase beta-subunit (HUMFPTB), glutamine-fructose-6-phosphate amidotransferase (GFAT) (M90516). The function of heterodimeric enzyme farnesyl: protein transferase (FPTase) 3 is to transfer of a 15-carbon isoprenoid moiety to a C-terminal cysteine of many cellular proteins. The inhibition of farnesyl protein transferase has effect for the prevention of proper functioning of the Ras protein that leads to oncogenesis or cancer. In fact when prostate cancer cell line treated with farnesyl-protein transferase inhibitor, elimination of cancer were increased [94]. Glutamine:fructose-6-phosphate amidotransferase (GFA), the first and rate-limiting enzyme in the hexosamine biosynthetic pathway, transfers the amide group from glutamine to fructose-6-phosphate to form glucosamine-6-phosphate (GlcN-6-P), a precursor of uridine diphosphate-N-acetyl-glucosamine. It has been demonstrated that overexpression of GFA in Rat-1 fibroblasts causes insulin resistance [95].

4.5.4 DLBCL

The diffuse large B-cell lymphoma (DLBCL) data set contains 58 samples from DLBCL patients and 19 samples from follicular lymphoma. The gene expression profiles were analyzed using Affymetrix human 6800 oligonucleotide arrays. In the DLBCL data set, the hyper-box enclosure method is again the most accurate classifier among other classifiers with leave-one-out validation (Table 4.8). However, in ten-fold-cross validation, hyper-box enclosure method gives worse results than RBF network and logistic regression with an accuracy of 91.25%. GESSES method predicts all samples correctly with different random numbers of starting top genes (from 77 to 130) [53]. The final predictors ranged in number of genes, from four to twelve. Statnikov et al. [54] reaches the accuracy of 97.50 (10-CV) with many methods without gene selection. Zhang and Deng [88] report a classification accuracy of 92.71% (LOOCV) using kNN (k=5) with 8 genes.

Table 4.8 Classification results of DLBCL data set

Classifier	10-CV	**LOOCV**
Hyper-box enclosure	91.25	**96.10**
Bayes Net	89.61	89.61
LibSVM	84.42	84.42
SMO	90. 91	89.61
Logistic Regression	92.21	89.61
RBF Network	**94.81**	93.51
IBk	89.61	88.31
J48	90.91	89.61
Random Forest	89.61	89.61

We have reached the maximum classification accuracy with 6 genes which are DNA replication licensing factor CDC47 homolog (D55716_at), gamma-interferon-inducible protein IP-30 precursor (J03909_at), LDHA lactate dehydrogenase A (X02152_at), CD69 antigen (Z30426_at), SLC (AB002409_at), Rad2 (HG4074-HT4344_at). When the function of proteins and their relation with DLBCL were searched, proteins plays significant role in the progression of DLBCL. DNA replication licensing factor CDC47 is a factor that helps the DNA to undergo a single round of replication per cell cycle. CDC47 is not only necessary for DNA replication and cell proliferation, but also for S-phase checkpoint activation upon UV-induced damage [96]. The function of LDHA lactate dehydrogenase is to catalyze the conversion of L-lactate and NAD to pyruvate and NADH in the last step of anaerobic glycolysis. It is proved that mutations in lactate dehydrogenase A gene causes the exertional myoglobinuria [97]. CD69, known as early T cell activation antigen, is expressed on a variety of immune cells, including T- and B- lymphocytes, NK cells, monocytes/macrophages, and granulocytes [98]. Secondary lymphoid-tissue chemokine (SLC) is a cytokine belonging to the CC chemokine family and it is

constitutively expressed in a variety of lymphoid tissues. This gene is a potent and specific chemoattractant for lymphocytes [99]. Rad 2 (Flap endonuclease 1) is a member of the XPG/RAD2 endonuclease family and is involved in DNA repair [100].

Table 4.9 Performance evaluation of HBE with 10-CV for DLBCL data set

Fold	TP	FP	FN	TN	TPR	FPR	ACC
1	6	1	0	1	0	0.5	0.88
2	6	1	0	1	1	0	0.88
3	5	0	1	2	0.83	0	0.88
4	6	0	0	2	1	0	1
5	5	0	1	2	0	0	0.88
6	6	1	0	1	1	0.5	0.88
7	7	0	0	1	0	0	1
8	7	0	0	1	0	0	1
9	6	0	0	2	0	0	1
10	5	0	0	3	1	0	1

4.5.5 Lymphoma

There are 47 samples, 24 of them are referred to as germinal center B-like group and 23 are activated B-like group. This gene expression data contains 4026 genes. In the lymphoma data set, hyper-box enclosure method is the most accurate classifier using both ten-fold-cross-validation and leave-one-out cross validation with the accuracy of 96.29% and 97.87%, respectively (Table 4.10). Also, prediction accuracies of HBE at each fold is given in Table 4.11. Support vector machine algorithm SMO is the second classifier with the accuracy of 95.75% in both validation methods. Hewett and Kijsanayothin [76] have obtained a classification accuracy of 97.87% (10-CV) with SVM and Bayesian network

methods. Dettling and Buhlmann [85] reach the accuracy of 100% (LOOCV) with 10 gene clusters (min: 1 and max: 16 of genes) using nearest neighbor method.

Table 4.10 Classification results of lymphoma data set

Classifier	10-CV	**LOOCV**
Hyper-box enclosure	**96.29**	**97.87**
Bayes Net	95.75	93.62
LibSVM	93.62	93.62
SMO	95.75	95.75
Logistic Regression	95.75	91.49
RBF Network	95.75	95.75
IBk	93.62	95.75
J48	82.98	87.23
Random Forest	89.36	89.36

As the best result in the literature including this presented paper, Zhang and Deng [88] again report 100% accuracy (LOOCV) using SVM with 3 genes. However, these three genes are not reported in their study.

Our algorithm gives the following genes: Deoxycytidylate deaminase (19408), lymphoid-restricted membrane protein (JAW1) (16886), PKU-beta=KIAA0137=protein kinase (20423), T-cell protein-tyrosine phosphatase (17140), TTG-2 Rhombotin-2 (19238), stress-activated protein kinase (JNK3) (19384), and unknown labeled genes with ids 19288, 19274, 13394. The following gene expressions are currently being used as markers for lymphoma in clinical diagnosis. Deoxycytidylate deaminase (dCMPase) hydrolyzes dCMP into dUMP, and it is suggested that this gene is a marker of the aggression of human lymphoid malignancies [101]. Jaw1, also known as lymphoid-restricted membrane protein (LRMP), is an endoplasmic reticulum-associated protein. It is known that the expression of

Jaw1/LRMP mRNA is high in germinal centre B-cells and in diffuse large B-cell lymphomas of 'germinal centre' subtype [102].

Table 4.11 Performance evaluation of HBE with 10-CV for lymphoma data set

Fold	TP	FP	FN	TN	TPR	FPR	ACC
1	2	0	0	3	0	0	1
2	2	0	0	3	1	0	1
3	4	0	0	1	1	0	1
4	1	0	0	4	1	0	1
5	2	0	0	3	0	0	1
6	4	0	0	1	1	0	1
7	3	0	0	2	0	0	1
8	1	0	0	4	0	0	1
9	5	0	0	0	0	0	1
10	3	0	0	2	1	0	1

In addition, following genes were selected and their expression pattern was reported to be greatly changed in lymphoma. Among these genes, PKU-beta, a serine/threonine protein kinase, has role in chromatin remodeling, DNA replication and mitosis [103]. T-cell protein tyrosine phosphatases, phospho tyrosine-specific protein phosphatase, nuclear dephosphorylation of phospho-STAT6 (pSTAT6) was observed in activated-B-cell (ABC)-like tumors. Their expression profile was quite different [104]. Moreover, TTG-2 Rhombotin-2 is a cysteine rich protein with LIM motif and immunohistologic analysis show that LMO2 protein is expressed as a nuclear marker in normal germinal-center (GC) B cells and GC-derived B-cell lines and in a subset of GC-derived B-cell lymphomas [105]. Finally, stress-activated protein Jun N-terminal kinase (JNK3) is a member of mitogen-activated protein kinase (MAPK) superfamily and it plays an important role in signaling

pathways of critical physiological processes, including apoptosis, differentiation and proliferation. It is known that the activation of JNK leads to the interferon-alpha-induced apoptosis in B-cell lymphoma [106].

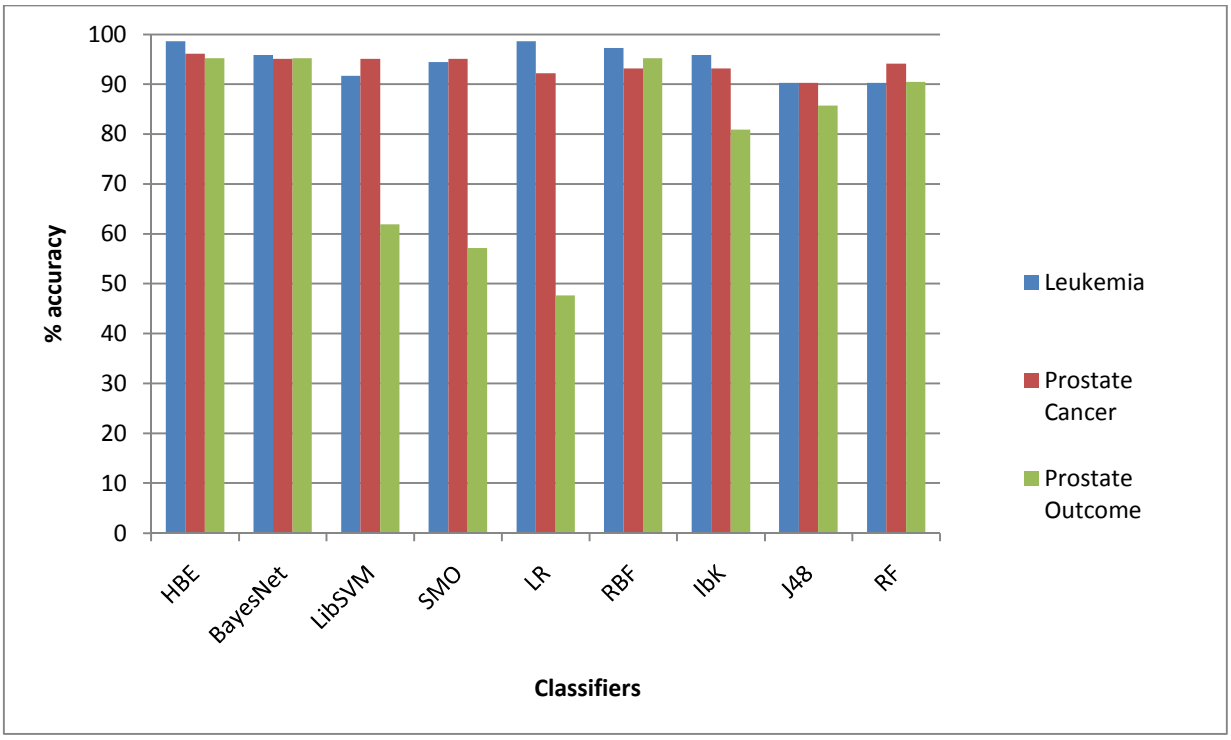

Figure 4.1 The comparison of results among all classifiers for leukemia, prostate cancer, and prostate cancer outcome data sets.

4.5.6 Small round blue cell tumors

Small round blue cell tumors (SRBCT) of childhood are diagnosed using single layer neural network [52]. In this work, they reduced the gene number to 96 to predict the classes of the test data perfectly. There are four different SRBCT in the data set: Ewing family tumor (EWS), Burkitt lymphoma (BL), neuroblastoma (NB) and rhabdomysarcoma (RMS). The training set contains 63 samples and the test set contains 20 samples. The cDNA microarrays comprise 2308 genes.

Table 4.12 Classification results of SRBCT data set

Classifier	Test Set	10-CV	LOOCV
Hyper-box enclosure	**100**	**98.33**	**96.39**
Bayes Net	85	91.57	95.18
LibSVM	90	84.34	84.34
SMO	95	92.77	93.98
Logistic Regression	80	92.77	91.57
RBF Network	90	91.57	93.98
IBk	90	91.57	92.77
J48	90	84.34	91.57
Random Forest	95	86.75	92.77

Table 4.12 reports that hyper-box enclosure method overperforms other classifiers using all validation methods. It gives perfect classification on test set with 5 genes. Moreover it has 98.33% using ten-fold-cross validation and 96.39% with leave-one-out-cross validation. Comparing to other studies in the literature, Dettling and Buhlmann [21] has obtained 100% (LOOCV) with 1 gene cluster (minimum: 1 gene maximum: 14 genes) using nearest neighbor method. Deutsch [8] predicts all test samples when 100 predictors were used, where the average number of genes in a predictor was 12.7. Statnikov et al. [54] obtain 100 accuracy using ten-fold-cross validation with many methods without gene

selection. Finally, Shen et al. [100] perfectly classifies all samples in test using 10 genes with SVM and kernel Fisher discriminant analysis. Considering these studies, hyper-box enclosure method is the most robust method, since it has highest accuracy with the least number of genes on not only test set but also using other types of validations including ten-fold and leave-one-out cross validation.

Table 4.13 Selected SRBCT genes overlapping with genes selected by other groups

Gene	Reference
FCGRT	[56, 70-71, 77-78]
Transmembrane protein	[78]
Fibroblast growth factor receptor	[71]
ESTs	[70, 77-79]
Recoverin	[78]

The selected genes with their gene ids in SRBCT classification are Fc fragment of IgG, receptor, transporter, alpha (FCGRT) (70394), transmembrane protein (812105), fibroblast growth factor receptor 4 (784224), ESTs (295985), recoverin (383188). FCGRT encodes a receptor binding the Fc region of monomeric immunoglobulin G. This protein both helps to transfer of immunoglobulin G antibodies from mother to fetus across the placenta, and binds to immunoglobulin G to prevent the antibody degradation [107]. Growth factor receptors (FGFRs) bind fibroblast growth factors which play key roles in proliferation and differentiation of different type of cells and tissues [108]. Recoverin is neuronal calcium-binding protein that plays a role in the inhibition of rhodopsinosopsin kinase which is a regulator in the phosphorylation of rhodopsin [109]. Zhoua et al. [78] also select FCGRT, transmembrane protein, ESTs, recoverin in their significant gene pool (Table 4.13). Chen et al. [71] selects FCGRT and fibroblast growth factor receptor.

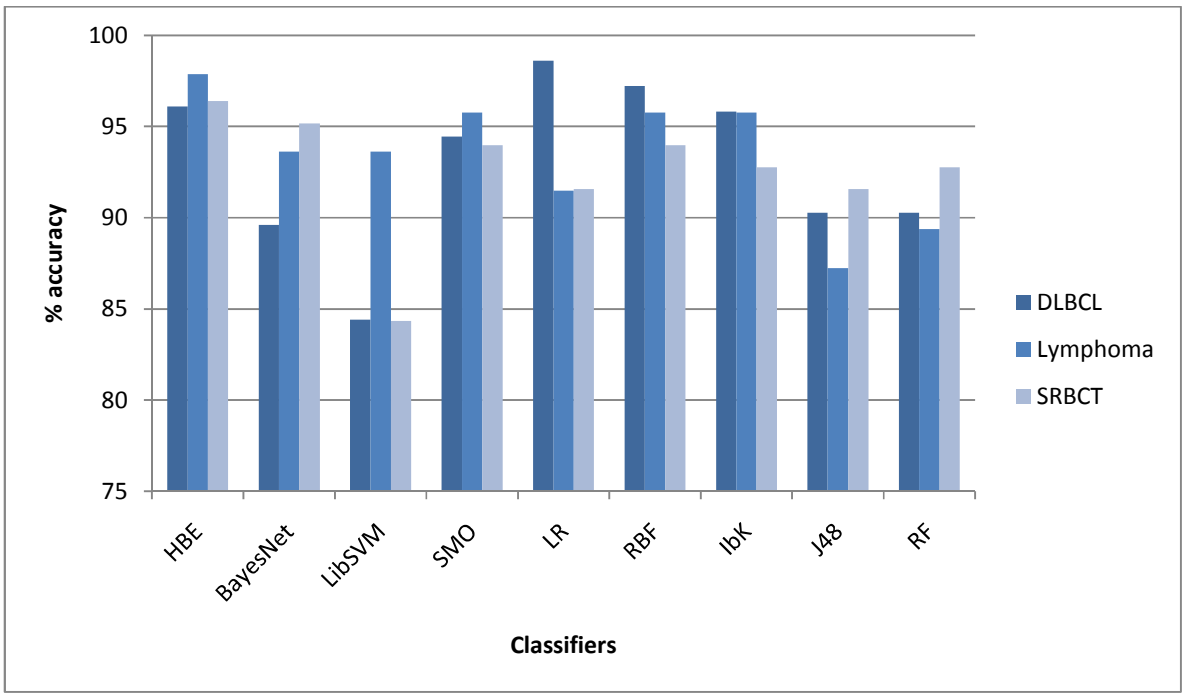

Figure 4.2 The comparison of results among all classifiers for DLBCL, lymphoma, and SRBCT data sets.

Chapter 5

CONCLUSIONS

Machine learning, especially classification is one the main tools in computational biology and bioinformatics that employ rapidly increasing biological data for the analysis and prediction purposes. A large number of data classification methods have been developed, however the majority of these algorithms, especially efficient ones require optimizations to obtain more accurate results depending on data type. Therefore, in this thesis, a recently developed optimization based classification algorithm hyper box enclosure (HBE) method is investigated.

HBE starts with the identification of problematic samples between classes. Then pure integer programming based algorithm, seed finding is used to find representative samples from each class. The objective in seed finding part is that the seeds for each class must be chosen to ensure that seeds are separated well from each other as well as being a good example of the group of instances in the same class. Next, hyper-boxes are constructed using these seeds by following the aim of enclosing all problematic instances. Depending on the complexity of data, more than one hyper-box can be built. If there are intersections among the boxes, these are eliminated using intersection elimination algorithm. Finally test samples are assigned to predefined classes built in the training part and the performance of the classification is evaluated.

One of the most important features of HBE is that this mixed-integer programming based classification method allows the use of hyper-boxes for defining boundaries of the classes that enclose all or some of the points in that set. HBE can be used for both binary and multi-class cases without any modifications. Hence, the same mathematical model can

be employed for data classification problems without any change. HBE does not require parameter optimization during the training of the model.

The main objective of this thesis is to apply HBE method for tumor classification and gene selection. The contributions of this application are two-fold. The first contribution is that we implement an effective optimization based classifier that gives very high performance and valuable insight into different type of cancer data sets. HBE approach does not require parameters to optimize in order to obtain high classification accuracies. This method can be used for any type of data without any modification or addition. The second contribution is finding of optimal predictor genes that give the highest accuracy in classification. This effort can provide to develop antibody assays for the diagnosis of specific types of cancer and to provide accurate diagnostics by only measuring expression of few genes. We have applied our algorithm on publicly available data sets including leukemia data set, two prostate cancer data sets, two lymphoma data sets and SRBCT data set. In conclusion, mixed-integer programming based hyper-box enclosure approach is robust and effective method for microarray analysis.

As a recommendation and future work, HBE method should be developed to reduce the computational complexity which causes a time problem in large data sets. Sophisticated preprocessing algorithms such as clustering can be developed before the building model by hyper-boxes. Additionally, HBE algorithm was successfully employed for protein folding problem and drug classification before. Therefore, we are planning to use HBE method in some other important applications including virtual screening, prediction of protein-protein interactions.

BIBLIOGRAPHY

1. Fix E, Hodges JJ: Discriminatory analysis: non-parametric discrimination: Consistency properties. *USAF School of Aviation Medicine,Randolph Field, TX* 1951, Report No. 4.

2. Cover TM, Hart PE: Nearest Neighbor Pattern Classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory* 1967, 13:21-27.

3. McCulloch WS, Pitts W: A logical calculus of ideas imminet in nervous activity. *Bulletin of Mathematical Biophysics* 1943, 5:115-133.

4. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*. Second edn: Springer; 2008.

5. Rosenblatt F: *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.*: Spartan Books; 1962.

6. Cybenko G: Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems* 1989, 2:303–314.

7. Rumelhart DE, Hinton GE, Williams RJ: Learning internal representations by backpropagation errors. *Nature* 1986, 323:533-536.

8. White H: Learning in artificial neural networks: a statistical perspective. *Neural Computing* 1989, 1:425-464.

9. Weiss SM, Kulikowski CA: *Computer systems that learn: classification and prediction methods from statistics, neural networks, machine learning and expert systems.* San Mateo, CA; 1991.

10. Breiman L, Friedman JH, Olsen RA, Stone CJ: *Classification and regression trees.* New York: Wadsworth and Brooks; 1984.

11. Roiger RJ, Geatz MW: *Data mining-a tutorial based primer.* Addison Wesley Press; 2003.

12. Breiman L: Random forests. *Mach Learn* 2001, 45:5-32.

13. Papagelis A, Kalles D: Breeding Decision Trees Using Evolutionary Techniques. In *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001: 393-400.

14. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles V: Machine learning in bioinformatics. *Brief Bioinform* 2006, 7:86-112.

15. Kleinbaum DG, Kupper LL, Chambless LE: Logistic-Regression Analysis of Epidemiologic Data - Theory and Practice. *Commun Stat a-Theor* 1982, 11:485-547.

16. Vapnik V: The Nature of Statistical Learning. *Sprinter-Verlag* 1995.

17. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank. *Nucleic Acids Research* 2008, 36.

18. GenBank. [http://www.ncbi.nlm.nih.gov/genbank/]

19. Mathe C, Sagot MF, Schiex T, Rouze P: Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research* 2002, 30:4103-4117.

20. Salzberg S: Localing protein coding regions in human
DNA using a decision tree algorithm. *Journal of Computational Biology* 1995, 2:473-485.

21. Castelo R, Guigo R: Splice site identification by idlBNs. *Bioinformatics* 2004, 20:i69-76.

22. Degroeve S, De Baets B, Van de Peer Y, Rouze P: Feature subset selection for splice site prediction. *Bioinformatics* 2002, 18 Suppl 2:S75-83.

23. Saeys Y, Degroeve S, Aeyels D, Rouze P, Van de Peer Y: Feature selection for splice site prediction: a new method using EDA-based feature ranking. *Bmc Bioinformatics* 2004, 5:64.

24. Carter RJ, Dubchak I, Holbrook SR: A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res* 2001, 29:3928-3938.

25. Lopez-Bigas N, Ouzounis CA: Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* 2004, 32:3108-3114.

26. Bao L, Cui Y: Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* 2005, 21:2185-2190.

27. Phuong TM, Lee D, Lee KH: Regression trees for regulatory element identification. *Bioinformatics* 2004, 20:750-757.

28. Selbig J, Mevissen T, Lengauer T: Decision tree-based formation of consensus protein secondary structure prediction. *Bioinformatics* 1999, 15:1039-1046.

29. Cai CZ, Han LY, Ji ZL, Chen YZ: Enzyme family classification by support vector machines. *Proteins* 2004, 55:66-76.

30. Rost B, Sander C: Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993, 232:584-599.

31. Kloczkowski A, Ting KL, Jernigan RL, Garnier J: Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins* 2002, 49:154-166.

32. Yang C, Dobbs D, Honavar V: A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics* 2004, 20:i371–378.

33. Kao KC, Huang JY: Accurate and fast computational method for identifying protein function using protein-protein interaction data. *Mol Biosyst* 2010, 6:830-839.

34. Qi Y, Bar-Joseph Z, Klein-Seetharaman J: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* 2006, 63:490-500.

35. Mohamed TP, Carbonell JG, Ganapathiraju MK: Active learning for human protein-protein interaction prediction. *Bmc Bioinformatics* 2010, 11 Suppl 1:S57.

36. Huang Y, Li Y: Prediction of protein subcellular locations using fuzzy k-NN mathos. *Bioinformatics* 2004, 20:21-28.

37. Dale JM, Popescu L, Karp PD: Machine learning methods for metabolic pathway prediction. *Bmc Bioinformatics* 2010, 11:15.

38. You ZH, Yin Z, Han K, Huang DS, Zhou X: A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *Bmc Bioinformatics* 2010, 11:343.

39. Plaimas K, Mallm JP, Oswald M, Svara F, Sourjik V, Eils R, Konig R: Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC Syst Biol* 2008, 2:67.

40. Craig RA, Liao L: Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. *Bmc Bioinformatics* 2007, 8:6.

41. Tan AC GD: Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics* 2003, 2:75-83.

42. Kubinyi H: QSAR and 3D QSAR in drug design .1. methodology. *Drug Discov Today* 1997, 2:457-467.

43. Yap CW, Chen YZ: Prediction of cytochrome p450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J Chem Inf Model* 2005, 45:982-992.

44. Yao XJ, Liu HX, Zhang RS, Liu MC, Hu ZD, Panaye A, Doucet JP, Fan BT: QSAR and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines. *Mol Pharmaceut* 2005, 2:348-356.

45. Neugebauer A, Hartmann RW, Klein CD: Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods. *J Med Chem* 2007, 50:4665-4668.

46. Slonim DK: From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* 2002, 32:502-508.

47. Schwarz G: Estimating the dimension of a model. *Annals of Statistics* 1976, 6:461-464.

48. Kohavi R, John GH: Wrappers for feature subset selection. *Artif Intell* 1997, 97:273-324.

49. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KFX, Mewes HW: Gene selection from microarray data for cancer classification - a machine learning approach. *Comput Biol Chem* 2005, 29:37-46.

50. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 1999, 286:531-537.

51. Hastie T TR, Eisen MB, Alizadeh A, Levy R, Staudt L, Chen WC, Botstein D, Brown P: "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* 2000, 1:1-21.

52. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001, 7:673-679.

53. Deutsch JM: Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics* 2003, 19:45-52.

54. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 2005, 21:631-643.

55. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, et al: Multiclass cancer diagnosis using tumor gene expression signatures. *P Natl Acad Sci USA* 2001, 98:15149-15154.

56. Antonov AV, Tetko IV, Mader MT, Budczies J, Mewes HW: Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics* 2004, 20:644-U145.

57. Diaz-Uriarte R, de Andres SA: Gene selection and classification of microarray data using random forest. *Bmc Bioinformatics* 2006, 7:-.

58. Yuksektepe FU, Yilmaz O, Turkay M: Prediction of secondary structures of proteins using a two-stage method. *Comput Chem Eng* 2008, 32:78-88.

59. Dagliyan O, Kavakli IH, Turkay M: Classification of Cytochrome P450 Inhibitors with Respect to Binding Free Energy and pIC(50) Using Common Molecular Descriptors. *J Chem Inf Model* 2009, 49:2403-2411.

60. Armutlu P, Ozdemir ME, Uney-Yuksektepe F, Kavakli IH, Turkay M: Classification of drug molecules considering their IC50 values using mixed-integer linear programming based hyper-boxes method. *Bmc Bioinformatics* 2008, 9:-.

61. Kahraman P, Turkay M: Classification of 1,4-dihydropyridine calcium channel antagonists using the hyperbox approach. *Ind Eng Chem Res* 2007, 46:4921-4929.

62. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, et al: Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002, 1:203-209.

63. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JG, Sabet H, Tran T, Yu X, et al: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000, 403:503-511.

64. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, et al: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 2002, 8:68-74.

65. Hall M FE, Holmes G, Pfahringer B, Reutemann P, Witten IH: The WEKA Data Mining Software: An Update; SIGKDD Explorations. *SIGKDD Explorations* 2009, 11.

66. Kononenko I: Estimating attributes: analysis and extensions of RELIEF. *Conference on Machine Learning* 1994:171-182.

67. Hall MA, Smith LA: Feature subset selection: A correlation based filter approach. *Progress in Connectionist-Based Information Systems, Vols 1 and 2* 1998:855-858 1372.

68. Dettling AC, Feldon J, Pryce CR: Early deprivation and behavioral and physiological responses to social separation/novelty in the marmoset. *Pharmacol Biochem Be* 2002, 73:259-269.

69. Nguyen DV, Rocke DM: Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 2002, 18:39-50.

70. Lee Y, Lee CK: Classification of multiple cancer types by tip multicategory support vector machines using gene expression data. *Bioinformatics* 2003, 19:1132-1139.

71. Chen PC, Huang SY, Chen WJ, Hsiao CK: A new regularized least squares support vector regression for gene selection. *Bmc Bioinformatics* 2009, 10:-.

72. Brennan M, Penn MS, Van Lente F, Nambi V, Shishehbor MH, Aviles RJ, Goormastic M, Pepoy ML, McErlean ES, Topol EJ, et al: Prognostic value of myeloperoxidase in patients with chest pain. *New Engl J Med* 2003, 349:1595-1604.

73. Freeman SD, Kelm S, Barber EK, Crocker PR: Characterization of Cd33 as a New Member of the Sialoadhesin Family of Cellular Interaction Molecules. *Blood* 1995, 85:2005-2012.

74. Estey EH, Giles FJ, Beran M, O'Brien S, Pierce SA, Faderl SH, Cortes JE, Kantarjian HM: Experience with gemtuzumab ozogamycin ("mylotarg"), and all-trans retinoic acid in untreated acute promyelocytic leukemia. *Blood* 2002, 99:4222-4224.

75. Yang AJ, Song XY: Bayesian variable selection for disease classification using gene expression data. *Bioinformatics* 2010, 26:215-222.

76. Hewett R, Kijsanayothin P: Tumor classification ranking from microarray data. *Bmc Genomics* 2008, 9:-.

77. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: Tissue classification with gene expression profiles. *Journal of Computational Biology* 2000, 7:559-583.

78. Zhoua X LK, Wong STC: Cancer classification and prediction using logistic regression with Bayesian gene selection. *J Biomed Inform* 2004, 37:249-259.

79. Bicciato S, Pandin M, Didone G, Di Bello C: Pattern identification and classification in gene expression data using an autoassociative neural network model. *Biotechnol Bioeng* 2003, 81:594-606.

80. Voronova A, Baltimore D: Mutations That Disrupt DNA-Binding and Dimer Formation in the E47 Helix-Loop-Helix Protein Map to Distinct Domains. *P Natl Acad Sci USA* 1990, 87:4722-4726.

81. Okuya M, Kurosawa H, Kikuchi J, Furukawa Y, Matsui H, Aki D, Matsunaga T, Inukai T, Goto H, Altura RA, et al: Up-regulation of Survivin by the E2A-HLF Chimera Is Indispensable for the Survival of t(17;19)-positive Leukemia Cells. *J Biol Chem* 2010, 285:1850-1860.

82. Brambillasca F, Mosna C, Colombo M, Rivolta A, Caslini C, Minuzzo M, Giudici G, Mizzi L, Biondi A, Privitera E: Identification of a novel molecular partner of the E2A gene in childhood leukemia. *Leukemia* 1999, 13:369-375.

83. Wong ETL, Jenne DE, Zimmer M, Porter SD, Gilks CB: Changes in chromatin organization at the neutrophil elastase locus associated with myeloid cell differentiation. *Blood* 1999, 94:3730-3736.

84. Sakhinia E, Farahangpour M, Tholouli E, Yin JAL, Hoyland JA, Byers RJ: Comparison of gene-expression profiles in parallel bone marrow and peripheral blood samples in acute myeloid leukaemia by real-time polymerase chain reaction. *J Clin Pathol* 2006, 59:1059-1065.

85. Dettling M BP: Supervised clustering of genes. *Genome Biology* 2002, 3:research0069.0061{0069.0015.

86. Fort G, Lambert-Lacroix S: Classification using partial least squares with penalized logistic regression. *Bioinformatics* 2005, 21:1104-1111.

87. Xiong HL, Chen XW: Kernel-based distance metric learning for microarray data classification. *Bmc Bioinformatics* 2006, 7:-.

88. Zhang JG, Deng HW: Gene selection for classification of microarray data based on the Bayes error. *Bmc Bioinformatics* 2007, 8:-.

89. Klezovitch O, Chevillet J, Mirosevich J, Roberts RL, Matusik RJ, Vasioukhin V: Hepsin promotes prostate cancer progression and metastasis. *Cancer Cell* 2004, 6:185-195.

90. Jin RJ, Lho Y, Wang YQ, Ao MF, Revelo MP, Hayward SW, Wills ML, Logan SK, Zhang P, Matusik RJ: Down-regulation of p57(Kip2) induces prostate cancer in the mouse. *Cancer Res* 2008, 68:3601-3608.

91. Hokaiwado N, Takeshita F, Naiki-Ito A, Asamoto M, Ochiya T, Shirai T: Glutathione S-transferase Pi mediates proliferation of androgen-independent prostate cancer cells. *Carcinogenesis* 2008, 29:1134-1138.

92. Petrash JM: All in the family: aldose reductase and closely related aldo-keto reductases. *Cell Mol Life Sci* 2004, 61:737-749.

93. Chu W, Ghahramani Z, Falciani F, Wild DL: Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics* 2005, 21:3385-3393.

94. Sepp-Lorenzino L, Tjaden G, Moasser MM, Timaul N, Ma Z, Kohl NE, Gibbs JB, Oliff A, Rosen N, Scher HI: Farnesyl : protein transferase inhibitors as potential agents for the management of human prostate cancer. *Prostate Cancer P D* 2001, 4:33-43.

95. Hebert LF, Daniels MC, Zhou JX, Crook ED, Turner RL, Simmons ST, Neidigh JL, Zhu JS, Baron AD, McClain DA: Overexpression of glutamine:fructose-6-phosphate amidotransferase in transgenic mice leads to insulin resistance. *J Clin Invest* 1996, 98:930-936.

96. Tsao CC, Geisen C, Abraham RT: Interaction between human MCM7 and Rad17 proteins is required for replication checkpoint signaling. *Embo J* 2004, 23:4660-4669.

97. Chung FZ, Tsujibo H, Bhattacharyya U, Sharief FS, Li SSL: Genomic Organization of Human Lactate Dehydrogenase-a Gene. *Biochem J* 1985, 231:537-541.

98. Ziegler SF, Ramsdell F, Alderson MR: The Activation Antigen Cd69. *Stem Cells* 1994, 12:456-465.

99. Yoshida R, Nagira M, Kitaura M, Imagawa N, Imai T, Yoshie O: Secondary lymphoid-tissue chemokine is a functional ligand for the CC chemokine receptor CCR7. *J Biol Chem* 1998, 273:7118-7122.

100. Shen BH, Nolan JP, Sklar LA, Park MS: Essential amino acids for substrate binding and catalysis of human flap endonuclease 1. *J Biol Chem* 1996, 271:9173-9176.

101. Ellims PH, Medley G: Deoxycytidylate Deaminase Activity in Lymphoproliferative Disorders. *Leukemia Res* 1984, 8:123-128.

102. Tedoldi CL, Bandinelli E, Barreto SSM, Manfroi WC: Heart failure as a strong risk factor for venous thromboembolism in pregnancy. *Eur Heart J* 2006, 27:296-296.

103. Hashimoto M MT, Iwabuchia K, Datea T: JPKU-beta/TLK1 regulates myosin II activities, and is required for accurate equaled chromosome segregation. *Mutation Research* 2008, 657:63-67.

104. Lu XQ, Chen J, Sasmono RT, Hsi ED, Sarosiek KA, Tiganis T, Lossos IS: T-cell protein tyrosine phosphatase, distinctively expressed in activated-B-cell-like diffuse large B-cell lymphomas, is the nuclear phosphatase of STAT6. *Mol Cell Biol* 2007, 27:2166-2179.

105. Natkunam Y, Zhao SC, Mason DY, Chen J, Taidi B, Jones M, Hammer AS, Dutoit SH, Lossos IS, Levy R: The oncoprotein LMO2 is expressed in normal germinal-center B cells and in human B-cell lymphomas. *Blood* 2007, 109:1636-1642.

106. Ying J, Li H, Cui Y, Wong AHY, Langford C, Tao Q: Epigenetic disruption of two proapoptotic genes MAPK10/JNK3 and PTPN13/FAP-1 in multiple lymphomas and carcinomas through hypermethylation of a common bidirectional promoter. *Leukemia* 2006, 20:1173-1175.

107. Ghetie V, Ward ES: Multiple roles for the major histocompatibility complex class I-related receptor FcRn. *Annu Rev Immunol* 2000, 18:739-766.

108. Ornitz DM IN: Fibroblast growth factors. *Genome Biology* 2001, 2:Reviews 3005.

109. Murakami A, Akaki Y, Ara F, Duval E, Inana G: Isolation of New Candidate Genes for Human Retinal Diseases - Differential Cloning Approach. *Invest Ophth Vis Sci* 1992,33:791-791.

VITA

Onur Dağlıyan was born in Istanbul, Turkey, on December 3, 1984. He is a graduate of Kadıköy Anadolu Lisesi and he received his B.Sc. Degree in Chemical and Biological Engineering from Koç University, Istanbul, Turkey in 2008. He was both research and teaching assistant at Koç University from September 2008 to August 2010.

He is a member of both *Molecular Biology & Biochemistry Laboratory* of Dr. Halil Kavaklı and *SystemsLab* of Dr. Metin Türkay. His research interests include computational biology, bioinformatics, machine learning, drug discovery and design.

He will continue his education as a Ph.D. candidate in Biological and Biomedical Sciences Program, School of Medicine, The University of North Carolina at Chapel Hill, NC, USA.