Artificial Intelligence in Medicine (2004) 31, 183–196

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Data mining and genetic algorithm based gene/SNP selection

Shital C. Shah, Andrew Kusiak*

Intelligent Systems Laboratory, MIE, 2139 Seamans Center, The University of Iowa, Iowa City, IA 52242-1527, USA

Received 31 August 2003; received in revised form 7 February 2004; accepted 3 April 2004

KEYWORDS
Single nucleotide polymorphisms (SNPs); Genes; Feature selection; Data mining; Genetic algorithm; Intersection approach; Drug effectiveness

Summary *Objective:* Genomic studies provide large volumes of data with the number of single nucleotide polymorphisms (SNPs) ranging into thousands. The analysis of SNPs permits determining relationships between genotypic and phenotypic information as well as the identification of SNPs related to a disease. The growing wealth of information and advances in biology call for the development of approaches for discovery of new knowledge. One such area is the identification of gene/SNP patterns impacting cure/drug development for various diseases. *Methods:* A new approach for predicting drug effectiveness is presented. The approach is based on data mining and genetic algorithms. A global search mechanism, weighted decision tree, decision-tree-based wrapper, a correlation-based heuristic, and the identification of intersecting feature sets are employed for selecting significant genes. *Results:* The feature selection approach has resulted in 85% reduction of number of features. The relative increase in cross-validation accuracy and specificity for the significant gene/SNP set was 10% and 3.2%, respectively. *Conclusion:* The feature selection approach was successfully applied to data sets for drug and placebo subjects. The number of features has been significantly reduced while the quality of knowledge was enhanced. The feature set intersection approach provided the most significant genes/SNPs. The results reported in the paper discuss associations among SNPs resulting in patient-specific treatment protocols.

1. Introduction

Genomic studies provide enormous volume of data for thousands of genes (segment of DNA that encodes RNA). The technology used to produce genomic expression data or single nucleotide polymorphisms (SNPs) is expensive. Associating sequence variations with heritable phenotypes is a key facet in genetic research [1]. The most widespread variations are single base pair differences, i.e., SNPs occurring approximately once every 100–300 bases.

The human genome is estimated to contain 10 million SNPs of which around 300,000 have significant genetic variations [2]. They are primarily responsible for the variation between humans as they determine among others, person's skin color, hair, immune response, and adverse effects due to drugs. They promise to significantly advance our ability to understand and treat diseases [3]. Genetic information in the DNA is transcribed to RNA and then translated to proteins, thus genetic polymorphisms indirectly affect the metabolism and

*Corresponding author. Tel.: +1 319 335 5934; fax: +1 319 335 5669.
E-mail address: andrew-kusiak@uiowa.edu (A. Kusiak).
URL: http://www.icaen.uiowa.edu/~ankusiak.

0933-3657/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2004.04.002

disposition of many medications. SNPs may be shared among groups of people with harmful but unknown mutations and serve as markers for them. Such markers help unearth the mutations and accelerate efforts to find therapeutic drugs. Thus polymorphisms in genes encoding (the receptors—targets of medications) can alter the pharmacodynamics of the drug response by changing receptor sensitivity.

Genetic profile of each individual (subject) can be assembled using the data produced by SNP-mapping technologies. Analysis of such data may lead to genes/SNP patterns that may be responsible for common diseases as well as genetic risk. Due to high cost, a typical data set (containing as many as 300,000 SNPs) is available for limited number of subjects (500—1000 patients). To handle such data sets there is a need to select the most informative genes/SNPs [4] for further analysis. Removal of uninformative genes/SNPs decreases the noise, confusion, and complexity [5], and increases chances for identification of most informative genes, classification of diseases, and prediction of various outcomes, e.g., effectiveness of a cancer therapy.

With the advancing technology, molecular pharmacology, and functional relationship of polymorphisms, there is a need for computational tools to determine drug responses [6]. These tools are needed to discover associations among alleles (the chemical bases such as adenine, guanine, thymine, cytosine) at different SNPs and between phenotypic and genotypic features [7].

This paper focuses on feature reduction approaches that can be effectively applied to SNP data sets. It discusses weighted decision-tree-based gene selection (WDTGS), genetic algorithm-based gene selection (GAGS), and feature set intersection approaches. The features derived from the data sets are evaluated in terms of the cross-validation accuracy, specificity, and number of features against the complete set of all features (baseline measurements).

2. Background

Clustering [8], data mining [9—11] gene identification [12], and gene regulatory network modeling [13,14] are used to perform DNA analysis. Data mining algorithms are commonly applied to analyze gene expression data. Data mining is the process of discovering interesting and previously unknown patterns in data sets [15]. The main emphasizes of data mining is on individual subject rather than the population, providing an avenue for personalization [16]. Several computational techniques have been applied for gene expression classification problems, including Fisher linear discriminant analysis [17], *k* nearest neighbor [18], decision-tree, multi-layer perceptron [19], support vector machines [20], self-organizing maps [4], hierarchical clustering [21], and graph theoretic approaches [22].

The goal of feature selection is to identify the minimum set of non-redundant features (e.g., SNPs, genes) that are useful in classification [5]. This can be achieved through various supervised and unsupervised methods such as neighborhood analysis [17], Pearson correlation, Spearman correlation, cosine coefficient, information gain, mutual information, and signal to noise ratio [8], clustering [5], principal component analysis, combining features (i.e., creating hybrid features), independent components analysis [23], supervised feature reduction by iteratively applying a supervised grouping (classification) algorithm, and eliminating the lowest weight features. DNA gene expression data sets are pruned by eliminating insignificant features. The results of the study performed to investigate the distribution of SNPs in CAPN10 gene in Chinese population and their impact on type two diabetes mellitus in Han people of Northern China is reported in [24]. The transmission-disequilibrium test (TDT) and sib transmission-disequilibrium test (STDT) was applied for analyzing the SNPs. They used statistical techniques to examine the SNPs and determined that there was no significant statistical difference between the two ethnic groups based on the CAPN10 gene. They examined a pre-selected gene and corresponding SNPs rather than investigating all potential genes/SNPs.

There is a need to develop a procedure that begins with the collection of sequences and ends with the creation of SNP data sets. Several strategies both experimental and based on computational intelligence have been devised for SNP discovery and mapping [25]. Experimental SNP discovery requires arduous, intricate, and expensive experimental procedures. The four main experimental SNP discovery methods are identification of single strand conformation polymorphisms (SSCPs), heteroduplex analysis, direct DNA sequencing, and variant detector arrays (VDAs) [26].

Computational intelligence-based discovery uses large-scale data sets with SNP information that might have been generated for other purposes, e.g., routine clinical studies. Feature selection approaches such as principal component analysis, information gain, clustering algorithms, and regression can be implemented but may not provide the best solution. To identify the most informative SNPs, there is a need for a global search mechanism

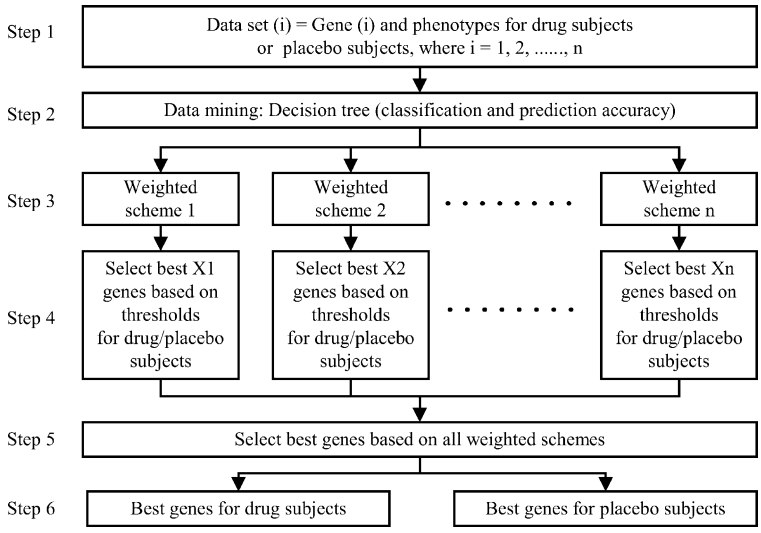

Figure 1 Weighted decision-tree-based selection (WDTGS) of genes/SNPs.

(genetic algorithms (GA) [27–30]) coupled with decision trees [31], domain experts, and multi-angle identification process. The domain experts can provide the essential knowledge for screening of genes/SNPs while the multi-angle identification process utilizes the computational and experimental models to identify and validate the significant genes/SNPs. The proposed approach provides such functionality.

3. Proposed approach

A typical data set in pharmaceutical industry includes data for drug and placebo subjects, normal and abnormal subjects, and genotypic and phenotypic data. The purpose of the research reported in this paper is to derive the most significant features that reflect the best interactions between genotypic and phenotypic data, drug effectiveness, and natural recovery (placebo related improvement) genes/SNPs. Three main approaches are proposed, namely the WDTGS, GAGS, and the feature set intersection approach.

3.1. Weighted decision-tree-based gene selection (WDTGS)

Partitioning a data set into drug and placebo subjects, application of data mining algorithms, and applying various weighted schemes initiates the WDTGS approach. This leads to the identification of significant genes/SNPs set per weighted scheme. Finally, the set of most significant genes is determined by intersecting all significant genes/SNPs sets.

The data sets are initially partitioned in the placebo and drug category. The analysis is performed independently for each set with the decision variable as a measure (e.g., test scores, lab findings, etc.) of improvement over time (i.e., decision value = final measure − initial measure) (see Step 1 of Fig. 1). The discretized decisions (Good and Bad) for placebo and drug sets are considered.

In Step 1 (Fig. 1) a data set is formed for each gene with more than three to five SNPs and the decision feature. The decision-tree algorithm [31] is applied in Step 2 (Fig. 1), which produces rules in the following format:

IF USP_SNP6 = T_T **AND** USP_SNP2
= T_T **AND** USP_SNP4 = A_G **AND** USP_SNP5
= G_G **THEN** Decision = D_BAD

Classification and prediction accuracy are used in data mining as the quality metrics [32]. In this paper, classification accuracy is defined as the ability of a gene to best explain the ''training'' data set. While the prediction accuracy (with the result of 10-fold cross-validation) is defined as ability of a gene to accurately predict a ''test'' data set. For example, for the data set containing the USP gene the classification accuracy is 71.49%, while the prediction accuracy is 56.09% (see Table 1).

Maximizing classification accuracy may lead to overfitting of the data and decreasing the prediction accuracy (attained by cross-validation). Thus a balance between classification and prediction accuracy needs to be maintained. It is accomplished by employing multiple user-defined weighting schemes (Step 3, Fig. 1) as illustrated next.

$$\text{weighted accuracy(i)} = \{A_i \times \text{classification accuracy} + B_i \times \text{prediction accuracy}\} \quad \text{with} \quad \{A_i + B_i = 1\} \tag{1}$$

Table 1 Example of classification and prediction measurements

Classification		Prediction	
Gene	Correct	Gene	Correct
USP	71.49	USP	56.09
BRH	59.32	BRH	49.24
NBC	64.07	NBC	53.04

Table 2 Ranking of combined weighted accuracy

WS1			WS2		
Rank	Gene	Correct	Rank	Gene	Correct
1	USP	66.87	1	USP	60.71
2	NBC	60.76	2	NBC	56.35
3	BRH	56.30	3	BRH	52.26

A_i and B_i are the weights of the ith weighting scheme. For example, if the classification is more critical than prediction, then the scheme can be 0.7 $\times$ classification accuracy + 0.3 $\times$ prediction accuracy, i.e. $(0.7 \times 71.49) + (0.3 \times 56.09) = 66.87\%$ for the USP gene (Table 2).

The results from each weighted scheme are ranked in the descending order of the weighted accuracy (Table 2). To select the number of genes for further analysis, two criteria are applied a threshold on the number of selected genes and a threshold combined weighted accuracy. For example, the threshold values can be 15 genes and 60% combined weighted accuracy. Ranked weighted accuracy for each weighted scheme will lead to potentially different genes set (Step 4, Fig. 1). If the threshold values for Table 2 are 2 genes and 60% combined weighted accuracy, then ranked weighted scheme 1 (WS1) will choose USP and NBC genes, while the ranked weighted scheme 2 (WS2) will choose only USP gene.

To obtain a final significant gene set (Step 5, Fig. 1), the intersection of all ranked gene sets is generated (Fig. 1). The selected genes satisfy various weighted schemes and form a multi-objective solution. The intersection of WS1 and WS2 weighted schemes results in the USP gene (Table 2).

The same procedure is repeated for placebo data sets. There are two sets of significant genes, one each of placebo and drug subjects (Step 6, Fig. 1).

3.2. Genetic algorithm-based gene selection (GAGS)

Partitioning the data into drug and placebo sets initiates the GAGS. The genetic algorithm-based feature-selection mechanisms such as a correlation-based heuristic and decision-tree wrapper approach are independently used to evaluate the quality of the genes/SNPs. The analysis of the outputs, i.e., frequency, results in the identification of the significant genes/SNPs for both drug and placebo sets. A brief introduction to the algorithms used by GAGS is presented next.

GA [27–30] is a search algorithm using the concepts from biology. A GA is initiated with a set of solutions (represented by chromosomes) called the population. Each solution in the population is evaluated in terms of its fitness. Solutions chosen to form new chromosomes (offspring) are selected according to their fitness, i.e., the more suitable they are the higher likelihood they will reproduce. This is repeated until a stopping condition (for example, the number of populations or improvement of the best solution) is satisfied. GA searches the solution space without following crisp constraints and potentially samples the entire feasible solution region. This provides a chance of visiting the previously unexplored space and there is a high possibility of achieving overall optimal/near-optimal solution, making the GA a global search mechanism.

With the GA as a global search tool, feature selection can be performed using two approaches, namely filter and wrapper search [33–35]. The wrapper search uses machine-learning algorithm, decision tree wrapper (DTW), to evaluate the GA solutions [34,36,37]. The filter approach evaluates the features using heuristic-based characteristics (e.g., correlation) of the data. Correlation-based feature selection (CFS) filter is a fast and effective way for feature selection [35]. It selects a feature if it correlates with the decision outcome but not to any other feature that has already been selected.

Partitioning the data into drug and placebo sets along with the decision forms the initial step of the GAGS approach (Step 1, Fig. 2). The drug data set with n features (all SNPs for all genes) and m observations (subjects) is evaluated using GAGS (i.e., GA–CFS and GA–DTW) approaches (Step 2, Fig. 2).

To avoid local optima, the GA–CFS approach (Step 3a, Fig. 2) applies the correlation-based heuristic n times ($n = 10$–30) to each (drug and placebo) data set. The output provides the frequency, i.e., number of times the feature was selected (Table 3). A higher value of the frequency indicates superior quality of the selected feature. The frequency is sorted in the descending order for easy identification of the quality features. A threshold on number of features selected as well as the threshold frequency can be set for inclusion in the final feature

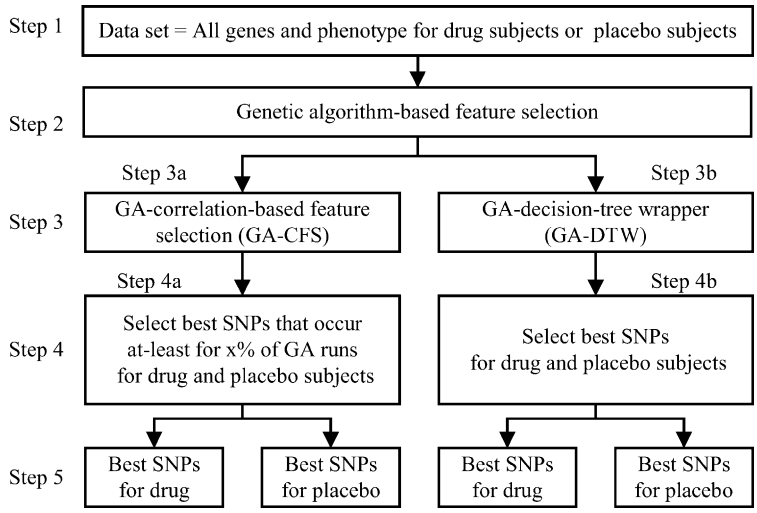

Figure 2 Genetic algorithm-based feature selection.

set (Step 4a, Fig. 2). For example, the threshold frequency can be set as = 80% (i.e., 8 out of 10 times). The features selected for this threshold are ABC_SNP6, CBS_SNP9, CBS_SNP1, and PAT_SNP4 (Table 3).

The GA—DTW approach (Step 3b, Fig. 2) is computationally intensive as it builds decision trees for each solution investigated by the GA. To gain confidence in the selected features, the GA—DTW approach is replicated (n times, where $n = 10$—30). The computational time is justified as it is performed only ones. In absence of multiple replications, all the outputted features form the final features set (Step 4b, Fig. 2). This single replication feature set is still reliable as it was obtained through global GA search supported by DTW with five-fold cross-validation.

The GA—DTW and GA—CFS approaches provide a set of potentially high quality features (Step 5, Fig. 2). The number of selected features is substantially reduced from that of the original data set. The same procedure is applied to the placebo data sets. Thus there are four features sets (two data sets multiplied by two GA approaches).

3.3. Feature set intersection approach

To further reduce the number of features, two or more feature sets (obtained in the previous section) are combined (Fig. 3). The generated intersection provides important features as they are selected by more than one approach, while the union may provide knowledge that may have been missed by one of the approaches. For example, the intersection of the GA—CFS and WDTGS for drug data sets has been performed (see Table 4). The same procedure is performed for the placebo data set.

Table 3 GA—CFS: frequency output with feature quality

No.	Unsorted		Sorted		SNP quality
	Frequency	Attribute	Frequency	Attribute	
1	10	ABC_SNP6	10	ABC_SNP6	Good
2	1	AUS_SNP3	10	CBS_SNP9	
3	9	CBS_SNP1	9	CBS_SNP1	
4	10	CBS_SNP9	8	PAT_SNP4	
5	3	CDE_SNP4	6	WXY_SNP1	Moderate
6	8	PAT_SNP4	3	CDE_SNP4	Bad
7	2	TUV_SNP1	2	TUV_SNP1	
8	0	USP_SNP2	1	AUS_SNP3	
9	6	WXY_SNP1	0	USP_SNP2	No use
10	0	WXY_SNP3	0	WXY_SNP3	

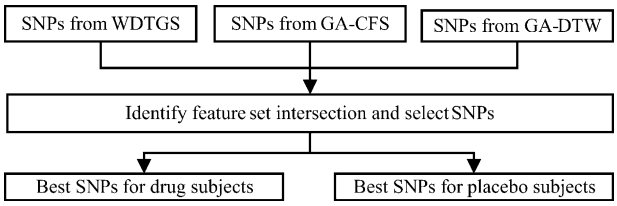

Figure 3 Identification of the feature set intersection.

In the first iteration, the WDTGS, GAGS, and the feature set intersection approaches reduce the number of features. To further reduce the number of features, the above approaches can be re-applied (iteratively) to each reduced data set (Fig. 4). The iterative process is terminated, if the 10-fold cross-validation accuracy deteriorates or the feature set remains static.

3.4. Evaluating selected feature sets

The WDTGS, GAGS, and the feature set intersection approaches provide four feature sets, each for the drug and placebo data set (i.e., eight feature sets in total). To evaluate the quality of each feature set, baseline accuracy (10-fold cross-validation) and specificity (true-negative rate, i.e., predicting improvement when given that individuals have had improvement due to drug/placebo treatment) is used. The baseline accuracy and specificity are obtained by performing data mining on all features for the drug and placebo data set. A decision-tree algorithm with default values and 10-fold cross-validation can be applied. All other feature sets (generated from various proposed approaches) are also mined (with 10 fold-cross-validation) using the same decision-tree algorithm with same default values as the baseline. This forms the bases for a fair comparison. The next quality measure of the feature set is the number of features pruned, while maintaining or improving the cross-validation accuracy. A separate analysis of the drug and placebo data was performed using the information gain (IG) and standard regression (REG) (with the-best-first search [33]) approaches that are reported in the literature. This provided an additional quality measure. Even, if the feature set accuracy did not increase, dealing with smaller number of features is advantageous. Ideally, the percentage reduction in the number of features should be meaningful.

The benefits of the above feature selection approaches are that they consider the training (classification) as well as testing (prediction) accuracy. The GA-based approach selects combinations of genes/SNPs, which are not likely to be selected by traditional approaches due to local optima. The proposed approach has a potential of identifying best performing set of genes/SNPs for drug effectiveness. The complex interactions and associations between genes/SNPs can be conveniently explained by the decision rules in IF-THEN format.

4. Application to a genetic data set

4.1. Data set

The data set used in this paper emulates a standard genetic data set. The naming convention for the genes and SNPs is arbitrary. The genes/SNPs selected for the analyses were based on domain knowledge, nature of the disease, drug structure, pharmacodynamics, pharmacokinetics, molecular pharmacology, etc. Pre-screening of genes/SNPs by above methods narrows the search space, reduces computational effort and allows targeted analysis. The data set (Table 5) consists of five phenotype features, 32 genes with a total of 172 SNPs (Tables 6 and 7) and the number of subjects affected by a disease is 1000.

The data set was divided into two parts similar to the actual clinical trial and one set-representing drug-treated subjects and the other placebo subjects. The decision for each data set was formulated. Thus a subject in the drug set was labeled with Decision = D_GOOD, if the difference between test scores was above 25, else, the Decision = D_BAD. Similarly, for placebo set Decision = P_GOOD, if the difference between test scores was above 12 else the Decision = P_BAD.

4.2. Weighted decision-tree-based gene selection (WDTGS)

Mining is performed on the data for each gene (i.e., 32 runs of the decision-tree algorithm) for drug

Table 4 Feature set for GA—CFS, WDTGS, and GA—CFS—WDTGS

	GA—CFS feature set			WDTGS feature set		GA—CFS—WDTGS feature set	
1	CBS_SNP9	KQE_SNP3	ORH_SNP2	CRH	NBC	CRH_SNP2	WXY_SNP4
2	CRH_SNP2	KQE_SNP7	ORH_SNP5	KQE	WXY	KQE_SNP3	WXY_SNP5
3	JGT_SNP2	NBC_SNP4	WXY_SNP4		KQE_SNP7		
4	JGT_SNP6	NOP_SNP1	WXY_SNP5		NBC_SNP4		

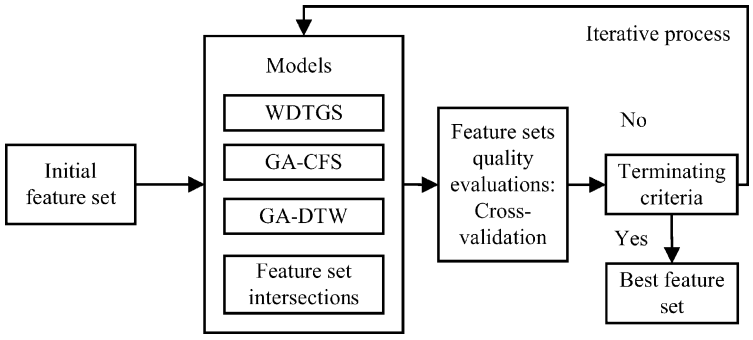

Figure 4 Iterative process for best feature set.

subjects (Fig. 1). The results of classification and prediction accuracy are presented in Table 8.

The weighted scheme 1 is defined as 0.7 × classification accuracy + 0.3 × prediction accuracy. The second weighted scheme is 0.3 × classification accuracy + 0.7 × prediction accuracy. The resulting combined weighted accuracy for each gene, for every weighted scheme is presented in Table 9. The threshold for inclusion of a gene was set to 12 genes with combined weighted accuracy = 55% for each scheme. To obtain a final significant drug gene set (Tables 10 and 12), an intersection of all ranked gene sets was performed (Fig. 1).

The same procedure was performed for the placebo set. The ranked genes for both weighted schemes are shown in Table 11, while the final significant genes are shown in Table 12.

The approach identified 10 and 8 significant genes for the drug and placebo data set, respectively. It can be observed that the sets selected for drug and placebo subjects have common genes (Table 12). These common genes may be indicative of natural improvement of the subjects.

4.3. Genetic algorithm-based gene selection (GAGS)

4.3.1. GA-CFS approach

The drug data set with all features was used to perform GA—CFS (Fig. 2). The 10 fold cross-validation provided results similar to those of Table 3. The values of GA parameters employed in this approach are: 100 GA runs, 100-population size, 0.6-crossover rate, and 0.033 mutation rate. The computational time on a standard PC (Pentium 4) was 130 s. The threshold frequency of 60% (i.e., 6 out of 10 runs) was set for selection of SNPs for the drug set. The set of 63 selected SNPs are provided in Table 13. The GA—CFS approach for placebo set yielded 59 SNPs.

4.3.2. GA-DTW approach

The drug data set with all features was used to execute the GA—DTW based feature selection (Fig. 2). The GA—DTW approach performed single replication with the GA parameters similar to the GA—CFS approach. A five-fold cross-validation with the decision-tree algorithm was used by the wrapper approach. The total number of decision trees built by this approach was [100 (GA runs) × 100 (population size) × 5 (DT five-fold cross-validation)] 50,000 decision trees. Building 50,000 decision trees is a slow and tedious process requiring approximately 82 h of computational time on a standard PC (Pentium 4). The set of 72 selected SNPs for drug set are provided in Table 14. Processing the placebo set with the same approach yielded 90 SNPs.

4.4. The feature-set intersection approach

To further reduce the number of features the intersection of significant genes/SNPs list produced by

Table 5 Symbolic representation of the data set

No.	ABC_SNP1	ABC_SNP2	ABC_SNP3	Gender	Age	Race	Weight	Height	Decision
1	C_G	A_A	C_G	Female	46	1	83.86	164.50	D_BAD
2	G_G	A_G	C_C	Male	46	1	62.14	176.71	D_BAD
3	G_G	G_G	C_G	Female	56	1	80.19	186.41	D_BAD
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
998	G_G	G_G	C_G	Male	69	1	58.31	168.28	P_GOOD
999	C_C	G_G	C_G	Female	49	3	89.77	177.48	P_GOOD
1000	C_C	G_G	C_C	Female	66	1	47.19	153.82	P_GOOD

Table 6 Phenotypic features

No.	Gene	Number of SNPs
1	ABC	6
2	AUS	6
3	BRH	6
4	CBS	9
5	CDE	5
6	CRH	6
7	CRS	5
8	DSS	4
9	EFG	6
10	GRE	4
11	HIJ	5
12	IND	5
13	JGT	6
14	JIT	5
15	KLM	7
16	KQE	7
17	NBC	5
18	NOP	4
19	NOR	5
20	NPR	3
21	ORH	8
22	ORS	9
23	OST	4
24	PAT	6
25	QRS	5
26	QTS	3
27	RHN	5
28	RHP	7
29	STP	6
30	TUV	5
31	USP	8
32	WXY	7

WDTGS and GA—CFS (i.e., GA—CFS—WDTGS) was performed. The resulting set of 26 SNPs for drug set is shown in Table 15. The same approach applied to the placebo data set yielded 21 significant SNPs.

4.5. Evaluating selected feature sets

4.5.1. Baseline measurements

Data mining of a drug set with all features was used to compute the baseline measurements. The baseline cross-validation accuracy and specificity for drug set were 48.85 and 50.76%, respectively. Various algorithms such as support vector machine, clustering algorithm, neural network, and regression produced similar results to that of decision trees. The possibility of noisy data sets, incorrect decision assignment, complex interaction of various human physiological processes, and the interaction between various diseases can explain some aspects of the poor baseline results. As the purpose of the proposed approaches is to enhance the knowledge base (i.e., rules/information that represent the group of individuals) over the existing ones (i.e., obtained from original data set or domain experts), the poor cross-validation and specificity results can still act as the baseline. Thus relative increase in cross-validation accuracy, specificity, and reduction in features are of prime importance. IG and REG approaches were selected to determine the significant gene/SNPs. All the approaches discussed in this paper were compared with IG and REG approaches with respect to quality measures discussed above.

Table 7 Example of genes and SNPs

Gene	SNP
ABC	ABC_SNP1
	ABC_SNP2
	ABC_SNP3
	ABC_SNP4
	ABC_SNP5
	ABC_SNP6
AUS	AUS_SNP1
	AUS_SNP2
	AUS_SNP3
	AUS_SNP4
	AUS_SNP5
	AUS_SNP6
IND	IND_SNP1
	IND_SNP2
	IND_SNP3
	IND_SNP4
	IND_SNP5
JGT	JGT_SNP1
	JGT_SNP2
	JGT_SNP3
	JGT_SNP4
	JGT_SNP5
	JGT_SNP6
BRH	BRH_SNP1
	BRH_SNP2
	BRH_SNP3
	BRH_SNP4
	BRH_SNP5
	BRH_SNP6
JIT	JIT_SNP1
	JIT_SNP2
	JIT_SNP3
	JIT_SNP4
	JIT_SNP5

4.5.2. Drug set

The WDTGS approach increased cross-validation accuracy over the baseline by 4.69% and reduced the number of features by 60.47%. The quality measures for GA—CFS and GA—DTW approaches are provided in Table 16. The best approach was

Table 8 Classification and prediction accuracy

Genes	CA	PA	Genes	CA	PA	Genes	CA	PA	Genes	CA	PA
ABC	60.71	49.24	EFG	59.32	47.72	NBC	64.07	53.04	QRS	57.23	46.20
AUS	65.40	49.43	GRE	60.71	53.42	NOP	59.32	50.00	QTS	57.04	48.29
BRH	59.32	49.24	HIJ	52.85	48.29	NOR	55.14	43.35	RHN	65.21	50.95
CBS	65.40	48.86	IND	52.85	46.96	NPR	59.13	55.52	RHP	70.73	50.00
CDE	61.79	47.77	JGT	60.84	45.25	ORH	58.18	49.24	STP	60.27	48.29
CRH	68.25	50.19	JIT	61.98	48.29	ORS	69.02	50.00	TUV	60.84	49.43
CRS	56.85	51.71	KLM	63.50	54.76	OST	56.66	44.11	USP	71.49	56.09
DSS	54.00	48.86	KQE	69.58	48.67	PAT	71.87	48.67	WXY	67.87	51.34

CA: classification accuracy; PA: prediction accuracy.

the intersection approach of WDTGS and GA–CFS (i.e., GA–CFS–WDTGS features set). This approach has increased the cross-validation accuracy over baseline by 8.58% with 84.88% reduction in the number of features. The specificity had increased by 3.29% over the baseline. Thus these approaches perform better than the baseline and also steadily improve over each other (see Fig. 5a and Table 16).

The IG approach had cross-validation accuracy of 50.19%, a 2.74% increase over the baseline. Though the IG approach resulted in a 63.37% reduction of the number of features, the specificity decreased by 2.28%. Similarly, the REG approach had cross-validation accuracy of 52.66% with the elimination of 133 features over the baseline. The GA–CFS–WDTGS approach performed much better than IG and REG approaches on all quality measures (Table 16). Although the cross-validation accuracy of the REG and GA–CFS–WDTGS approaches differed by less than 0.5%, the number of features reduced by GA–CFS–WDTGS approach was considerably higher. The GA–CFS–WDTGS approach identified 3 and 14 better gene/SNPs than IG and REG, respectively (Table 17). Thus the GA was able to uniquely identify some drug gene/SNPs that were not identified by the traditional approaches.

Table 9 Weighted schemes

Gene	WS1	WS2	Gene	WS1	WS2	Gene	WS1	WS2	Gene	WS1	WS2
ABC	57.27	52.68	EFG	55.84	51.20	NBC	60.76	56.35	QRS	53.92	49.51
AUS	60.61	54.22	GRE	58.52	55.61	NOP	56.52	52.80	QTS	54.42	50.91
BRH	56.30	52.26	HIJ	51.48	49.66	NOR	51.60	46.88	RHN	60.93	55.23
CBS	60.44	53.82	IND	51.08	48.73	NPR	58.05	56.60	RHP	64.51	56.22
CDE	57.58	51.98	JGT	56.16	49.93	ORH	55.50	51.92	STP	56.68	51.88
CRH	62.83	55.61	JIT	57.87	52.40	ORS	63.31	55.71	TUV	57.42	52.85
CRS	55.31	53.25	KLM	60.88	57.38	OST	52.90	47.88	USP	66.87	60.71
DSS	52.46	50.40	KQE	63.31	54.94	PAT	64.91	55.63	WXY	62.91	56.30

Weighted scheme $= A \times$ classification $+ B \times$ prediction. WS1: weighted scheme 1 ($A = 0.7$ and $B = 0.3$); WS2: weighted scheme 2 ($A = 0.3$ and $B = 0.7$).

Table 10 Ranked genes for drug subjects

Weighted scheme 1						Weighted scheme 2					
Rank	Gene	Correct	Rank	Gene	Correct	Rank	Gene	Correct	Rank	Gene	Correct
1	**USP**	66.87	7	**CRH**	62.83	1	**USP**	60.71	7	**ORS**	55.71
2	**PAT**	64.91	8	**RHN**	60.93	2	**KLM**	57.38	8	**PAT**	55.63
3	**RHP**	64.51	9	**KLM**	60.88	3	NPR	56.60	9	**CRH**	55.61
4	**ORS**	63.31	10	**NBC**	60.76	4	**NBC**	56.35	10	GRE	55.61
5	**KQE**	63.31	11	AUS	60.61	5	**WXY**	56.30	11	**RHN**	55.23
6	**WXY**	62.91	12	CBS	60.44	6	**RHP**	56.22	12	**KQE**	54.94

Genes marked in bold are repeated for both weighted schemes.

Table 11 Ranked genes for placebo subjects

Weighted scheme 1						Weighted scheme 2					
Rank	Gene	Correct	Rank	Gene	Correct	Rank	Gene	Correct	Rank	Gene	Correct
1	**RHP**	65.36	7	**CBS**	63.31	1	**AUS**	59.14	7	**ORS**	56.61
2	**KQE**	63.72	8	**WXY**	63.31	2	**CRH**	58.04	8	**KQE**	56.12
3	**CRH**	63.69	9	**NBC**	62.49	3	**CBS**	58.00	9	GRE	56.05
4	**ORS**	63.44	10	PAT	61.95	4	NOR	57.26	10	STP	55.89
5	**AUS**	63.44	11	RHN	61.12	5	**WXY**	57.15	11	**RHP**	55.74
6	USP	63.42	12	JGT	59.96	6	**NBC**	56.93	12	CRS	55.45

Genes marked in bold are repeated for both weighted schemes.

Table 12 Significant genes

No.	Drug gene	Placebo gene	Common gene
1	**CRH**	AUS	**CRH**
2	KLM	CBS	**KQE**
3	**KQE**	**CRH**	**NBC**
4	**NBC**	**KQE**	**ORS**
5	**ORS**	**NBC**	**RHP**
6	PAT	**ORS**	**WXY**
7	RHN	**RHP**	
8	**RHP**	**WXY**	
9	USP		
10	**WXY**		

Genes marked in bold are common to drug and placebo subjects.

Some prominent sample rules for drug set are as follows:

RULE 1: **IF** STP_SNP6 = C_C **AND** NBC_SNP1 = C_T **AND** KLM_SNP1 = G_G **AND** USP_SNP7 = C_C **AND** CRH_SNP4 = C_C **AND** KQE_SNP7 = C_C **THEN** Decision = D_GOOD

RULE 2: **IF** STP_SNP6 = C_T **AND** WXY_SNP2 = C_C **AND** CRH_SNP3 = C_T **AND** RHP_SNP2 = C_C **THEN** Decision = D_BAD

Cross-testing of the knowledge obtained from the drug set was performed by testing the rules against the placebo set. A lower cross-validation accuracy (of placebo set cross testing on drug knowledge, i.e., rule sets) may indicate significant drug-related genes (Fig. 5a). The reason for improved placebo cross-testing accuracy (Fig. 5a) in this analysis can

Table 13 The GA—CFS—based significant SNPs

Significant drug SNPs				Significant placebo SNPs			
1	ABC_SNP2	HIJ_SNP4	ORS_SNP6	1	ABC_SNP2	KLM_SNP1	OST_SNP3
2	ABC_SNP6	JGT_SNP2	OST_SNP4	2	ABC_SNP4	KLM_SNP5	PAT_SNP4
3	AUS_SNP1	JGT_SNP6	PAT_SNP1	3	AUS_SNP2	KLM_SNP6	PAT_SNP6
4	AUS_SNP2	JIT_SNP3	PAT_SNP2	4	AUS_SNP3	KQE_SNP2	QRS_SNP3
5	AUS_SNP5	JIT_SNP4	PAT_SNP4	5	AUS_SNP4	KQE_SNP4	QRS_SNP4
6	BRH_SNP2	KLM_SNP1	QRS_SNP2	6	AUS_SNP5	KQE_SNP5	QTS_SNP3
7	BRH_SNP6	KLM_SNP2	QRS_SNP4	7	BRH_SNP6	NBC_SNP1	RHN_SNP1
8	CBS_SNP1	KLM_SNP6	QRS_SNP5	8	CBS_SNP4	NBC_SNP2	RHN_SNP2
9	CBS_SNP2	KQE_SNP3	QTS_SNP1	9	CBS_SNP5	NBC_SNP4	RHN_SNP4
10	CBS_SNP9	KQE_SNP7	RHN_SNP1	10	CBS_SNP6	NOP_SNP1	RHP_SNP2
11	CRH_SNP2	NBC_SNP1	RHP_SNP2	11	CDE_SNP2	NOP_SNP2	RHP_SNP3
12	CRH_SNP3	NBC_SNP4	STP_SNP6	12	CDE_SNP5	NOP_SNP3	STP_SNP4
13	CRH_SNP4	NOP_SNP1	TUV_SNP2	13	CRS_SNP3	NOP_SNP4	STP_SNP6
14	CRH_SNP5	NOP_SNP2	TUV_SNP3	14	CRS_SNP4	NOR_SNP2	TUV_SNP1
15	CRH_SNP6	NPR_SNP1	TUV_SNP4	15	DSS_SNP3	NOR_SNP3	TUV_SNP5
16	CRS_SNP3	NPR_SNP3	USP_SNP4	16	EFG_SNP1	ORH_SNP4	USP_SNP4
17	DSS_SNP4	ORH_SNP2	USP_SNP5	17	EFG_SNP4	ORS_SNP4	WXY_SNP2
18	EFG_SNP4	ORH_SNP5	USP_SNP7	18	GRE_SNP3	ORS_SNP5	
19	GRE_SNP4	ORH_SNP6	WXY_SNP2	19	IND_SNP2	ORS_SNP7	
20	HIJ_SNP1	ORS_SNP3	WXY_SNP4	20	IND_SNP5	ORS_SNP8	
21	HIJ_SNP2	ORS_SNP4	WXY_SNP5	21	JIT_SNP5	ORS_SNP9	

Table 14 The GA—DTW-based significant SNPs

Significant drug SNPs				Significant placebo SNPs				
1	ABC_SNP2	IND_SNP2	ORH_SNP7	1	ABC_SNP4	GRE_SNP1	NOP_SNP4	RHP_SNP2
2	ABC_SNP3	IND_SNP3	ORS_SNP1	2	AUS_SNP1	GRE_SNP2	NOR_SNP3	RHP_SNP3
3	ABC_SNP5	JGT_SNP2	ORS_SNP3	3	AUS_SNP2	GRE_SNP3	NPR_SNP3	RHP_SNP6
4	ABC_SNP6	JGT_SNP4	ORS_SNP4	4	AUS_SNP3	GRE_SNP4	ORH_SNP1	RHP_SNP7
5	AUS_SNP2	JGT_SNP5	ORS_SNP7	5	BRH_SNP4	HIJ_SNP1	ORH_SNP2	STP_SNP4
6	AUS_SNP3	JGT_SNP6	OST_SNP2	6	BRH_SNP5	HIJ_SNP3	ORH_SNP3	STP_SNP5
7	AUS_SNP4	JIT_SNP1	OST_SNP4	7	BRH_SNP6	IND_SNP2	ORH_SNP4	STP_SNP6
8	AUS_SNP5	JIT_SNP4	PAT_SNP2	8	CBS_SNP1	IND_SNP3	ORH_SNP5	TUV_SNP2
9	AUS_SNP6	JIT_SNP5	QRS_SNP1	9	CBS_SNP4	IND_SNP5	ORS_SNP1	TUV_SNP3
10	BRH_SNP1	KLM_SNP1	RHN_SNP1	10	CBS_SNP5	JGT_SNP3	ORS_SNP2	TUV_SNP4
11	BRH_SNP2	KLM_SNP3	RHN_SNP4	11	CDE_SNP1	JGT_SNP5	ORS_SNP6	TUV_SNP5
12	CBS_SNP7	KQE_SNP3	RHN_SNP5	12	CDE_SNP3	JGT_SNP6	ORS_SNP7	USP_SNP1
13	CDE_SNP2	KQE_SNP5	RHP_SNP1	13	CDE_SNP4	JIT_SNP1	ORS_SNP8	USP_SNP2
14	CDE_SNP4	KQE_SNP7	RHP_SNP6	14	CRH_SNP2	JIT_SNP4	ORS_SNP9	USP_SNP3
15	CDE_SNP5	NBC_SNP1	STP_SNP2	15	CRH_SNP3	KLM_SNP1	OST_SNP1	USP_SNP4
16	CRH_SNP1	NBC_SNP3	STP_SNP6	16	CRH_SNP6	KLM_SNP5	OST_SNP3	USP_SNP7
17	CRS_SNP1	NBC_SNP4	TUV_SNP2	17	CRS_SNP2	KLM_SNP6	PAT_SNP1	WXY_SNP4
18	CRS_SNP3	NBC_SNP5	TUV_SNP4	18	CRS_SNP3	KLM_SNP7	PAT_SNP3	WXY_SNP5
19	CRS_SNP4	NOP_SNP3	USP_SNP5	19	CRS_SNP5	KQE_SNP3	PAT_SNP4	
20	DSS_SNP2	NOP_SNP4	USP_SNP6	20	DSS_SNP1	KQE_SNP4	QRS_SNP3	
21	DSS_SNP4	NPR_SNP3	USP_SNP7	21	DSS_SNP2	KQE_SNP5	QTS_SNP3	
22	EFG_SNP2	ORH_SNP1	WXY_SNP1	22	DSS_SNP4	KQE_SNP6	RHN_SNP1	
23	GRE_SNP1	ORH_SNP4	WXY_SNP3	23	EFG_SNP1	NBC_SNP2	RHN_SNP5	
24	HIJ_SNP4	ORH_SNP6	WXY_SNP4	24	EFG_SNP4	NBC_SNP3	RHP_SNP1	

Table 15 The GA—CFS—WDTGS-based significant SNPs

Significant drug SNPs				Significant placebo SNPs		
1	CRH_SNP2	NBC_SNP4	USP_SNP7	1	AUS_SNP2	NBC_SNP2
2	CRH_SNP3	ORS_SNP3	WXY_SNP2	2	AUS_SNP3	NBC_SNP4
3	CRH_SNP4	ORS_SNP4	WXY_SNP4	3	AUS_SNP4	ORS_SNP4
4	CRH_SNP5	ORS_SNP6	WXY_SNP5	4	AUS_SNP5	ORS_SNP5
5	CRH_SNP6	PAT_SNP1		5	CBS_SNP4	ORS_SNP7
6	KLM_SNP1	PAT_SNP2		6	CBS_SNP5	ORS_SNP8
7	KLM_SNP2	PAT_SNP4		7	CBS_SNP6	ORS_SNP9
8	KLM_SNP6	RHN_SNP1		8	KQE_SNP2	RHP_SNP2
9	KQE_SNP3	RHP_SNP2		9	KQE_SNP4	RHP_SNP3
10	KQE_SNP7	USP_SNP4		10	KQE_SNP5	WXY_SNP2
11	NBC_SNP1	USP_SNP5		11	NBC_SNP1	

Table 16 Improvement in quality measures for the drug data set

Feature list	Accuracy (%)	Percent increase in accuracy	Number of features	Percent reduction in features	Specificity (%)	Percent increase in specificity
ALL	48.85	0.00	172	0.00	50.76	0.00
WDTGS	51.14	4.69	68	60.47	49.43	−2.63
GA—DTW	51.33	5.08	72	58.14	54.17	6.70
GA—CFS	53.23	8.97	63	63.37	52.57	3.57
GA—CFS—WDTGS	53.04	8.58	26	84.88	52.43	3.29
IG	50.19	2.74	63	63.37	49.61	−2.28
REG	52.66	7.80	39	77.33	52.06	2.55

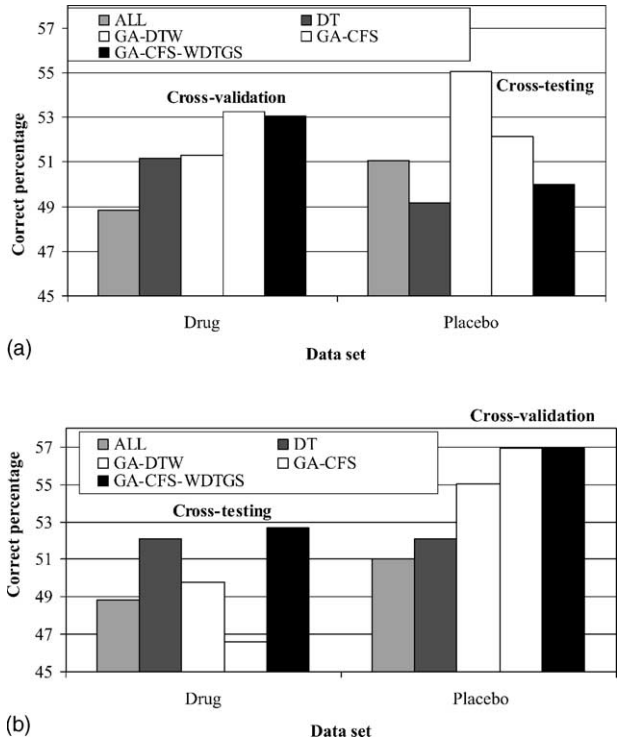

Figure 5 Cross-validation and cross-testing results using (a) drug knowledge and (b) placebo knowledge.

be explained by common significant genes/SNPs for both placebo as well as drug subjects.

4.5.3. Placebo set

Similarly, the best approach for the placebo set was the intersection approach of WDTGS and GA—CFS (i.e., GA—CFS—WDTGS features set). The increase in cross-validation accuracy over baseline for this approach was 11.58% and the number of features was reduced by 87.79% (Table 18). The specificity had increased by 3.22% over the baseline. Table 18 and Fig. 5b explain the quality of each approach for the placebo set.

The IG approach had cross-validation accuracy of 53.38%, which was between the WDTGS approach and GA—DTW approach. The REG approach performed worst than the baseline cross-validation accuracy due to the elimination of quality features. The specificity decreased by 7.38% over the baseline measurement. The GA—CFS—WDTGS approach performed far superior than both IG and REG approaches in terms of all quality measures (Table 18). The GA—CFS—WDTGS approach identified 4 and 8 more interesting gene/SNPs than IG and REG, respectively (Table 17). Thus the GA was able to uniquely identify some placebo gene/SNPs

Table 17 Uniquely identified drug and placebo genes/SNPs by the GA—CFS—WDTGS approach

No.	Drug data set: SNPs uniquely identified by GA—CFS—WDTGS over				Placebo data set: SNPs uniquely identified by GA—CFS—WDTGS over		
	IG	REG			IG	REG	
1	ORS_SNP4	CRH_SNP3	ORS_SNP3	USP_SNP4	AUS_SNP2	AUS_SNP2	ORS_SNP4
2	USP_SNP4	CRH_SNP6	ORS_SNP4	USP_SNP7	ORS_SNP5	KQE_SNP2	ORS_SNP7
3	USP_SNP7	KLM_SNP1	PAT_SNP1	WXY_SNP2	ORS_SNP7	KQE_SNP5	ORS_SNP8
4		KQE_SNP7	PAT_SNP4	WXY_SNP5	RHP_SNP2	NBC_SNP4	RHP_SNP3
5		NBC_SNP1	RHN_SNP1				

Table 18 Improvement in quality measures for the placebo data set

Feature list	Accuracy (%)	Percent increase in accuracy	Number of features	Percent reduction in features	Specificity (%)	Percent increase in specificity
ALL	51.05	0.00	172	0.00	55.69	0.00
WDTGS	52.11	2.08	57	66.86	50.41	−9.48
GA—DTW	55.06	7.86	90	47.67	57.08	2.51
GA—CFS	56.96	11.58	59	65.70	57.92	4.01
GA—CFS—WDTGS	56.96	11.58	21	87.79	57.48	3.22
IG	53.38	4.56	59	65.70	54.84	−1.52
REG	50.63	−0.82	41	76.16	51.57	−7.38

that were not identified by the traditional approaches.

Two important sample rules for the placebo set are as follows:

RULE 1: **IF** AUS_SNP4 = C_C **AND** ORS_SNP5 = C_T **THEN** Decision = P_GOOD
RULE 2: **IF** KQE_SNP2 = C_C **AND** CBS_SNP6 = C_C **AND** KQE_SNP5 = G_G **AND** ORS_SNP5 = T_T **THEN** Decision = P_BAD

5. Conclusion

Three different approaches for selection of significant genes/SNPs were presented. The identified significant genes may lead to improvement of drug effectiveness. For the data sets considered in this paper, the number of features was reduced by 85% and the cross-validation accuracy was increased by 10% over the baseline measurements. The specificity increased by 3.2%. The proposed approach has substantially enriched the knowledge base. Bagging, boosting, meta-decision-making, and other approaches can be used to further increase the cross-validation accuracy and specificity. The GA—CFS—WDTGS approach performed far better than the IG and REG approach in terms of all three-quality measures, i.e., cross-validation accuracy, specificity, and the number of significant genes/SNPs. The GA—CFS—WDTGS approach uniquely identified some gene/SNPs that could not be identified by the IG and REG approaches.

Incorporating traditional feature selection approaches could further enhance the significant feature set. A modification of the inclusion procedure of features in the significant feature set is needed, e.g., weights (% decline/increase of the accuracy and specificity and the reduction in the number of features) could be used.

Various drug and diseases related analyses would benefit from the proposed approaches. They will ultimately lead to customized treatment protocols and medications.

References

[1] NCBI-single nucleotide polymorphism, DbSNP overview—a database of single nucleotide polymorphisms, NCBI. Available at http://www.ncbi.nlm.nih.gov/SNP/get_html.cgi?whichHtml=overview. Accessed on 30 July 2003.

[2] Herrera S. With the race to chart the human genome over, now the real work begins. Red Herring magazine. 1 April 2001. Available at http://www.redherring.com/mag/issue95/1380018938.html. Accessed on 30 July 2003.

[3] SNP Consortium, single nucleotide polymorphisms for biomedical research. The SNP Consortium Ltd. Available at http://www.snp.cshl.org/. Accessed on 30 July 2003.

[4] Golub TR, Slonim DK, Tamayo P, Huard C, GaasenBeek M, Mesirov JP et al. Molecular classification of cancer: class discovery and class prediction by gene-expression monitoring. Science 1999;286:531—7.

[5] Raychaudhuri S, Sutphin PD, Chang JT, Altman RB. Basic microarray analysis: grouping and feature reduction. Trends Biotechnol 2001;19(5):189—93.

[6] Johnson JA, Evans WE. Molecular diagnostics as a predictive tool: genetics of drug efficacy and toxicity. Trends Mol Med 2002;8(6):300—5.

[7] NHGRI, Executive summary of the SNP meeting, National Human Genome Research Institute. Available at http://www.genome.gov/10001884. Accessed on 30 July 2003.

[8] D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 2000;16:707—26.

[9] Kirschner M, Pujol G, Radu A. Oligonucleotide microarray data mining: search for age-dependent gene expression. Biochem Biophys Res Commun 2002;298(5):772—8.

[10] Ponomarenko J, Merkulova T, Orlova G, Fokin O, Gorshkov E, Ponomarenko M. Mining DNA sequences to predict sites which mutations cause genetic diseases. Knowl-based Syst 2002;15(4):225—33.

[11] Oliveira G, Johnston DA. Mining the schistosome DNA sequence database. Trends Parasitol 2001;17(10):501—3.

[12] Fuhrman S, Cunningham MJ, Wen X, Zweiger G, Seilhamer J, Somogyi R. The application of Shannon entropy in the identification of putative drug targets. Biosystems 2000;55:5—14.

[13] Arkin A, Shen P, Ross J. A test case of correlation metric construction of a reaction pathway from measurements. Science 1997;277:1275—9.

[14] Cho SB, Won HH. Machine learning in DNA Microarray analysis for cancer classification. In: Yi-Ping Phoebe Chen, editors. Proceedings of the First Asia-Pacific Bioinformatics Conference. Australian Computer Society; 2003. p. 189—98, ISBN: 0909925976.

[15] Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining. Cambridge, MA: AAAI/MIT Press; 1995.

[16] Kusiak A, Kern JA, Kernstine KH, Tseng TL. Autonomous decision-making: a data mining approach. IEEE Trans Inf Technol Biomed 2000;4(4):274—84.

[17] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576. Department of Statistics, University of California, Berkeley, CA; 2000.

[18] Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001;17(12):1131—42.

[19] Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7(6):673—9.

[20] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler andD. validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16(10):906—14.

[21] Eisen MB, Spellman, PT, Brown PO, Bostein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 1998;95(25):14863—8.

[22] Hartuv E, Schmitt A, Lange J, Meier-Ewert S, Lehrach H, Shamir R. An algorithm for clustering cDNA fingerprints. Genomics 2000;66(3):249—56.
[23] Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw 2000;13: 411—30.
[24] Sun HX, Zhang KX, Du WN, Shi JX, Jiang ZW, Sun H et al. Single nucleotide polymorphisms in CAPN10 gene of Chinese people and its correlation with type 2 diabetes mellitus in Han people of northern China. Biomed Environ Sci 2002;15(1):75—82.
[25] Useche F, Gao G, Hanafey M, Rafalski A. High-throughput identification, database storage and analysis of SNPs in EST sequences. Genome Inform 2001;12:194—203.
[26] Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 2000;9(16):2403—8.
[27] Goldberg DE. Genetic algorithms in search, optimization, and machine learning. New York: Addison Wesley Longman Inc.; 1989.
[28] Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. Cambridge, MA: MIT Press; 1975.
[29] Michalewicz Z. Genetic algorithms + data structures = evolution programs. Berlin: Springer-Verlag; 1992.
[30] Lawrence D. Handbook of genetic algorithms. New York: Van Nostrand Reinhold; 1991.
[31] Quinlan R. C 4.5 programs for machine learning. San Meteo CA: Morgan Kaufmann; 1992.
[32] Witten I, Frank E. Data mining: practical machine learning tools and techniques with java implementations. San Francisco, CA: Morgan Kaufmann; 2000.
[33] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97(1—2):273—324.
[34] John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In Cohen WW, Hirsh H, editors. In: Proceedings of the 11th International Conference on Machine Learning ICML94. San Francisco, CA: Morgan Kaufmann; 1994. p. 121—9.
[35] Hall MA, Smith LA. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In: Kumar A, Russell I, editors. Proceedings of the Florida Artificial Intelligence Research Symposium, Orlando, Florida. Menlo Park, CA: AAAI Press; 1999. p. 235—239. ISBN: 1577350804.
[36] Vafaie H, DeJong K. Genetic algorithms as a tool for restructuring feature space representations, In: Proceedings of the Seventh International Conference on Tools with Artificial Intelligence. Los Alamitos, CA: IEEE Computer Society Press; 1996. p. 8—11. ISBN: 0818673125.
[37] Zhang L, Zhao Y, Yang Z, Wang J. Feature selection in recognition of handwritten Chinese characters. In: Proceedings of the 2002 International Conference on Machine Learning and Cybernetics. Piscataway, NJ: IEEE; 2002. p. 1158—62. ISBN: 0780375084.