REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

META-ANALYSIS OF MICRORNA AND GENE SELECTION USING MACHINE LEARNING

ELNAZ PASHAEI

Ph.D. THESIS
DEPARTMENT OF COMPUTER ENGINEERING
PROGRAM OF COMPUTER ENGINEERING

ADVISER
PROF. DR. NİZAMETTİN AYDIN

İSTANBUL, 2017

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

META-ANALYSIS OF MICRORNA AND GENE SELECTION USING MACHINE LEARNING

A thesis submitted by Elnaz PASHAEI in partial fulfillment of the requirements for the degree of **DOCTER OF SCIENCE** is approved by the committee on 08.06.2017 in Department of Computer Engineering. Program of Computer Engineering.

Thesis Adviser

Prof. Dr. Nizamettin AYDIN

Yıldız Technical University

Approved By the Examining Committee

Prof. Dr. Nizamettin AYDIN

Yıldız Technical University _____

Prof. Dr. Banu DİRİ, Member

Yıldız Technical University _____

Assist. Prof. Dr. Arzucan ÖZGÜR, Member

Bogazici University _____

Assoc. Prof. Dr. Songul ALBAYRAK, Member

Yıldız Technical University _____

Prof. Dr. Zümray Dokur ÖLMEZ, Member

Istanbul Technical University _____

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere appreciation and thanks to my advisor Prof. Nizamettin Aydin for the continuous support of my Ph.D. study and related research, for his patience, motivation, encouragement, and consultation. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Words cannot express the feelings I have for my parents and my in-laws for their constant unconditional support - both emotionally and financially. My hard-working parents have sacrificed their lives for us and provided unconditional love and care. I love them so much, and I would not have made it this far without them. Thank you.

June, 2017

Elnaz PASHAEI

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AUC	Area under ROC Curve
BBHA	Binary Black Hole Algorithm
BPSO	Binary Particle Swarm Optimization
CV	Cross Validation
DT	Decision Tree
FS	Feature Selection
GA	Genetic Algorithm
GEO	Gene Expression Omnibus
NCBI	National Centers for Biotechnology Information
PCa	Prostate Cancer
RF	Random Forest
SA	Simulated Annealing

LIST OF FIGURES

Page

LIST OF TABLES

ABSTRACT

META-ANALYSIS OF MICRORNA AND GENE SELECTION USING MACHINE LEARNING

Elnaz PASHAEI

Department of Computer Engineering

Ph.D. Thesis

Adviser: Prof. Dr. Nizamettin AYDIN

The DNA microarray technology allows for monitoring and measuring the expression level of a great number of genes in tissue samples simultaneously. In microarray datasets, the number of samples is much smaller than the number of genes. The classification of such data resulting in the known problem of "curse of dimensionality" and data overfitting. For a successful disease diagnosis, it is necessary to select a small number of discriminating genes that are relevant for classification. Gene selection in microarray data analysis not only increases the classification accuracy but also decreases the processing time in the clinical setting. Therefore, it is quite important to determine a minimum subset of genes to develop a successful disease diagnostic system. In this thesis, two approaches for selecting highly discriminating genes in cancer classification based on a hybrid of nature-inspired optimization algorithms and different classifiers are proposed. In the first proposed approach, Black Hole Algorithm is, for the first time, being used to solve a feature selection (FS) problem. By applying the hyperbolic tangent function, a new binary version of BHA called BBHA is utilized to solve FS in the text, image, and biomedical data. Two classifiers (RF and NB) serve as the evaluators of our proposed algorithm. Experimental results show that BBHA wrapper-based feature selection method is superior to BPSO, GA, SA, and CFS in terms of all criteria. BBHA gives a significantly better performance than the BPSO and GA in terms of CPU Time, the number of parameters for configuring the model, and the number of chosen optimized features. Also, BBHA has competitive or better performance than the other methods in the literature.

In the second proposed approach, we improve the performance of Binary Particle Swarm Optimization (BPSO) and help it to avoid being trapped in a local optimum by

applying BBHA as the local optimizer for BPSO. Experimental results and statistical analysis on four clinical datasets demonstrate that the proposed method yields very small subsets of informative genes, while achieving significantly better classification performance than other approaches such as Firefly, ant colony, bat search, genetic algorithm, harmony search, Fast Correlation-Based Filter (FCBF), and Correlation-based Feature Subset Selection (CFS). Moreover, It was also shown that applying BBHA as the local optimizer for BPSO can significantly improve the performance of BPSO and help it to avoid being trapped in a local optimum.

Several studies on miRNA expression datasets have been conducted in prostate cancer recurrence. However, the results have varied among different studies. By integrating the individual studies the statistical power is increased and more reliable conclusions and new biological insights can be drawn. In this thesis, we conducted a meta-analysis on six available miRNA expression datasets for prostate cancer recurrence after radical prostatectomy and identified a potentially significant list of differentially expressed microRNA genes. We did gene ontology enrichment, KEGG analysis, and common pathway analysis to identify the molecular pathways in which the identified microRNA genes participate and reveal new directions for drug treatments of recurrent prostate cancer.

MiR-145, an important tumor suppressor microRNA, has shown to be downregulated in many cancer types and has crucial roles in tumor initiation, progression, metastasis, invasion, recurrence, and chemoradioresistance. In this thesis by meta-analysis of eight GEO datasets, we investigated potential common target genes of miR-145 to help to understand the underlying molecular pathways of tumor pathogenesis in association with those common target genes.

Keywords: Feature Selection, Black Hole Optimization Algorithm, Decision Tree Algorithms, Particle Swarm Optimization, Gene expression, prostate cancer recurrence, meta-analysis, Mir-145.

ÖZET

MAKİNA ÖĞRENMESİ KULLANARAK MICRORNA META-ANALİZİ VE GEN SEÇİMİ

Elnaz PASHAEI

Bilgisayar Mühendisliği Anabilim Dalı

Doktora Tezi

Tez Danışmanı: Prof. Dr. Nizamettin AYDIN

DNA mikrodizi teknolojisi doku örneklerinde çok sayıda genin ifade düzeyini aynı anda izlemeyi ve ölçmeyi mümkün kılar. Mikrodizi veri setlerinde örnek sayısı gen sayısından çok daha azdır. Bu tür verilerin sınıflandırılması bilinen "boyutsallık belası (curse of dimentionality)" ve veri aşırı uyumluluk problemiyle sonuçlanır. Başarılı bir hastalık teşhisi için, sınıflandırma ile alakalı az sayıda ayrımcı gen seçmek gerekir. Mikrodizi veri analizinde gen seçimi sadece sınıflandırma doğruluğunu arttırmakla kalmaz, aynı zamanda klinik ortamda işleme süresini azaltır. Bu nedenle, başarılı bir hastalık teşhis sistemi geliştirmek için genlerin minimum bir alt kümesini belirlemek oldukça önemlidir. Bu tezde, melez doğadan esinlenmiş optimizasyon algoritmaların ve farklı sınıflayıcılara dayanan kanser sınıflandırmasında yüksek derecede ayırıcı gen seçimi için iki yaklaşım önerilmiştir. İlk önerilen yaklaşımda Kara Delik Algoritması, ilk defa bir özellik seçimi (FS) problemini çözmek için kullanılmaktadır. Hiperbolik teğet fonksiyonunu uygulayarak, metin, görüntü ve biyomedikal verilerin FS'sini çözmek için BHA'nın BBHA adlı yeni bir iki tabanlı biçimi kullanılır. İki sınıflayıcı (RF ve NB) önerilen algoritmamızın değerlendiricileri olarak görev yapmaktadır. Deneysel sonuçlar BBHA sarmalayıcı (wrapper) temelli özellik seçim yönteminin tüm kriterler açısından BPSO, GA, SA ve CFS'den üstün olduğunu göstermektedir. BBHA, CPU Zamanı, modeli yapılandırma parametrelerinin sayısı ve seçilen en iyileştirilmiş özelliklerin sayısı açısından BPSO ve GA'ya göre önemli ölçüde daha iyi bir performans sunar. Ayrıca, BBHA, literatürdeki diğer yöntemlere kıyasla rekabetçi veya daha iyi bir performansa sahiptir. Önerilen ikinci yaklaşımda, İkili Parçacık Sürüsü Optimizasyonunun (BPSO) performansını iyileştiriyoruz ve BPSO için yerel iyileştirici olarak BBHA uygulayarak yerel bir optimumda sıkışmayı önlemeye yardımcı oluyoruz. Dört klinik veri kümesindeki deneysel sonuçlar ve istatistiksel analiz, önerilen

yöntemin, ateş böceği, karınca koloni, yarasa arama, genetik algoritma, armoni araştırması, hızlı korelasyon tabanlı süzgeç ve korelasyon tabanlı özellik alt küme seçimi gibi yaklaşımlara göre önemli derecede daha iyi sınıflandırma performansı elde ederken çok küçük bilgi grubu genleri ürettiğini göstermektedir. Dahası, BPSO için yerel iyileştirici olarak BBHA'nın uygulanmasının BPSO'nun performansını belirgin bir şekilde artırabileceği ve yerel optimumda sıkışmayı önlemesine yardımcı olacağı da gösterildi.

Prostat kanseri reküransında çeşitli miRNA ifade veri setleri yapılmıştır. Bununla birlikte, sonuçlar farklı çalışmalar arasında çeşitlilik göstermektedir. Bireysel çalışmaları entegre ederek istatistiksel güç artar ve daha güvenilir sonuçlar ve yeni biyolojik bilgiler elde edilebilir. Bu tezde, radikal prostatektomiden sonra prostat kanseri reküransı için altı mevcut miRNA ifade veri seti üzerinde bir meta-analiz yaptık ve potansiyel olarak farklı olarak eksprese edilen mikroRNA genlerinin önemli bir listesini tespit ettik. Tanımlanmış mikroRNA genlerinin katıldığı moleküler yolakları tanımlamak ve nükseden prostat kanseri üzerinde ilaç tedavileri için yeni yönergeler ortaya çıkarmak için gen ontolojisi zenginleştirmesi, KEGG analizi ve ortak yolakanalizi yaptık.

Önemli bir tümör baskılayıcı mikroRNA olan MiR-145, birçok kanser çeşidinde downregüle edildiğini ve tümörün başlatılması, progresyonu, metastazı, invazyonu, reküransı ve kemoradyolojik direncinde önemli rollere sahip olduğunu göstermiştir. Sekiz GEO veri kümesinin meta-analizi ile bu tezde, ortak hedef genlerle bağlantılı olarak tümör patogenezinin altında yatan moleküler yolaklarının anlaşılmasına yardımcı olmak için, miR-145'in potansiyel ortak hedef genlerini araştırdık.

Anahtar Kelimeler: Özellik Seçimi, Kara Delik Optimizasyonu Algoritması, Karar Ağacı Algoritmaları, Parçacık Swarm Optimizasyonu, Gen ifadesi, prostat kanseri rekürrensi, meta-analiz, Mir-145.

CHAPTER 1

INTRODUCTION

This chapter introduces this thesis. It starts with the literature review, then outlines the research goals, the major contributions (hypothesis) and the organization of the thesis.

1.

1.1 Literature Review

1.1.1 Classification Algorithms

The advance of gene expression microarrays makes it possible to take a genome wide approach for disease prognosis, diagnosis, and prediction of therapeutic responsiveness. Classification or prediction methods can be generally summarized into two branches: supervised learning and unsupervised learning. Here, we will focus on the supervised learning where the class labels are known beforehand. High dimension and small size of microarray samples facilitate new developments in not only gene selection techniques but also classifier design. With supervised learning methods for gene expression data, various classifiers with promising performance have been constructed. Two most commonly used classifiers, which will be used in this thesis, are reviewed in this section. They are Sparse Partial Least Squares Discriminant Analysis (SPLSDA) and Decision Tree (DT) classifiers including Random Forest (RF); C4.5; C5.0; Boosted C5.0; Bagging, Classification and Regression tree (CART).

Sparse Partial Least Squares Discriminant Analysis (SPLSDA)

Partial least squares (PLS) is a widely used regression method in high-dimensional genomic data. SPLSDA is based on PLS for discrimination analysis, but a Lasso penalization has been added to select features. Some significant advantage of SPLSDA, compared to other classifiers is that SPLSDA is statistically very efficient,

computationally very fast with a tunable sparsity parameter, can be applied to various types of data of any dimensionality (especially appropriate for small sample data with many genes), easy to implement, and is able to automatically perform feature selection.

Decision Tree Classifiers (DT)

A greedy algorithm is a basic method for DT algorithms, which are based on top-down recursive divide-and-conquer manner. Greedy strategies are preferred to utilize as they are easy and efficient to implement. At each node of the tree in DT algorithms, all possible splits are evaluated. Each split has own information gain. If an information gain of the split is highest among the others, it should be chosen as a divider of data into binary parts. The algorithm runs until the stop condition is met. Iterative Dichotomiser 3 (ID3) is the first series of algorithms created by Ross Quinlan based on greedy strategies to generate DTs. Finding the optimal size of the final tree in a DT algorithm is known as the horizon effect problem. A common strategy for solving this problem is to grow the tree until each node contains a small number of samples then use pruning to replace irrelevant branches with leaf nodes. Pruning reduce the size of a DT without reducing predictive accuracy. It removes nodes that do not provide additional information.

An ensemble of unpruned DTs using bagging and bootstrap techniques is known as RF, which was introduced by Leo Breiman in 2001 [1]. RF constructs multiple DTs with randomly selected features and samples. The final classification of RF can be obtained by combining the classification results from the individual DTs. No needing for pruning trees, automatically generation of accuracy and variable importance, being robust to overfitting and outliers, high speed even in prediction, are some of the properties of RF [2]. As demonstrated by several bioinformatics studies, RF is well suited for high-dimensional data and have been increasingly applied for gene selection and classification [3]. For instance, in [4] RF was utilized as fitness function of GA-tuned PSO and was applied for prediction of o-glycosylation sites in proteins. The GA-tuned PSO has achieved higher classification accuracy in terms of AUC with comparison to PSO-RF and several tools for predicting the O-glycosylation sites in proteins. RF is used in [5] against nine multi-class microarray data sets as a gene selection method and it has yielded very small sets of genes while preserving predictive accuracy. Also, it has shown high performance compared to Direct Linear Discriminant Analysis (DLDA), K-Nearest Neighbors (KNN), and SVM classifiers.

In [6] a new algorithm based on RF namely Balanced Iterative Random Forest (BIRF) is proposed to select informative genes from four imbalanced microarray data sets. BIRF has ability in handling the class-imbalanced data and has outperformed the predictive performance of SVM-Recursive Feature Elimination (SVM-RFE), Multi-class SVM-RFE, RF and Naive Bayes (NB) classifiers. Bagging was proposed by Breiman in 1996 [7]. Bagging averages the predictions of classification trees over a group of bootstrap samples. It helps to avoid overfitting. Bagging improves stability and accuracy of DT methods by reducing variance. In [8] a new method based on hybrid of gene selection and bagging classifier namely select-bagging is proposed for classification of high-dimensional and balanced datasets in bioinformatics. A C4.5 DT is an improved form of the ID3 algorithm and was introduced in [9]. The performance of C4.5 is high. In main memory algorithms, training data are completely loaded into main memory, and are thus severely limited in the number of examples they can learn from. Classical tree-based models such as ID3, Classification and Regression tree (CART), LDA, and quadratic discriminant analysis (QDA) are some example of main memory algorithms. C4.5 Comparing to the main memory algorithms is quicker [10]. In [11] the C4.5 was used as fitness function of PSO namely PSODT on 5 small medical datasets from UCI Machine Learning Repository. The C4.5 classifier is also adopted as a fitness function of PSO for gene selection in [12] against eleven benchmark gene expression microarrays. By improving the algorithm of C4.5, Ross Quinlan introduced C5.0 [13]. Missing value and numeric attributes can be handled by C5.0. Lower error rate, high speed, less memory, and support for boosting can be mentioned as characteristics of C5.0 [14]. In BoostedC5.0, Ada-boost algorithm can be used to improve the accuracy of C5.0 [15]. In [16, 17] Boosted C5.0 was used as fitness function of PSO and was applied on small medical datasets and some benchmark microarrays to improve the performance of PSODT. Leo Breiman in 1984 introduced CART classifier [18]. It uses Gini index for node impurity, allows only binary outcomes, and prunes tree based on the complex model [19]. In [20] the researchers proposed a new approach based on CART algorithm namely Sequential CART (S-CART) for gene selection on binary microarrays. They proved that the performance of S-CART is better than Stochastic Search Variable Selection (SSVS) and RF classifiers in terms of speed and accuracy. Previous research all indicates that DTs which are listed in the top 10 most influential data mining algorithms [21] in combination with

optimization algorithms and on their own are promising to solve the feature (gene) selection and classification problem.

1.1.2 Feature Selection Algorithms

Feature (gene) selection on high throughput biological data, such as gene expression data (microarrays) is a key issue in the domain of bioinformatics. Gene expression data usually has a small number of samples and large number of genes which most of them are irrelevant and redundant. Irrelevant and redundant genes deteriorate the performance of the learning models (classifiers). Therefore, selecting high discriminative gene subsets from microarray data helps to save computational costs by reducing dimensionality and improve the prediction accuracy of classifiers. In fact the aim of gene selection is to find a small fraction of gene subset that has the most discriminative power to improve predictive performance with robustness. This technology help physicians in clinical practice to have efficient diagnosing as well as effective treatments. In general, three types of gene selection methods have been developed; filter, embedded, and wrapper methods. Embedded and wrapper methods utilize gene selection as a part of training the learning model, whereas filter methods choose genes independently from a classification model. Wrapper and Embedded methods require to utilize nested cross validation and involve fitting more hyper parameters [22]. Decision tree approaches are the most typical embedded based gene selection algorithms. Ranking and space searching are two categories for filter based gene selection algorithm. Chi-square and t-test are commonly used ranking methods on microarray data. Correlation-based Feature Selection (CFS) and Minimum Redundancy Maximum Relevance (MRMR) are space searching based filter methods which have been suggested to remove the redundancy during gene selection.

1.1.3 Gene Selection Using Optimization Algorithms

Finding an optimal gene subset for a given problem with N number of genes requires evaluating all 2^N possible subsets. In fact, the size of search space for finding optimal gene subset grows exponentially with regard to the number of genes. Finding the optimal subset of genes is an NP-hard problem. Therefore, an efficient global search algorithm is necessary for solving gene selection problems. Metaheuristic algorithms are well-known for their global search ability. These algorithms are capable of handling

high dimensional optimization problems with satisfactory solutions within a reasonable time. For solving gene selection problem and biomarker discovery many metaheuristic algorithms have been conducted on microarray datasets. Four most commonly used optimization algorithms, which will be used in this thesis, are reviewed in this section. They are Genetic Algorithm (GA) [23], Binary Particle Swarm Optimization (BPSO) [24, 12], Simulated Annealing.

Genetic Algorithm

FS with GA needs to consider the process of FS as an optimization problem and then mapping it to the genetic structure of stochastic variation and natural selection. In the first step of GA algorithm, a primary population of chromosomes is generated randomly. The chromosomes are modeled as the binary vectors. Then, fitness value of each chromosome is evaluated by using a classifier. Two chromosomes with the best fitness value are selected. Then, for these chromosomes, a split point is chosen randomly. In the next step the front of one chromosome is mapped to the back of the other (and vice versa) in order to generate two offspring chromosomes with combined genes. In the last step these two offspring chromosomes are mutated randomly according to predetermined probability. The process would end if maximum iteration was met.

This algorithm works based on two genetic operators; crossover and mutation. GA has the ability to solve complex and non-linear problems. An important disadvantage of GA is its unguided mutation which is the only reason of a very slow convergence of GA. It also has a lot of parameters for tuning [23].

Binary Particle Swarm Optimization

PSO is an evolutionary algorithm based on population which was originally introduced as an optimization technique for real-number spaces (Eberhart R, 1995). The binary version of PSO which can be used for gene selection was conducted by the same author in 1997. Its general steps are described as a follows. Initially, the position of each particle (gene subset) is initialized randomly in binary coding format; the bit value 0 and 1 indicates the gene is discarded and is selected, respectively. Then, the fitness values for all particles are evaluated by predictive accuracy of specific classifier. The best personal memories of each particle is $pbest_{id}^{old}$ (i is number of particles and d is number of genes) and the globally best particle in the whole swarm is $gbest_{d}^{old}$. Each

particle has two parameters; velocity and position. In each iteration each particle adjusts its velocity and position to follow the features of $pbest_{id}^{old}$ and $gbest_{d}^{old}$ particles. The velocity and position of all particles are updated by using Eq. (1.1), (1.2), (1.3), and (1.4).

$$v_{id}^{new} = w \times v_{id}^{old} + c_1\, r_1\left(pbest_{id}^{old} - x_{id}^{old}\right) + c_2 r_2\left(gbest_{d}^{old} - x_{id}^{old}\right) \tag{1.1}$$

$$\textbf{If } v_{id}^{new} \notin (v_{min}, v_{max}) \textbf{ then} v_{id}^{new} = max(min(v_{max}, v_{id}^{new}), v_{min}) \tag{1.2}$$

$$sigmoid(v_{id}^{new}) = \frac{1}{1+e^{-v_{id}^{new}}} \tag{1.3}$$

$$\textbf{if } sigmoid(v_{id}^{new}) > r_3 \textbf{ then } x_{id}^{new} = 1 \textbf{ else } x_{id}^{new} = 0 \tag{1.4}$$

$$x_{id}^{new} = x_{id}^{old} + v_{id}^{new} \tag{1.5}$$

In these equations, w denotes the inertia weight to control the impact of the last velocity to the current velocity, $c1$ and $c2$ are acceleration (learning) constants, and r_1, r_1, $r_3 \in uniform$ [0,1]. $v_{id}^{new}$and x_{id}^{old} stand for the updated velocity and current position of the ith particle. v_{min} and v_{max} are minimum and maximum velocity which are user specified parameters. The whole procedure is repeated until the maximum number of iterations is met.

One of drawbacks of PSO algorithm in feature selection is that only considered classification accuracy. To solve this problems PSOPG1 algorithm which considers both the classification performance and the number of features was introduced by Bing Xue in 2013.

PSOPG1 follows the basic steps of standard PSO in updating particle's position. At the standard PSO particles update their position by using Eq. (1.5) instead of equations (1.3), (1.4) and a threshold θ is utilized to determine whether a feature is chosen or not. If $x_{id}^{old} > \theta$ the dth feature is selected ($\theta = 0.6$). PSOPG1 uses a new $pbest$ and $gbest$ updating mechanism. In this new updating mechanism $pbest$ are refreshed (replaced by the particle's new position) when the fitness value of particle's new position is better than $pbest$ or fitness value of them are the same but the number of features for particle's new position is smaller. After updating the $pbest$ of each particle, $gbest$ is updated in the same way. PSOPG1 has been applied on problems with a few hundreds of features and shown high performance. In chapter 3, binary version of PSOPG1

(BPSOPG1) which updates the position of each particle according to the Eq. (1.3) and (1.4) is utilized to selects gene subset from high-dimensional gene expression data.

PSO algorithm have been utilized by several study as gene selector. Ref. [12] proposed a wrapper approach based on PSO and decision tree classifier (C4.5) for cancer classification on ten benchmark and one clinical gene expression data. In ref. [25] a novel hybrid framework for gene selection and classification of high-dimensional microarray data, which combines k-means clustering, Signal to Noise Ratio (SNR), PSO, and three classifiers, has been proposed. In this framework, firstly k-means clustering has been used to eliminate redundant genes. Then, for each cluster SNR has been applied. Finally, top genes from each cluster has been gathered and used as input to PSO. The SVM, KNN, and Probabilistic neural network (PNN) have been adapted as the classifiers. Similar to this work has been done in ref. [17] which, uses Random Forest Ranking (RFR) instead of SNR and adopts boostedC5.0 as the classifier. It was tested on ten benchmark gene expression datasets. In [26] a new approach based on improved form of binary PSO and SVM classifier has been suggested to solves gene selection problem. In ref. [27] a modified PSO namely, Geometric PSO, has been proposed for gene selection and cancer classification of high-dimensional microarray data. Geometric PSO has been implemented in Waikato Environment for Knowledge Analysis (WEKA).

The PSO algorithm is more likely to be caught in a local optimum. In order to overcome the local optimum problem, ref. [28] proposed an improved form of PSO (PSO-RG) by developing a new $gbest$ updating mechanisms. In PSO-RG $gbest$ will be restarted whenever it is not improved in a number. This simple but effective method help PSO to avoid being trapped in a local optimum, achieve superior classification result, and reduce number of selected genes. In ref. [29] an improved form of PSO-RG namely PSO-LSRG has been developed. In PSO-LSRG approach reset mechanism is applied to $gbest$ to avoid stagnation in local optima and a new local search method is applied to $pbest$ in order to exploit better solutions.

Several studies have combined Genetic Algorithm (GA) with PSO to benefit the useful advantages of both of them and covered their problems. Ref. [30] proposed a gene selection method based on hybrid of PSO/GA and utilized SVM as the classifier. Another study used hybrid of PSO/GA for designing a fuzzy based expert system [31]. The experiments have been done on six gene expression data sets and shown high

performance compared to other approaches. In ref. [32] a hybrid PSO/GA algorithm and Artificial Neural Network (ANN) has been proposed for biomarker discovery on microarray datasets. Ref. [33] suggested a hybrid method of BPSO and a combat genetic algorithm (CGA) in order to reduce the number of genes expression and achieves a low classification error rate. In this method the CGA was embedded in the BPSO and the LOOCV classification accuracy of K-NN classifier served as a fitness function.

Simulated Annealing

The trope of SA comes from the annealing specifications in metal processing. The annealing process contains the control of heat and its cooling rate. Compared to other optimization algorithms, SA has an advantage of being able to avoid the algorithm from being stuck at the local minimum. SA uses random chain in terms of Markov chain. In the FS based on SA, a primary solution is chosen randomly and it is supposed to be the optimal solution. Afterward, the value of the primary solution is calculated using the fitness function. Whereas heat T does not meet the end condition, a neighboring solution of the current optimal solution is chosen and its fitness value is calculated. If the fitness value of the freshly chosen neighboring solution is greater than or equal to the current optimal solution, the current optimal solution is substituted with a freshly chosen neighbor solution. If the fitness value of the neighboring solution is less than the current optimal solution, a random number is generated in the range of (0, 1). In this situation, the substitution of the optimal solution is allowed only if a generated random number is less than Eq. (1.6). Then the heat is reduced by Eq. (1.7). The process would finish if maximum iteration was met.

$$e^{-\frac{cost\ (\text{neighbor solution})-cost(\text{optimal solution})}{T}} \tag{1.6}$$

$$T \leftarrow r \times T \tag{1.7}$$

GA and SA are more or less equivalent with respect to the quality of the solutions. The SA is not efficient in exploring large solution spaces because of randomly seeding and needs large number of parameters for tuning [34].

1.1.4 Problem Statement

1.1.4.1 Feature Selection

Biological data, such as microarrays can contain many irrelevant and redundant features (genes). These features may cause misleading in the modeling of algorithms for cancer classification and overfitting with long training times. In order to obtain optimal performance with short training times and reduce memory requirement, the Feature Selection (FS) process should be considered to use as a pre-process step in machine learning before applying classifiers to a dataset [35].

There are many studies based on FS methods. The FS algorithms are broadly categorized into three groups: filter, wrapper, and embedded approaches. This categorization is based on whether or not they are combined with a specific learning algorithm (classifier).

Filter based FS approaches consider the features independently and remove irrelevant features according to the statistical characteristics of the data. The t-test, chi-squared test, information gain, and Correlation based FS (CFS) are some well-known filter approaches [36, 37].

Wrapper based FS methods apply a specific machine learning algorithm to evaluate the score of selected feature subsets. These methods utilize Cross-Validation (CV) schema to train learning algorithm [38]. Comparing the wrapper methods to the filter approaches, wrapper methods are more accurate than the filter approaches because of considering the interactions among the features. However, they are computationally more expensive and the performances of them strongly depend on the given learning algorithm. Embedded based FS methods are special cases of wrapper methods that are characterized by a deeper interaction between the construction of the learning algorithm and the FS. In these methods the FS algorithm is always regarded as a component in the learning model. The Decision Tree (DT) algorithms, such as C4.5 [9] and Classification and Regression tree (CART) [18] are known as the most typical embedded besed FS approaches.

FS is known as an NP-hard and combinatorial problem. Hence, meta-heuristic methods are more appropriate to untie this laborious problem because of their population-based characteristics. Various stochastic global search algorithms have been used to solve FS problem on medical datasets, such as, Genetic Algorithm (GA) [23], Binary Particle

Swarm Optimization (BPSO) [11, 24], hybrid of GA and PSO algorithms [30], and Simulated Annealing (SA) [34].

The congenital drawbacks of the mentioned optimization algorithms still puzzle themselves. Therefore, to better address FS problems, a simple and efficient global search technique is needed. The Black Hole Algorithm (BHA) is one of the newest meta-heuristic methods based on the swarm intelligence [39]. This algorithm was discovered by simulating the behavior of black hole in outer space. In the real world black hole is an object of extreme density with intense gravitational attraction. The black hole's gravitational attraction swallows all objects if they come near enough. Because of BHA's characteristics including powerful optimal performance, single parameter, and fast convergence, the BHA has been used for solving a number of problems such as clustering [39], multi-objection reactive power dispatch problem [40], optimization problem [41], spam detection [42], and optimal coordination of digital overcurrent relays problem [43]. A comprehensive study of black hole approach and its applications in different research fields is provided in [44]. However, to the best of our knowledge there is no reported research related to FS using the BHA in literature.

1.1.4.2 Drawbacks of BPSO

BPSO is a population-based stochastic optimization algorithm which tries to solve the FS problem by simulating the social behavior of fish schooling or bird flocking. BPSO is more computationally efficient than GA and mostly provides better solution. The main disadvantage of PSO is that it is more likely to fall into local optimum. Currently, there is no research outcomes on hybrid of BBHA and BPSO in order to help BPSO to avoid being trapped in a local optimum in the field of feature (gene) selection on microarray datasets.

1.1.4.3 Meta Analysis of Recurrent PCa

Prostate cancer (PCa) is the most diagnosed malignancy and the second most reason of cancer-related death for the men over the age of 50 in the western countries [45]. The prostate-specific antigen (PSA) is the most reliable biomarker for PCa, which is helpful for diagnosis, screening, and follow-up after surgery. For treatment of PCa, two treatment methods, radiation therapy or radical prostatectomy (RP) and hormone ablation therapy are used. Yet, these methods do not provide enhanced survival rates

and nearly 30% of patients experience a biochem¬ical recurrence with enhanced PSA levels after curative treatment of RP [46]. Moreover, metastatic and advanced tumors of PCa respond very poorly to chemotherapy [47]. All these facts emphasize the significance of developing early diagnostic biomarkers for PCa progression. Identifying effective predictors of tumor recurrence after the surgical operation to determine whether treatment is required or not is a main challenge in the PCa research. To predict biochemical recurrence (BCR) of PCa after RP and develop effective predictors of tumor recurrence, multiple studies have been conducted for gene expression profiling [48-50]. Recently, numerous studies have been published which show that the alterations in microRNAs are associated with PCa initiation and progression [51-53].

The miR-1, miR-133b, miR-519d, and miR-647 are new biomarkers with prognostic and diagnostic value for recurrence of PCa, which have been identified through miRNA expression profiling [54, 55]. The miR-449b, miR-21, miR-141 and miR-221 are also known as putative prognostic or predictive markers in PCa recurrence after RP [56-58].

Meta-analysis utilizes statistical methods to contrast and combines results from multiple studies in the hope of increasing the statistical power and reproducibility over individual studies and identifying patterns across studies [59]. A limited number of studies [35, 54-58, 60, 61] has been conducted on microRNA expression profiles to distinguish recurrent from non-recurrent prostate tumor tissues and to identify novel biomarkers for prediction of PCa progression. The average differential expression level (fold change) and some level of significance as measured by the t-test are common procedures for identifying the biomarkers. These miRNA microarray data sets provide a rich resource for genome-wide information on PCa progression and make an ideal chance to perform a meta-analysis study. We assumed that a meta-analysis of some miRNA expression datasets of PCa progression can give a potentially significant list of co-deregulated miRNAs in PCa progression, which is important to specify pathways in which the miRNAs of interest and their target genes are involved.

1.1.4.4 Meta Analysis of Mir-145 Target Genes

MiR-145 is a well-studied miRNA, which is located at 5p32 chromosomal region and its expression is controlled by p53 and some other transcriptional factors like RREB1, FoxO, and C/EBP- β [62]. Mir-145 acts as a tumor suppressor and has been shown to be downregulated in several cancer types including prostate, head and neck, pancreatic

ductal adenocarcinoma, lung, breast, colorectal, bladder, and gastric cancer. It promotes apoptosis in the growing cells by silencing MYC (MYC-c), PPP3CA, EGFR, NUDT1, TNSF10, SWAP70, DEFA, CBFB, CLINT1, and RTKN [63]. MiR-145 has been found to be associated with tumorigenesis via suppressing the expression of several genes such as Insulin-like growth factor 1 in colorectal cancer [64], c-Myc and Cyclin-dependent kinase 6 (Cdk6) in oral squamous cell cancer [65], ER-α in breast cancer, SOX2 in larynx and prostate cancer, and several other genes in distinct cancer types [66]. However, to the best of our knowledge there is no a meta-analysis study investigating mir-145 targets and this is the first study, which combines and correlates miR-145 and mRNA microarray data in the literature.

1.2 Objective of the Thesis

One of the goals of this thesis is to investigate/improve the capability of BHA for gene selection and propose some new wrapper approaches to the use of BHA for gene selection in classification problems to reduce the number of genes and achieve better classification performance than other optimization algorithms. Second goal of this thesis is to improve the performance of BPSO and help it to avoid being trapped in a local optimum. Third goal of this thesis is to do meta-analysis on specific disease and combine results from multiple studies in the hope of increasing the statistical power and reproducibility over individual studies and identifying patterns across studies.

To achieve these goals, a set of research objectives have been established to guide this research, which can be seen as follows.

1. Develop a Binary version of Black Hole Algorithm called BBHA for solving feature selection problem in biological data.
2. Develop a new wrapper approach that hybridizes the modified version of Binary Particle Swarm Optimization (BPSOPG1) and the Binary Black Hole Algorithm (BBHA), called BPSOPG1-BBHA, for solving gene selection problem.
3. perform a meta-analysis of 6 available miRNA expression datasets on recurrent PCa and identified a panel of co-deregulated miRNA genes.
4. perform meta-analysis of 8 available microarray datasets (consider samples for mir-145) and identified a panel of co-deregulated genes upon mir-145 over expression in prostate, breast, esophageal, bladder cancer, and head and neck squamous cell carcinoma.

1.3 Hypothesis

This thesis makes the following major contributions.

1. In this thesis, the BHA is, for the first time, being used to solve a feature selection problem. By applying the hyperbolic tangent function, a new binary version of BHA called BBHA is used to solve FS problem in text, image, and biomedical data. The BBHA is an extension of existing BHA through appropriate binarization. Moreover, the performances of six well-known decision tree classifiers (Random Forest (RF), Bagging, C5.0, C4.5, Boosted C5.0, and CART) are compared to employ the best one as an evaluator of proposed algorithm. The performance of the proposed algorithm is tested upon eight publicly available biological datasets and is compared with Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Simulated Annealing (SA), and Correlation based Feature Selection (CFS) in terms of accuracy, sensitivity, specificity, Matthews' Correlation Coefficient (MCC), and Area Under the receiver operating characteristic (ROC) Curve (AUC). In order to verify the applicability and generality of the BBHA, it was integrated with Naive Bayes (NB) classifier and applied on further datasets on the text and image domains. The experimental results confirm that the performance of RF is better than the other decision tree algorithms and the proposed BBHA wrapper based feature selection method is superior to BPSO, GA, SA, and CFS in terms of all criteria. BBHA gives significantly better performance than the BPSO and GA in terms of CPU Time, the number of parameters for configuring the model, and the number of chosen optimized features. Also, BBHA has competitive or better performance than the other methods in the literature.
   Part of this contribution has been published in:

   Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, "An application of black hole algorithm and decision tree for medical problem", Proceedings of 2015 IEEE International Conference on Bioinformatics and Bioengineering (BIBE). Belgrade, Serbia. 2-4 Nov 2015. IEEE Press. pp. 1-6.

Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, "Biomarker discovery based on BHA and AdaboostM1 on microarray data for cancer classification", Proceedings of 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Orlando, FL, USA. 16-20 Aug 2016. IEEE Press. pp. 3080-3083.

Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, "Gene selection and classification approach for microarray data based on Random Forest Ranking and BBHA", Proceedings of 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). Las Vegas, NV, USA. 24-27 Feb 2016. IEEE Press. pp. 308-311.

Elnaz Pashaei, Nizamettin Aydin, "Binary black hole algorithm for feature selection and classification on biological data", Applied Soft Computing, Vol. 56 (2017) 94-106.

2. This thesis proposes a new wrapper approach that hybridizes the modified version of Binary Particle Swarm Optimization (BPSOPG1) and the Binary Black Hole Algorithm (BBHA), called BPSOPG1-BBHA, for solving gene selection problem. In the proposed approach, BBHA is embedded in the BPSOPG1 and plays the role of a local optimizer for each iteration. Three classifiers including Sparse Partial Least Squares Discriminant Analysis (SPLSDA),), *k*-nearest neighbor, and Naive Bayes methods with Leave-One-Out-Cross-Validation (LOOCV) schema are adopted as fitness function of hybrid BPSOPG1-BBHA. The performance of the proposed method was evaluated on four clinical datasets. In addition, comparative studies were provided between the proposed method and eight well-known gene selection approaches such as firefly, ant colony, bat search, genetic algorithm, harmony search, Fast Correlation-Based Filter (FCBF), and Correlation-based Feature Subset Selection (CFS). Experimental results and statistical analysis demonstrate that the proposed method yields very small sub sets of informative genes while achieving significantly better classification performance than other approaches.

Part of this contribution has been published in:

Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, "A Novel Gene Selection Algorithm for cancer identification based on Random Forest and Particle Swarm Optimization", Proceedings of 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Niagara Falls, Canada. 12-15 Aug 2015. IEEE Press. pp. 1-6.

Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, "Improving Medical Diagnosis Reliability Using Boosted C5.0 Decision Tree empowered by Particle Swarm Optimization", Proceedings of 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Milano, Italy. 25-29 Aug 2015. IEEE Press. PP. 7230-7233.

Elnaz Pashaei, Nizamettin Aydin, "Gene selection using hybrid binary black hole algorithm and binary particle swarm optimization", Genomics (under revision).

3. This thesis performs a meta-analysis on 6 available miRNA expression datasets for recurrent PCa and identified a panel of co-deregulated miRNA genes. Meta-analysis of six miRNA datasets revealed miR-125A, miR-199A-3P, miR-28-5P, miR-301B, miR-324-5P, miR-361-5P, miR-363*, miR-449A, miR-484, miR-498, miR-579, miR-637, miR-720, miR-874 and miR-98 are commonly upregulated miRNA genes, while miR-1, miR-133A, miR-133B, miR-137, miR-221, miR-340, miR-370, miR-449B, miR-489, miR-492, miR-496, miR-541, miR-572, miR-583, miR-606, miR-624, miR-636, miR-639, miR-661, miR-760, miR-890, and miR-939 are commonly downregulated miRNA genes in recurrent PCa samples in comparison to non-recurrent PCa samples. The network-based analysis showed that some of these miRNAs have an established prognostic significance in other cancers and can be actively involved in tumor growth. Gene ontology enrichment revealed many target genes of co-deregulated miRNAs are involved in "regulation

of epithelial cell proliferation" and "tissue morphogenesis". Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis indicated that these miRNAs regulate cancer pathways. The PPI hub proteins analysis identified CTNNB1 as the most highly ranked hub protein. Besides, common pathway analysis showed that TCF3, MAX, MYC, CYP26A1, and SREBF1 significantly interact with those DE miRNA genes. The identified genes have been known as tumor suppressors and biomarkers which are closely related to several cancer types, such as colorectal cancer, breast cancer, PCa, gastric, and hepatocellular carcinomas. Additionally, it was shown that the combination of DE miRNAs can assist in the more specific detection of the PCa and prediction of biochemical recurrence (BCR).

This contribution has been published in:

Elnaz Pashaei, Elham Pashaei, Maryam Ahmady, Mustafa Ozen, Nizamettin Aydin. "Meta-analysis of miRNA Expression Profiles for Prostate Cancer Recurrence following Radical Prostatectomy", PLoS ONE. Vol. 12, Issue 6, Jun 2017. doi:10.1371/journal.pone.0179543.

4. This thesis performs a meta-analysis on 8 available microarray datasets (consider samples for mir-145) and identified a panel of co-deregulated genes upon mir-145 over expression in prostate, breast, esophageal, bladder cancer, and head and neck squamous cell carcinoma. Meta-analysis of different GEO datasets showed that UNG, FUCA2, DERA, GMFB, TF, and SNX2 were commonly downregulated genes, whereas MYL9 and TAGLN were found to be commonly upregulated upon mir-145 over expression in prostate, breast, esophageal, bladder cancer, and head and neck squamous cell carcinoma. Biological process, molecular function, and pathway analysis of these potential targets of mir-145 through functional enrichments in PPI network demonstrated that those genes are significantly involved in telomere maintenance, DNA binding and repair mechanisms.

   This contribution has been published in:

Elnaz Pashaei, Esra Guzel, Mete Emir Ozgurses, Goksun Demirel, Nizamettin Aydin, Mustafa Ozen, “A Meta-Analysis: Identification of Common Mir-145 Target Genes that have Similar Behavior in Different GEO Datasets”, PLOS ONE. Vol. 11, Issue 9, September 2016. doi:10.1371/journal.pone.0161491.

1.4 Organization of the Thesis

The remainder of this thesis is organized as follows. The main contributions of the thesis are presented in Chapters 2-5, which can be seen in Figure 1.1. Each chapter addresses one of the research objective. Chapter 6 concludes the thesis.

Chapter 2 propose a Binary version of Black Hole Algorithm called BBHA for solving feature selection problem in biological data.

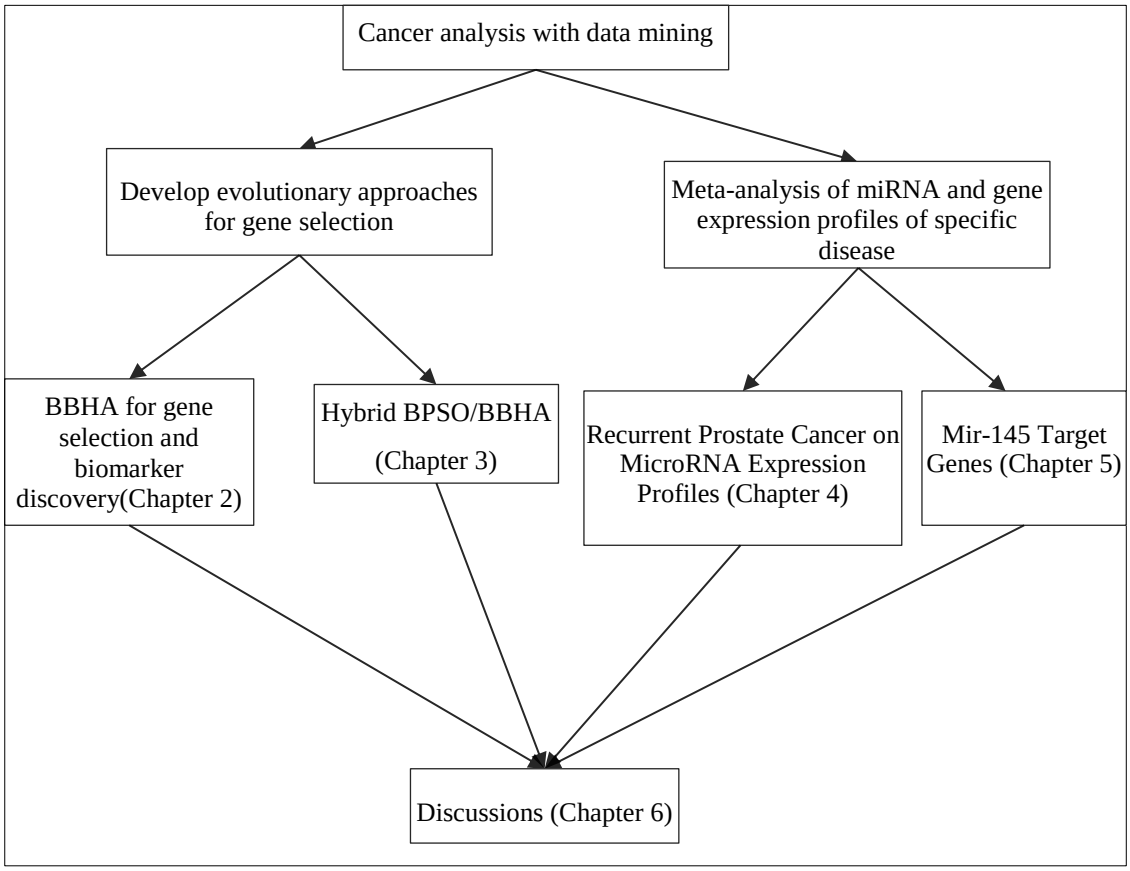


Figure 1.1 The overall structure of the contributions.

Chapter 3 proposes a new wrapper approach based on hybrid of BBHA and BPSOPG1 to address the problem of gene selection on high-dimensional gene expression data in order to significantly reduce the number of genes and increase the classification performance.

Chapter 4 analyzes miRNA expression profile in PCa progression considering 5 studies (6 datasets), in order to increase the probability of revealing truly significant deregulated miRNA genes, which should have higher potentials to be utilized as new biomarkers for the disease.

Chapter 5 performs a meta-analysis on target genes of miR-145 in several cancer types including prostate, breast, esophageal, bladder, head, and neck squamous cell carcinoma cancer, in order to unravel the underlying molecular pathways associated with mir-145 in tumor pathogenesis.

Chapter 6 summaries the work and draws overall conclusions of the thesis. It also suggests some possible future research directions.

1.5 Benchmark and Clinical Datasets

Throughout this thesis, the proposed algorithms are evaluated on a number of clinical and benchmark datasets. The datasets are summarized in Table 1.1, Table 1.2, Table 1.3 and Table 1.4.

Table 1.1 Benchmark datasets

Dataset	Number of Features	Number of data objects	Number of class	Domain
Chess	36	3196 (1669, 1527)	2	Text
Email word subject	242	64 (35, 29)	2	Text
WraoAR10P	2400	130	10	Image, face
WrapPIE10P	2420	210	10	Image, face
Wisconsin diagnostic breast cancer	31	569(357,212)	2	Life
Parkinson's	22	195(48,147)	2	Life
Heart-Statlog	13	270(150,120)	2	Life
Colon Tumor	2000	62(40,22)	2	Microarray
Central Nervous System	7129	60(39,21)	2	Microarray
ALL-AML (Leukemia)	7129	72(47,25)	2	Microarray
Breast Cancer	24481	97(51,46)	2	Microarray
Ovarian Cancer	15154	253(91,162)	2	Microarray

Table 1.2 Characteristics of each gene expression datasets

GEO Accession	Type of Platform	# of samples (BCR+, BCR−)*	# of genes	References
GSE25136	GPL96	79 (40, 39)	22283	[49]
GSE70769	GPL10558	90 (41, 49)	47323	[48]
GSE70768	GPL10558	43 (34, 9)	47323	[48]
GSE31684	GPL570	93 (54,39)	54675	[67]

Table 1.3 Characteristics of each miRNA Expression datasets

GEO Accession	Platform of dataset	Type of Platform	#of samples (BCR+, BCR−)*	# of miRNAs	References	Model for generating expression summaries
GSE55323	GPL10701	Agilent	40 (20, 20)	15744	[54]	log2 transformed and quantile normalized
GSE26245	GPL11350	Illumina	71 (29, 42)	733	[55]	quantile-normalized expression signal
GSE26247	GPL11350	Illumina	82 (29, 53)	1145	[55]	quantile-normalized expression signal
GSE65061	GPL17537	nCounter Human miRNA Expression Assay, V2	43 (19, 24)	800	[61]	normalized data
GSE62610	GPL18942	microRNA Card A+B Set v3.0	36 (22, 14)	536	[56]	normalized data
GSE46738	GPL8786	Affymetrix	51 (34, 17)	847	[58]	log scale RMA generated

Table 1.4 Summary of GEO datasets

GEO Accession	platform	Type of Platform	Samples Containing Mir-145	Cancer Type
GSE47657	GPL13607	Agilent	GSM1154161(PC3), GSM1154163(DU145), GSM1154165 (LNCap)	PCa
GSE24782	GPL10332	Agilent	GSM610397(PC3), GSM610398(DU145)	PCa
GSE58295	GPL4133	Agilent	GSM1406126 (PC3-8h), GSM1406127 (PC3-16h), GSM1406128 (PC3-24h)	PCa
GSE37119	GPL10332	Agilent	GSM911053(HNSCC, IMC-3)	Head and neck squamous cell carcinoma
GSE18625	GPL570	Affymetrix	GSM462902, GSM462903, GSM462904, GSM462905	Colon cancer (Exclude)
GSE19737	GPL570	Affymetrix	GSM492843, GSM492844, GSM492845	Breast cancer
GSE20028	GPL4133	Agilent	GSM500946 (TE2), GSM500948 (TE13)	Esophageal squamous cell carcinoma
GSE19717	GPL4133	Agilent	GSM492573 (KK47), GSM492575 (T24)	Bladder cancer

CHAPTER 2

BLACK HOLE ALGORITHM FOR FEATURE SELECTION

2.

2.1 Introduction

Biological data often consist of redundant and irrelevant features. These features can lead to misleading in modeling the algorithms and overfitting problem. Without a feature selection method, it is difficult for the existing models to accurately capture the patterns on data. The aim of feature selection is to choose a small number of relevant or significant features to enhance the performance of the classification. Existing feature selection methods suffer from the problems such as becoming stuck in local optima and being computationally expensive. To solve these problems, an efficient global search technique is needed. Black Hole Algorithm (BHA) is an efficient and new global search technique, inspired by the behavior of black hole, which is being applied to solve several optimization problems. However, the potential of BHA for feature selection has not been investigated yet. Therefore we proposes a Binary version of Black Hole Algorithm called BBHA for solving feature selection problem in biological data.

2.1.1 Chapter Goals

The goal of this chapter is to propose a Binary version of Black Hole Algorithm called BBHA for solving feature selection problem in biological data. The BBHA is an extension of existing BHA through appropriate binarization. Moreover, the performances of six well-known decision tree classifiers (Random Forest (RF), Bagging, C5.0, C4.5, Boosted C5.0, and CART) are compared in this chapter to employ

the best one as an evaluator of proposed algorithm. Specifically, the highlights of this chapter are:

- We propose a binary version of the Black Hole Algorithm called BBHA based on hyperbolic tangent function for solving discrete problems.
- We apply the proposed BBHA as a wrapper based feature selection method.
- We test the effectiveness of BBHA wrapper based feature selection method with two classifiers, Random Forest and Naive Bayes, on twelve benchmark datasets from different domains (biological, text, and image).
- We compare the performance of six popular decision tree algorithms (Random Forest, Bagging, C5.0, Boosted C5.0, C4.5, and CART) to select the robust and best of them as a fitness function of optimization algorithms.
- We compare the performance of the proposed BBHA wrapper based approach with the Genetic Algorithm (GA), Binary Particle Swarm Optimization (BPSO), Simulated Annealing (SA), and Correlation based Feature Selection (CFS) algorithms in terms of eight evaluation criteria.
- Experimental results demonstrate that Random Forest is the best decision tree algorithm and the proposed BBHA wrapper based feature selection approach outperforms the performances of BPSO, GA, SA, and CFS in terms of all criteria.
- It is also shown that the proposed method performed much faster, needs single parameter for configuring the model, and is simple to understand.

2.1.2 Chapter Organization

The remainder of this chapter is organized as follows. In section 2.2 the methods used are introduced. It includes continuous BHA, proposed binary version of BHA, proposed wrapper approach for FS based on BBHA. The characteristics of datasets, experimental design and setting are described in section 2.3 The experimental results on twelve datasets from different domains and summary of discussion are presented in section 2.4. Finally, section 2.5 provides a summary of this chapter.

2.2 The Proposed Algorithms

In this section, we give a detailed description of the continuous Black Hole Algorithm (BHA), proposed Binary Black Hole Algorithm (BBHA) and BBHA wrapper based approach, which is introduced for solving discrete problems including Feature Selection (FS).

2.2.1 Continuous Black Hole optimization Algorithm (BHA)

The black hole optimization algorithm is a robust stochastic optimization technique based on simulation of the behavior of black hole in outer space.

The below steps explain manner of simulating BHA from black hole phenomenon:

Step 1: Outer space is full of known and unknown stars. In real space black hole is formed by collapsing individual stars so BHA begins with the population of stars that located arbitrarily in the explore space. In BHA each star has a fitness value, which is evaluated by a fitness function to be optimized. The best star that has the best fitness value is selected as the black hole. It is called "black" because it absorbs all the light and reflects nothing. Figure 2.1 shows BHA schema. The black circle is the black hole and green circles are stars. They placed randomly in the search space.

Step 2: In the real space, a black hole is an object of extreme density with an intense gravitational attraction. This leads to a great amount of gravitational force pulling stars around it. BHA has followed the same behavior. By Eq. (2.1) all the stars began moving toward the black hole.

Step 3: The sphere shaped bound of a black hole in outer space is known as the event horizon. The event horizon radius is called as the Schwarzschild radius. The red circle in Figure 2.1 shows the event horizon of black hole. In the real space the Schwarzschild radius is computed by Eq. (2.2) and in BHA is computed by Eq. (2.3).

Step 4: Because of extreme density and strong gravitational attraction of black hole when a star crosses the event horizon, it will be swallowed by the black hole and disappear. In the region of event horizon the escapee speed is tantamount to the speed of the light, so nothing can get away from within the event horizon. In BHA, the Euclidean distance between black hole and star is

computed. If this distance is less than Schwarzschild radius, substitute it with a fresh star in the random location in the search space.

Step 5: In BHA if a star reaches a location with lower cost than the black hole, in that case theirs locations should be replaced [39, 68, 70].

$$X_i(t+1) = X_i(t) + rand \times \left(X_{BH} - X_i(t)\right) \quad i = 1,2,\dots,N \tag{2.1}$$

$$R = 2GM/C^2 \tag{2.2}$$

$$R = \frac{f_{BH}}{\sum_{i=1}^{N} f_i} \tag{2.3}$$

where $X_i\ (t)$ and $X_i\ (t+1)$ signify the locations of the i^{th} star at *iterations* t and $t+1$, respectively. *Rand* indicates uniform distribution with a range from 0 to 1. N denotes the number of stars. X_{BH} points the location of the black hole in the exploration space. M, G, and C signify the mass of the black hole, the gravitational constant, and the speed of light respectively. f_i denotes the fitness value of the ith star and f_{BH} indicates the fitness value of the black hole.

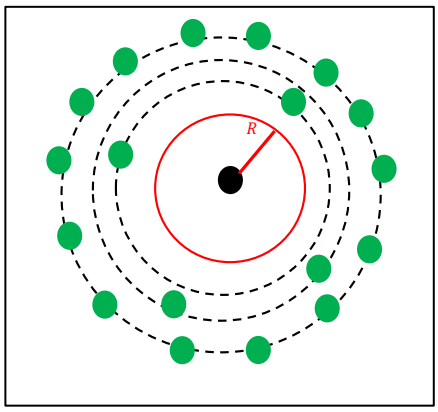


Figure 2.1 Black Hole Schema

Based on the above explanation the framework of the BHA method is presented in Algorithm 1.

01	**Input**
02	number of stars(N), number of iteration
03	**Output**
04	Black hole
05	The fitness value of black hole
06	**Begin**
07	Initialize a population of stars
08	**For** $j = 1$ to numbers of stars
09	calculate the objective function of the star(j) and save in fitness array(f)

10 **Next** *j*
11 The star with the most remarkable fitness value is chosen as the black hole
12 **While** (max iteration or convergence criteria is not met) do
13 **For** $a = 1$ to numbers of stars
14 $X_a^{new} = X_a^{old} + rand \times (X_{BH} - x_a^{old})$
15 Evaluate fitness value of the star(X_a)
16 **if** fitness of X_a > fitness of X_{BH} **Then**
17 $X_{BH} = X_a$
18 **End if**
19 Replace the new fitness value of the star (X_a) with the previous value
20 Update fitness array (f) and Calculate : $R = \frac{f_{BH}}{\sum_{i=1}^{N} f_i}$
21 **if** $\sqrt{(X_{BH} - X_a)^2}$ <R **Then**
22 replace X_a with a new star in an optional location in the search scope
23 **End if**
24 **next** a
25 **end while**
26 **End**

Algorithm 1. The continuous black hole algorithm

2.2.2 The proposed Binary Black Hole Algorithm (BBHA)

The BHA was originally developed for continuous valued spaces. But there exist a number of discrete combinatorial optimization problems, such as FS, in which the values are not continuous numbers but rather discrete binary integers. For this reason, we have introduced binary version of BHA and called it BBHA. Binarization techniques can be categorized into two groups: two steps binarization and continuous-binary operator transformation. Our proposed binarization technique belongs to the first group. In the first group without any modifications in the operators, only two steps is added after the continuous iteration.

In solving FS problem the search space must be modeled as a d-dimensional Boolean lattice, where the i^{th} star moves around the d-dimensional space.

Since the problem is to select or not select of a given feature, the position of a star only takes the values 1 or 0. Therefore, a transfer function is needed to forces stars to move in a binary space. Transfer functions define the probability of changing position's elements from 0 to 1 and vice versa. In the proposed approach, Hyperbolic Tangent function is utilized to modify the position of stars as in the Eq. (2.4) and (2.5).

$$S\left(X_{id}\,(t+1)\right) = abs(\tanh\left(X_{id}\,(t+1)\right)) \tag{2.4}$$

$$X_{id}\,(t+1) = \begin{cases} 1 & \text{If } S\left(X_{id}\,(t+1)\right) > rand \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

where $rand$ is a uniform random number between 0 and 1. In Eq. (2.5), instead of $rand$ threshold 0.6 can also be considered. Hyperbolic Tangent function belongs to the group of v-shaped transfer functions. It has been used because it shows good performance compared to the other transference functions such as sigmoid function [71]. In addition, in the proposed algorithm we may face with situation that one star with small number of features has the same fitness value with black hole. In this situation we should change their positions.

In BBHA we only need to set number of stars. The proposed algorithm does not suffer from some of other optimization algorithms difficulties such as the slow convergence rate and adjusting several parameters. Compared with other optimization algorithms, BBHA is easier to implement, depend on a single parameter for configuring the model, requires much less memory, and converges more rapidly.

2.2.3 The proposed wrapper approach based on BBHA for FS

In this section, we present details on the process used to enable FS with BBHA. At the beginning of BBHA, the primary population of the star's position is initialized randomly. Each star encodes a candidate feature subset based on a bit string. The length of the string is equivalent to the total number of features in the dataset of interest. In the binary encoding, a bit of one implies the feature is chosen and a bit of zero means that the feature is not chosen. Similar to other optimization algorithms, the fitness value of each star is calculated by using an evaluator. Here, two classifiers; RF and NB serve as the evaluators of our proposed algorithm. For biological data accuracy of RF classifier and for text and image datasets accuracy of NB classifier are used. The proposed wrapper approach based on integration of BBHA with RF is called BBHA-RF and based on combination with NB classifier is called BBHA-NB.

In the part of evaluating fitness value of stars, when two founded stars have identical fitness value, the one with smaller number of features is chosen as the best star (black hole).

The procedure stops once stopping criteria (maximum number of iterations) is met. The parameters for BBHA specify 25 iterations of population consisting of 10 stars. At the end of the BBHA wrapper based FS algorithm, the star with the best performance is selected. The position of this star gives the selected features. By using the subset of data that contains these selected features, again the performance of RF or NB model is assessed using 10-fold-CV for five different metrics. In order to avoid producing random results and provide an assurance for impartial comparison of the classification performances, assessing the efficiency of RF or NB model for selected features by optimization algorithms is executed 100 times. Average across 100 times of run is considered as a last result of one execute of whole procedure. The whole procedure runs 5 times for biological data and 3 times for text and image data. In each time, different subsets of features are selected by optimization algorithms. Average of these 5 times or 3 times of runs for whole procedure is reported. Algorithm 2 illustrates the procedure of applying BBHA for FS.

Input
Rand, number of stars, number of iteration
Output
Black hole
The fitness value of black hole
Begin
Initialize a population of stars
For $j = 1$ to numbers of stars
    Evaluate fitness value of the star(j) by 10-fold-CV RF or NB and save in fitness array(f)
Next *j*
The star with the most remarkable fitness value is chosen as the black hole
While (max iteration or convergence criteria is not met) do
    For $a = 1$ to numbers of stars
        Evaluate fitness value of the star(X_a) by 10-fold-CV RF or NB
        if (fitness of X_a > fitness of X_{BH}) **Then**
            $X_{BH} = X_a$
        else if ((fitness of X_a == fitness of X_{BH}) and (| X_a|< |X_{BH}|)) **Then**
            $X_{BH} = X_a$

end if

Replace the new fitness value of the star (X_a) with the previous value

Update fitness array (f) and Calculate : $R = \frac{f_{BH}}{\sum_{i=1}^{N} f_i}$

if $\sqrt{(X_{BH} - X_a)^2} < R$ **Then**

replace X_a with a new star in an optional location in the search scope

end if

next a

For $i = 1$ to numbers of stars

For $d = 1$ to number of features

$X_{id}^{new} = X_{id}^{old} + rand \times (X_{BH\,d} - x_{id}^{old})$

if $abs(tanh(X_{id}^{new})) >$ rand **Then**

$X_{id}^{new} = 1$

Else

$X_{id}^{new} = 0$

end if

next d

next i

end while

End

Algorithm 2. Pseudo code of Binary BHA for FS

By following this algorithm, we attempt to find optimal feature subset, which could improve the classification accuracy of medical data.

2.3 Experiments

2.3.1 Experimental Design

We have tested the performances of different DT algorithms to find out which of them has higher performance than the others on medical data sets to choose it as fitness function of optimization algorithms. Then, we have examined the relative performance of the combined BBHA and RF method (best DT algorithm as fitness function) denoted as BBHA-RF for true classification of medical data, with a series of repeated 10-fold-

CV experiments (repeat for 5 times to avoid bias). We have compared the performance of the proposed BBHA-RF method with GA-RF, BPSO-RF, SA-RF, and CFS. The results are reported in terms of accuracy, sensitivity, specificity, MCC, and AUC. Larger values of these criteria represent good classification performance. These measures are defined as follows:

2.3.1.1 Accuracy

Accuracy represents the percentage of correct predictions. Let TN, TP, FN, FP denotes true negative, true positive, false negative and false positive, respectively. The Accuracy is:

$$Accuracy = (TP + TN)/(TP + TN + FN + FP) \quad (2.6)$$

2.3.1.2 Sensitivity & Specificity

To see how the accuracy is distributed over the classes, sensitivity and specific values are presented. Sensitivity is the ability of the classifier to find all the positive samples. Specificity is the ability of the classifier to find all the negative samples.

$$Sensitivity = TP/(TP + FN) \quad (2.7)$$

$$Specificity = TN/(TN + FP) \quad (2.8)$$

2.3.1.3 Matthews Correlation Coefficient

In machine learning, the quality of unbalance binary (two-class) classifications can be obtained by Matthew's correlation coefficient. If the classes are unbalanced (not equal), computing the MCC of classification system can be so much more appropriate than computing accuracy. The Matthews correlation coefficient (MCC) is:

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2.9)$$

The aim of Matthew's correlation coefficient is to measure the quantity of correlation between predictions and real target values. The answer is bounded between the range +1 and -1.

The answer in the range of (0, +1] shows that predictions are positively related to the target values. A zero shows that the prediction is completely random. The answer in the range of [-1, 0) shows that predictions are negatively related to the target values.

2.3.1.4 Area Under ROC Curve (AUC)

The impact of a threshold on the false positives and false negatives (FP/FN) tradeoff can be visualized by Receiver Operating Characteristic (ROC) curve. The functions of the threshold can be described by the coordinates of ROC curve points:

$$threshold = \theta \in \mathbb{R}, \text{here} \theta \in [0,1] \tag{2.10}$$

$$ROC_X(\theta) = FPR(\theta) = \frac{FP(\theta)}{FP(\theta) + TN(\theta)} = \frac{FP(\theta)}{\#N} \tag{2.11}$$

$$ROC_Y(\theta) = TPR(\theta) = \frac{TP(\theta)}{FN(\theta) + TP(\theta)} = \frac{TP(\theta)}{\#P} = 1 - \frac{FN(\theta)}{\#P}$$
$$= 1 - FNR(\theta) \tag{2.12}$$

No false positives and all true positives (FPR, TPR) = (0, 1) is the optimal point on the ROC curve. AUC can be achieved by computing the area of the convex shape under the ROC curve. When AUC reaches to 1 this means that ROC is reached to the optimal point of perfect prediction.

2.3.2 Dataset

To evaluate the performance of the proposed method twelve datasets, which belong to completely different domains, are employed. These domains are biological (life and microarray), text, and image.

Text datasets are Chess and Email word subject which are two classes and obtained from UCI Machine Learning Repository at the website: http://www.ics.uci.edu/-mlearn/MLRepository.html. Image datasets are warpAR10P and warpPIE10P which are multiclass and taken from the http://featureselection.asu.edu/datasets.php. The summary of these data sets are given in Table 1.1 on Page 18.

For the biological datasets, three small medical datasets with a variety of complexity and five widely-used binary microarrays were used. Small medical datasets are Wisconsin diagnostic breast cancer, Parkinson's, and Heart Statlog, which are obtained from UCI Machine Learning Repository. Colon Tumor, Central Nervous System, Leukemia, Breast Cancer and Ovarian Cancer are microarray datasets that are available for download at the website: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html. The characteristics of these medical data sets are shown in Table 1.1. The datasets are

diverse in terms of the number of samples and features. The number of classes is two for all biological datasets.

2.3.3 Experimental Setup

Our experiment results consist of two parts. Firstly, we compared the performance of BBHA with BPSO, GA and CFS on the text and image datasets. We used NB classifier with 10-fold-CV as a fitness function and feature subset evaluator on these datasets. Then the performances of six well-known DT classifiers (Random Forest, Bagging, C5.0, Boosted C5.0, C4.5, and CART) are compared with each other to identify the best one of them. Finally, we considered the best DT (RF) classifier with 10-fold-CV as fitness function and gene subset evaluator of BBHA. Then, the proposed BBHA is conducted on a set of eight well-known medical datasets and compared with BPSO, GA, SA, and CFS.

In the process of 10-fold-CV, the samples of data are divided into 10 equally subsets. Each time, 9 subsets are located next to each other to create the training set and the remained one subset is utilized as the test set. Then the average accuracy across all 10 trials is calculated. Since one try of the 10-fold CV is generally biased and in order to have statistically meaningful conclusion we repeated 10-fold-CV for 100 times and reported average and standard deviation of them. The experiments are carried out on a laptop with Windows 7, 2.40 GHz CPU and 4 GB of RAM, using R version 2.2.1. For RF classifier we used 'randomForest' package. For Bagging and C4.5 classifiers, 'RWekajars', 'rJava', and 'RWeka' packages have been used. 'C50' package has been utilized for C5.0 classifier and Boosted C5.0 (trials=10). For CART classifier 'rpart' package, for CFS filter approach 'FSelector' package, and for NB classifier 'e1071' package have been employed. Also, for GA and SA optimization algorithms 'caret' package has been used.

For all datasets, the number of particles for BPSO, the number of stars for BBHA, and the number of chromosomes for GA are set to 10. Most of the genes in the high dimensional data like microarrays are irrelevant and not useful for classification problems. Selection of top ranked genes as a preparation step for microarrays by removing a large number of irrelevant, redundant and noisy genes can provide a better classification accuracy [72]. We have selected 50 top ranked genes by Chi- Squared statistic with leave-one-out cross validation method. The parameters of Binary PSO are

adjusted as follow [12]; (v_{min}), (v_{max}), (c_1), (c_2) and (w) are set at -4, 4, 2, 2 and 0.4 respectively. The crossover and mutation probability of GA are set to 0.8 and 0.1 respectively. The process would stop if maximum iteration was met. Here, a maximum number of iteration is set at 25 because by increasing the number of iterations the improvement in results was insignificant.

2.4 Results and Analyses

2.4.1 Experimental Results on Text and Image Datasets using NB Classifier

The summary of four text and image datasets considered here are given in Table 1.1. NB classifier with 10-fold-CV was considered as a fitness function of BBHA, BPSO, and GA optimization algorithms and also for feature subset evaluation of CFS. The average classification accuracy of BBHA-NB, BPSO-NB, GA-NB and CFS on the text and image datasets are reported in Table 2.1. In particular, the number of selected features and CPU times are tabulated. Due to the stochastic nature of BBHA, GA, and BPSO the average FS results of them for three independent runs are reported. Here, the maximum iteration is set to 30 and as a pre-process step 250 top ranked features are chosen by Chi- Squared statistic for image datasets. In Table 2.1, among the four algorithms on each dataset the best average classification accuracies are highlighted in bold typeface. Table 2.1 shows that BBHA-NB outperforms the other three algorithms in terms of accuracy on all datasets except the Email word subject dataset. GA-NB gives better accuracy than BBHA-NB for this dataset. The accuracies of BBHA-NB and GA-NB are not significantly different from each other on 3 out of four datasets. BBHA-NB performs significantly better than GA-NB only on the WraoAR10P dataset. With regard to the number of selected features, CFS chooses the least number of features but it does so at the cost of low classification accuracy. After CFS, BBHA-NB chooses fewer number of features but with a classification accuracy that is superior to the BPSO-NB, GA-NB, and CFS. CFS filter method uses less time than the other 3 wrapper approaches. BBHA-NB is the second approach which costs less time than BPSO-NB and GA-NB.

Table 2.1 Average accuracy, number of features, and computational efficiency of each wrapper method for 3 independent runs

Dataset	Criteria	BPSO- NB	GA-NB	CFS-NB	BBHA-NB
Chess	*Accuracy*	93.93 ± 0.04	94. 33 ± 0.03	90.42	**94.66** ± 0.017
	# of features	14 ± 1.41	14 ± 2.82	**3**	**6** ± 1
	CPU time	1090.22	3233.11	2.31	510.27
Email word subject	*Accuracy*	91.23 ± 4.57	**93.37** ± 0.01	92.18	92.28± 4.05
	# of features	118.5 ± 7.77	111.66 ± 3.21	**2**	**25.66** ±7.09
	CPU time	591.11	2102.99	6.17	127.33
WraoAR10P	*Accuracy*	74.63 ± 1.69	76.40 ± 0.44	74.93	**80.46** ± 1.41
	# of features	124 ± 1.41	93.66 ± 18.82	19	**14.33** ± 2.51
	CPU time	1091.71	3473.11	130.87	396.33
WrapPIE10P	*Accuracy*	92.32 ± 0.95	93.96 ± 0.72	91.29	**94.91** ± 0.89
	# of features	126.5 ± 6.36	120 ± 8.18	**32**	37 ± 2.64
	CPU time	1282.07	3561.11	463.21	422.34

2.4.2 Experimental Results on Biological Datasets using RF Classifier

In order to determine which DT algorithm is more robust and has higher performance than the others to be used as fitness function of optimization algorithms, we have compared six well-known DT classifiers on eight medical datasets. The computational results of this experiment are shown in Table 2.2. The classification accuracy, sensitivity, specificity, MCC, and AUC along with standard deviations of them are presented in this Table. According to Table 2.2, RF classifier has higher performance in almost all datasets. Beside the high performance, the robustness is an important factor in evaluating a classifier. The standard deviation of all criteria for RF in all datasets is small. This shows that RF is a robust classifier.

The Figure 2.2 displays average AUC, classification accuracy, and MCC of six DT classifiers on eight medical datasets, respectively. As can be seen from this figure, it is clearly found that the performance of RF is better than other classifiers. Therefore, we choose this classifier as a fitness function of four optimization algorithms (i.e. BBHA, BPSO, GA, and SA). After RF, Boosted C5.0 has higher performance than the others. Bagging, C5.0, and C4.5 are placed in the next positions, respectively. CART is the classifier which has worse performance than the others.

To evaluate the effectiveness of our proposed method, we compare the results of BBHA-RF with BPSO-RF, GA-RF, SA-RF, and CFS. The average AUC, classification accuracy, sensitivity, specificity and MCC along with standard deviations of them for 100 independent runs of the selected features in 5 executions of the whole procedure

and CPU time are presented in Table 2.3. In order to illustrate the good performance of the proposed FS method, Table 2.4 reports an average number of the selected features from the entire data set by BBHA-RF, BPSO-RF, GA-RF, SA-RF, and CFS.

As can be seen from Table 2.3, the proposed BBHA-RF outperformed BPSO-RF, GA-RF, SA-RF, and CFS in terms of all criteria. Table 2.4 demonstrates that the proposed approach is significantly better than all wrapper approaches and CFS filter approach in term of the number of selected optimized features. BBHA-RF selects features approximately 5 times fewer than BPSO-RF and GA-RF. Compared with SA-RF and CFS the proposed method selects 3 times less features.

The computational efficiency of BBHA-RF is comparable to SA-RF and is better than BPSO-RF and GA-RF. BBHA-RF converges approximately 3 times faster than BPSO-RF and approximately 6 times faster than GA-RF. For all biological datasets BBHA-RF can achieve high performance with least number of features in short time.

Figure 2.3 shows the average solution quality of BBHA-RF, BPSO-RF, GA-RF, SA-RF, and CFS on eight biological datasets. We can observe that the proposed wrapper approach (BBHA-RF) compared with other FS algorithms maximizes solution quality while using fewer features. All medical datasets except heart-statlog and breast cancer microarray are unbalanced. So consideration of MCC is more appropriate than accuracy. Average MCC of BBHA-RF is significantly better than all mentioned FS algorithms. Figure 2.4 displays the computational time in seconds for each of the filter and wrapper approaches. The speed of CFS is much higher than mentioned wrapper approaches. The rapidity of BBHA-RF is approximately similar to SA-RF. The proposed approach converges much faster than BPSO-RF and GA-RF.

In the following, the performance of BBHA-RF on microarray datasets is compared with eight state-of-the-art methods from literature. Table 2.5 reports the results of BBHA and different feature (gene) selection methods for five microarrays. The classification accuracy (first value in every table cell) and the number of selected features (the value in parenthesis) are used as criteria for comparing the performances of methods.

From the results of Table 2.5, one observed that BBHA-RF except breast cancer microarray gives highly competitive results compared with these reference methods. The most remarkable result for BBHA-RF concerns the ovarian cancer microarray. We obtain 99.82 % accuracy with average 2.8 genes while the previous methods reach a prediction rate no greater than 99.44% with at least 4 genes.

For Wisconsin diagnostic breast cancer dataset 97.38 % accuracy with average 6.4 features is obtained by BBHA-RF. The proposed method outperformed the results of the literature in [73-75]. For this dataset 100% classification accuracy with only 3 features is obtained by a method based on modified correlation rough set FS and MLP classifier with 80-20 train-test scheme [76].

For Parkinson's dataset BBHA-RF obtained 94.20 % accuracy with average 4 features. The proposed method gave better performance than other approaches reported in [77-79]. For this dataset the highest classification accuracy (98.12%) with 11 features is obtained by a method based on minimum redundancy maximum relevance FS and complex-valued artificial neural network classifier with 10-fold-CV scheme [80].

For Heart-Statlog dataset, the best accuracy (89.96%) with only 3 features is obtained by a method which uses self-regulated learning PSO as FS and extreme learning machine as a classifier with 70-30 train-test scheme [81]. BBHA-RF obtained 85.40 % accuracy with average 4.8 features. Our proposed algorithm performs better than the results of the literature in [82-84].

Table 2.2 Solution quality of each decision tree classifier on medical data sets

Dataset	Criteria	RF	Bagging	C5.0	Boosted C5.0	C4.5	CART
Wisconsin diagnostic breast cancer	*Accuracy*	**96.08 ± 0.31**	94.66 ± 0.57	93.71± 0.79	95.93 ± 0.53	93.39± 0.74	92.41± 0.71
	Sensitivity	**93.75± 0.67**	91.90 ± 0.98	90.59± 1.42	93.38 ± 0.94	90.54± 1.42	89.46± 1.55
	Specificity	**97.48± 0.34**	96.37± 0.66	95.38± 0.80	97.46 ± 0.58	95.36± 0.92	94.13± 0.82
	MCC	**91.60± 0.84**	88.73 ± 1.27	86.49± 1.63	91.26 ± 1.13	86.18± 1.45	83.82± 1.54
	AUC(ROC)	**99.07± 0.13**	98.51 ± 0.28	96.24± 0.64	98.97 ± 0.25	92.6 ± 1.27	93.80± 0.87
Parkinson's disease	*Accuracy*	**90.70± 0.88**	87.73 ± 1.45	84.16± 2.05	89.61 ± 1.49	85.07± 1.95	86.13± 2.08
	Sensitivity	**72.70± 3.95**	65.52 ± 5.93	69.64± 6.55	71.82 ± 4.50	69.62± 6.02	65.44± 5.89
	Specificity	**96.90±** 0.75	94.88 ± 1.35	89.14± 2.14	95.73 ± 1.26	89.34± 2.36	93.63± 1.65
	MCC	**74.16±** 3.87	63.25 ± 5.01	56.55± 5.76	70.25 ± 4.95	59.39± 4.95	60.77± 6.29
	AUC(ROC)	**96.09± 1**.92	92.74 ± 2.56	82.02± 3.68	93.98 ± 2.86	80.19± 3.74	84.28± 3.42
Heart-Statlog	*Accuracy*	**82.95± 1.02**	81 ± 1.29	77.96± 1.59	79.94 ± 1.48	78.15± 1.57	80.27± 1.57

Table 2.2 (cont'd)

	Sensitivity	**77.56± 1.34**	74.80 ± 2.38	73.14± 2.68	76.05 ± 2.26	72.77± 2.74	74.64± 2.39
	Specificity	**87.53± 1.42**	85.60 ± 1.84	82.41± 2.26	83.88 ± 2.01	82.01± 2.45	85.21± 2.03
	MCC	**65.66± 1.86**	61.45 ± 2.86	55.65± 4.17	60.25 ± 3.31	56.93± 3.52	60.01± 3.38
	AUC(ROC)	**90.34± 0.74**	88.31 ± 1.29	82.08± 1.59	87.78 ± 1.11	78.28± 2.04	82.4 ± 1.83
Colon Tumor	*Accuracy*	**85.58± 1.46**	81.35 ± 3.96	81.91± 3.41	82.08 ± 2.63	83.03± 3.19	76.09± 3.96
	Sensitivity	87.72± **2.38**	**90.47** ± 4.62	88.93± 4.13	88.17 ± 3.01	88.65± 3.55	84.81± 4.84
	Specificity	**82.35± 5.68**	66.03 ± 7.76	70.74± 8.65	69.73 ± 7.47	74.88± 6.97	61.97± 8.76
	MCC	**65 ±** 7.71	52.71 ± **7.54**	56.73± 10.5	56.12 ± 8.15	57.87± 9.27	44.68± 8.85
	AUC(ROC)	**86.61±** 7.29	81.70 ± 7.36	75.96± 7.97	83.56 ± 7.66	76.35± 7.84	68.48± **6.38**
Central Nervous System	*Accuracy*	**84.86± 1.93**	76.96 ± 4.20	78.93± 4.06	79.61± 3.67	74.13± 4.31	71.46± 4.35
	Sensitivity	**92.49± 2.07**	90.09 **±** 4.18	87.30± 6.01	90.61 ± 3.69	82.85± 5.83	83.44± 6.24
	Specificity	**71.99± 5.93**	53.34± 10.8	60.78± 10	59.71 ± 9.49	57.51± 9.88	50.6 ± 10.3
	MCC	**62.03± 7.68**	44.14 ± 10.1	48.66± 11.6	48.95 ± 10.1	40.39± 10.8	30.97± 11.1
	AUC(ROC)	**87.61± 6.95**	78.90 ± 8.43	69.40± 7.98	79.08 ± 7.92	64.13± 7.63	62.70± 7.40
ALL-AML (Leukemia	*Accuracy*	**97.98± 0.86**	94.7 ± 1.61	86.07± 2.84	92.01 ± 2.56	85.08± 2.68	83.71± 2.60
	Sensitivity	**99.41± 0.92**	96.68 ± 2.32	86.40± 4.01	95.12 ± 2.93	86.61± 3.68	85.71± 3.92
	Specificity	**95.53± 2.31**	90.31 ± 5.02	84.98± 6.74	87.50 ± 5.54	82.88± 6.36	80.58± 6.16
	MCC	**91.28± 6.58**	82.84 ± 7.97	67.68± 8.22	79.16 ± 6.33	65.09± 7.13	62.28± 7.97
	AUC(ROC)	**96.02**± 5.24	95.71± **4.86**	82.47± 6.18	90.28 ± 6.44	80.9 ± 6.35	78.23± 6.43
Breast Cancer	*Accuracy*	**80.04± 2.02**	74.52 ± 3.36	67.67± 4.17	75.80 ± 3.07	69.05± 3.92	65.97± 4.21
	Sensitivity	**81.57± 3.58**	76.48 ± 4.76	71.57± 5.91	76.89 ± 5.08	71.30± 6.58	68.15± 6.11
	Specificity	**79.40± 3.51**	74.54 ± 4.95	66.79± 6.65	75.34 ± 5.29	67.89± 7.14	63.50± 7.24
	MCC	**60.19± 4.72**	49.86± 7.03	36.97± 8.33	51.15 ± 6.83	37.86± 9.45	31.42± 8.80

Table 2.2 (cont'd)

Ovarian Cancer	*AUC(ROC)*	**90.07± 3.20**	83.37 ± 3.68	67.92± 5.17	83.55 ± 4.10	68.47± 5.53	68.13± 5.11
	Accuracy	**99.12**± 0.40	97.45 ± 0.44	98.19± 0.43	98.64 ± 0.59	98.06± 0.65	97.14± 0.39
	Sensitivity	**99.54** ± 0.45	98.56 ± 0.40	98.15± 0.74	98.90 ± 0.60	97.53± 0.67	98.03± 0.53
	Specificity	98.20± **0.95**	95.23 ± 1.16	98.01± 1.12	98.38 ± 1.17	**98.86**± 1.26	95.67± **0.68**
	MCC	**98.05± 0.90**	94.51 ± 1.25	95.93± 1.19	97.16 ± 1.19	95.96± 1.39	93.68± 1
	AUC(ROC)	**99.94 ± 0.08**	99.18 ± 0.55	98.46± 0.43	99.40 ± 0.43	98.14± 0.77	96.86± 0.42

Table 2.3 Best, average solution quality and computational efficiency of each wrapper method for 5 independent runs

Dataset	Criteria	BPSO-RF	GA-RF	SA-RF	CFS-RF	BBHA-RF	Best BBHA (# features)
Wisconsin diagnostic breast cancer	*Accuracy*	96.92 ± 0.39	96.21 ± 0.33	95.90± 0.30	95.86	**97.38 ± 0.28**	97.85 (5)
	Sensitivity	94.66 ± 0.66	93.65 ± **0.65**	93.2 ± 0.68	93.50	**95.79** ± 0.87	96.71
	Specificity	98.41 ± 0.33	97.62 ± 0.36	97.43± **0.28**	97.27	**98.57** ± 0.30	99.421
	MCC	93.11 ± 0.75	91.75 ± 0.65	91.25± 0.64	91.11	**93.85 ± 0.71**	95.53
	AUC(ROC)	99.30 ± 0.11	99.08 ± 0.13	98.84± 0.15	98.93	**99.47 ± 0.09**	99.71
	CPU Time	4212.114	6226.58	800.28	**2.16**	2079.33	2070
Parkinson's	*Accuracy*	92.30 **± 0.77**	93.37 ± 0.86	92.98± 0.80	91.20	**93.91** ± 0.77	95.78 (3)
	Sensitivity	78.37 ± 2.80	79.44 ± 4.35	80.54± 3.89	76.28	**84.79 ± 2.76**	88
	Specificity	97.33 ± **0.67**	**98.10** ± 0.64	97.19± 0.53	96.23	97.95 ± 1.13	98.61
	MCC	80.27 ± 16.27	81.28 ± 3.64	79.79± **3.36**	75.07	**82.61±** 3.52	89.89
	AUC(ROC)	96.15 ± 2.46	97.08 ± 2.07	95.91± 2.15	95.19	**97.22 ± 1.80**	99.02
	CPU Time	649.53	2401.26	115.66	1.03	373.35	320
Heart-Statlog	*Accuracy*	84.44 ± 1.03	83.54 ± 0.99	83.57± 0.95	81.17	**85.75 ± 0.44**	86.29 (3)
	Sensitivity	**80.88** ± 1.66	80.19 ± 1.29	78.88± 1.62	75.13	80.74 **± 0.65**	82.06

Table 2.3 (cont'd)

	Specificity	87.67 ± 1.44	86.03 ± 1.10	87.46± 1.26	86.20	**89.71** ± 0.89	91.40
	MCC	67.03 ± 2.05	66.55 ± 2.04	66.46± 2.11	61.92	**71.09** ±1.07	73.55
	AUC(ROC)	**90.75** ± 0.69	90.37 ± 0.69	89.25± 0.81	88.82	88.55 **± 0.67**	90.23
	CPU Time	628.69	517.4	132.46	0.36	314.34	300 5, 11
Colon Tumor	*Accuracy*	86.40 ± 1.63	86.56 ± 0.92	85.70± 1.33	88.08	**91.41 ± 1.3**	93.33 (3)
	Sensitivity	87.85± 2.11	88.69 ± 1.28	89.43± 0.99	90.05	**95.57 ± 1.91**	100
	Specificity	83.65 ± 3.69	83.86 ± 0.77	80.18± 3.59	84.51	**85.90** ± 4.83	96.29
	MCC	65.77 ± 4.22	66.48 ± 2	64.36± 3.19	68.39	**76.68 ± 3.97**	92.5
	AUC(ROC)	86.71 ± 1.43	86.08± 1.78	85.34± 1.48	**89.32**	87.68 ± 1.94	100
	CPU Time	361.42	776.88	76.27	2.52	114.48	102
Central Nervous System	*Accuracy*	90.27 ± 1.97	87.96 ± 1.76	84.47± 4	87.93	**91.85** ± 1.96	93.33 (5)
	Sensitivity	96.26 ± 0.46	94.93 ± 0.52	91.8 ± 3.03	96.83	**96.90** ± 1.26	98.33
	Specificity	80.26 ± 4.49	76.77 ± 4.86	71.72± 6.53	73.10	**83.85** ± 5.27	93.51
	MCC	72.45 ± 8.11	66.83 ± 5.09	60.49± 7.63	66.29	**75.41 ± 7.50**	89.14
	AUC(ROC)	89.44 ± 2.75	88.19 ± 1.47	86.24± 3.39	89.27	**93.06 ± 3.2**	100
	CPU Time	333.05	1009.03	139.93	7.22	107.47	116.68
ALL-AML (Leukemia)	*Accuracy*	98.55 ± 1.20	98.11 ± 0.73	96.59± 0.73	98	**98.61** ± 1.23	100 (2)
	Sensitivity	**99.93** ± 0.37	99.63 ± 0.68	99.01± 0.72	99.10	98.88 ± 1.43	100
	Specificity	96.56 ± 2.83	95.29 ± 1.18	92.55± 1.95	95.47	**98.77** ± 1.99	100
	MCC	92.47 ± 1.73	91.57 ± 1.23	88.74± 1.11	91.68	**92.69** ± 1.34	100
	AUC(ROC)	96.30 ± 1.44	95.72 ± 0.52	96.15± 0.49	95.79	**96.38** ± 1.38	100
	CPU Time	268.39	1013.63	87.47	1.95	101.84	91.71
Breast Cancer	*Accuracy*	83.94 ± 1.85	83.72 ± 0.98	79.56± 2.25	84.22	**87.77** ± 1.78	91.11 (6)
	Sensitivity	84.84 ± 3.53	86.24 ± 2.01	80.84± 1.47	87.57	**87.74** ± 3.18	94.66

Table 2.3 (cont'd)

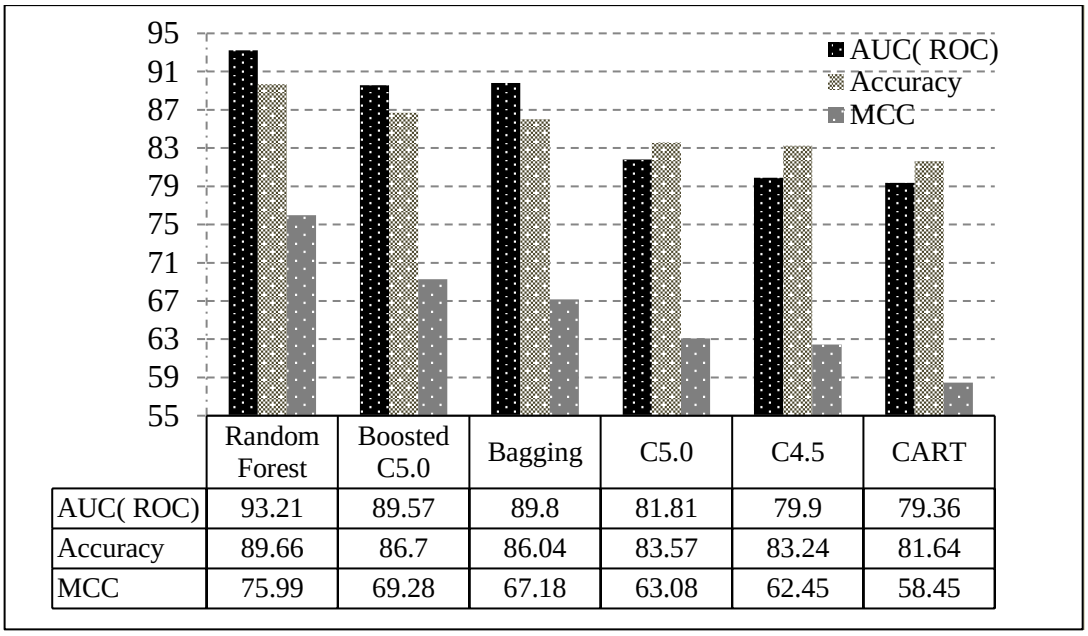
	Specificity	83.72 ± 3.67	81.96 ± 1.77	79.60± 3.55	82.17	**88.49** ± 3.17	94.66
	MCC	68.61 ± 2.04	67.93 ± 2	59.99± 3.90	67.46	**75.45** ± 3.83	83.85
	AUC(ROC)	91.23 ± 2.62	90.66 ± 1.04	88.25± 1.50	92.50	**93.47** ± 2.83	97.38
	CPU Time	407.37	909.25	105.89	6.02	129.26	102
Ovarian Cancer	*Accuracy*	99.64 ± 0.35	99.52 ± 0.11	98.58± 0.96	98.63	**99.82** ± 0.34	100 (3)
	Sensitivity	99.62 ± 0.43	99.66 ± 0.20	99.17± 0.80	98.77	**100** ± 0.0	100
	Specificity	**99.77** ± 0.51	99.31 ± 0.68	97.57± 1.11	98.30	99.58± 0.85	100
	MCC	99.29 ± 0.82	99 ± 0.24	96.85± 2.01	97.16	**99.69**± 0.61	100
	AUC(ROC)	99.99 ± 0.009	99.97 ± 0.03	99.91± 0.11	99.95	**100** ± 0.0	100
	CPU Time	743.96	1086.28	170.17	1.99	266.60	100

Table 2.4 Average number of selected feature

Dataset	BPSO-RF	GA-RF	SA-RF	CFS	BBHA-RF
Wisconsin diagnostic breast cancer	13.5 ± 3.53	25± 0.70	13 ± 1.41	9 ± 0.0	5.4 ± 2.40
Parkinson's	9±1.41	10 ± 1	5.5± 0.7	9 ± 0.0	4 ± 0.70
Heart-Statlog	8.2± 0.83	9 ± 2.08	7 ± 2.94	6± 0.0	4.8 ± 1.78
Colon Tumor	20.8± 1.30	18.6± 2.96	11.8 ± 1.92	14 ± 0.0	3.4 ± 0.54
Central Nervous System	24.4± 5.45	27 ± 5.14	14.02 ± 1.92	28 ± 0.0	8.4 ± 3.20
ALL-AML (Leukemia)	26 ± 2.16	16.4± 9.20	16 ± 1.87	7 ± 0.0	5.6 ± 2.70
Breast Cancer	24.2 ± 2.16	22 ± 8.15	14.02 ± 1.30	26 ± 0.0	6.2 ± 1.78
Ovarian Cancer	24.8 ± 1.92	22.5 ± 7.18	14.25 ± 4.71	9± 0.0	2.8 ± 0.44

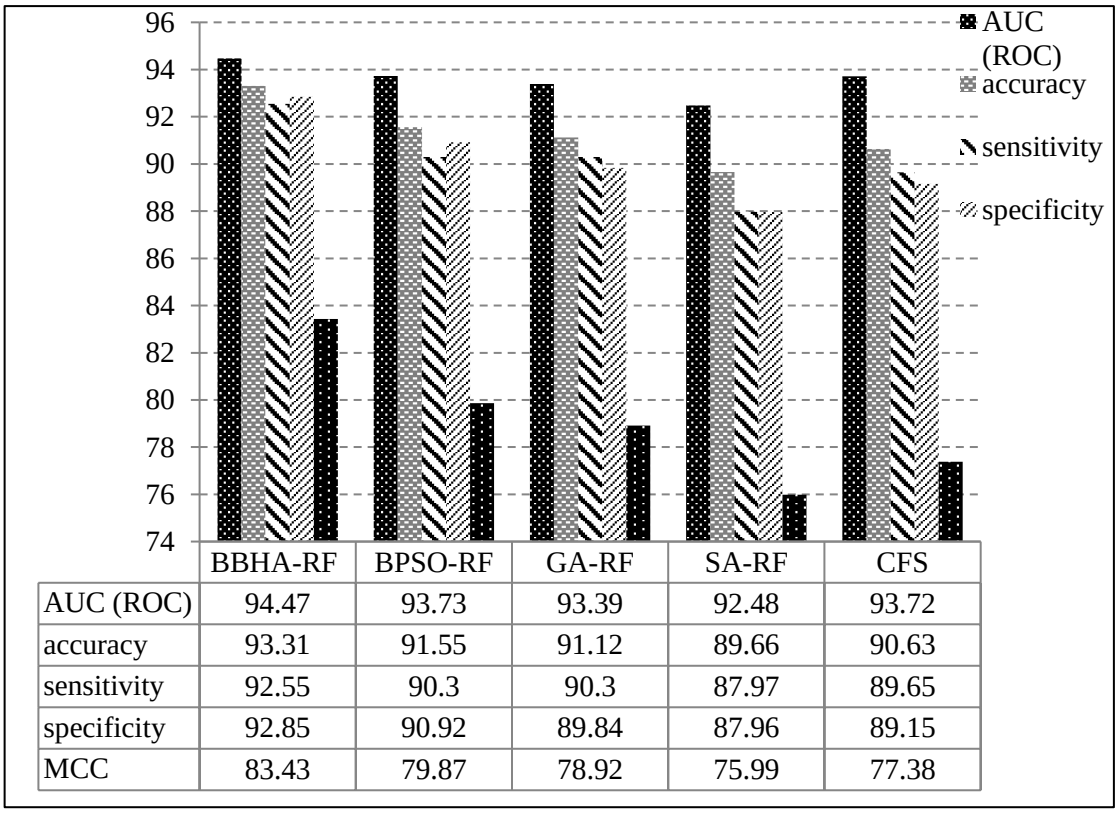
Table 2.5 Comparison of relevant works on cancer classification with our proposed method BBHA-RF

Dataset	BBHA-RF	[88]	[30]	[85]	[89]	[32]	[25]	[86]	[87]
Colon Tumor	91.41 (3.4)	**100 (2)**	91.9 (18.0)	93.32 (8)	93.55 (6)	96.67 (20)	99.44 (5)	90 (2)	93.5 (9)
Central Nervous System	**91.85 (8.4)**	-	-	-	-	-	-	90 (2)	86.6 (7)
ALL-AML(Leukemia)	98.61 (5.6)	97.38 (3)	97.2 (18.7)	98.61 (7)	98.74 (4)	**100 (17)**	99.10 (10)	-	100 (5)
Breast Cancer	87.77 (6.2)	95.86 (4)	93.4 (26.9)	-	-	96 (12)	**100 (20)**	-	-
Ovarian Cancer	**99.82 (2.8)**	99.44 (4)	-	-	-	-	-	-	98.8 (19)



	Random Forest	Boosted C5.0	Bagging	C5.0	C4.5	CART
AUC(ROC)	93.21	89.57	89.8	81.81	79.9	79.36
Accuracy	89.66	86.7	86.04	83.57	83.24	81.64
MCC	75.99	69.28	67.18	63.08	62.45	58.45

Figure 2.2 Average Classification AUC (ROC), Accuracy, and MCC of 6 Well-known Decision Tree Classifiers on 8 Biological Datasets



	BBHA-RF	BPSO-RF	GA-RF	SA-RF	CFS
AUC (ROC)	94.47	93.73	93.39	92.48	93.72
accuracy	93.31	91.55	91.12	89.66	90.63
sensitivity	92.55	90.3	90.3	87.97	89.65
specificity	92.85	90.92	89.84	87.96	89.15
MCC	83.43	79.87	78.92	75.99	77.38

Figure 2.3 Average Solution Quality of One Filter and Four Wrapper Approaches on 8 Medical Datasets

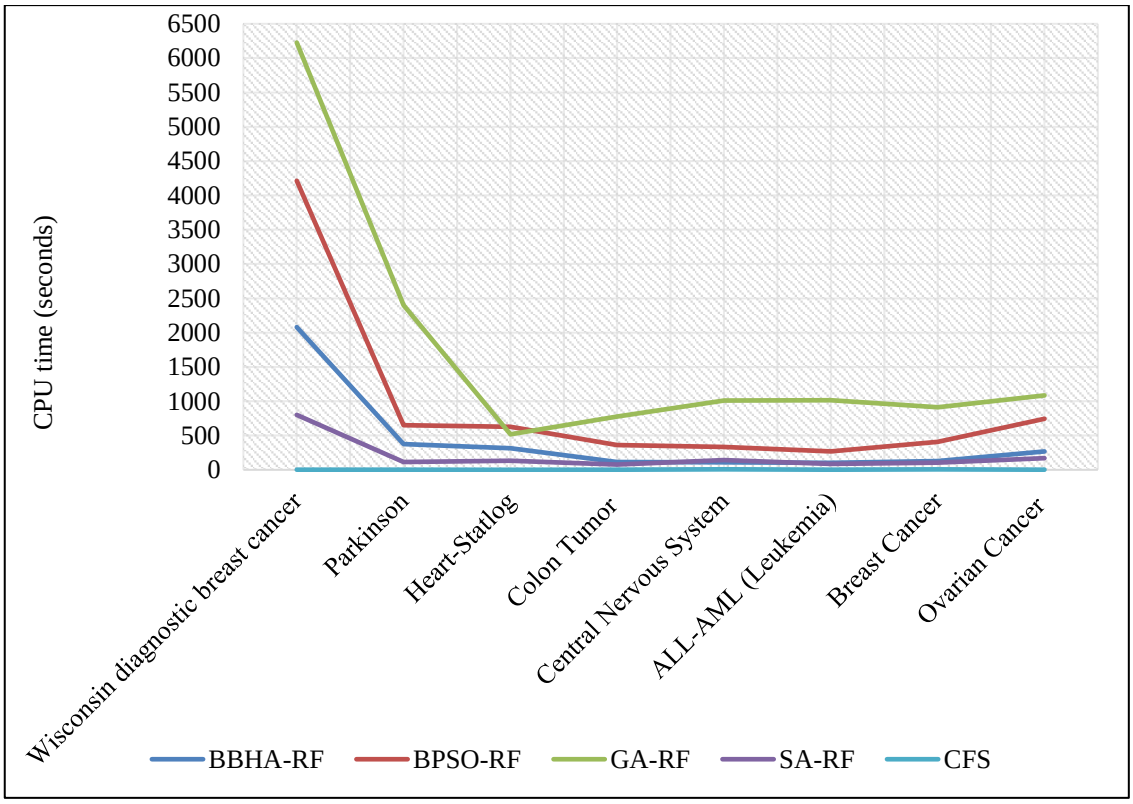


Figure 2.4 Computational efficiency of one filter and four wrapper approaches on 8 medical datasets

2.4.3 Discussion

In summary, from the experimental results on text and image data sets it is worth noting that CFS selects the lowest number of features on 3/4 datasets in the least amount of running time but suffers in terms of classification accuracy. GA-NB and BPSO-NB obtains good classification accuracy but require more running time and select many more features. Compared to mentioned methods, BBHA with NB is able to find significantly least number of features and to provide better classification performance in a sensible CPU time. Also, the computational results on medical datasets confirmed that RF gives better performance compared to other five DT classifiers therefore is chosen as a fitness function of optimization algorithms and evaluator of feature subsets. From the experimental results on biological data sets, it is inferred that BBHA-RF approach not only improves classification accuracy of RF by selecting the most informative features, but also obtains better performance in terms of all eight evaluation criteria when compared to BPSO-RF, GA-RF, SA-RF, and CFS filter method. The eight evaluation criteria are classification accuracy, MCC, AUC (ROC), sensitivity, specificity, the number of selected features, CPU time, and robustness. Because 6/8

biological datasets are unbalanced, considering MCC criteria is more suitable. Average MCC of BBHA-RF is significantly better than all mentioned algorithms and selects significantly fewer feature subsets on all datasets in a reasonable time. BBHA-RF converges much faster than GA-RF and BPSO-RF. The speed of proposed approach is comparable to SA-RF. Moreover, the standard deviations of the computational results are relatively small, indicating that the repeated 10-fold-CV is reliable and appropriate for classification of medical data. The comparison of BBHA-RF with other approaches in the literature suggests that BBHA-RF has competitive or better performance. Ultimately, a summary of the best subsets of genes found for each microarray by BBHA-RF is listed in Table 2.6.

Table 2.6 Best subsets of genes which found by BBHA-RF

Dataset	Accuracy (# of genes)	Name of genes
Colon Tumor	93.33 (3)	X12671, M16937, M91463
Central Nervous System	93.33 (5)	S71824_at,D83542_at,AF002020_at, HG2417-HT2513_at, U43747_s_at
ALL-AML(Leukemia)	100 (2)	L09209_s_at, M92287_at
Breast Cancer	91.11 (6)	Contig24311_RC, Contig7258_RC, NM_005192, Contig38726_RC, Contig14882_RC, NM_003450
Ovarian Cancer	100 (3)	MZ2.8234234, MZ418.49538, MZ435.46452

2.5 Chapter Summary

Feature selection is an important approach that used before applying classifiers to a data set in order to select informative features. A good FS method by selecting significant features helps to successfully and meaningfully modeling with low computational cost and high classification accuracy. During the past years, several metaheuristic algorithms such as GA, firefly, PSO, binary bat algorithm and ant colony algorithm make an effort to design the FS as a combinatorial optimization problem. However, almost all of the existing methods have a lot of parameters for configuring the model and are computationally expensive. Therefore, proposing a FS approach with a few parameters, high computational speed, and simplicity are necessary for classification.

This chapter presents the first study on using the BHA for solving FS problem. By applying the hyperbolic tangent function, a new binary version of BHA called BBHA is proposed to solve FS problem in text, image, and biomedical data. Two classifiers (RF and NB) serve as the evaluators of our proposed algorithm. In addition, to confirm that

RF is the best DT classifier, the performances of six popular DT algorithms were compared in this chapter.

Experimental results demonstrate that RF is the best DT algorithm and the proposed BBHA wrapper based FS approach outperforms the performances of BPSO, GA, SA, and CFS in terms of AUC, accuracy, MCC, sensitivity, specificity, and the number of selected optimized features. Furthermore, if the computational cost is taken into account, BBHA wrapper approach performs much faster than BPSO and GA. BBHA only needs a single parameter for configuring the model and is simple to understand.

CHAPTER 3

WRAPPER BASED HYBRID APPROACH FOR GENE SELECTION

3.

3.1 Introduction

Gene selection from high throughput technologies, such as microarrays, which is a well-known NP-hard problem, is a difficult task because of gene interactions and the large search space. Gene selection aims to find the smallest possible set of relevant genes that could obtain the optimal performance. Although many gene selection approaches have been introduced most of them still suffer from the problems of becoming stuck in local optima and excessive computational cost. To solve these problems an efficient global search technique is required. Metaheuristic approaches are powerful global search algorithms, which are capable of handling high dimensional optimization problems with satisfactory solutions within a reasonable time. Hence, we propose a new wrapper approach that hybridizes the modified version of Binary Particle Swarm Optimization (BPSOPG1) and the Binary Black Hole Algorithm (BBHA), called BPSOPG1-BBHA, for solving gene selection problem. In Chapter 2, we have shown that the simplicity, lower computational cost, fast convergence, and single parameter are some advantages of BBHA when compared with other meta-heuristic algorithms. Therefore, BBHA is embedded in the BPSO to overcome the drawbacks of PSO. Since BBHA and PSOPG1 are easy to implement, have fewer parameters, and both of them can converge more quickly, hybridization of them which was assumed to make a powerful approach for solving gene selection problem was considered in this chapter. Combining BPSOPG1 with BBHA has not been considered yet.

3.1.1 Chapter Goals

The overall goal of this chapter is to present a new wrapper approach based on hybrid of BBHA and BPSOPG1 to address the problem of gene selection on high-dimensional gene expression data in order to significantly reduce the number of genes and increase the classification performance. To achieve this goal, BBHA is embedded in the BPSOPG1 and plays the role of a local optimizer for each iteration. The SPLSDA method with LOOCV schema served as fitness function of hybrid BPSOPG1/BBHA. SPLSDA has been found useful in handling classification tasks in the case of the high dimensionality and small sample data. The performance of our proposed method was evaluated on four Gene Expression Omnibus (GEO) datasets taken from the National Centers for Biotechnology Information (NCBI). In addition, the performance of our proposed approach was compared with three classifiers, and eight gene selection approaches; six well-known optimization algorithms and two filter approaches. Moreover, the optimal subset of genes in each GEO dataset were found and Fuzzy Unordered Rule Induction Algorithm (FURIA) was used to find the relation between candidate genes. Specifically, the highlights of this chapter are:

- We proposed a new wrapper approach based on hybrid modified version of Binary Particle Swarm Optimization (BPSOPG1) and the Binary Black Hole Algorithm (BBHA), called BPSOPG1-BBHA, for solving gene selection problem.
- We tested the effectiveness of BPSOPG1-BBHA wrapper based gene selection method with three classifiers, Sparse Partial Least Squares Discriminant Analysis (SPLSDA), *k*-nearest neighbor, and Naive Bayes, on four clinical datasets from NCBI GEO (GSE25136, GSE70769, GSE70768, and GSE31684).
- We statistically compared the performance of the proposed BPSOPG1-BBHA wrapper based approach with the firefly, ant colony, bat search, genetic algorithm, harmony search, Fast Correlation-Based Filter (FCBF), and Correlation-based Feature Subset Selection (CFS) algorithms in terms of three evaluation criteria (number of genes, accuracy, and AUC) .
- We found the optimal subset of genes in each GEO dataset and used Fuzzy Unordered Rule Induction Algorithm (FURIA) to find the relation between candidate genes for biological point of view

- The experimental results and statistical analysis have demonstrated that BPSOPG1-BBHA/SPLSDA compare with many other methods, leads to a better performance in term of accuracy, AUC, and number of selected genes. The obtained results indicate that the BPSOPG1-BBHA/SPLSDA is a useful tool for selecting marker genes in clinical datasets.
- It was also shown that applying BBHA as the local optimizer for BPSOPG1 can significantly improve the performance of BPSOPG1 and help it to avoid being trapped in a local optimum.

3.1.2 Chapter Organization

The rest of this chapter is organized as follows. Section 3.2 presents algorithms of the proposed BPSOPG1-BBHA-SPLSDA wrapper approach for gene selection. Section 3.3 describes the characteristics of each dataset and parameter settings. Section 3.4 discusses the experimental results on 4 GEO datasets. Section 3.5 presents a discussion of the results and, finally, Section 3.6 provides a summary of this chapter.

3.2 Proposed Approach Based on Hybrid of BPSOPG1 and BBHA

The main idea of hybrid BPSOPG1/BBHA algorithm is to apply BBHA as a local optimizer for BPSOPG1 algorithm in order to improve the performance of it and minimize the number of genes. Figure 3.1 shows the flowchart of the hybrid BPSOPG1-BBHA.

Firstly, a pre-process step based on a simple backwards selection, a.k.a. recursive feature selection (RFE) was done to take the advantages of it. In backwards selection the genes are ranked and the less important ones are sequentially eliminated prior to modeling. Here, random forest is selected as the model [90].

Then $I*D$ population was generated using optimal feature subset from backwards selection by a binary system; I stands for the number of particles in a swarm, and D is the dimension of the microarray data. The LOOCV classification accuracy of a SPLSDA was used to measure the fitness of particles and each star. The BPSOPG1-BBHA algorithm is presented as below:

Step.1 Generate $I*D$ initial population for stars using the optimal feature subset from backwards selection.

Step.2 Perform BBHA process. Stars begin to move around black hole (best star) and update their positions via equations (3.1), (3.2), and (3.3).

$$X_{id}(t+1) = X_{id}(t) + rand \times \left(X_{BH} - X_{id}(t)\right) \quad i = 1,2,\dots,N \tag{3.1}$$

$$S\left(X_{id}(t+1)\right) = abs\left(\tanh\left(X_{id}(t+1)\right)\right) \tag{3.2}$$

$$X_{id}(t+1) = \begin{cases} 1 & \text{If } S\left(X_{id}(t+1)\right) > 0.6 \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

Step.3 Check termination. If termination condition is satisfied (maximum number of iteration=10), go to Step 4. Otherwise go to Step 2.

Step.4 The position of new optimized stars are passed to the BPSOPG1 process as population of particles.

Step.5 Compute fitness value of all particles and update $pbest$ and $gbest$.

Step.6 Perform BPSOPG1 operators. Each particle updates its velocity and position according to equations (3.4), (3.5), (3.6) and (3.7).

$$v_{id}^{new} = w \times v_{id}^{old} + c_1\, r_1\left(pbest_{id}^{old} - x_{id}^{old}\right) + c_2 r_2\left(gbest_{d}^{old} - x_{id}^{old}\right) \tag{3.4}$$

$$\textbf{if} \quad v_{id}^{new} \notin (v_{min}, v_{max}) \quad \textbf{then} \quad v_{id}^{new} = max(min(v_{max}, v_{id}^{new}), v_{min}) \tag{3.5}$$

$$sigmoid(v_{id}^{new}) = \frac{1}{1+e^{-v_{id}^{new}}} \tag{3.6}$$

$$\textbf{if} \quad sigmoid(v_{id}^{new}) > r_3 \ \textbf{then} \quad x_{id}^{new} = 1 \ \textbf{else} \quad x_{id}^{new} = 0 \tag{3.7}$$

Step.7 Judge termination. If satisfied, output the final solution. Otherwise go to Step 2.

The BPSOPG1 was configured to contain 10 particles and depend on the dataset different number of iterations was considered as the termination criterion of it. The number of stars in the BBHA was equal to the number of particles. After each iteration of the BPSO, the BBHA was run 10 times.

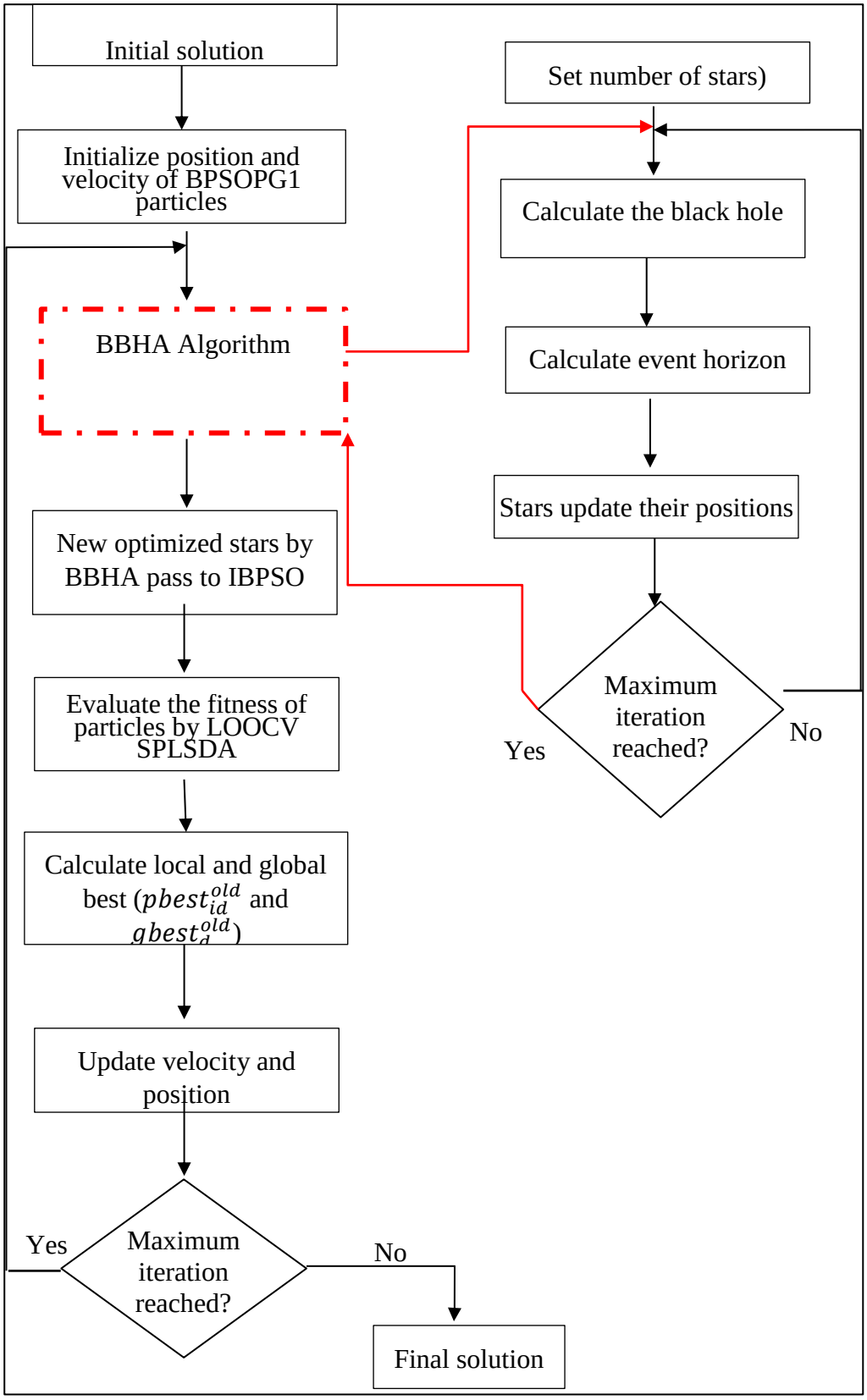


Figure 3.1 Flowchart of Hybrid BPSOPG1/BBHA

3.3 Datasets and Parameter Settings

3.3.1 Datasets

A set of experiments have been conducted on 4 GEO datasets from NCBI. Table 1.2 shows a summary description of these four datasets. GSE25136 as gene expression profile data associated with prostate cancer (PCa) recurrence and obtained from 79 cases, 39 of which were classified as having disease recurrence. This dataset has been used in the derivation of molecular signatures for recurrent PCa [49].

GSE70769 and GSE70768 contain expression profiling for PCa tissue samples that obtained from patients undergoing radical prostatectomy. These datasets include missing values and few labels are undetermined. We replace missing values with means of data and remove unlabeled column then combine these GEO datasets. Because the platform of these two datasets are the same there is no needing for removing batch effects. These datasets have been utilized in the identification of molecular profiles for recurrent PCa [48].

GSE31684 is a microarray data associated with bladder cancer recurrence that obtained from 93 bladder cancer patients managed by radical cystectomy. This dataset has been utilized for prediction of survival in high risk bladder cancer [67]. All GSE series matrix files, platform sets, and annotations file were downloaded and parsed by GEO query package in Bioconductor on R.

3.3.2 Parameter Setting

In order to examine the performance of the proposed approach, eight gene selection approaches are utilized, which are firefly, ant colony, bat search, genetic algorithm, harmony search, FCBF, and CFS. As a wrapper approach, the proposed hybrid algorithm requires a classifier to evaluate the fitness of the selected gene subsets. Three simple and commonly used classification algorithms are used here. There are Naive Bayes, KNN (K=1), and SPLSDA. For Naive Bayes classifier “e1071” package and for KNN and SPLSDA classifiers “caret” packages on R language have been utilized. SPLSDA has two key tuning parameters, thresholding parameter (eta) and number of hidden components (K).

For each dataset the tuning parameters for SPLSDA have been chosen by 10-fold cross-validation. For this aim the “cv.splsda” function from “spls” package have been used.

Since the datasets include a small number of samples, LOOCV classification accuracy of a classifier is utilized as evaluator of each selected gene subset. In LOOCV schema, one sample is evaluated as the test set while the remaining samples are used as the train set.

The experiments of firefly, ant colony, bat search, genetic algorithm, harmony search, FCBF, and CFS are conducted using Weka. Except number of swarms and iterations for optimization algorithms which are equal with proposed algorithm, all the settings are kept to the defaults because they can achieve good performance. The detailed settings are shown in Table 3.1. For each dataset, the experiments of each algorithm has been conducted for 10 independent runs to avoid bias. The non-parametric statistical significance test namely Wilcoxon test is done between the Area Under the ROC Curve (AUC) of different algorithms. 0.05 is chosen as the significance level (or confidence interval is 95%).

3.4 Results and Analyses

The LOOCV classification accuracy and AUC of several classifiers by using all genes are summarized in Table 3.2. This table demonstrate that without using the gene selection approaches, classifiers are not able to capture the pattern on data. In order to accelerate the speed of convergence, alleviate the burden of computation, and improve the performance of BPSOPG1/BBHA for gene selection a pre-process step based on RF-RFE is done. The RF-RFE finds an optimal subset of 1100, 600, and 850 genes for GSE25136, (GSE70769, GSE70768) and GSE31684, respectively. Figure 3.2 shows the resampling results for the candidate subset sizes evaluated during the RF-RFE process. The performance of different classifiers on these optimal genes are shown in Table 3.3. From Table 3.3, it is not difficult to gain a fact that gene selection on these GEO datasets can improve the performance of classifiers. Then we compared the performance of hybrid BPSOPG1/BBHA proposed in this chapter with the single BPSOPG1 and single BBHA algorithms with the SPLSDA classifier. The corresponding parameter settings of the two algorithms are the same as the hybrid BPSOPG1/BBHA method. Table 3.4 shows the classification accuracy, AUC (%), and the number of genes that obtained by the three methods with SPLSDA classifier for the best random seed. From Table 3.4, we can see the perfect performance is obtained on GSE25136 and GSE31684 using the hybrid BPSOPG1/BBHA methods. For the datset2, an accuracy of 98.49% is

obtained by the proposed method, which is significantly better than the results obtained by single BPSOPG1 or BBHA method. Also, the variation curves (number of iterations vs. classification accuracy and number of iterations vs. number of selected genes) of these three methods are described in Figure 3.3. This figure indicates that the convergence speed of BPSOPG1 is slower compared with BBHA and hybrid BPSOPG1/BBHA for all datasets. Until the end cycle, it can only find a not excellent enough solution with highest number of genes. In contrast, hybrid algorithm proposed in this chapter rapidly converge to an excellent solution with least number of genes in the 5th cycle for all datasets. We can draw a conclusion that BPSOPG1 algorithm with BBHA operators integrated in has excellent performance for gene selection compared to single BBHA and especially BPSOPG1.

A summary of the best subsets of genes found for each GEO dataset by the proposed approach is listed in Table 3.5. Furthermore, to see the relation between founded genes Table 3.6 shows the AUC and extracted rules by FURIA. A Heat map of two-way hierarchical clustering based on correlation distance and average linkage for these significant genes are shown in Figure 3.4.

In order to illustrate the efficient performance of the proposed gene selection approach, Table 3.7 reports accuracy, AUC, and number of genes selected by Naive Bayes, KNN, and SPLSDA combined with eight well-known gene selection algorithms, which all using LOOCV evaluation. In Table 3.7, "T" shows the result of the Wilcoxon test, where "-" indicates that the classification performance (AUC) of BPSOPG1/BBHA/SPLSDA is significantly better than other approaches. The best result are shown in bold type. As shown in Table 3.7, for the GSE25136, our proposed hybrid algorithm achieves the perfect performance using SPLSDA with an average of 40.66 genes. The combination of proposed hybrid approach with KNN classifier is the second highest ones. The BPSOPG1/BBHA hybrid approach with NB classifier yields to averagely 26.25 number of genes which is the least one compare to other approaches. For the (GSE70769, GSE70768) and GSE31684, the best performance is also obtained by the proposed gene selection approach with SPLSDA classifier. The combination of proposed approach with KNN and NB classifiers are placed in the next best ones.

From Table 3.7, it can be seen that our proposed BPSOPG1/BBHA algorithm selects a smaller gene subset with better performance (LOOCV classification accuracy and AUC)

than many other methods in all GEO datasets. Therefore, our proposed algorithm are more effective for optimal gene subset selection and pattern classification.

Table 3.1 Parameters used for experiments

BPSO/BBHA parameters	GSE25136	(GSE70769, GSE70768)	GSE31684
Population	10	10	10
Individual length	1100	600	850
Termination iterations	20	30	30
Inertia weight (ω)	0.4	0.4	0.4
Acceleration constants ($c1 = c2$)	2	2	2
k	4	7	5
eta	0.1	0.1	0.1

Table 3.2 Prediction results of the 6 well-known classifiers on data without gene selection

dataset	metric	*KNN*	*Random Forest*	*Naive Bayes*	*Simple logistic*	*SGD*	*Logit boost*
GSE25136	*accuracy*	59.49	59.49	64.55	68.35	65.82	70.88
	AUC	59.40	66.80	64.50	74.20	65.70	72.50
(GSE70769, GSE70768)	*accuracy*	57.89	69.92	59.39	64.66	66.91	63.15
	AUC	51.70	69.8	63.20	69.60	65.70	65.20
GSE31684	*accuracy*	51.61	60.21	55.91	44.08	54.83	40.86
	AUC	51.90	48.80	51.40	44.10	55.10	34.90

Table 3.3 Prediction results of the 10 well-known classifiers on selected gene subsets by RF-RFE

dataset	metric	*KNN*	*SMO*	*Random Forest*	*Hoeffding Tree*	*Naive Bayes*	*Simple logistic*	*SGD*	*Logit boost*	*Bayes Net*	*SPLSDA*
GSE25136	*accuracy*	78.48	89.87	78.48	75.94	77.21	72.15	84.81	72.15	84.81	**93.67**
	AUC	78.50	89.80	92.40	78.80	79.40	82.90	84.70	80.80	89.60	**95.60**
(GSE70769, GSE70768)	*accuracy*	69.92	83.45	83.45	56.39	59.39	74.43	**86.46**	76.69	63.90	81.20
	AUC	67.10	82	85.20	67.20	73.50	77.10	85.10	83.20	79.60	**87.90**
GSE31684	*accuracy*	63.44	76.34	78.49	75.26	76.34	68.81	75.26	60.21	62.36	**78.49**
	AUC	61	75	85.70	80.50	80.20	78.90	73.40	63.30	64.90	**90.30**

Table 3.4 The performance of three methods with SPLSDA classifier for the best random seed

Method	GSE25136			(GSE70769, GSE70768)			GSE31684		
	accuracy	*AUC*	*#Genes*	*accuracy*	*AUC*	*#Genes*	*accuracy*	*AUC*	*#Genes*
Single BPSOPG1	96.20	95.1	556	87.96	89.10	292	90.32	92.53	394
Single BBHA	98.73	97.6	303	92.48	93.66	189	94.62	95.5	150
Hybrid BPSOPG1/ BBHA	**100**	**100**	**14**	**98.49**	**98.36**	**42**	**100**	**100**	**24**

Table 3.5 The best subsets of genes/probe set IDs which found by BPSOPG1/ BBHA/ SPLSDA approach

dataset	AUC (#of genes)	Gene symbol
GSE25136	100 (14)	RS1, ZNF407, LOC101928625 /// MED21, LRRC59, EVC, ICAM1, HGD, IGHG1 ///MIR8071-2, MAP2K5, ZSCAN12, NPAP1, COX16 /// SYNJ2BP-COX16, LOC101060373 /// NOMO1 /// NOMO2 /// NOMO3, EMC10
(GSE70769, GSE70768)	98.80 (42)	NRN1, HP, BG214874, CACNG5, LOC100129222, RNASE6, RAB6A, GPBP1L1, SNX10, ATP6V0D2, TTC3, POLR1C, LOC100128525, NDUFA5, LOC651503, SDHALP1, KLC4, RPESP, APOE, FBXO42, FAM179B, SLC43A3, BID, TJP2, LOC652640, LOC642342, LOC391358, TPM2, DPYSL4, LYSMD1, SNRK, SLC26A5, KIAA1012, AY375451, CD245475, TTTY17A, STXBP6, CDKN1B, LOC648695, LOC100127884, C16orf5, AI792205
GSE31684	100 (24)	MPDZ, TMEM245, EPB41L3, 236041_at, LPAR2, SBNO1, BCAS4, YIPF1, DDX41, SOX6, SYTL4, CNIH2, ZNF395, CAPN13, UBE2W, LOC100505902, 210848_at, FAM115A /// LOC100294033, LAMP1, MAP2K5, ZMIZ2, JMJD7 /// JMJD7-PLA2G4B, RPAP2, NDUFAF5

Table 3.6 Extracted Rules by FURIA for the best subsets of genes

dataset	AUC	Extracted Rules with FURIA
GSE25136	**80.9**	**1**) **if** (RS1≤ 651.6) **and** (EMC10 ≥1355.8) **and** (LOC101928625(MED21) ≤ 1785.8) => class =Non-R (CF = 0.95) **2**) **if** (ICAM1≤ 27.6, 29.8]) and (ZSCAN12 ≥ 73.6) => class =Non-R (CF = 0.93) **3**) **if** (RS1≤ 529.5) **and** (MAP2K5≤ 39.7) **and** (LRRC59≤ 356.8) => class =Non-R (CF = 0.94) **4**) **if** (LOC101928625(MED21)≤ 1128.9) **and** (HGD ≥ 761.5) => class =Non-R (CF = 0.92) **5**) **if** (RS1 ≥ 660.4) **and** (ICAM1 ≥ 11.4) => class =R (CF = 0.96)

Table 3.6 (cont'd)

(GSE70769, GSE70768)	**79.60**	**1**) **if** (ATP6V0D2 ≤ 6.255025) **and** (CACNG5 ≤ 5.724875) **and** (DPYSL4 ≤ 8.592791) => class=R (CF = 0.95) **2**) **if** (BG214874 ≤ 6.485483) **and** (LOC391358 ≤ 5.804915) => class=R (CF = 0.93) **3**) **if** (KLC4 ≤ 5.684083) **and** (LOC100128525 ≤ 5.960904) => class=R (CF = 0.9) 4) **if** (ATP6V0D2 > 6.268148) **and** (SDHALP1≤6.733884) **and** (FAM179B≤7.130155) **and** (RNASE6≥6.56459) => class=Non-R (CF = 0.98) **5**) **if** (SLC43A3 ≤6.045351) **and** (CACNG5≥ 5.768709) and (RAB6A≥ 6.145977) => class=Non-R (CF = 0.96) **6**) **if** (KLC4 ≥ 5.717375) **and** (NDUFA5 ≤6.091466) => class=Non-R (CF = 0.92) **7**) **if** (NRN1 ≥ 6.420847) **and** (DPYSL4 ≥7.825856) => class=Non-R (CF = 0.96) **8**) **if** (LOC648695 ≤ 8.100189) **and** (GPBP1L1 ≤7.026667) => class=Non-R (CF = 0.97)
GSE31684	**74.70**	**1**) **if** (MPDZ≤ 5.993157) **and** (LPAR2≤ 4.790078) **and** (YIPF1≤ 8.503948) **and** (SBNO1≥5.080016) => class=Non-R (CF = 0.98) **2**) **if** (NDUFAF5≥3.908629) **and** (DDX41≤ 8.596171) => class= Non-R (CF = 0.94) **3**) **if** (EPB41L3≤ 3.768221) **and** (236041_at ≥3.039413) => class= Non-R (CF = 0.93) **4**) **if** (ZMIZ2≤ 4.526462) **and** (BCAS4≥ 2.376569) **and** (YIPF1≤ 8.082038) => class= Non-R (CF = 0.92) **5**) **if** (MPDZ ≥5.980839) **and** (NDUFAF5<3.91553) **and** (EPB41L3>3.768221) => class=R (CF = 0.94) **6**) **if** (EPB41L3≥6.690668) **and** (FAM115A(LOC100294033) ≥ 8.021252) **and** (BCAS4≤ 2.521673) **and** (DDX41≥8.403737) => class=R (CF = 0.93)

Table 3.7 The performance of eight gene selection approaches with three classifiers on selected gene subsets by RF-RFE

		GSE25136								(GSE70769, GSE70768)								GSE31684							
Classifier	Selection algorithm	*Accuracy*				*AUC*				*Accuracy*				*AUC*				*Accuracy*				*AUC*			
		Size	Best	Mean	SD	Best	Mean	SD	T	Size	Best	Mean	SD	Best	Mean	SD	T	Size	Best	Mean	SD	Best	Mean	SD	T
NB	*Fire Fly*	453.75	91.13	89.55	1.89	90.5	88.8	1.36	-	71.75	69.17	63.34	4.52	77.9	73.8	3.43	-	399.2	92.47	90.31	2.90	92.4	90.37	2.43	-
	Ant colony	429.5	92.4	89.86	1.79	94.5	89.9	3.34	-	60.25	66.91	63.15	3.57	79.3	76.67	3.21	-	392.7	92.47	89.78	2.55	92.3	90.17	2.16	-
	Bat search	348	91.13	88.91	1.59	91	88.75	2.46	-	57.75	66.91	63.34	3.02	80.1	75.02	4.68	-	394.7	91.39	87.89	2.68	93.8	90.22	2.80	-
	GA	435.5	91.13	86.70	3	91.6	86.25	3.71	-	41.25	69.17	63.52	4.93	79.4	76.27	3.09	-	377	90.32	88.97	1.02	90.2	87.65	2.36	-
	harmony search	201.25	89.87	87.97	2.19	93.4	90.25	3.13	-	25.5	70.67	67.47	2.24	78.7	76.15	3.03	-	191	91.39	86.55	3.97	95.5	89.27	4.45	-
	Geometric PSO	347.75	91.13	89.86	1.78	92.7	90.25	2.88	-	74.5	68.42	64.28	4	79.1	76.2	3.61	-	375.7	94.62	92.73	1.61	96.6	93.75	1.94	-
	FCBF	60	91.13			96.2			-	29	60.90			84.2			-	**19**	76.34			83.1			-
	CFS	75	89.87			94.50			-	38	60.15			83.3			-	33	82.79			91.2			-
	BPSOPG1/BBHA	**26.25**	97.46	95.13	1.63	97.37	95.64	2.24	-	**10**	82.70	78.37	3.15	81.80	78.1	2.87	-	33.5	95.69	94.34	1.02	94.1	93.61	0.49	-
KNN	*Fire Fly*	529	88.6	88.28	0.63	88.6	88.07	0.61	-	248	85.71	83.64	1.42	83.6	81.2	1.62	-	317.7	81.72	80.10	1.39	80	78.6	1.31	-
$(k=1)$	*Ant colony*	509.25	88.6	88.28	0.63	88.6	88.27	0.65	-	213	84.96	80.63	3.03	83.5	78.12	3.73	-	299.5	81.72	80.91	1.03	80	79.02	0.98	-
	Bat search	485.5	87.34	85.44	1.63	87.3	85.37	1.63	-	252.7	80.45	78.38	1.42	78	76.02	1.39	-	259	81.72	76.33	3.72	80.3	74.72	3.73	-
	GA	513.25	89.87	88.6	1.03	89.9	88.6	1.06	-	268.7	81.95	79.88	2.40	79.8	77.62	2.51	-	267.2	80.64	78.49	1.75	78	76.42	1.93	-
	harmony search	266	87.34	85.12	2.16	87.2	85	2.18	-	91	80.45	78.38	1.66	77.6	76.37	1.50	-	133.7	77.41	77.14	0.53	75.9	75.15	0.58	-
	Geometric PSO	487.25	89.87	88.6	1.03	89.8	88.55	1.02	-	236.7	84.21	81.57	1.99	82.6	79.45	2.49	-	309	83.87	80.90	2.96	82.2	79.12	3.52	-
	FCBF	60	77.20			77.21			-	29	75.93			71.8			-	19	68.81			66.4			-
	CFS	75	79.74			79.70			-	38	75.18			70.9			-	33	73.11			70.4			-
	BPSOPG1/BBHA	84.25	98.73	96.19	2.30	97.2	95.4	2.32	-	52.5	90.22	89.04	0.8	89.10	87.52	1.33	-	72.75	90.32	89.24	1.96	88.1	87.05	1.46	-
SPLSDA	*BPSOPG1/BBHA*	40.66	**100**	**100**	0	**100**	**100**	0		58	**98.49**	**96.98**	1.22	**98.80**	**96.70**	1.46		38.75	**100**	**99.19**	1.03	**100**	**99.25**	**0.67**	

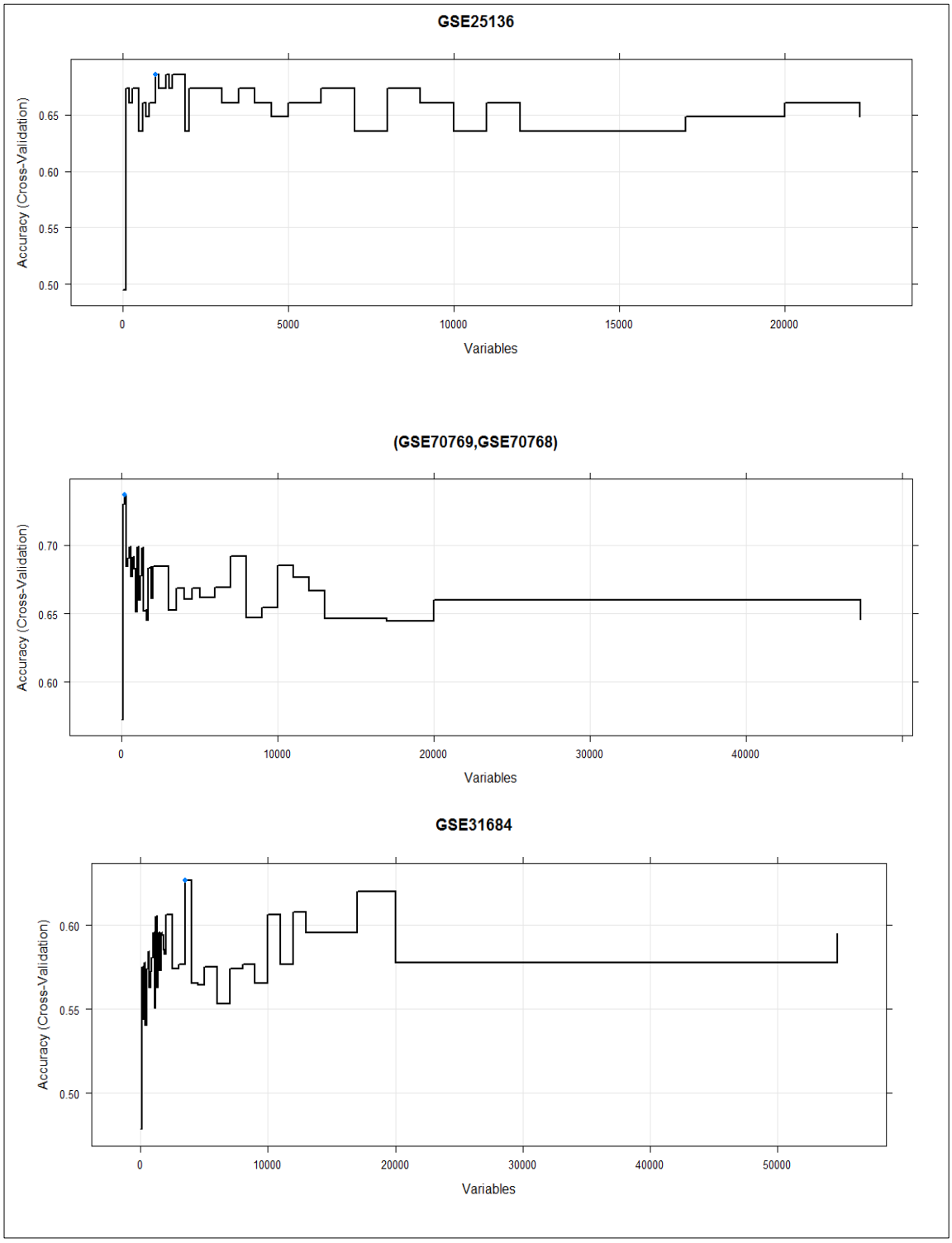


Figure 3.2 Choosing Size of the Signature by the RF-RFE Process

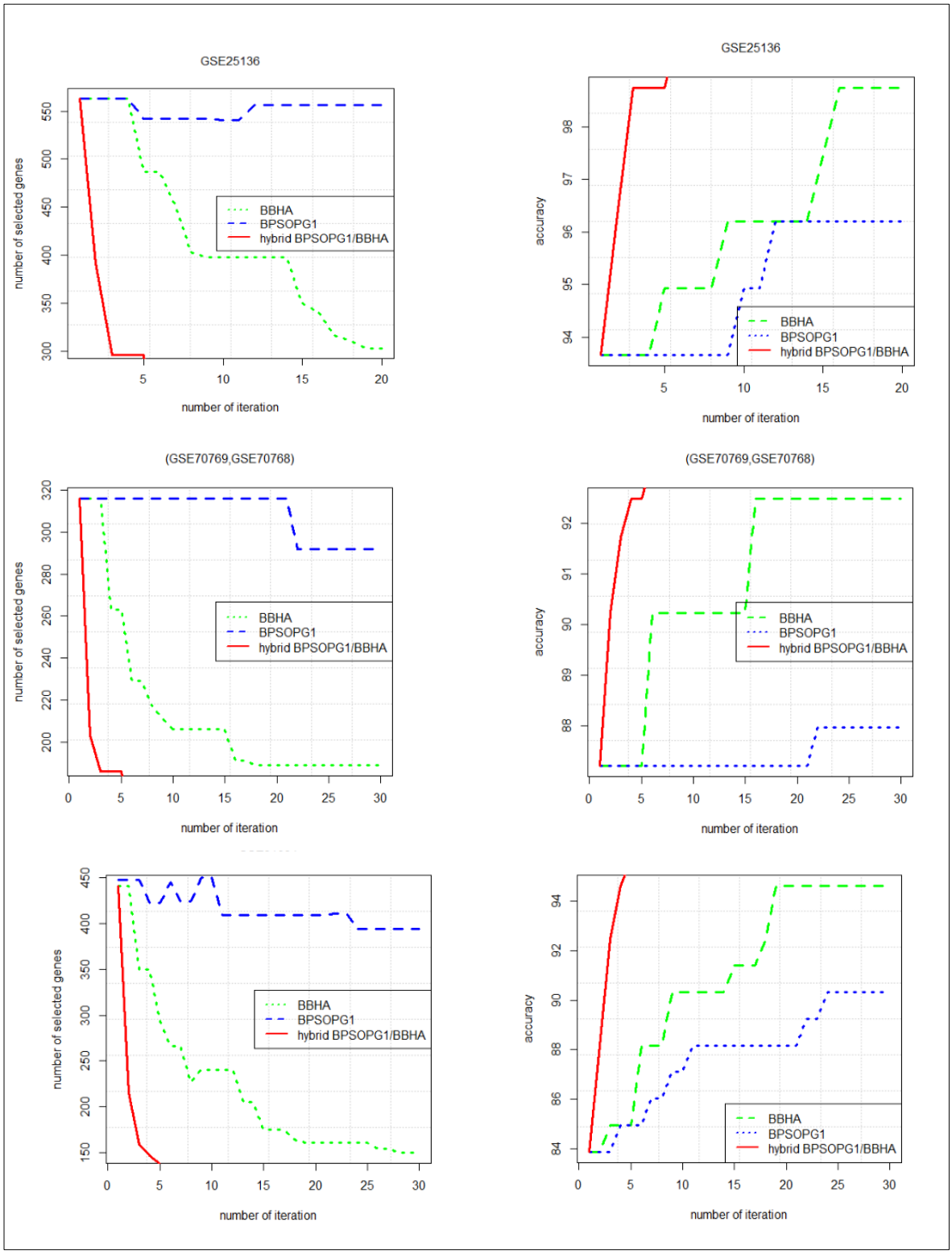


Figure 3.3 Variation Curves of Classification Accuracy and Number of Optimized Genes for BBHA, BPSOPG1, and hybrid Approach

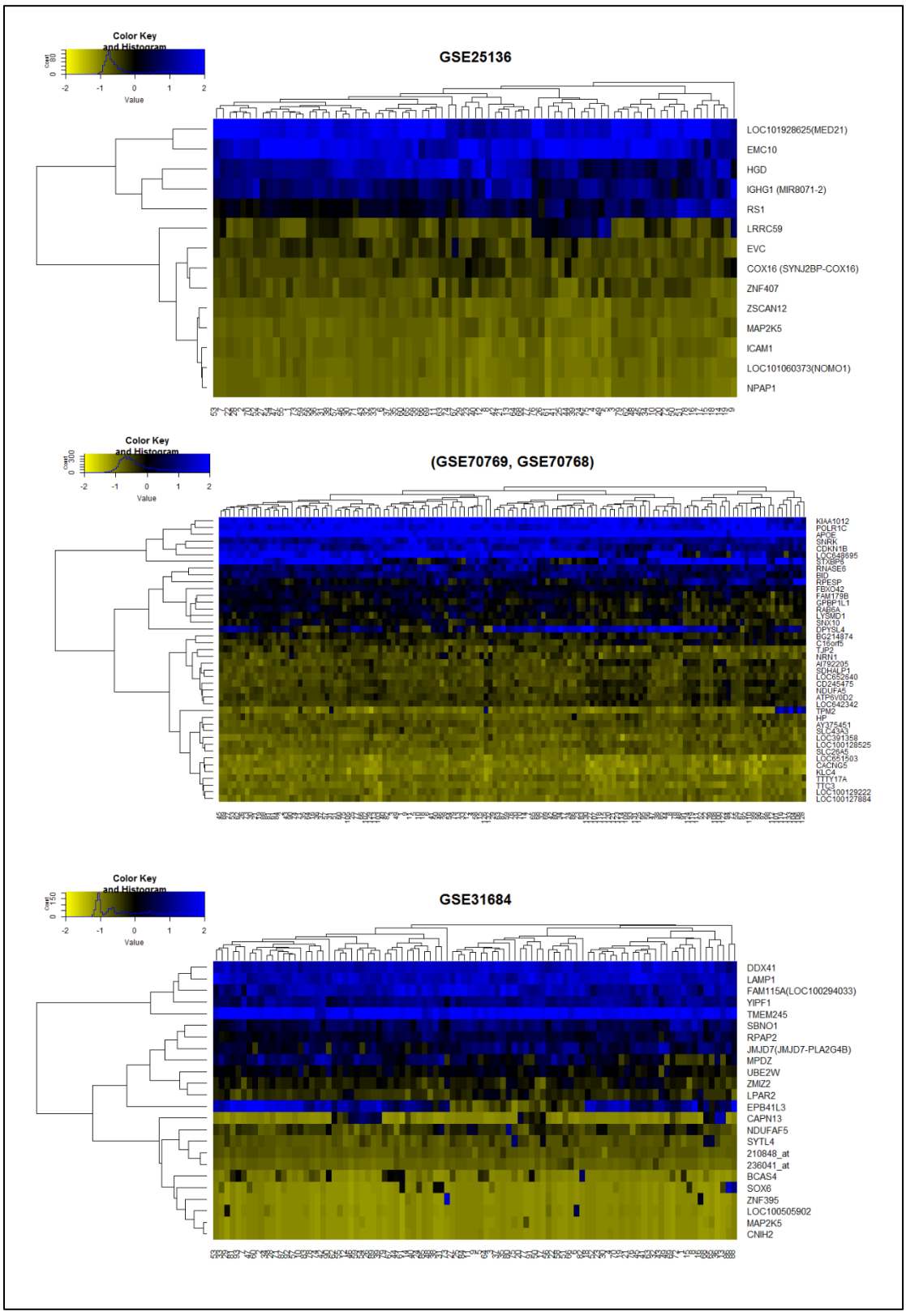


Figure 3.4 Heat Map Representation with Two-Way Hierarchical Clustering Based on Correlation Distance and Average Linkage for Optimal Gene Subsets.

3.5 Discussion

Because of more emphasizing on presented PSOPG1/BBHA/SPLSDA hybrid algorithm, we perform further analysis with more details on these results. The optimal gene signatures generated by training with the GSE25136, (GSE70769, GSE70768) and GSE31684 datasets contained 14, 42, and 24 genes, respectively. In this section, we individually examine these genes for relevance in the diagnosis of recurrent PCa (GSE25136, GSE70769, and GSE70768) and recurrent bladder cancer (GSE31684).

Of the fourteen markers of the prognostic gene signature derived from the GSE25136, RS1was placed on the top of ranked genes with corrected P-value <0.004 (using "limma" package in R). For GSE70769, GSE70768, and GSE31684, none of the genes were identified as differentially expression genes (DEGs).

The role of identified optimal gene signatures in the PCa development are described as follows.

The RS1gene: RS1 protein levels has been reported to increase in patients with recurrence PCa within 5 years, which is negatively correlated with AR expression, in addition a meta-analysis showed that the RS1 gene was amplified in up to 32% of castration-resistant prostate tumours. The ICAM1 gene: It has been reported that a region on chromosome 19p13.2 containing the genes ICAM1, ICAM4 and ICAM5 influences breast and PCa risk. ICAM1 as the target gene of microRNA-296-3p promotes metastasis of PCa by possible enhancing survival of natural killer cell-resistant circulating tumour cells. It is also known as a potential target for PCa therapy and prognosis. Suppression of IGHG1 gene expression by siRNA leads to growth inhibition and apoptosis induction in human PCa cell. MAP2K5 (MEK5) is known as a differentially methylated promoters in PCa cell lines. MEK5 overexpression is associated with metastatic PCa, and stimulates proliferation, MMP-9 expression and invasion. ZSCAN12 as the prostate tumor DNA methylation is associated with cigarette smoking and adverse prostate cancer outcomes . Leucine-rich repeat-containing protein 59 (LRRC59) mediates nuclear import of cancerous inhibitor of PP2A in PCa cells . The expression of this gene associates with PCa recurrence, metastases, and/or PCa-specific death after radical prostatectomy. The TJP2 gene is known as a differentially methylated promoter in PCa cell lines. It has been reported that rearrangement of TTC3 gene occurs in PCa. The expression of TPM2 are downregulated in PCa and there is a significant negative correlation between the expression of the TPM2 and the prognosis

of PCa. LOC391358 has been reported as the top 500 most abundant (by RPKM) in one or more LuCaP in. SLC43A3, CACNG5, and HGD have been reported as DEGs in neuroendocrine vs adenocarcinoma LuCaPs (n=24) FDR < 0.01. The mRNA level and genetic polymorphism of haptoglobin (HP) is related to the prognosis of advanced castration-resistant PCa patients treated with peptide vaccination. Apolipoprotein E (APOE) gene polymorphism influences aggressive behavior in PCa cells by deregulating cholesterol homeostasis . BID as the anti-apoptotic member of the Bcl-2 family proteins is overexpressed in PCa and is promising molecular target for modulating chemo resistance of PCa. CDKN1B also known as p27Kip1 is a useful prognostic marker for PCa. The use of this protein in clinical practice can improve prognosis prediction, disease screening and treatment response of PCa. Also, a polymorphism in the CDKN1B gene is associated with increased risk of hereditary PCa.

We apply FURIA on optimal subset of genes which obtained by the introduced hybrid approach in Table 3.6 to use for extracting rules between them. We find 5 rules with 80.9% AUC using LOOCV for GSE25136. In GSE25136, classification is performed using eight genes (RS1, EMC10, LOC101928625 (MED21), ICAM1, ZSCAN12, MAP2K5, LRRC59, and HGD).

The rest of the Table 3.6 shows the rules for the (GSE70769, GSE70768) and GSE31684. The obtained AUC for these GEO datasets on best subset of genes are 79.60% with 8 rules and 74.70% with 6 rules, respectively. Finally, we compared the performance of our predictive model derived from the 3 GEO datasets with those of the original studies that were previously derived from the same datasets [48, 49, 67]. For GSE25136, our signature performed better than the predictive nomogram (AUC: 0. 86, sensitivity: 90, specificity:73), genetic prognostic signature (AUC: 0. 90, sensitivity: 90, specificity:85) , and the hybrid predictive model (combination of genetic and nomogram data) (AUC: 0. 96, sensitivity: 90, specificity:95) [49], in terms of AUC, sensitivity, and specificity. In [49], LOOCV AUC of LDA classifier has been used. We utilized SPLSDA classifier and achieved (AUC: 100, sensitivity: 97.5, specificity: 100) with only 14 genes. None of candidate genes are common with original study.

In original study [48] for (GSE70769, GSE70768) datasets, based on statistical analysis, 100 discriminating genes were identified to consistently predict biochemical relapse for PCa. The performance of 100-gene set in predicting relapse has been reported as the log rank p-value (0.0330). However, we combined these two datasets and by applying soft

computing technique identified a high-performance signature (AUC: 98.80) consisted of only 42 genes. None of candidate genes are common with original study.

In original study [67] for GSE31684 dataset, based on statistical analysis (corrected P-value) for patients who died of bladder cancer which labeled as a (recurrence/death of disease), no significantly DEGs were identified to consistently predict biochemical relapse for bladder cancer. Therefore, the restricting were performed on samples and only the patients with pathologically organ-confined MI (pT2N0) tumors were considered for identifying DEGs in the reference paper. However, we did not perform any restricting on samples and reached perfect performance with only 24 genes for predicting recurrence of bladder cancer. The biologically relevant of these informative genes which are extracted by the proposed method to build high performance prediction model is remains unknown.

3.6 Chapter Summary

Gene selection plays a crucial role in developing a successful disease diagnostic system for microarray data. In order to identify the most beneficial genes for classification, this chapter proposed a hybrid approach based on BPSOPG1 and BBHA algorithm which is combined with SPLSDA classifier.

The experimental results running on four GEO datasets and statistical analysis have demonstrated that the proposed approach compare with many other methods, leads to a better performance in term of accuracy, AUC, and number of selected genes. The proposed method not only effectively reduced the number of genes, but also obtained a high classification accuracy. The obtained results indicate that the BPSOPG1-BBHA/SPLSDA is a useful tool for selecting informative genes in clinical datasets. Moreover, It was also shown that applying BBHA as the local optimizer for BPSOPG1 can significantly improve the performance of BPSOPG1 and help it to avoid being trapped in a local optimum.

CHAPTER 4

META ANALYSIS OF MIRNA EXPRESSION PROFILES FOR PROSTATE CANCER RECURRENCE

4.

4.1 Introduction

Prostate cancer (PCa) is a leading reason of death in men and the most diagnosed malignancy in the western countries at the present time. It is more widespread in older men (above 65 years old). After radical prostatectomy (RP), nearly 30% of men develop clinical recurrence with high serum prostate-specific antigen levels. An important challenge in PCa research is to identify effective predictors of tumor recurrence.

Alterations in microRNAs equally contribute to PCa initiation and progression. Several miRNA microarray studies have been conducted in recurrence PCa, but the results varies among different studies.

Meta-analysis utilizes statistical methods to contrast and combine results from multiple studies in the hope of increasing the statistical power and reproducibility over individual studies and identifying patterns across studies. Therefore, meta-analysis of some miRNA expression datasets of PCa progression can give a potentially significant list of co-deregulated miRNAs in PCa progression, which is important to specify pathways in which the miRNAs of interest and their target genes are involved.

4.1.1 Chapter Goals

The goal of this chapter is to analyze miRNA expression profile in PCa progression considering 5 studies (6 datasets), in order to increase the probability of revealing truly

significant deregulated miRNA genes, which should have higher potentials to be utilized as new biomarkers for the disease. This meta-analysis increases the significance of results. Specifically, the highlights of this chapter are:

- We conducted a meta-analysis of 6 available miRNA expression datasets and identified a panel of co-deregulated miRNA genes.
- Co-deregulated miRNAs were investigated for network interrelation by MIROB.
- For TF and target genes of co-deregulated miRNAs Gene Ontology (GO), KEGG pathways, and Pathway Commons Analysis were applied.
- A ROC test was performed for the candidate miRNAs in PCa recurrence using the extracted normalized expression signal values of each miRNA in each GEO datasets
- The best combination of candidate co-deregulated miRNAs that predicted BCR of PCa after RP with very high accuracy was identified using a soft computing technique (PSO/ multinomial logistic regression) for each GEO dataset.
- A comparison between expression of the co-deregulated microRNAs in recurrent vs. non-recurrent PCa was done by plotting boxplots

4.1.2 Chapter Organization

The remainder of this chapter is organized as follows. The section 4.2 presents the design of the experiments. The results are presented in section 4.3. The section 4.4 presents the discussions .The section 4.5 provides a summary of this chapter.

4.2 Design of Experiments

4.2.1 Literature Analysis

There are a limited number of reports in the literature studied miRNAs in PCa progression. We systematically queried for these studies from PubMed database.

The following Medical Subjective Heading (MeSH) and Embase tree were used: “recurrence” or “recurrence” and “prostatic neoplasms” or “prostate cancer” and “micrornas” or “microRNA” and “gene expression” or “expression”. In addition, publicly available microRNA data sets were searched by “RISmed” package in R to ensure no relevant studies were missed. Through database searching, a total of 24 studies was identified. Of these, 19 studies were retained after rejecting repetition.

According to the title and abstract, a total of 14 studies was excluded. Review, case report, animal experiment, no association with PCa, and experiment on DNA microarray were the reasons for excluding these articles. The full-text articles were evaluated for the remaining 5 studies, and all of them (6 datasets) were retained in the final meta-analysis. These miRNA data sets were obtained from the National Centers for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/).

4.2.2 MiRNA Microarray Datasets

In this study, a total of six microRNA datasets related to the recurrent PCa after RP (GSE55323 [54], GSE26245 and GSE26247 [55], GSE65061 [61], GSE62610 [56], and GSE46738 [58]) met the inclusion criteria and were selected for meta-analysis.

In the GSE55323, a total of 41 recurrent and 41 non-recurrent tumors after RP, which have been obtained from Baylor College of Medicine Prostate Cancer program, have been considered for performing miRNA profiling. Recurrence has been defined as a two consecutive serum PSAs greater than 0.2 ng/ml. To carry out microarray analysis, 20 samples from each group have been profiled using miRNA microarray chips.

In GSE26245 and GSE26247, total RNA from 71 formalin-fixed-paraffin-embedded (FFPE) specimens with known long-term outcome have been used for performing DASL expression profiling with a custom-designed panel of 522 PCa relevant genes. Recurrence has been defined as a two consecutive serum PSAs greater than 0.2 ng/ml. In the GSE26245, samples from 71 patients (29 with BCR and 42 without BCR) and in the GSE26247, samples from 82 patients (29 with BCR and 53 without BCR) have been used. In this study, the samples with unknown BCR have been removed.

For the GSE65061, total RNA has been extracted from tumor-enriched 1mm cores from 43 RP paraffin tissue blocks. Tissue isolated at the time of RP has been utilized for miRNA profiling. Thirty-six months has been considered as the cutoff, as it was near the median time to recurrence. From 43 patients, 19 were labeled as the samples with BCR ($\leq$ 36 months) and 24 as the samples without BCR ($>$ 36 months).

In the GSE62610, total RNA has been taken from tumor-enriched 1.5 mm cores in diameter from 36 formalin fixed paraffin embedded (FFPE) specimens. Then biochemical failure has been defined as two consecutive measurements of PSA > 0.2

ng/ml. From 36 patients 22 has been classified as the samples with BCR and 14 as the samples without BCR. In the GSE62610 most of microRNAs have null expression. After excluding miRNAs with no expression in any of the samples, 536 miRNAs have been kept for further analysis.

For GSE46738, total RNA has been taken from tumors from the 51 patients that underwent an RP by the same surgeon to treat localized PCa. In the GSE46738, the BCR status of samples is not mentioned explicitly. In the present study, according to the expression level of the miRNAs with greater statistical power, which has been reported in Table 3 of the study [58], tumors were divided into the positive BCR and the negative BCR by using clustering techniques. In GSE46738, from 51 samples, 34 were classified as the samples with BCR and 17 as the samples without BCR. Table 1.3 has provided detailed information of each dataset.

The microRNA microarray datasets were obtained from GEO NCBI. All GEO series matrix files (GSE), platform sets, and annotation files were downloaded and parsed using 'GEOquery' package of Bioconductor 3.2 in R version 3.2.2. To identify Differentially Expressed (DE) miRNAs in each individual dataset, moderated t-test was used.

4.2.3 Statistical Analysis

The meta-analysis was performed using the 'MetaDE' package in R. The moderated t-statistic was utilized to identify DE miRNAs in each individual dataset. The Fisher and AW were used to combine the p-values from moderated t-test for meta-analysis. Fisher's method is a summation of –log (p-value) across studies. An adjusted p-value of < 0.05, based on the False Discovery Rate (FDR) using the Benjamini–Hochberg procedure [91] was used to select DE microRNA genes.

4.3 Results

4.3.1 Identification of Candidate Prostate Cancer Recurrence Markers for Pathway Analysis

To identify a common DE microRNAs for PCa recurrence, six miRNA studies (Table 4.1) were analyzed using "MetaDE" package in R. First, individual analysis was performed and the moderated t - test was used to calculate the p-values which frequently used in meta-analysis. Then, AW and Fisher's method were utilized to combine the p-

values and find miRNAs that were differentially expressed between samples with recurrence and non- recurrence (+/− BCR) across all studies. From miRNA microarray meta-analysis, we identified a total of 37 DE miRNAs including 15 overexpressed and 22 under expressed microRNAs across at least two datasets under the significance threshold of adjusted p-value < 0.05. Figure 4.1 shows the number of DE microRNAs against FDR obtained from individual analysis as well as meta-analysis. It is clearly seen that the meta-analysis has detected more candidate markers. Figure 4.2 shows the heat map of those 37 microRNAs. A complete list of DE microRNAs has been provided in Table 4.2. The miR-449A, miR-484, and miR-579 were among the most significant overexpressed genes, while miR-449B, miR-1, miR-137, miR-370, miR-375 were the most under expressed genes across all miRNA datasets (See Table 4.2).

4.3.2 MiRNA Genes Network

MIROB tool was used to perform regulatory microRNA network analysis to identify regulators responsible for the observed patterns in miRNA meta-analysis studies. The interaction network was constructed between DE microRNAs, TF and target genes associated with the complete set of DE (Figure 4.3). Twenty four of DE miRNAs were found in the network. The details of those miRNA gene networks have been given in Table 4.3. Key targets, ontology information on target genes, TF and a descriptive analysis of expression of the DE miRNAs have been summarized in this table. In addition, it shows that DE miRNAs are highly associated with colorectal, PCa, breast, and gastric cancer.

4.3.3 Further Enrichment Analysis

We performed gene set enrichment analysis by EnrichR tool, using the complete list of key targets and TF of DE miRNAs. GO terms and biological pathways were significantly overrepresented in the gene list if they showed an adjusted p-value < 0.05. Results for gene ontology and enriched biological pathways (KEGG, Reactome) have been shown in Table 4.4, Table 4.5, and Table 4.6, respectively. DE microRNAs in meta-analysis results were associated with the enriched pathways with adjusted p-value < 0.05, including “MicroRNAs in cancer (hsa05206)”, “Pathways in cancer (hsa05200)”, “Proteoglycans in cancer (hsa05205)”, “PI3K-Akt signaling pathway (hsa04151)”, “Prostate cancer (hsa05215)” and “Signal Transduction (R-HSA-

162582)". The most important GO terms associated with key targets and TF of DE miRNA genes included "regulation of epithelial cell proliferation (GO: 0050678)", "tissue morphogenesis (GO: 0048729)", "regulation of cellular response to stress (GO: 0080135)", and "positive regulation of cellular component movement (GO: 0051272)".

To further investigate the function of DE miRNAs, we mapped them to the KEGG database. Eleven of them (miR-1, miR-125A, miR-133A, miR-133B, miR-137, miR-199A, miR-221, miR-28, miR-324, miR-363 and miR-449A) were found in the "miRNAs in cancer" pathway (KEGG-ID: hsa05206; Figure 4.4) with adjusted *P*-value of 7.554e-15 (Table 4.5). Moreover, common pathway analysis revealed that TCF3, MYC, MAX, CYP26A1, and SREBF1 significantly interact with DE miRNAs (Figure 4.5).

4.3.4 Diagnostic Performance

We assessed the diagnostic potential of the 37-miRNA signature identified by meta-analysis. ROC curve analysis gave AUCs from 0.55–0.84 for miRNAs set in each GEO dataset (See Figure 4.6). To investigate whether a miRNA signature may increase diagnostic accuracy over 37-miRNA signature, we employed a soft computing technique (PSO/ logistic regression) and trained and tested on miRNA expression profiles. The best subset of DE miRNAs was identified in each GEO dataset and shown in Table 4.7.

Notably, the discriminating power of the identified signatures in each GEO dataset is higher than the case where 37-miRNA classifier was considered. For the best subset of DE miRNAs in each GEO dataset, the ROC curve analysis gave AUCs from 0.75–0.97 (See Figure 4.7). The highest diagnostic accuracy (97%) was given for GSE55323 with 11-miRNAs. Moreover, in order to correctly classify BCR+ vs. BCR- samples, simple rules were extracted using a decision tree classifier (Table 4.7). Among six GEO datasets, rules with high diagnostic potentials were extracted for GSE46738 and GSE26247.

Finally, a comparison between the expressions of co-deregulated microRNAs in BCR+ vs. BCR- was done by plotting boxplots (Figure 4.8). The boxplots were drawn for co-deregulated microRNAs that are involved in the PCa pathway.

Table 4.1 The 37 shared significantly deregulated miRNAs identified in the meta-analysis.

	GSE55323		GSE26245, GSE26247		GSE65061		GSE62610		GSE46738		Merged data		
Down regulated	***P*-Value**	**FC**	***P*-value**	**FC**	***P*-value**	**FC**	***P*-value**	**FC**	***P*-value**	**FC**	**Meta. Stat**	**Meta. *P* value**	**Meta. FDR**
miR-1	0.0039	-1.77	0.0872	-1.2	0.0125	1.72	0.799	-1.08	0.7681	1.07	25.6944	0.0039	0.0342
miR-133A	0.0256	-1.19	0.01529	-1.22	0.02678	1.62	0.5653	-1.17	0.1863	1.35	24.066	0.00245	0.040833
miR-133B	0.0041	-1.41	0.3924	1.1	0.0188	-1.23	0.9675	-1.01	0.5129	1.16	22.2085	0.0089	0.0294
miR-137	0.3072	1.08	0.032	-1.18	0.0091	-2.69	0.0129	-11.49	0.1852	-1.08	30.7251	0.0005	0.019
miR-221	0.00065	-1.51	0.7294	1.03	5.40E-05	2.33	0.506	1.22	0.099	-1.29	19.26	0.00074	0.0477
miR-340	0.8665	1.01	0.8959	1	0.6935	-1.04	0.31227	-1.31	<0.001	-1.89	20.7435	0.00044	0.04
miR-370	0.395	-1.12	0.1995	-1.26	0.1671	-1.3	0.0422	-2.15	0.0037	1.85	26.1757	0.0033	0.0342
miR-449B	0.0485	-1.2	0.2516	-1.14	NA	NA	0.00318	-3.97	0.0676	1.39	22.501	0.000381	0.044
miR-489	0.2688	-1.08	0.9699	-1	0.6031	1.04	0.0164	-1.67	0.0179	-1.54	19.9117	0.0074	0.0455
miR-492	0.8683	1.11	0.0269	-1.07	0.0001	-1.4	0.8009	-7.16	0.7347	1.09	26.547	0.0012	0.025
miR-496	0.001	-1.69	0.1743	-1.17	0.6391	-1.05	0.04696	-3.48	0.05464	-1.28	26.5035	0.0003	0.008
miR-541	0.4518	1.06	NA	NA	0.0042	-1.51	NA	NA	<0.001	-1.69	21.215	0.00032	0.05
miR-572	0.2212	-1.21	0.2326	1.08	0.005	-1.31	0.0531	-1.41	0.3638	1.25	24.4192	0.00446	0.0416
miR-583	0.5071	1.07	0.442	1.1	0.0089	-1.42	NA	NA	0.00061	-1.51	27.205	0.00061	0.048
miR-606	0.3955	1.09	0.2885	-1.21	0.1067	-1.22	0.9715	-1.03	0.001	-1.78	22.7104	0.0042	0.05
miR-624	0.1552	1.12	0.6498	1.07	0.05	-1.26	0.0296	-1.98	0.0002	-1.48	20.58	0.00039	0.038
miR-636	0.6497	1.03	0.4884	1.05	0.8496	-1.03	<0.001	-2.06	0.5493	-1.09	95.9233	<0.001	<0.001
miR-639	0.004	-1.16	0.85501	1.04	0.2339	-1.16	0.3246	-1.2	0.1879	1.13	19.5331	0.0082	0.0455
miR-661	0.9746	1	0.11	-1.11	0.04517	-1.29	0.00053	-1.32	0.27696	1.06	23.87	0.00125	0.028
miR-760	0.4702	1.13	0.2285	-1.21	0.0003	-1.46	0.2971	-1.3	0.1529	1.19	24.3088	0.0022	0.035
miR-890	0.489	-1.14	NA	NA	0.0442	-1.24	NA	NA	0.0002	-1.86	23.07	0.00014	0.013
miR-939	0.8377	1.03	NA	NA	0.0085	-1.32	NA	NA	0.0288	1.46	16.61	0.0023	0.049
Up regulated	***P*-value**	**FC**	***P*-value**	**FC**	***P*-value**	**FC**	***P*-value**	**FC**	***P*-value**	**FC**	**Meta. Stat**	**Meta. P value**	**Meta. FDR**
miR-125A-5P	0.24	-1.13	0.632	1.17	0.0011	1.58	0.06081	-1.46	NA	NA	22.9155	0.0028	0.038
miR-199A-3P	0.7639	-1.05	NA	NA	0.00172	1.78	0.344	-1.3	NA	NA	0.0016	0.00274	0.042
miR-28-5P	0.6761	-1.05	0.9039	-1.02	0.00041	1.47	0.3982	-1.22	NA	NA	24.05	0.0024	0.04
miR-301B	0.7513	-1.01	NA	NA	0.0049	1.59	0.0164	-1.76	0.717	-1.02	20.0917	0.0066	0.0455

Table 4.1 (cont'd)

miR-324-5P	0.147	-1.13	0.1263	-1.13	0.0001	1.55	0.6594	-1.13	NA	NA	32.2291	0.00065	0.01625
miR-361-5P	0.3474	-1.08	0.3478	1.21	0.00083	1.76	0.5897	-1.14	NA	NA	0.00077	0.00122	0.038
miR-363*	0.1773	1.14	0.2258	1.41	NA	NA	0.00038	-1.95	NA	NA	22.176	0.0005	0.044
miR-449A	0.0332	1.35	0.5059	1.08	0.6952	1.05	0.0007	-5.43	0.2048	1.38	26.5308	0.0031	0.0342
miR-484	0.152	1.3	0.2685	1.09	0.0049	1.19	0.1188	-1.42	0.1252	1.33	25.4578	0.0043	0.0342
miR-498	0.7157	1.03	0.2734	1.08	0.0147	1.37	NA	NA	0.0013	1.66	25.0151	0.0019	0.035
miR-579	0.1908	-1.1	<0.001	1.38	0.0338	1.23	0.6918	-1.16	0.8592	1.01	29.3	0.00025	0.01625
miR-637	0.5443	1.07	0.6948	1.01	0.2487	-1.17	NA	NA	0.0001	2.22	20.69	0.00055	0.0375
miR-720	0.0175	-1.69	NA	NA	0.0008	1.76	NA	NA	0.0043	1.96	25.13	7.30E-05	0.0125
miR-874	0.1547	-1.18	NA	NA	0.00321	1.55	0.00017	-2.05	0.30732	1.29	34.7751	0.000178	0.0208
miR-98	0.80911	-1.02	0.6266	1.07	0.00016	1.77	NA	NA	0.8112	1.03	23.903	0.0007	0.04625

"*", denotes the mature miRNA sequence. ''-'', represents ''not available''.

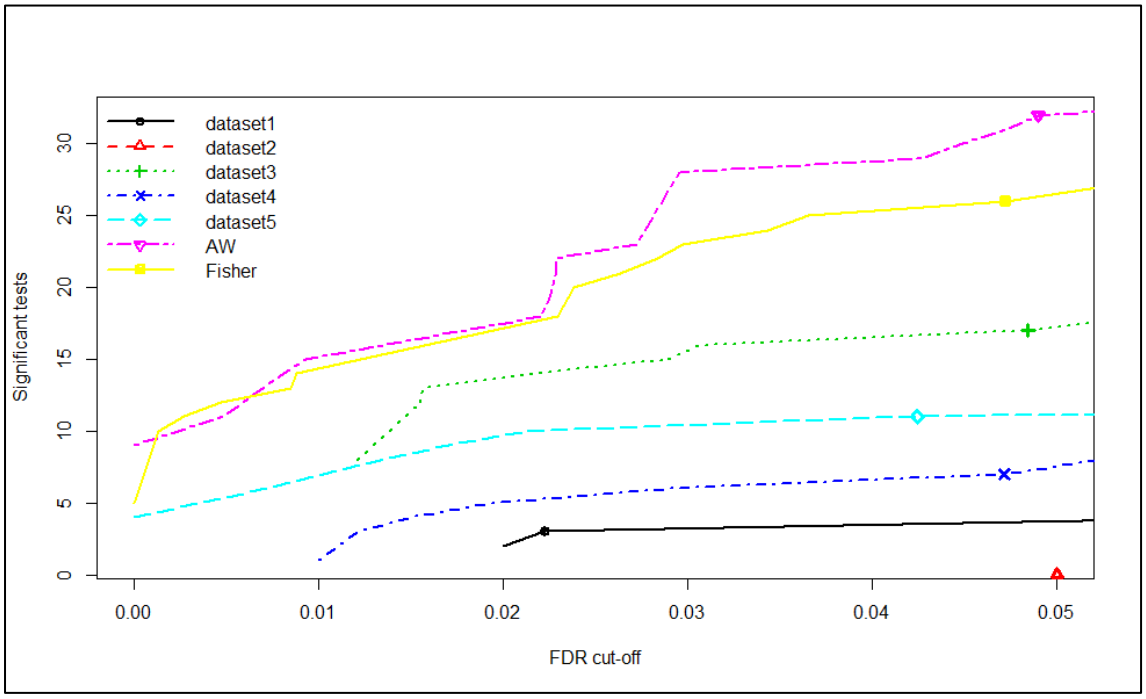


Figure 4.1 P-value (or FDR) vs number of detected miRNAs for individual analysis as well as meta-analysis. In each individual dataset, moderated-t statistics was used to generate p-values while adaptive weight and Fisher's methods were utilized to combine these p-values for meta-analysis. This figure is generated using the "MetaDE" package in R.

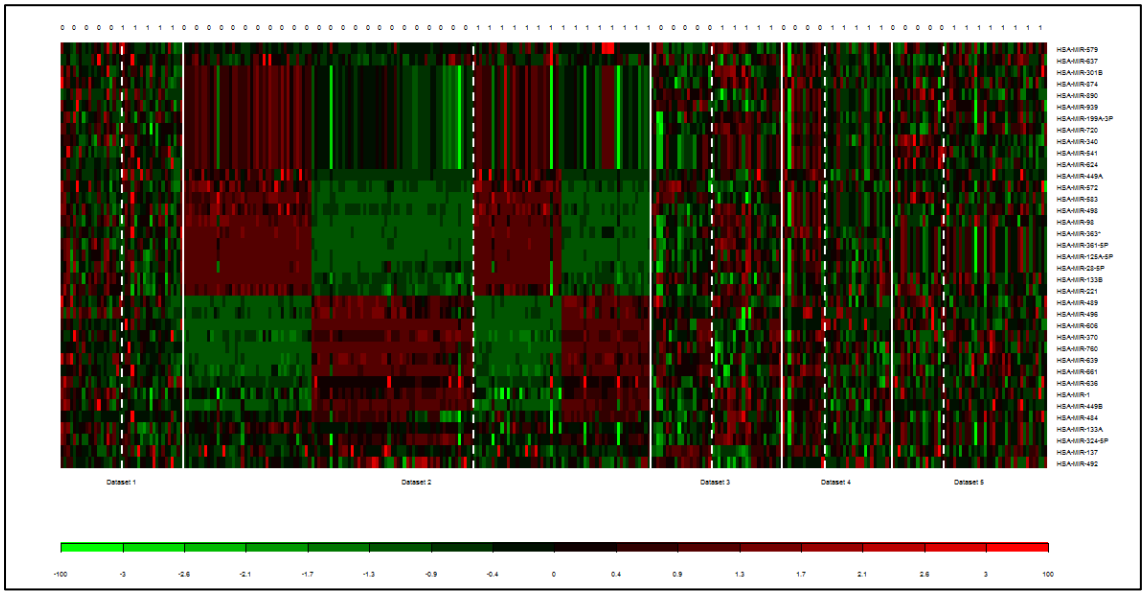


Figure 4.2 The heat map of the actual expression profiles for the 15 up- and 22 downregulated DE microRNAs obtained from the meta-analysis across at least two studies. The heat map is generated using the "MetaDE" package in R. The expression profiles greater than the mean are colored in red and those below the mean are colored in green. 0: Non-recurrence; 1: Recurrence.

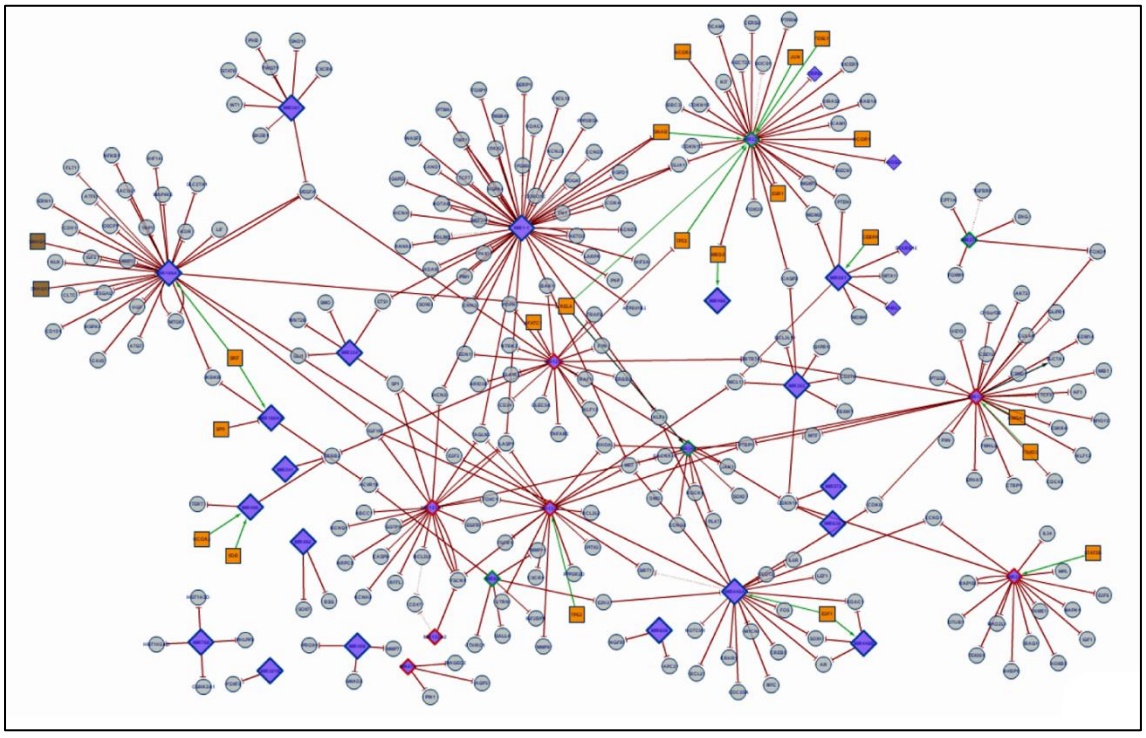


Figure 4.3 Network interrelation of DE microRNAs identified in the meta-analysis. Orange squares show TF. The circles show the targets of DE microRNAs. Green and red lozenges show up regulated and down regulated microRNAs in various types of diseases. The network was generated using a MIROB web tool to explore DE microRNAs relationships and collective functions.

Table 4.2 The details of 37 DE miRNAs that are involved in the interaction network, which has been drawn by MIROB.

MicroRNAs	Transcription Factors	Target genes	Disease influence (expression)	pathogenesis of a disease
miR-1	SNAI2	FOXP1, HDAC4, PDLIM5, PIM1, CCND2, CXCL12, PNP, LASP1, SNAI2, PAX7, KLF4, MET, FN1, PTMA, TAGLN2, PAX3, GJA1, SOX6, ATP6V1B2, LARP4, CNN3, HSPD1, HSPA4, POGK, PGM2, SERP1, NETO2, Srxn1, CAND1, ADAR, KIF2A, G6PD, MEF2A, KCNJ2, PPP2R5A, HCN2, TWF1, HCN4, KCNE1, ANXA2, ETS1	-	Metastasis, Angiogenesis, growth, Proliferation, Invasion, migration, Apoptosis, cell cycle arrest, differentiation, WNT signaling.
miR-125A	NFATC1, TP53	RHOA, FYN, CDKN1A, EDN1, BAK1, ARID3B, CD34, ERBB2, ERBB3, NTRK3, ELAVL1, TNFAIP3, PDPN, KLF13, CLEC5A, TRAF6, RAF1, ZBTB7A, VEGFA	Colorectal cancer (down)	Proliferation, Invasion, migration, differentiation, cell cycle arrest, Angiogenesis, survival, Sorafenib resistance, myeloid, differentiation
miR-133A	-	CD47, LASP1, GSTP1, FSCN1, ARPC5, TAGLN2, CASP9, KCNH2, CACNA1C, HCN2, KCNQ1, EGFR,IGF1R, RFFL, SP1, ABCC1, FOXC1, BCL2L1	**Prostate Cancer** (down)	Proliferation, Invasion, migration, Apoptosis, cell cycle arrest, colony formation, ERK pathway (MAPK pathway), Liver metastasis, Lung metastasis, **tumor growth**, Adriamycin (Adr) resistance, 5-fluorouracil resistance, cisplatin resistance
miR-133B	TP63	BCL2L2, MCL1, FGFR1, FSCN1, MET, PITX3, IGF1R, CXCR4, UTRN, SP1, RHOA, MMP9, EGFR, TAGLN2, LASP1, SIRT1, PPP2R2D, FOXC1, PTBP1	Colorectal cancer (up) **Prostate Cancer** (down) Gastric (down)	Proliferation, Invasion, migration, Apoptosis, cell cycle arrest, WNT signaling, **tumor growth**, cisplatin resistance, Cell growth
miR-137	FOXD3, HMGA1	CDK6, CDC42, SLC7A1, KDM1A, CSMD1, C10orf26, CACNA1C, TCF4, ESRRA, CTBP1, FMNL2, MIB1, GLIPR1, CSE1L, PTGS2, MITF, PXN, PTBP1, NF1, EPHA7, AKT2, ZBTB7A, HEY2, KLF12, MYO1C, CUL4A, FOXO1, CDK6.	Colorectal cancer (down), Gastric (down)	Metastasis, Angiogenesis, growth, colony formation, Proliferation, Invasion, migration, Apoptosis, **tumor growth**, Cell growth, cell cycle arrest, Stemness, cell viability, aerobic glycolysis, cell cycle
miR-199A1	SRF, SPI1, SNHG12, SNHG1, RELA	ST6GAL1, HSPA5, ATF6, ERN1, IKBKB, CACUL1, CAV2, MTOR, LIF, RELA, NFKB1, ATG7, CLTC, NLK, CDH1, SLC27A1, MAP4K3, CD151, YAP1, OSCP1, HIF1A, VEGFA, IGF1R, IGF2, FLT1, KDR, HGF, MMP2, E2F3, ACVR1B	-	Proliferation, Invasion, migration, Apoptosis, cell cycle arrest, Angiogenesis, colony formation, ERK pathway (MAPK pathway), **tumor growth**, cisplatin resistance, cell viability, Chemoresistance,

Table 4.2 (cont'd)

miR-221	FOSL1, SNAI2, RELA, JUN, ESR1, NCOR2, NCOR1, TP53	CERS2, TRPS1, DICER1, KIT, NOS3, BBC3, MBD2, CDKN1C, GJA1, ICAM1, CDKN1B, DIRAS3, RAB1A, HECTD2, TICAM1, PTPRM, MGMT, FOXO3, RECK, MDM2, PTEN, SOCS1, CASP3	Breast cancer (up), Colorectal cancer (up) Gastric (up)	Proliferation, Invasion, migration, Apoptosis, cell cycle arrest, Metastasis, Cell growth, motility, cell cycle progression, Chemoresistance, doxorubicin resistance, Radioresistance, survival, Sorafenib resistance
miR-28	STAT5B	STAT5B, CDKN1A, CCND1, HOXB3, NME1, N4BP1, OTUB1, TEX261, MAPK1, E2F6, MPL, BAG1, MAD2L1, RAP1B, IL34, IGF1	Colorectal cancer (down)	Proliferation, Invasion, migration, Apoptosis, cell cycle arrest, Metastasis, ERK pathway (MAPK pathway), P38 signaling, AKT signalling, PI3K signaling
miR-301B	-	FOXF2	-	
miR-324	-	SMO, GLI1, WNT2B, ETS1, SP1	-	Proliferation, Invasion, migration, cell cycle arrest, Metastasis, Radioresistance
miR-340	RELA	RELA, MET, ROCK1, PTBP1, SOX2, MITF, RHOA, PLAT, DMD, JAK1, CCNG2	Gastric(up)	Proliferation, Invasion, migration, differentiation, cell cycle arrest, Metastasis, **tumor growth**, Cell growth, stemness, aerobic glycolysis, cell viability, cell cycle progression, Senescence, JAK/STAT signaling
miR-361	-	STAT6, VEGFA, TWIST1, WT1, SH2B1, CXCR6, SND1, PHB	-	Proliferation, Invasion, migration, Apoptosis, Metastasis, colony formation, **tumor growth**, Cell growth, stemness
miR-363	-	CDKN1A, S1PR1, BCL2L11, CASP3, CD276, FBXW7, MCL1	-	Proliferation, Apoptosis, cisplatin resistance, cell viability, Chemoresistance, survival
miR-370	-	CPT1A, TGFBR2, FOXM1, FOXO1, ENG	Gastric(up)	colony formation, Proliferation, Apoptosis Chemoresistance, colony
miR-449A	E2F1, EZH2, MYCN	E2F3, CDC25A, MET, SIRT1, CDK6, BCL2, CCND1, CRHR1, LEF1, KLF4, NOTCH1, HDAC1, AR, IL6R, SOX4, CREB5, FOS, MYC.	**Prostate** Cancer (down) Gastric (down)	Metastasis colony, formation, Proliferation, Invasion, migration, Apoptosis, motility, EMT, cell cycle arrest, cisplatin resistance, differentiation, Cell growth, cell viability, Radioresistance, Senescence, Antiapoptosis

Table 4.2 (cont'd)

miR-449B	E2F1, AR	CDK6 CDC25A, HDAC1, SOX4	-	Proliferation, migration Apoptosis, Cell growth colony formation, cell viability
miR-489	-	SMAD3, MMP7, PROX1	-	Proliferation, Invasion, migration, Lung metastasis, Adriamycin (Adr) resistance, EMT
miR-492	-	BSG, SOX7	-	Proliferation, Oxaliplatin, resistance
miR-498	VDR, NCOA3	TERT, ERBB2	-	Apoptosis, **tumor growth**, Cell growth
miR-661	CEBPA	STARD10, PVRL1, MTA1, MCL1, MDM2, MDM4, PTEN	-	Proliferation, Invasion, migration, cell cycle arrest, Metastasis**, tumor growth**, motility, EMT
miR-760	-	CSNK2A1, HIST1H3D, HIST1H2AD, PHLPP2	-	Proliferation, colony formation, Senescence
miR-874	-	AQP3, PIN1, MAGEC2	Gastric (down)	Proliferation, Invasion, Apoptosis, colony formation, Cell growth, mTOR signaling
miR-939	-	APC2, NGFR	-	Proliferation, WNT signaling
miR-98	EZH2	ACVR1B, MMP11, EZH2, SALL4, IGF2BP1, CTHRC1	Gastric(up)	Angiogenesis, growth, Proliferation, Invasion, migration, Apoptosis, EMT, cell cycle arrest, WNT signaling,

Table 4.3 Top enriched Gene Ontology (GO) biological process identified by functional analysis of the target genes and TFs of the DE microRNAs in the meta-analysis. Gene sets functional analysis was performed using extended libraries of the EnrichR tool

GO-ID	**Description**	**Overlap***	**Adjusted P-value**
GO:0050678	regulation of epithelial cell proliferation	32/258	7.430E-18
GO:0048729	tissue morphogenesis	35/358	1.191E-16
GO:0080135	regulation of cellular response to stress	36/404	4.884E-16
GO:0051272	positive regulation of cellular component movement	31/296	1.209E-15
GO:0070482	response to oxygen levels	29/259	2.118E-15
GO:2001233	regulation of apoptotic signaling pathway	33/356	2.466E-15
GO:2000147	positive regulation of cell motility	30/287	2.699E-15
GO:0040017	positive regulation of locomotion	30/304	1.184E-14

* Overlap: indicates the number of hits from the meta-analysis compared to each curated gene set library.

Table 4.4 Top enriched KEGG pathways identified by functional analysis of the target genes and TFs of the DE microRNAs in the meta-analysis. Gene sets functional analysis was performed using extended libraries of the EnrichR tool.

Pathway ID	Name	Overlap	Adjusted P-value
hsa05206	MicroRNAs in cancer	56/297	3.476E-45
hsa05200	Pathways in cancer	55/397	4.152E-37
hsa05205	Proteoglycans in cancer	33/203	4.675E-24
hsa04151	PI3K-Akt signalling pathway	38/341	7.293E-22
hsa05215	Prostate cancer	23/89	1.259E-21
hsa05212	Pancreatic cancer	20/66	2.252E-20
hsa05218	Melanoma	19/71	2.982E-18
hsa05220	Chronic myeloid leukemia	19/73	4.676E-18
hsa04520	Adherens junction	19/74	5.524E-18
hsa04933	AGE-RAGE signalling pathway in diabetic complications	21/101	7.221E-18

Table 4.5 Top enriched Reactome pathways identified by functional analysis of the target genes and TFs of the DE microRNAs in the meta-analysis. Gene sets functional analysis was performed using extended libraries of the EnrichR tool.

Pathway ID	**Name**	**Overlap**	**Adjusted P-value**
R-HSA-162582	Signal Transduction	100/2465	4.220E-22
R-HSA-1266738	Developmental Biology	46/786	2.574E-14
R-HSA-1236394	Signalling by ERBB4	29/330	5.348E-13
R-HSA-166520	Signalling by NGF	33/450	8.395E-13
R-HSA-180292	GAB1 signalosome	19/125	1.065E-12
R-HSA-198203	PI3K/AKT activation	19/125	1.065E-12
R-HSA-5654695	PI-3K cascade:FGFR2	18/122	4.702E-12
R-HSA-1257604	PIP3 activates AKT signalling	18/122	4.702E-12

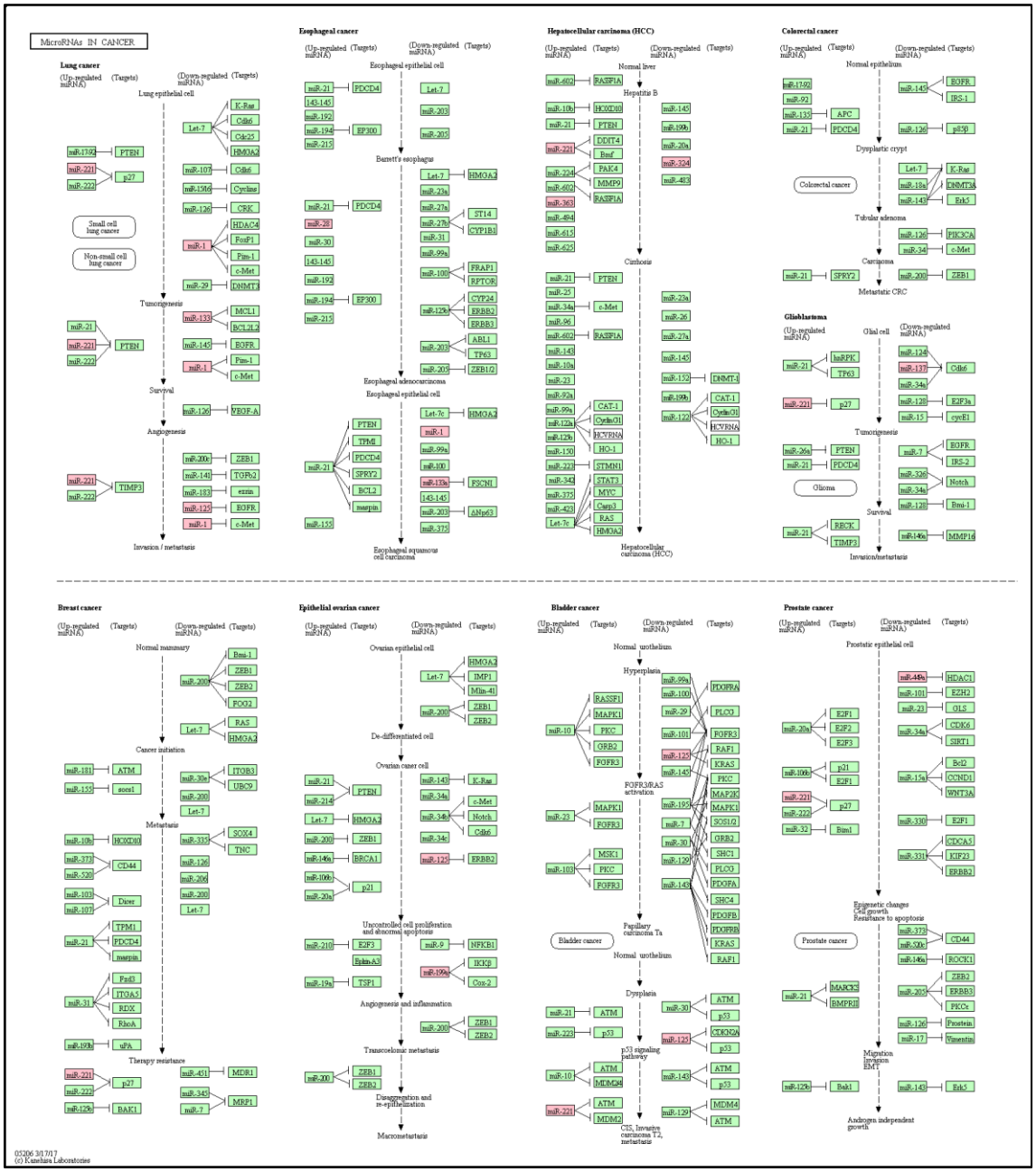


Figure 4.4 The most significant enriched KEGG pathway for the DE microRNAs identified from meta-analysis. The microRNAs in the red box indicates co-deregulated microRNA genes in our list. The DE microRNAs identified from meta-analysis were mapped to "microRNAs in cancer" pathway (KEGG-ID: hsa05206) by using the KEGG mapper web tool

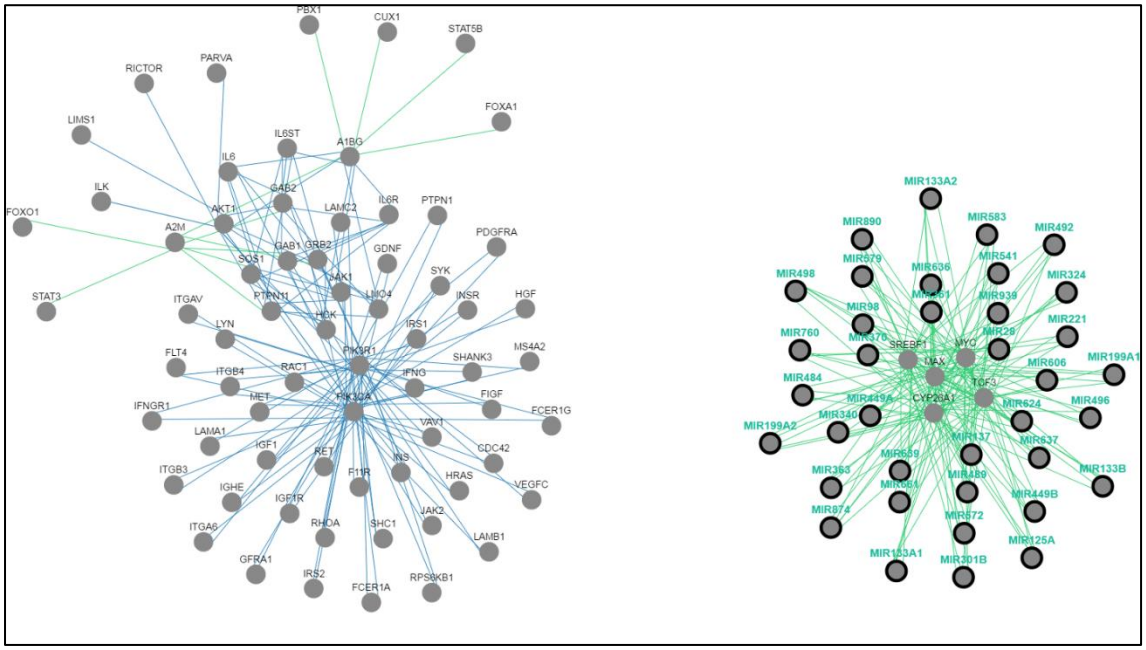


Figure 4.5 Common pathway analysis for DE microRNAs identified from meta-analysis. This analysis revealed that TCF3, MYC, MAX, CYP26A1 and SREBF1 are significantly interacting with candidate miRNA genes.

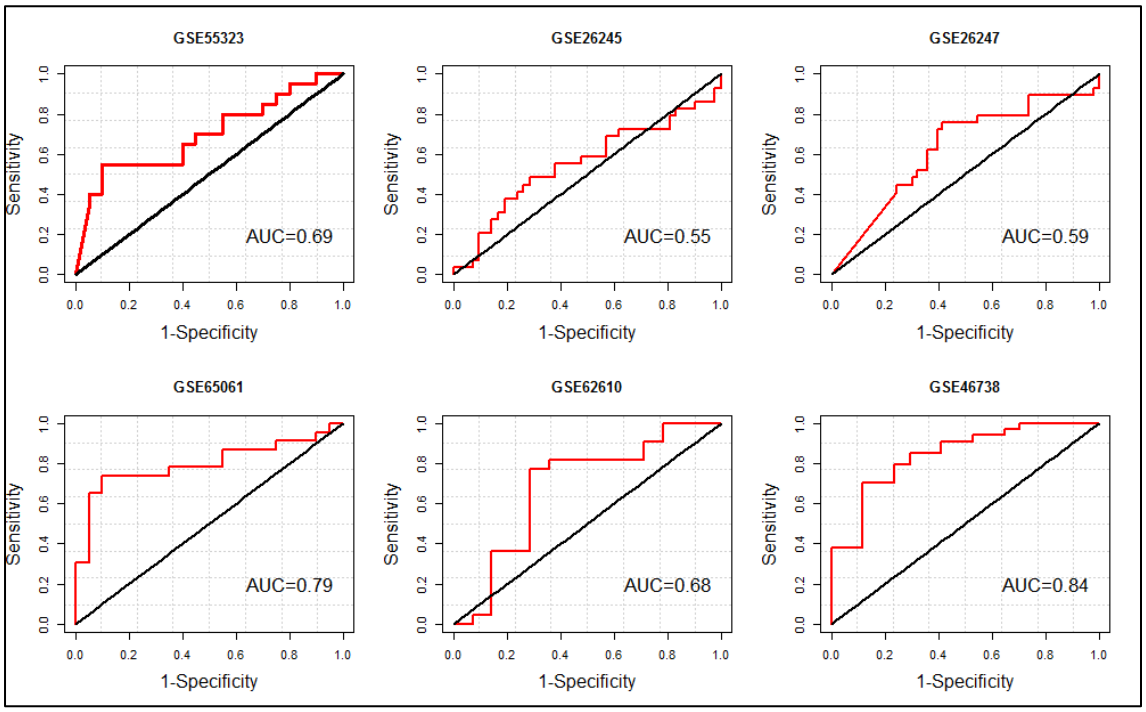


Figure 4.6 Receiver operating characteristics (ROC) analysis of 37-miRNA signature in biochemical disease recurrence vs. the non-recurrence samples using each GEO datasets. The DE miRNAs are depicted in Table 2. AUC; area under the ROC curve.

Table 4.6 Best subset, PART's decision rules and diagnostic potentials for the DE microRNAs identified from meta-analysis in 6 GEO datasets.

	Best subset	Extracted rules by PART	PART's AUC (95%CI↑)	PART's F-measure
GSE55323	miR-1, miR-221, miR-28-5P, miR-301B, miR-324-5P, miR-370, miR-449A, miR-606, miR-624, miR-661, miR-98	**1) IF** miR-496 > 8.13 **AND** mir-1 >9.67 **AND** : BCR- (11.0) **2) IF** miR-137 > 6.36 **AND** miR-449A ≤ 7.10 **AND** mir-137 ≤ 6.82: BCR- (11.0/2.0)	0.75	0.72
GSE26245	miR-370, miR-492, miR-579, miR-639, miR-98	**1) IF** miR-579 ≤ 9.401: **AND** miR-639 ≤ 9.120: BCR- (46.0/10.0) **2) IF** miR-324-5P > 12.032: BCR+ (17.0/1.0) **3) IF** mir-639 >9.212 : BCR- (5.0)	0.60	0.78
GSE26247	miR-1, miR-133A, miR-137, miR-363*	**1) IF** miR-363* ≤ 8.89 **AND** miR-636 ≤ 9.34: BCR- (43.0/4.0) **2) IF** miR-363* > 8.44 **AND** miR-661 > 12.95: BCR+ (20.0)	0.804	824
GSE65061	miR-1, miR-221-3P, miR-301B, miR-489, miR-637, miR-939, miR-98	**1) IF** miR-221-3P ≤ 6.97 **AND** miR-489 > 5.36 **AND** miR-98 ≤ 7.84 **AND** miR-939> 3.87 **AND** miR-637 ≤ 4.17 : BCR- (14.0) **2) IF** miR-301B > 3.97 **AND** miR-221-3P > 6.029: BCR+ (21.0) **3) IF** miR-1 > 5.23: BCR- (6.0)	0.734	0.744
GSE62610	miR-449A, miR-496, miR-636, miR-492	**1) IF** miR-449A > 16.80 **AND** miR-636 ≤ 17.73 **AND** miR-496 ≤ 20.304 : BCR+ (12.0/2.0) **2) IF** miR-449A >16.804 :BCR- (13.0/1.0)	0.763	0.833
GSE46738	miR-340,miR-541, miR-624	**1) IF** miR-340 ≤ 2.13 **AND** miR-541 ≤ 3.107: BCR+ (33.0/1.0) **2) IF** miR-541 > 1.374 : BCR- (16.0)	0.8823	0.865

↑ CI: confidence interval.

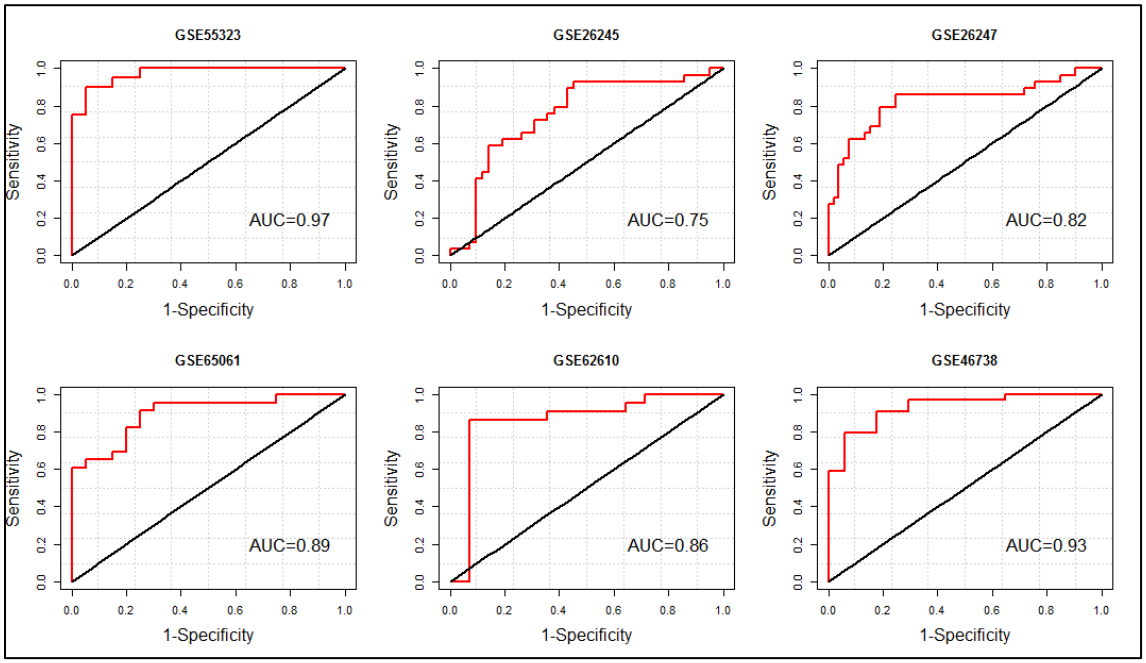


Figure 4.7 ROC analysis of the best subset of the DE miRNAs in biochemical disease recurrence vs. the non-recurrence samples using each GEO datasets. The best subset of DE miRNAs is shown in the first column of Table 3 which has been found by using soft computing technique (PSO/ logistic regression).

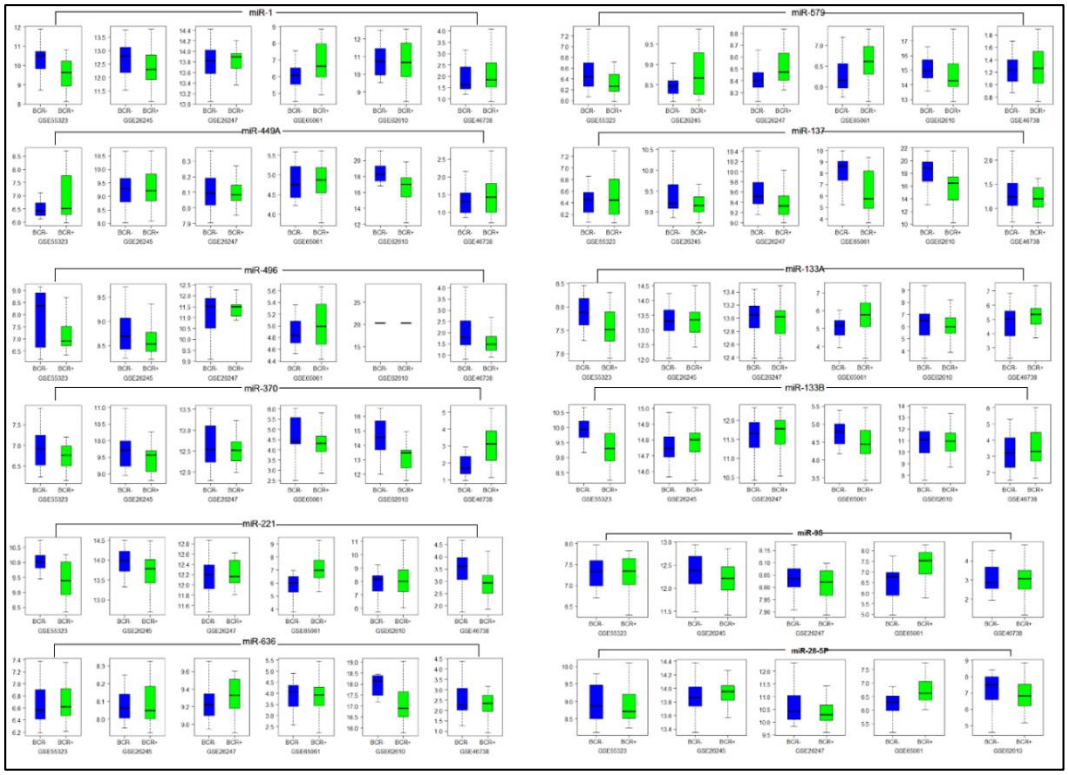


Figure 4.8 A comparison between expression of co-deregulated microRNAs in recurrent vs. non-recurrent PCa samples. Those miRNAs that were selected for analysis are depicted above the box plots (Table 3). Lines within the boxes indicate median values; whiskers - min and max for miRNA values. BCR+/ -, biochemical disease recurrence status (positive, negative).

4.4 Discussion

Various miRNAs are DE in individuals with recurrent PCa, and identifying the most important miRNAs and pathways associated with the disease is very important. A meta-analysis of multiple miRNA datasets combines the generated p-values of individual studies, making the identification of DE microRNA genes more reliable.

In this study, we attempted to identify common miRNAs underlying recurrent PCa using meta-analysis of six publicly available microRNA datasets to focus deeply on identifying DE microRNA genes and risk factors shared between them.

By meta-analysis of six published miRNA expression datasets of recurrent PCa, we identified a common signature of a total of 37 DE microRNAs including 15 overexpressed and 22 under expressed microRNA genes across at least two datasets under the significance threshold of adjusted p-value < 0.05 in recurrence compared to non-recurrence samples. The identified 37 microRNAs in this meta-analysis were discovered as DE microRNAs in at least one dataset in the prior individual analysis. Of the 37 DE miRNAs associated with BCR after RP (Table 4.1), all except miR-606 have been reported to be associated with cancer in general. Fifteen miRNAs (miR-1, miR-133A, miR-133B, miR-449A, miR-137, miR-370, miR-221, miR-449B, miR-125A-5P, miR-199A-3P, miR-301B, miR-340, miR-361, miR-363, miR-98) have been previously linked to PCa and of those, miR-1, miR-133B, miR-449B, and miR-221 have been described as predictive markers in PCa recurrence after RP.

Among the overexpressed DE microRNAs, miR-449A and miR-579 had high combined P-values across all studies.

Tumor-suppressive miR-449A targets HDAC1 and induces growth arrest in Pca. It also causes Rb-dependent cell cycle arrest and senescence in PCa cells (Table 4.2). For a previously poorly characterized miRNA, namely miR-579, no PCa related functions have been reported. MiR-579-3p is only known as a master regulator of melanoma progression and drug resistance.

Among the under expressed DE microRNAs mir-496, miR-137, miR-1, and miR-370 had the highest combined P-values across all studies.

MiR-496 is also a previously poorly characterized miRNA, which has no functions in PCa. Methylated DNA binding domain protein 2 (MBD2) is known as the only TF of

miR-496, which coordinately silences gene expression through activation of the miR-496 promoter in breast cancer cell line.

Methylated mir-137 host gene is promising diagnostic and/or prognostic biomarker of PCa. The epigenetic silencing of miR137 is an important event in promoting androgen signaling during prostate carcinogenesis and progression. MiR-137 suppresses cell growth in several cancers such as ovarian, colorectal, and gastric.

MiR-1 is known as a biomarker of recurrence PCa, which is in agreement with the findings in present meta-analysis study. MiR-1 functions as a tumor suppressor which suppresses cancer cell proliferation, metastasis, angiogenesis, invasion, cell cycle arrest, WNT signaling and promotes apoptosis by ectopic expression. This miRNA is a potential prognostic biomarker of hepatocellular carcinoma (HCC) and colorectal cancer. The expression of miR-1 alters in several cancers such as lung, gastrointestinal, prostate, bladder, head and neck, and renal cancer.

MiR-370 plays an important role in the proliferation of human PCa cells by directly suppressing the tumor suppressor FOXO1.

PPI Hub Proteins analysis of the TF and target genes of DE MicroRNAs was conducted for prioritization of the most important hub genes using the EnrichR web tool. CTNNB1 was the most important hub genes among TF and target genes of DE microRNAs across six microarray studies.

CTNNB1 (Catenin Beta 1) functions as a Key downstream component of the canonical WNT signaling pathway. WNTs and their downstream effectors have crucial roles in the regulation of various processes that are important for cancer progression, including tumor growth, tumor initiation, differentiation, cell senescence, cell death, differentiation and metastasis. Nuclear accumulation and abnormal stabilization of CTNNB1 as a consequence of missense mutations occurs at a high frequency in a variety of epithelial cancers such as colorectal cancer, medulloblastoma, ovarian cancer, and pilomatrixoma. Upregulation of CTNNB1 is also associated with PCa.

To elucidate the role of DE microRNAs obtained from the meta-analysis, we performed pathway analysis and gene set enrichment analysis for TF and target genes of DE miRNAs using the EnrichR web tool. The most enriched pathway and Gene Ontology (GO) term among the TF and target genes of DE miRNAs were “MicroRNAs in cancer (hsa05206)”, “Pathways in cancer (hsa05200)”, Signal Transduction (R-HSA-162582)”,

“regulation of epithelial cell proliferation (GO: 0050678)” and “tissue morphogenesis (GO: GO:0048729)”.

Common pathway analysis revealed that TCF3, MYC, MAX, CYP26A1 and SREBF1 were the most significant proteins associated with DE miRNA genes. Of note, these proteins were not identified as TF and target genes of DE microRNAs.

Previous studies have reported that the diminished activity of TCF3 plays a role in lymphoid malignancies, and up-regulation of it is involved in the development and progression of colorectal cancer. TCF3 is regulated by androgens and acts as a tumor promoter in Pca.

Overexpression, Mutations, translocation and rearrangement of MYC is related to several cancers such as breast, PCa, gastrointestinal, melanoma, and small cell lung cancer.

MAX is known as a tumor suppressor in renal oncocytomas and small cell lung cancer. The mutation of it has been identified in gastrointestinal stromal tumors. High expression of CYP26A1 is associated with several cancers such as breast, head and neck, colorectal and ovarian. CYP26A1 is a methylation marker of PCs associated with ERG-positive cancers.

Sterol regulatory element-binding protein1 (SREBP1) is a key regulatory factor that controls lipid homeostasis. SREBP1 is a critical link between oncogenic signaling and tumor metabolism. The overexpression of SREBF1 is related to a variety of cancers such as PCa, breast, head and neck, colorectal, endometrial, glioblastoma, pancreatic, and ovarian.

To understand the association of the DE microRNAs list with the most significant target genes and transcription factors, we conducted a regulatory gene network analysis using the MIROB web tool. CDKN1A and LASP1 were amongst the most significant target genes associated with the DE microRNAs.

Cyclin-dependent kinase inhibitor 1 (CDKN1A) also is known as P21 is involved in p53/TP53 mediated inhibition of cellular proliferation in response to DNA damage and its overexpression results in cell cycle arrest and autophagy cell death. The expression of this gene is tightly controlled by the tumor suppressor protein p53 in a human brain tumor cell line. The CDKN1A genotypes CT and TT are associated with an increased

risk of advanced prostate carcinoma compared with the CC genotype. Elevated p21 levels are associated with higher Gleason score, and increased PCa recurrence.

LIM and SH3 protein 1 (LASP1), a promoter of cell proliferation and migration, play a significant role in cancer development and progression. LASP-1 is involved in numerous biological and pathological processes. It plays an important role in the regulation of dynamic actin-based and cytoskeletal activities. LASP-1 is highly expressed in the central nervous system and contributes to the formation and progression of prostate cancer through a NF-KB pathway.

RELA, SNAI2, and TP53 were among the most significant transcription factors associated with the DE microRNAs.

RELA also known as NF-kappa-B is a ubiquitous transcription factor involved in many biological processes such as immunity, inflammation, cell growth, differentiation, tumorigenesis and apoptosis. Zinc finger protein (SNAI2) is known as a transcriptional repressor that modulates both activator-dependent and basal transcription. SNAI2 regulates cell proliferation and invasiveness of metastatic PCa cell lines. Cellular tumor antigen p53 (TP53) acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type.

Moreover, network analysis showed that ten of the 37 DE miRNAs (miR-125A, miR-133B, miR-137, miR-221, miR-28, miR-340, miR-370, miR-449A, miR-874, and miR-98) have an established prognostic significance in other cancers such as colorectal, gastric, and breast. This network also indicated that eight of 37 DE miRNAs (miR-133A, miR-133B, miR-137, miR-199A1, miR-340, miR-361, miR-498, and miR-661) can be actively involved in tumor growth.

In this study, we also built new miRNA diagnostic classifiers in each GEO datasets based on best subset of DE miRNAs in the meta-analysis. These classifiers predicted BCR after RP with very high accuracy. The highest diagnostic accuracy (97%) was given for GSE55323 with 11-miRNAs. The performance of our 11-miRNA diagnostic classifier (97%) exceeded that of a 2-miRNA classifier (miR-1+miR-133B; AUC: 71%) developed earlier by Karatas et al. [54]. One miRNA (miR-1) is shared between these classifiers, further supporting the validity of our findings.

Briefly, we used “MetaDE” package to perform a meta-analysis, which provides options for gene matching across studies, gene filtering before meta-analysis and functions for

conducting several major meta-analysis methods such as Fisher and AW for differential expression analysis. Then performed the GO enrichment analysis, pathway analysis, network analysis, and ROC analysis.

In conclusion, this is the first report that provides biological insights on common microRNA expression signatures for recurrent PCa after RP. The candidate miRNAs are worthy to be validated in the wet lab.

4.5 Chapter summary

The goal of this chapter was to do meta-analysis for recurrent PCa on miRNA expression profile in order to increase the probability of revealing truly significant deregulated miRNA genes, which should have higher potentials to be utilized as new biomarkers for the disease

Meta-analysis of six miRNA datasets revealed miR-125A, miR-199A-3P, miR-28-5P, miR-301B, miR-324-5P, miR-361-5P, miR-363*, miR-449A, miR-484, miR-498, miR-579, miR-637, miR-720, miR-874 and miR-98 are commonly upregulated miRNA genes, while miR-1, miR-133A, miR-133B, miR-137, miR-221, miR-340, miR-370, miR-449B, miR-489, miR-492, miR-496, miR-541, miR-572, miR-583, miR-606, miR-624, miR-636, miR-639, miR-661, miR-760, miR-890, and miR-939 are commonly downregulated miRNA genes in recurrent PCa samples in comparison to non-recurrent PCa samples. The network-based analysis showed that some of these miRNAs have an established prognostic significance in other cancers and can be actively involved in tumor growth.

Gene ontology enrichment revealed many target genes of co-deregulated miRNAs are involved in “regulation of epithelial cell proliferation” and “tissue morphogenesis”. Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis indicated that these miRNAs regulate cancer pathways. The PPI hub proteins analysis identified CTNNB1 as the most highly ranked hub protein. Besides, common pathway analysis showed that TCF3, MAX, MYC, CYP26A1, and SREBF1 significantly interact with those DE miRNA genes. The identified genes have been known as tumor suppressors and biomarkers which are closely related to several cancer types, such as colorectal cancer, breast cancer, PCa, gastric, and hepatocellular carcinomas. Additionally, it was shown that the combination of DE miRNAs can assist in the more specific detection of the PCa and prediction of biochemical recurrence (BCR).

We found that the identified miRNAs through meta-analysis are candidate predictive markers for recurrent PCa after radical prostatectomy.

CHAPTER 5

META ANALYSIS OF MIR145 TARGET GENES

5.

5.1 Introduction

MicroRNAs, which are small regulatory RNAs, post-transcriptionally regulate gene expression by binding 3'-UTR of their mRNA targets. Their deregulation has been shown to cause increased proliferation, migration, invasion, and apoptosis. MiR-145, an important tumor suppressor microRNA, has shown to be downregulated in many cancer types and has crucial roles in tumor initiation, progression, metastasis, invasion, recurrence, and chemoradioresistance. Our aim is to investigate potential common target genes of miR-145, and to help understanding the underlying molecular pathways of tumor pathogenesis in association with those common target genes.

5.1.1 Chapter Goals

The overall goal of this chapter is to show potential common target genes of miR-145 in several cancer types including prostate, breast, esophageal, bladder, head, and neck squamous cell carcinoma cancer, using GEO database and to unravel the underlying molecular pathways associated with mir-145 in tumor pathogenesis. To achieve this goal, eight published microarray datasets, where targets of mir-145 were investigated in cell lines upon mir-145 over expression, were included for meta-analysis. Inter group variabilities were assessed by box-plot analysis. Microarray datasets were analyzed using GEOquery package in Bioconductor 3.2 with R version 3.2.2 and two-way

Hierarchical Clustering was used for gene expression data analysis. Specifically, the highlights of this chapter are:

- We conducted a meta-analysis of 8 available microarray datasets (consider samples for mir-145) and identified a panel of co-deregulated genes upon mir-145 over expression in prostate, breast, esophageal, bladder cancer, and head and neck squamous cell carcinoma.
- Co-deregulated miRNAs were investigated for protein-protein interaction network by STRING.
- Biological process, molecular function, and pathway analysis are applied for identified potential targets of mir-145. These analyses demonstrated that identified genes are significantly involved in telomere maintenance, DNA binding and repair mechanisms.
- In addition, target analysis and miRNA target prediction are performed. In silico analysis tools predicted the identified genes as potential targets of miR-145.

5.1.2 Chapter Organization

The remainder of this chapter is organized as follows. The section 5.2 presents the design of the experiments. The results are presented in section 5.3. The section 5.4 presents the discussions .The section 5.5 provides a summary of this chapter.

5.2 Design of Experiments

5.2.1 Literature Search

A systematic review of the microarray literature from GEO database was documented to identify studies, where expression profiling was performed for miR-145 over-expressing cancer cell lines, published up to Jun 15, 2014.Medical subjective heading (MeSH) was "miR-145 in human cancer". A total of 55 studies were identified through GEO database searching. Of these, 20 studies were retained after rejecting replications. A total of 9 articles were excluded according to the title of samples (GSMs). The reason for the exclusion was the following: the title of samples no association with miR-145. The full-text articles were evaluated for the remaining 11 studies, and 7 were recruited in the final meta-analysis. The other 4 investigations were excluded for these reasons: the median-centered across samples is not zero (not suitable for comparison) and data

sets are containing null values. Remained 7 microarray datasets were obtained from National Centers for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/).The entire study selection process for meta-analysis is shown in Figure 5.1. All GSE Series Matrix files, platform sets and annotations files were downloaded and parsed by GEOquery package in Bioconductor 3.2 with R version 3.2.2. Box-plots were drawn for each selected GSE microarray data for visual comparison of inter-group variability in a statistical population. These box plots were used to display the statistical distribution of different data values of mir-145 GSMs.

5.2.2 Data Preparation and Statistical Analysis

Before being able to do statistical analysis, we need to be prepared to meet the following requirements: raw data must be log2- scaled and all datasets must exhibit the same data precision. In our dataset except GSE58295 all of GSEs were in the form of (log) 10 (ratio), therefore, we converted them to (log) 2 (ratio) format. All GSEs have 16 bit precision. Since the comparable data are provided, the analysis should be limited to genes that are expressed in all data sets. In 16 GSMs (samples) 17085 common genes were founded. Then, we checked for the cross platform bias (batch effect) by computing and plotting the Principal Components Analysis (PCA) for combined dataset. PCA plot (Figure 5.2, left panel) shows that sample of the each technology (GSE) clusters together. This means that we have batch effect in our data. We have removed batch effect by “removeBatchEffect” function in “limma” package. The right panel of Figure 5.2 shows PCA plot after removing the batch effect. Note that the expression values of combined data will change after removing batch effect. In the next step we try to design and make proper contrast matrix again by “limma” package. We have five groups (each disease) with eight replicates in group1, one replicate in group2, three replicates in group3, two replicates in group 4, and two replicates in group5. In order to test the interaction effect of the different disease, the meaningful contrast was considered to be group1—group2—group3 –group4 + control. Then, differentially expressed genes with P-value < 0.01 were selected as potential candidates in different cancer types.

5.3 Results

The aim of this chapter was to identify common target genes of tumor suppressor miR-145 via meta-analysis of microarray-based gene expression profiles in several human cancer types. In this chapter, we collected a total of 7 gene expression profile data set from previously published studies considering the inclusion criteria. We retrieved the GSE Series Matrix files, platform sets and annotations files from the Gene Expression Omnibus (GEO) including 16 array samples using GEOquery package in Bioconductor with R version. Then, did transformation on raw data, found common genes, and removed batch effects. Details of the each individual microarray studies are summarized in Table 1.4. Data set 1 (GSE47657) is obtained from a microarray analysis, which was performed in human prostate cancer cell lines including PC3, DU145, and LNCaP cells, treated with miR-145 to investigate the differentially expressed genes using Sure Print G3 Human GE 8×60K Microarray (Agilent Technologies, Santa Clara, CA, USA) containing 62,976 probes. In the data set 2 (GSE24782), The microarray analysis was generated from PC3 and DU145 human prostate cancer cell lines which were transfected withmiR-145 were using Agilent-026652Whole Human Genome Microarray 4x44K v2 with 44,495 probes.

Data set 3 (GSE58295) was generated from mir-145 transfected PC3 cells which were collected at 8, 16 and 24 hours after transfection along with un-transfected control PC3 cells using Agilent-014850Whole Human Genome Microarray 4x44K G4112F containing 45,015 probes. Besides, data set 4 (GSE37119) is obtained from a microarray analysis which was performed in human head and neck squamous cell carcinoma cell lines HNSCC and IMC3 transfected with miRNA 145 utilizing Agilent-026652Whole Human Genome Microarray 4x44K v2 to arrays spotted with 44,495 probes. Data set 5 (GSE18625) microarray analysis was performed using DLD-1, colon carcinoma cell line, transfected withmiR-145 and collected 24 hours after transfection. Gene expressions were profiled on Affymetrix Human Genome U133 Plus 2.0 Array containing 54675 probes.

Data set 6 (GSE19737) was generated from miR-145 or a negative control pre-miRNA Transfected MDA-MB-231 cell line which is most typical cell line with highly metastatic features in breast cancer using Affymetrix Human Genome U133 Plus 2.0 Array with 54,675 probes. Data set 7 (GSE20028) identifiedmiR-145 targets in squamous cell carcinoma. The aim of study was to explore of miR-145 target genes

using Agilent-014850Whole Human Genome Microarray 4x44K G4112F which includes 45,015 probes. Lastly, data set 8 (GSE19717) gene expression profiles of bladder cancer cell line KK-47, and urinary bladder cancer cell line T24 were investigated upon miR-145 transfection using Agilent-014850Whole Human Genome Microarray 4x44K G4112F with 45,015 probes. In order to clarify whether the microarray data were comparable, we initially prepared the Box plots representation of median-centered gene expression as provided in Figure 5.3. It shows the common target gene expression levels for probe set over all arrays. According to the result all GSE datasets were centered on zero except from GSE18625 (colon cancer). Therefore, we excluded this data set before the statistical analyses. Expression data in all GSEs were converted from (log10) to (log2) to eliminate variability among the datasets and batch effect was removed. Then, by making proper contras matrix significantly differentially expressed genes were found. As a result of the meta-analysis, we found that UNG, FUCA2, DERA, GMFB, TF, and SNX24 are significantly downregulated, and MYL9 and TAGLN are significantly upregulated in all

GSM data. As a result, we found eight common target genes of mir-145 that have similar behavior in different GEO datasets. A heat map representation of these genes is demonstrated in Figure 5.4. In silico analysis tools predicted these genes as potential targets of miR-145 (Table 5.1). Biological process (Table 5.2), molecular function (Table 5.3), cellular component (Table 5.4), and KEGG pathways (Table 5.5) analysis of these potential targets of mir-145 through functional enrichments in PPI network, demonstrated that those genes are significantly involved in telomere maintenance, DNA binding and repair mechanisms. Besides, PPI network of commonly deregulated mir-145 targets and pathway analysis of MYL9, UNG, TAGLN, FUCA2, DERA, GMFB, TF, and SNX24 are represented in Figures 5.5 and 5.6, respectively.

Table 5.1 Representation of the potential targets of mir-145 by in-silico analysis.

Gene	EntrezID	RefseqID	miRWalk	miRanda	RNA22	Targetscan	SUM
MYL9	10398	NM_006097	0	0	1	0	1
UNG	7374	NM_003362	1	0	1	0	2
TAGLN	6876	NM_001001522	0	0	0	0	0
FUCA2	2519	NM_032020	0	1	0	1	2
DERA	51071	NM_015954	0	0	0	0	0
GMFB	2768	XM_005267541	1	0	1	1	3
TF	7018	NM_001063	0	0	1	0	1
SNX24	28966	NM_014035	0	1	0	1	2

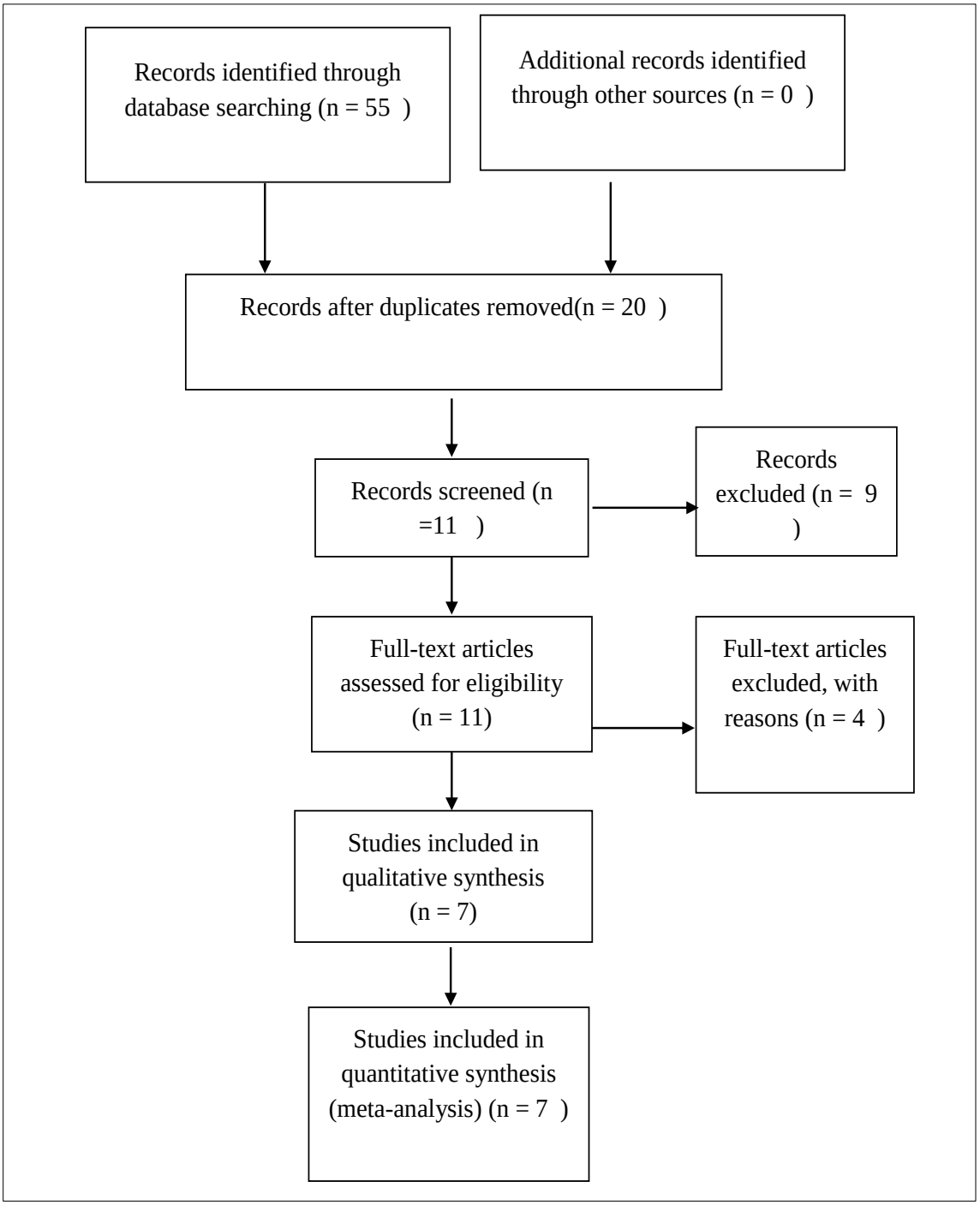


Figure 5.1 Flow Chart of Study Selection in the Meta-Analysis.

Table 5.2 Biological process (GO) of the potential targets of mir-145 by functional enrichments in PPI network.

Pathway ID	pathway description	count in network	FDR
GO:0032201	telomere maintenance via semi-conservative replication	8	1.14e-14
GO:0000722	telomere maintenance via recombination	8	1.79e-14
GO:0006284	base-excision repair	9	1.79e-14
GO:0033260	nuclear DNA replication	8	1.79e-14
GO:0006271	DNA strand elongation involved in DNA replication	8	8.3e-14

Table 5.3 Molecular function (GO) of the potential targets of mir-145 by functional enrichments in PPI network.

pathway ID	pathway description	count in network	FDR
GO:0003684	damaged DNA binding	5	5.91e-05
GO:0015091	ferric iron transmembrane transporter activity	2	0.00248
GO:0003689	DNA clamp loader activity	2	0.00446
GO:0003676	nucleic acid binding	15	0.00878
GO:0042623	ATPase activity, coupled	5	0.0109

Table 5.4 Cellular component (GO) of the potential targets of mir-145 by functional enrichments in PPI network.

pathway ID	pathway description	count in network	FDR
GO:0005663	DNA replication factor C complex	5	5.39e-12
GO:0005657	replication fork	6	3.13e-08
GO:0044427	chromosomal part	9	4.64e-05
GO:0001725	stress fiber	4	0.000123
GO:0005654	nucleoplasm	15	0.000123

Table 5.5 KEGG Pathways of the potential targets of mir-145 by functional enrichments in PPI network.

pathway ID	pathway description	count in network	FDR
03430	Mismatch repair	8	5.68e-16
03030	DNA replication	8	1.35e-14
03420	Nucleotide excision repair	7	1.97e-11
03410	Base excision repair	5	5.46e-08
04810	Regulation of actin cytoskeleton	6	1.84e-05

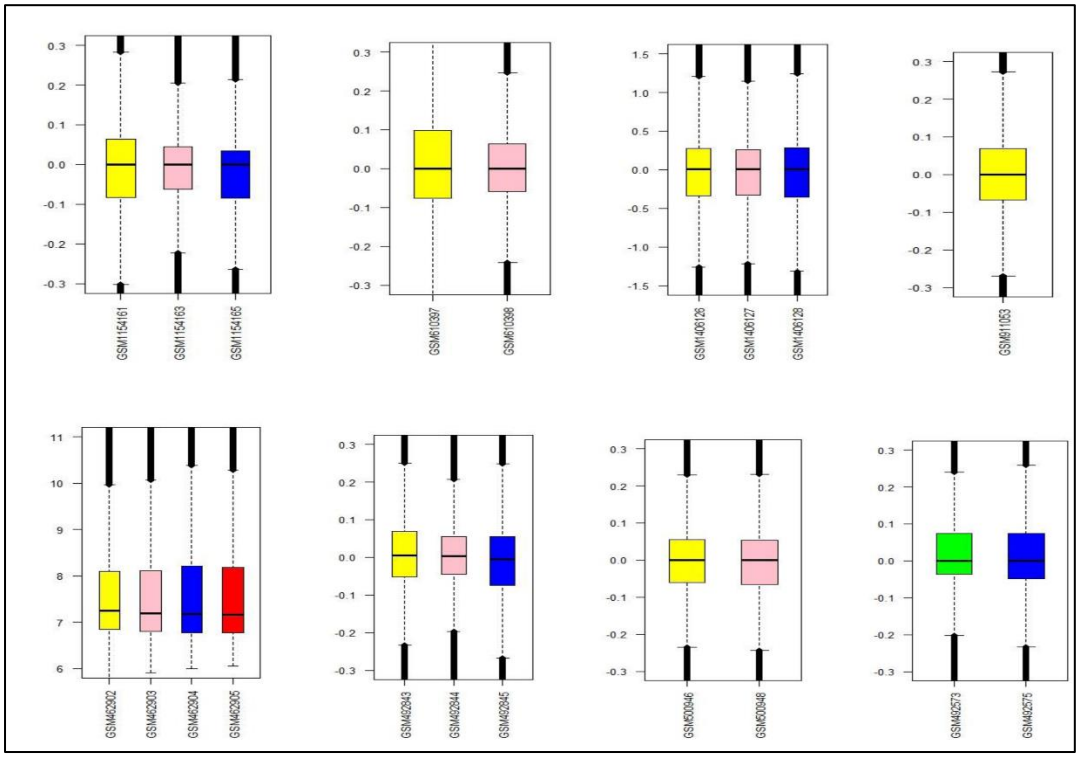


Figure 5.2 Box-plot representations of the GSM datasets.

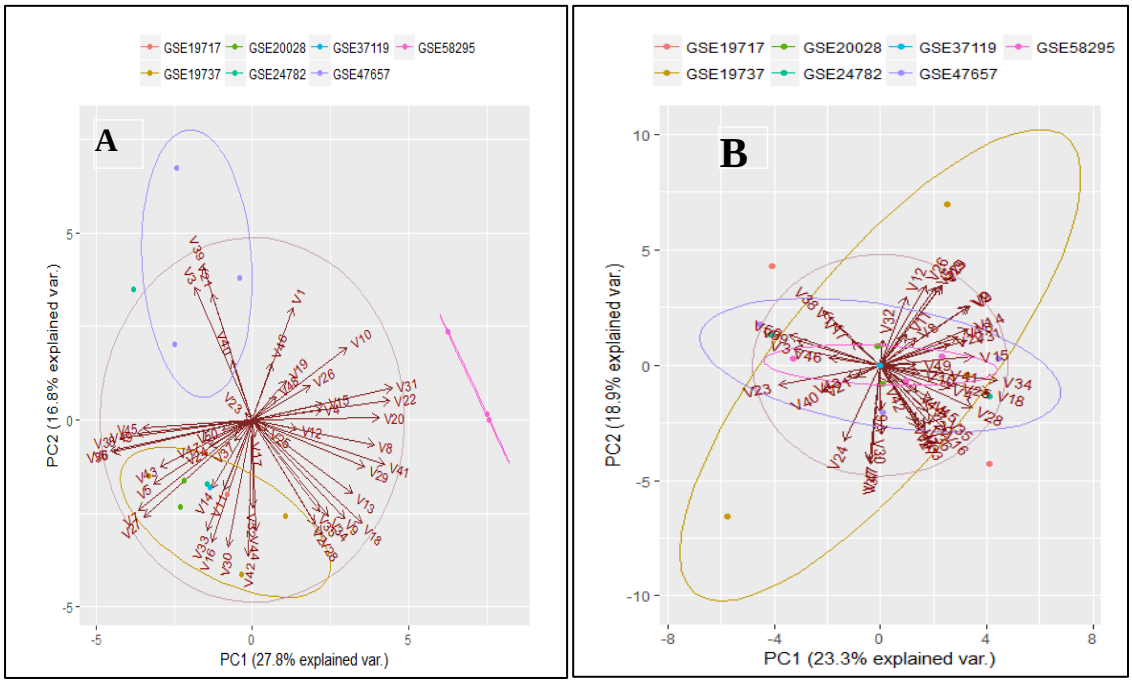


Figure 5.3 Principal Component Analysis (PCA) Plot for Combined Dataset Before (a) and After (b) Removing Batch Effect.

.

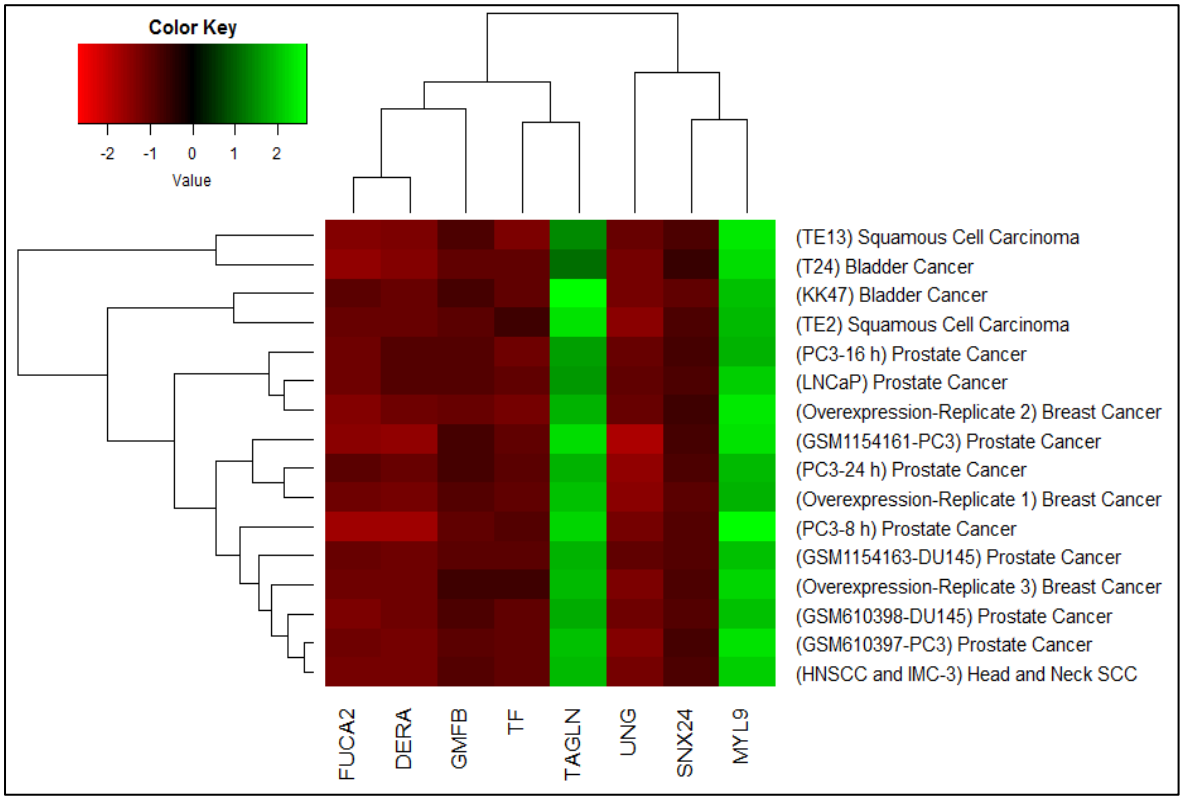


Figure 5.4 Heat Map Representation of Commonly Deregulated genes by mir-145 Overexpression in 5 Types of Cancer.

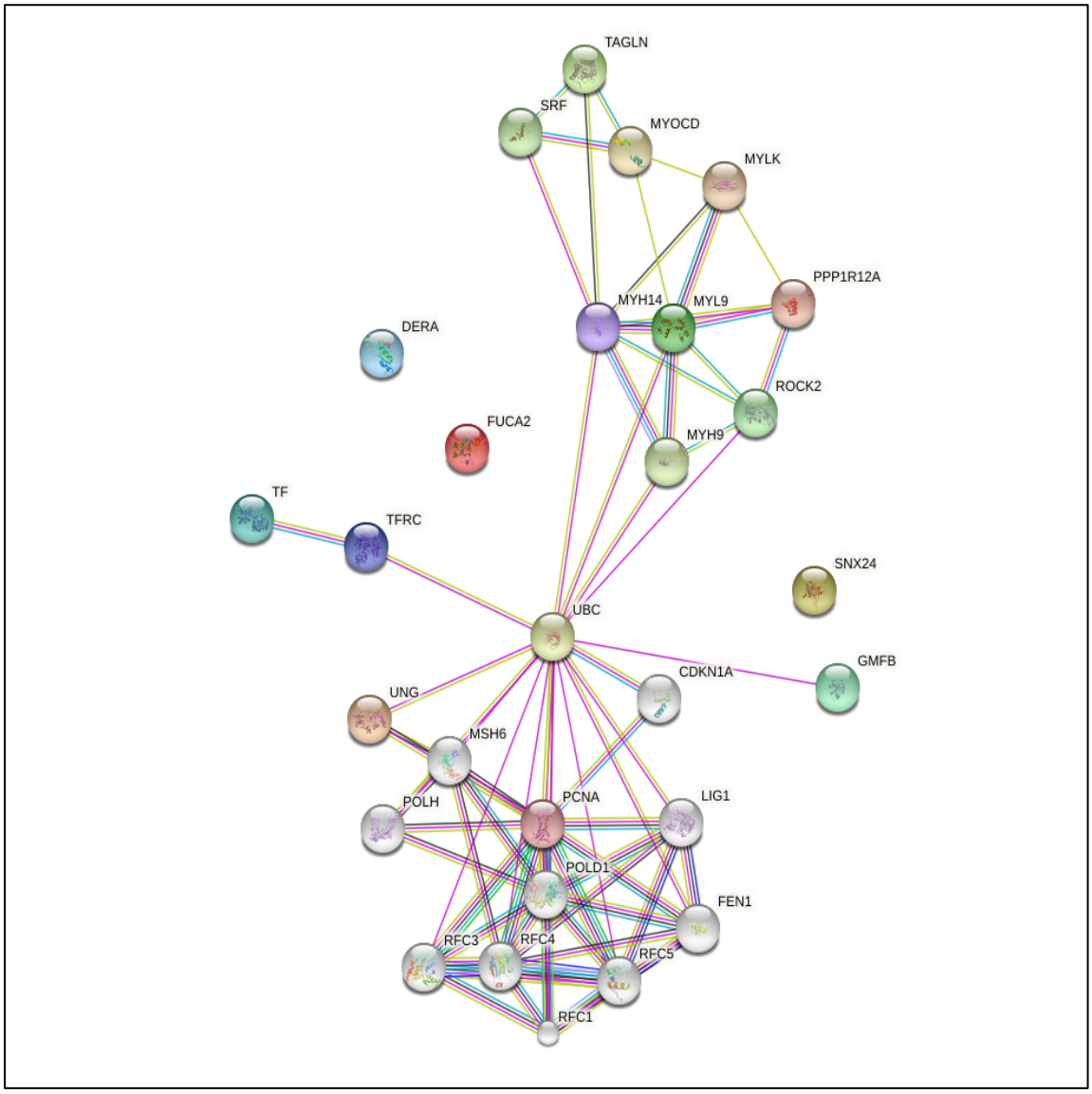


Figure 5.5 PPI Network of Commonly Deregulated Mir-145 Targets.

Pink: experimentally determined(known interactions), Blue: from curated databases (known interactions). Yellow: textmining, Green: gene neighborhood (Predicted interactions), Black: co-expression. The interaction score was set to high confidence (0.700) .

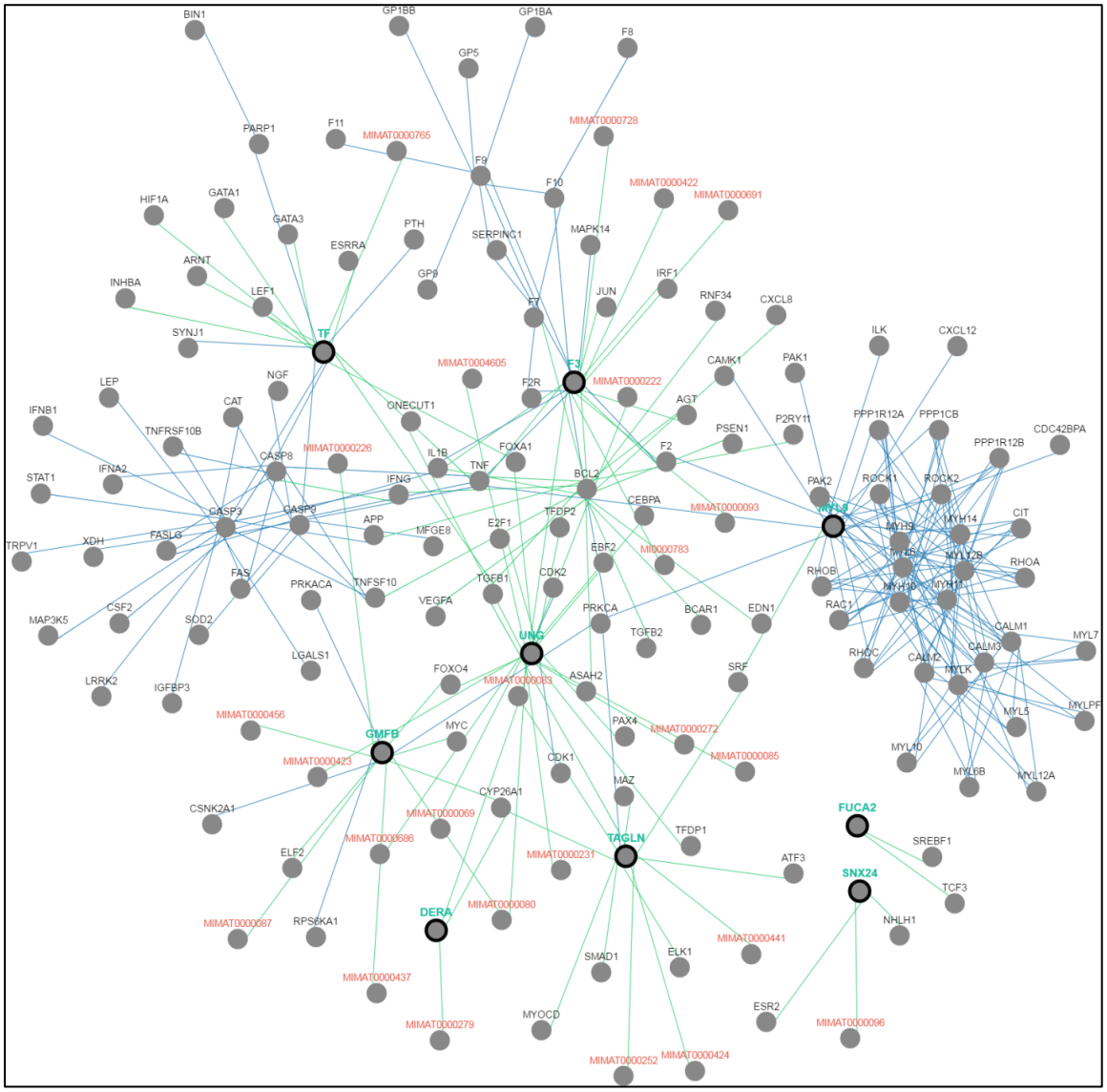


Figure 5.6 Pathway Analysis of MYL9, UNG, TAGLN, FUCA2, DERA, GMFB, TF, and SNX24. Green: control expression, Blue: controls state change.

5.4 Discussion

MiRNAs are frequently located in the cancer-associated genomic regions or in fragile sites of the genome. In addition to in vitro and in vivo tools, bioinformatics approaches are of paramount importance to evaluate their roles in the pathogenesis of different types of cancer [92]. Expression of miR-145 has been commonly identified as downregulated in several human cancer types. Several studies suggested that low levels of miR-145 might contribute to pathogenesis and progression of human tumors [93, 94]. MiR-145 is a well characterized tumor suppressor in human malignancies which targets various oncogenes in cancer cells. Functional analyses of target genes, which are repressed by mir-145, are crucial to explain their roles in cancer development. The aim

of the present study is to investigate the commonly targeted genes by mir-145 and relevant pathways through evaluating publicly available microarray datasets. We selected miR-145 in our meta-analysis due to its well-known function as a tumor suppressor and since its presence has been reported in a variety of cancers including prostate, esophageal, head and neck, breast, bladder cancer, and squamous cell carcinoma [95, 96]. In this study, we extracted 7 gene expression microarray datasets from GEO database (1 of the min Affymetrix Array, others Agilent), which are generated from cell line samples. We identified six genes including UNG, FUCA2, DERA, GMFB, TF, and SNX24 as significantly downregulated and two genes including MYL9 and TAGLN as significantly upregulated upon mir-145 over-expression in distinct cancer types. The tumorigenic potentials of those genes have not been studied extensively until now. Considering our results, we suggest the most important contributors in tumorigenesis, which should be demonstrated in in vitro and in vivo studies. UNG, FUCA2, DERA, GMFB, TF, and SNX24 as significantly downregulated upon mir-145 over-expression, are expected to have elevated expressions in tumor samples considering low levels of miR-145 in several cancer types. Among those genes, elevated expression of UNG, an essential enzyme for post-replicative repair of uracil in DNA [97], has been found to be associated with pemetrexed-resistance and present in cell lines derived from pemetrexed-resistant histologic subtypes [98]. Besides, high GMFB expression was related to poor disease-free survival and overall survival in patients with SOC (serous ovarian cancer) [99]. In order to minimize false positives, we used target prediction tools, including miRwalk, TargetScan, RNA22, and miRanda. The significantly downregulated genes that we found as a result of ourmeta-analysis were predicted to be targeted by mir-145 in at least one in-silico tools out of four. Interestingly, biological process, molecular function, cellular component, and KEGG pathways analysis of these potential targets of mir-145 through functional enrichments in PPI network showed that UNG, FUCA2, DERA, GMFB, TF, and SNX24 are significantly involved in telomere maintenance, DNA binding and repair mechanisms.

As a conclusion, our results pointed out the importance of mir-145 and its targets and suggested that they contribute to carcinogenesis in distinct tumor types. To the best of our knowledge, this is the first study retrieving gene expression data from several cancer types and investigating the common targets of mir-145 to help enlightening the roles of mir-145 in cancer pathogenesis.

5.5 Chapter Summary

The goal of this chapter was to show potential common target genes of miR-145 in several cancer types including prostate, breast, esophageal, bladder, head, and neck squamous cell carcinoma cancer, using GEO database and to unravel the underlying molecular pathways associated with mir-145 in tumor pathogenesis.

This chapter shows that UNG, FUCA2, DERA, GMFB, TF, and SNX2 were commonly downregulated genes, whereas MYL9 and TAGLN were found to be commonly upregulated upon mir-145 over expression in prostate, breast, esophageal, bladder cancer, and head and neck squamous cell carcinoma. Biological process, molecular function, and pathway analysis of these potential targets of mir-145 through functional enrichments in PPI network demonstrated that those genes are significantly involved in telomere maintenance, DNA binding and repair mechanisms.

CHAPTER 6

CONCLUSION AND FEATURE WORKS

This thesis presented new evolutionary-based FS techniques for analyzing biological datasets acquired via mass throughput technologies, in order to improve performance on selecting informative genes and increasing prediction ability. The microarray datasets are typically high dimensional with only a small number of samples making the task of their analysis especially challenging. Moreover, this thesis focuses on performing meta-analysis for recurrent prostate cancer on miRNA expression profiles and mir-145 target genes in order to draw more reliable conclusions and new biological insights.

Section 6.1 summarizes the key findings from each individual chapter and Section 6.2 outlines suggestions for future work.

6.

6.1 Conclusion

This thesis presents the first study on using the BHA for solving FS problem (chapter 2). By applying the hyperbolic tangent function, a new binary version of BHA called BBHA is proposed to solve FS problem in text, image, and biomedical data. Two classifiers (RF and NB) serve as the evaluators of proposed algorithm. In addition, to confirm that RF is the best DT classifier, the performances of six popular DT algorithms were compared in this chapter.

This thesis finds that RF is the best DT algorithm and the proposed BBHA wrapper based FS approach outperforms the performances of BPSO, GA, SA, and CFS in terms of AUC, accuracy, MCC, sensitivity, specificity, and the number of selected optimized features. Furthermore, if the computational cost is taken into account, BBHA wrapper

approach performs much faster than BPSO and GA. BBHA only needs a single parameter for configuring the model and is simple to understand.

In order to identify the most beneficial genes for classification, chapter 3 proposed a hybrid approach based on BPSOPG1 and BBHA algorithm which is combined with SPLSDA classifier. The study showed that the the proposed approach compare with many other methods, leads to a better performance in term of accuracy, AUC, and number of selected genes. The proposed method not only effectively reduced the number of genes, but also obtained a high classification accuracy. The obtained results indicate that the BPSOPG1-BBHA/SPLSDA is a useful tool for selecting informative genes in clinical datasets. Moreover, It was also shown that applying BBHA as the local optimizer for BPSOPG1 can significantly improve the performance of BPSOPG1 and help it to avoid being trapped in a local optimum.

Meta-analysis of different miRNA datasets in chapter 4 revealed that miR-125A, miR-199A-3P, miR-28-5P, miR-301B, miR-324-5P, miR-361-5P, miR-363*, miR-449A, miR-484, miR-498, miR-579, miR-637, miR-720, miR-874 and miR-98 are commonly upregulated miRNA genes, while miR-1, miR-133A, miR-133B, miR-137, miR-221, miR-340, miR-370, miR-449B, miR-489, miR-492, miR-496, miR-541, miR-572, miR-583, miR-606, miR-624, miR-636, miR-639, miR-661, miR-760, miR-890, and miR-939 are commonly downregulated miRNA genes in recurrent PCa samples in comparison to non-recurrent PCa samples. The network-based analysis showed that some of these miRNAs have an established prognostic significance in other cancers and can be actively involved in tumor growth. Gene ontology enrichment revealed many target genes of co-deregulated miRNAs are involved in “regulation of epithelial cell proliferation” and “tissue morphogenesis”. Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis indicated that these miRNAs regulate cancer pathways. The PPI hub proteins analysis identified CTNNB1 as the most highly ranked hub protein. Besides, common pathway analysis showed that TCF3, MAX, MYC, CYP26A1, and SREBF1 significantly interact with those DE miRNA genes. The identified genes have been known as tumor suppressors and biomarkers which are closely related to several cancer types, such as colorectal cancer, breast cancer, PCa, gastric, and hepatocellular carcinomas. Additionally, it was shown that the combination of DE miRNAs can assist in the more specific detection of the PCa and prediction of biochemical recurrence (BCR).

This thesis conducted a meta-analysis of 8 available microarray datasets (consider samples for mir-145) and identified a panel of co-deregulated genes upon mir-145 over expression in prostate, breast, esophageal, bladder cancer, and head and neck squamous cell carcinoma (chapter 5).

This thesis finds that UNG, FUCA2, DERA, GMFB, TF, and SNX2 were commonly downregulated genes, whereas MYL9 and TAGLN were found to be commonly upregulated upon mir-145 over expression in prostate, breast, esophageal, bladder cancer, and head and neck squamous cell carcinoma. Biological process, molecular function, and pathway analysis of these potential targets of mir-145 through functional enrichments in PPI network demonstrated that those genes are significantly involved in telomere maintenance, DNA binding and repair mechanisms.

6.2 Future Work

Future directions from this research could examine:

- Investigations and development of gene selection techniques that combines BBHA and Nearest Shrunken Centroid foe solving multiclass cancer diagnostic.
- Performing cross-platform merging and meta-analysis on Gene Expression profiles of recurrent prostate cancer.

REFERENCES

[1] Breiman, L., (2001). “Random forests”, Machine learning, 45:5-32.

[2] Biau, G.E., (2012). “Analysis of a random forests model”, The Journal of Machine Learning Research, 13:1063–1095.

[3] Qi, Y., (2012). Random Forest for Bioinformatics. Berlin: Springer.

[4] Hassan, H., Badr, A. and Abdelhalim, M., (2015). “Prediction of O-glycosylation Sites Using Random Forest and GA-Tuned PSO Technique”, Bioinformatics and Biology Insights, 9:103–109.

[5] Díaz-Uriarte R. and Alvarez de Andrés, D., (2006). “Gene selection and classification of microarray data using random forest”, BMC Bioinformatics.

[6] Anaissi, A., Kennedy, P.J., Goyal, M. and Catchpoole, D.R., (2013). “A balanced iterative random forest for gene selection from microarray data”, BMC Bioinformatics.

[7] Breiman, L., (1996). “Bagging predictors”, Machine Learning, 24:202-207.

[8] Dittman, D.J., Khoshgoftaar, T.M., Napolitano, A. and Fazelpour, A., (2014). “Select-Bagging: Effectively Combining Gene Selection and Bagging for Balanced Bioinformatics Data”, in IEEE International Conference on Bioinformatics and Bioengineering (BIBE), USA, 2014, 413–419.

[9] Quinlan, J.R., (1993). C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann.

[10] Lim, T.S., Loh, W.Y. and Shih, Y.S., (2000). “A comparison of prediction accuracy, complexity and training time of thirty-tree old and new classification algorithms”, Machine Learning, 40:203- 228.

[11] Tsai, M.C., Chen, K.H., Su, C.T. and Lin, H.C., (2012). “An Application of PSO Algorithm and Decision Tree for Medical Problem”, in 2nd International Conference on Intelligent Computational Systems (ICS'2012), Indonesia, 2012, 124-126.

[12] Chen, K.H., Wang, K.J., Tsai, M.L., Wang, K.M., Adrian, A.M., Cheng, W.C., et al. (2014). “Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm”, BMC Bioinformatics.

[13] Kuhn M. and Johnson, K., (2013) Applied Predictive Modeling: Springer, 2013.

[14] RuleQuestResearch. (2012). Is See5/C5.0 Better Than C4.5? Available: http://rulequest.com/see5-comparison.html.

[15] Schapire R.E. and Freund, Y., (2013). Boosting: Foundations and Algorithms.

[16] Pashaei, E., Ozen, M. and Aydin, N., "Improving Medical Diagnosis Reliability Using Boosted C5.0 Decision Tree empowered by Particle Swarm Optimization", in 37th Annual International Conference of the Engineering in Medicine and Biology Society, Milano.

[17] Pashaei, E., Ozen, M. and Aydin, A., (2015). "A Novel Gene Selection Algorithm for cancer identification based on Random Forest and Particle Swarm Optimization", presented at the Computational Intelligence in Bioinformatics and Computational Biology, Niagara Falls, Canada.

[18] Breiman, L., Friedman, H.J., Olshen, R. and Stone, C., (1984). Classification and Regression Trees: CRC Press.

[19] Loh, W.Y., (2011). "Classification and regression trees (overview)", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1:14-23.

[20] Bastian, C.D. and Rempala, G.A., (2011). "Gene Selection with Sequential Classification and Regression Tree Algorithm", Biostatistics, Bioinformatics and Biomathematics, 2:157–186.

[21] Wu, X., Kumar, V., Quinlan, J.R., Ghosh,J., Yang, Q., Motoda, H., et al., (2008). "Top 10 algorithms in data mining", Knowledge and Information System, 14:1–37.

[22] Saeys, Y., Inza, I., and Larrañaga, P., (2007). "A review of feature selection techniques in bioinformatics", Bioinformatics, 23:2507-2517.

[23] Liu, J.J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., et al. (2005). "Multiclass cancer classification and biomarker discovery using GA-based algorithms", Bioinformatics, 21:2691-2697.

[24] Harb, H.M. and Desuky, A.S., (2014). "Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization", International Journal of Computer Applications, 14-17.

[25] Sahua, B., Mishrab, D., (2012). "A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data", in: International Conference on Modeling Optimization and Computing (ICMOC-2012), 27-31.

[26] Mohamad, M.S., Omatu, S., Deris, S., Yoshioka, M., (2009). "Particle swarm optimization for gene selection in classifying cancer classes", Artificial Life and Robotics, 14:16-19.

[27] Moraglio, A., Chio, C.D., Togelius, J., Poli, R., (2008). "Geometric Particle Swarm Optimization", Journal of Artificial Evolution and Applications.

[28] Chuanga, L.Y., Changb, H.W., Tuc, C.J., Yang, C.H., (2008). "Improved binary PSO for feature selection using gene expression data", Computational Biology and Chemistry, 32:29–38.

[29] Tran, B., Xue, B., Zhang, M., (2014). "Improved PSO for Feature Selection on High-Dimensional Datasets", Springer International Publishing Switzerland, 503-515.

[30] Li, S., Wu, X. and Tan, M., (2008). "Gene selection using hybrid particle swarm optimization and genetic algorithm", Soft Computing (Springer), 12, 1039–1048.

[31] Kumar, P.G., Victoire, T.A., Renukadevi, P., Devaraj, D., (2012). "Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm", Expert Systems with Applications, 39:1811–1821.

[32] Moteghaed, N.Y, Maghooli, K., Pirhadi, S., Garshasbi, M., (2015). "Biomarker Discovery Based on Hybrid Optimization Algorithm and Artificial Neural Networks on Microarray Data for Cancer Classification", Journal of Medical Signals & Sensors, 5:88-96.

[33] Chuang, L., CHU, Y., Li, J.C., Yang, C., (2012). "A Hybrid BPSO-CGA Approach for Gene Selection and Classification of Microarray Data", Journal of computational biology, 19:68-82.

[34] Gonzalez, F. and Belanche, L.A., (2014). "Feature Selection for Microarray Gene Expression Data using Simulated Annealing guided by the Multivariate Joint Entropy", Comput Syst, 18:275-293.

[35] Cao, B., Shen, D., Sun, J.T., Yang, Q., and Chen, Z., (2007). "Feature selection in a kernel space", in International conference on machine learning (ICML), USA, 121–128.

[36] Maroño, N.S., Betanzos, A.A. and Sanromán, M.T., (2007). "Filter Methods for Feature Selection – A Comparative Study", in Intelligent Data Engineering and Automated Learning (IDEAL), 178-187.

[37] Hall, M.A., (1999). "Correlation-based Feature Selection for Machine Learning", New Zealand.

[38] Kohavi, R. and John, G.H., (1997). "Wrappers for feature subset selection", Artificial Intelligence - Special issue on relevance, 97: 273–324.

[39] Hatamlou, A., (2013). "Black hole: A new heuristic optimization approach for data clustering", Information sciences, 222:175-184.

[40] Lenin, K., Reddy, B.R. and Kalavathi, M.S., (2014). "Black Hole Algorithm for Solving Optimal Reactive Power Dispatch Problem", International Journal of Research in Management, Science and Technology, 2:2321-3264.

[41] Farahmandian M. and Hatamlou, A., (2015). "Solving optimization problem using black hole algorithm", Journal of Advanced Computer Science andTechnology, 4:68-74.

[42] Zhang, J., Liu, K., Tan, Y. and He, X., (2008). "Random black hole particle swarm optimization and its application", in IEEE International Conference on Neural Networks and Signal Processing, China, 359 - 365.

[43] Ghaffarzadeh N. and Heydari, S., (2015). "Optimal Coordination of Digital Overcurrent Relays using Black Hole Algorithm", TI Journals of World Applied Programming, 5:50-55.

[44] Kumar, S., Datta, D. and Singh, S.K., (2014). "Black Hole Algorithm and Its Applications", Computational Intelligence Applications in Modeling and Control, 575, 147-170.

[45] Fendler, A., Jung, M., Stephan, C., Honey, R.J., Stewart, R.J., et al. (2011). "miRNAs can predict prostate cancer biochemical relapse and are involved in tumor progression", Int J Oncol, 39(5):1183-92.

[46] Barron, N., Keenan, J., Gammell, P., Martinez, V.G., Freeman, A., et al. (2012). "Biochemical relapse following radical prostatectomy and miR-200a levels in prostate cancer", Prostate, 72: 1193–1199.

[47] Bhatnagar, N., Li, X., Padi, S.K., Zhang, Q., Tang, M.S., et al. (2010). "Downregulation of miR-205 and miR-31 confers resistance to chemotherapy-induced apoptosis in prostate cancer cells", Cell Death Dis, 1:e105.

[48] Ross-Adams, H., Lamb, A.D., Dunning, M.J., Halim, S., Lindberg, J., et al. (2015). "Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study", EBioMedicine, 2(9):1133-44.

[49] Sun, Y., Goodison, S., (2009). "Optimizing molecular signatures for predicting prostate cancer recurrence", prostate, 69(10):1119-27.

[50] Mortensen, M.M., Høyer, S., Lynnerup, A.S., Ørntoft, T.F., Sørensen, K.D., et al. (2015). "Expression profiling of prostate cancer tissue delineates genes associated with recurrence after prostatectomy", Scientific Reports, 5: 16018.

[51] Ozen, M., Creighton, C.J., Ozdemir, M., Ittmann, M., (2008). "Widespread deregulation of microRNA expression in human prostate cancer. Oncogene", 27:1788–1793.

[52] Schaefer, A., Jung, M., Mollenkopf, H.J., Wagner, I., Stephan, C., et al. (2010). "Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma", Int J Cancer, 126:1166–1176.

[53] Tong, A.W., Fulgham, P., Jay, C., Chen, P., Khalil, I., et al. (2009). "MicroRNA profile analysis of human prostate cancers", Cancer Gene Ther, 16:206–216.

[54] Karatas, O.F., Guzel, E., Suer, I., Ekici, I.D., et al. (2014). "miR-1 and miR-133b are differentially expressed in patients with recurrent prostate cancer", PLoS One, 9(6):e98675. PMID: 24967583.

[55] Long, Q., Johnson, B.A., Osunkoya, A.O., Lai, Y.H., Zhou, W., et al. (2011), "Protein-coding and microRNA biomarkers of recurrence of prostate cancer following radical prostatectomy", Am J Pathol, 179(1):46-54.

[56] Mortensen, M.M., Høyer, S., Orntoft, T.F., Sørensen, K.D., Dyrskjøt, L., Borre, M., (2014). "High miR-449b expression in prostate cancer is associated with biochemical recurrence after radical prostatectomy", BMC Cancer, 14:859. doi: 10.1186/1471-2407-14-859.

[57] Zheng, Q., Peskoe, S.B., Ribas, J., Rafiqi, F., Kudrolli, T., et al. (2014), "Investigation of miR-21, miR-141, and miR-221 expression levels in prostate adenocarcinoma for associated risk of recurrence after radical prostatectomy", Prostate. 74:1655–1662

[58] Leite, K.R., Reis, S.T., Viana, N., Morais, D.R., Moura, C.M., et al. (2015), "Controlling RECK miR21 Promotes Tumor Cell Invasion and Is Related to Biochemical Recurrence in Prostate Cancer", Journal of Cancer, 6(3):292-301. doi: 10.7150/jca.11038.

[59] Hong, F., Breitling, R., (2008). "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments", Bioinformatics, 24(3):374-82.

[60] Kristensen, H., Thomsen, A.R., Haldrup, C., Dyrskjøt, L., Høyer, S., et al., (2016). "Novel diagnostic and prognostic classifiers for prostate cancer identified by genome-wide microRNA profiling", Oncotarget. 7(21):30760-71. doi: 10.18632/oncotarget.8953.

[61] Bell, E.H., Kirste S., Fleming, J.L., Stegmaier, P., Drendel, V., et al., (2015), "A novel miRNA-based predictive model for biochemical failure following post-prostatectomy salvage radiation therapy", PLoS One, 10(3):e0118745. doi: 10.1371/journal.pone.0118745. e Collection 2015.

[62] Sachdeva, M., Liu, Q., Cao, J., Lu, Z., Mo, Y.Y., (2012). "Negative regulation of miR-145 by C/EBP-beta through the Akt pathway in cancer cells", Nucleic Acids Res, 40(14):6683-92.

[63] Avgeris, M., Stravodimos, K., Fragoulis, E.G., Scorilas, A., (2013). "The loss of the tumour-suppressor miR-145 results in the shorter disease-free survival of prostate cancer patients", Br J Cancer, 108(12):2573-81.

[64] Su, J., Liang, H., Yao, W., Wang, N., Zhang, S., Yan, X., et al. (2014). "MiR-143 and MiR-145 regulate IGF1R to suppress cell proliferation in colorectal cancer", PLoS One, 9(12):e114420.

[65] Shao, Y., Qu, Y., Dang, S., Yao, B., Ji, M., (2013). "MiR-145 inhibits oral squamous cell carcinoma (OSCC) cell growth by targeting c-Myc and Cdk6", Cancer Cell Int, 13(1):51.

[66] Spizzo, R., Nicoloso, M.S., Lupini, L., Lu, Y., Fogarty, J., Rossi, S., et al. (2010). "miR-145 participates with TP53 in a death-promoting regulatory loop and targets estrogen receptor-alpha in human breast cancer cells", Cell Death Differ, 17(2):246-54.

[67] Riester, M., Taylor, J.M., Feifer, A., Koppie, T., Rosenberg, J.E., Downey, et al. (2012). "Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer", Clin Cancer Res, 18:1323–1333.

[68] Kaper, L., Heuvel, E., Woudt, P. and Giacconi, R., (1999). "Black hole research past and future", in Black Holes in Binaries and Galactic Nuclei: Diagnostics, Demography and Formation, Berlin/Heidelberg, 3–15.

[70] Pickover, C.A., (1999). Black Holes: A Traveler's Guide. United States of America: John Wiley & Sons.

[71] Mirjalili, S. and lewis, A., (2013). "S-shaped versus V-shaped transfer functions for binary Particle Swarm Optimization", Swarm and Evolutionary Computation, 9:1–14.

[72] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., et al. (1999). "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", Science, 286:531 – 537.

[73] Lavanya, D. and Rani, K.U., (2011). "Analysis of feature selection with classification: Breast cancer datasets", Indian Journal of Computer Science and Engineering (IJCSE).

[74] Lavanya, D., (2012). "Ensemble decision tree classifier for breast cancer data", International Journal of Information Technology Convergence and Services, 2(1):17-24.

[75] Darzi, M., AsgharLiaei, A., Hosseini, M., Asghari, H., (2011). "Feature selection for breast cancer diagnosis: a case-based wrapper approach", Engineering and Technology International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering, 5(5).

[76] Sridevi, T. and Murugan, A., (2014). "A novel feature selection method for effective breast cancer diagnosis and prognosis", International Journal of Computer Applications, 88 (11):28-33.

[77] Caglar, M.F., Cetisli, B., Toprak, I.B., (2010). "Automatic recognition of parkinson's disease from sustained phonation tests Using ANN and adaptive neuro-fuzzy classifier", Journal of Engineering Science and Design, 1(2):59-64.

[78] Ozcift, A. and Gulten, A., (2011). "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms", Computer Methods and Programs in Biomedicin (elsevier), 104:443-451.

[79] Psorakis, I., Damoulas, T. and Girolami, M.A., (2010). "Multiclass relevance vector machines: sparsity and accuracy", IEEE Transactions on Neural Networks, 21(10):1588–1598.

[80] Peker, M., Şen, B. and Delen, D., (2015). "Computer-Aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm", Journal of Healthcare Engineering, 6(3):281–302.

[81] Subbulakshmi, C.V. and Deepa, S.N., (2015). "Medical dataset classification: a machine learning paradigm integrating Particle Swarm Optimization with extreme learning machine classifier", The Scientific World Journal (Hindawi Publishing).

[82] Mart´ınez-Estudillo, F.J., Herv´as-Mart´ınez, C., Guti´errez, P.A. and Mart´ınez-Estudillo, A.C., (2008). "Evolutionary product-unit neural networks classifiers", Neurocomputing, 72(1–3):548–561.

[83] Herv´as-Mart´ınez, C., Mart´ınez-Estudillo, F.J. and Carbonero-Ruz, M., (2008). "Multilogistic regression by means of evolutionary product-unit neural networks", Neural Networks, 21(7):951–961.

[84] Jaganathan, P. and Kuppuchamy, R., (2013). "A threshold fuzzy entropy based feature selection for medical database classification", Computers in Biology and Medicine, 43(12): 2222–2229.

[85] El Akadi, A., Amine, A., El Ouardighi, A. and Aboutajdine, D., (2011). "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper", Knowledge and Information Systems, 26(3):487–500.

[86] Wang, X. and Gotoh, O., (2009). "Microarray-based cancer prediction using soft computing approach", Cancer Informatics7, 123-139.

[87] Huerta, E., Duval, B. and Hao, J., (2008). "Gene selection for microarray data by a LDA-based Genetic Algorithm", Springer, 252–263.

[88] Alba, E., Garcıa-Nieto, J., Jourdan, L. and Talbi, E., (2007). "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms", IEEE Transactions on Evolutionary Computation, 284–290.

[89] Abdi, M., Hosseini, S. and Rezghi, M., (2012). "A novel weighted support vector machine based on Particle Swarm Optimization for gene selection and tumor classification", Corporation Computational and Mathematical Methods in Medicine (Hindawi Publishing).

[90] Deng, H. and Runger, G., (2012). "Feature selection via regularized trees", in: Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), IEEE.

[91] Benjamini, Y. and Hochberg, Y., (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", Wiley: Journal of the Royal Statistical Society. Series B (Methodological).

[92] Seven, M., Karatas, O.F., Duz, M.B., Ozen, M., (2014). "The role of miRNAs in cancer: from pathogenesis to therapeutic implications", Future Oncol; 10(6):1027-48.

[93] Ichimi, T., Enokida, H., Okuno, Y., Kunimoto, R., Chiyomaru, T., Kawamoto, K., et al. (2009). "Identification of novel microRNA targets based on microRNA signatures in bladder cancer". Int J Cancer. 125(2):345-52.

[94] Karatas, O.F., Suer, I., Yuceturk, B., Yilmaz, M., Hajiyev, Y., Creighton, C.J., et al. (2015). "The role of miR-145 in stem cell characteristics of human laryngeal squamous cell carcinoma Hep-2 cells", Tumour Biol.

[95] Das, A.V., Pillai, R.M., (2015). "Implications of miR cluster 143/145 as universal anti-oncomiRs and their dysregulation during tumorigenesis", Cancer Cell Int.15:92.

[96] Karatas, O.F., Yuceturk, B., Suer, I., Yilmaz, M., Cansiz, H., Solak, M., et al. (2016). "Role of miR-145 in human laryngeal squamous cell carcinoma", Head Neck. 38(2):260-6.

[97] Hegre, S.A., Saetrom, P., Aas, P.A., Pettersen, H.S., Otterlei, M., Krokan, H.E., (2013). "Multiple microRNAs may regulate the DNA repair enzyme uracil-DNA glycosylase", DNA Repair (Amst), 12(1):80-6.

[98] Weeks, L.D., Fu, P., Gerson, S.L., (2013). "Uracil-DNA glycosylase expression determines human lung cancer cell sensitivity to pemetrexed", Mol Cancer Ther, 12(10):2248-60.

[99] Li, Y.L., Ye, F., Cheng, X.D., Hu, Y., Zhou, C.Y., Lu, W.G., et al. (2010). "Identification of glia maturation factor beta as an independent prognostic predictor for serous ovarian cancer", Eur J Cancer, 46(11):2104-18.

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname : Elnaz PASHAEI

Date of birth and place :1985- Iran-Urmia

Foreign Languages : Persian (native)-English-Turkish

E-mail :Elnaz.pashaei@yildiz.edu.tr

EDUCATION

Degree	Department	University	Date of Graduation
Master	Computer Engineering	Qazvin Islamic Azad University, IRAN	2010 -2012
Undergraduate	computer Engineering	Khoy Islamic Azad University, IRAN	2005-2009

PUBLISHMENTS

Papers

1. Elnaz Pashaei, Nizamettin Aydin, “Binary Black Hole Algorithm for Feature Selection and Classification on Biological Data”, Applied Soft Computing. Vol.56, (2017) 94-106

2. Elnaz Pashaei, Esra Guzel, Mete Emir Ozgurses, Goksun Demirel, Nizamettin Aydin, Mustafa Ozen, “A Meta-Analysis: Identification of Common Mir-145 Target Genes that have Similar Behavior in Different GEO Datasets”, PLoS ONE. Vol. 11, Issue 9, September 2016. doi:10.1371/journal.pone.0161491.

3. Elnaz Pashaei, Elham Pashaei, Maryam Ahmady, Mustafa Ozen, Nizamettin Aydin. “Meta-analysis of miRNA Expression Profiles for Prostate Cancer Recurrence following Radical Prostatectomy”, PLoS ONE. Vol. 12, Issue 6, Jun 2017. doi:10.1371/journal.pone.0179543.

4. Elnaz Pashaei, Nizamettin Aydin, “Gene selection using hybrid binary black hole algorithm and binary particle swarm optimization”, Genomics (under revision).

Conference Papers

1. Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, “Improving Medical Diagnosis Reliability Using Boosted C5.0 Decision Tree empowered by Particle Swarm Optimization”, Proceedings of 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Milano, Italy. 25-29 Aug 2015. IEEE Press. PP. 7230-7233.

2. Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, “A Novel Gene Selection Algorithm for cancer identification based on Random Forest and Particle Swarm Optimization”, Proceedings of 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Niagara Falls, Canada. 12-15 Aug

2015. IEEE Press. pp. 1-6.

3. Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, “An application of black hole algorithm and decision tree for medical problem”, Proceedings of 2015 IEEE International Conference on Bioinformatics and Bioengineering (BIBE). Belgrade, Serbia. 2-4 Nov 2015. IEEE Press. pp. 1-6.

4. Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, “Gene selection and classification approach for microarray data based on Random Forest Ranking and BBHA”, Proceedings of 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). Las Vegas, NV, USA. 24-27 Feb 2016. IEEE Press. pp. 308-311.

5. Elnaz Pashaei, Mustafa Ozen, Nizamettin Aydin, “Biomarker discovery based on BHA and AdaboostM1 on microarray data for cancer classification”, Proceedings of 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Orlando, FL, USA. 16-20 Aug 2016. IEEE Press. pp. 3080-3083.