Exploring Genetic Variability in Drug Therapy by Selecting a Minimum Subset of the Most Informative Single Nucleotide Polymorphisms through Approximation of a Markov Blanket in a Kernel-induced Space

Qiang Lou[1], Henry P. Parkman[2], Michael R. Jacobs[3], Evgeny Krynetskiy[3,4], Zoran Obradovic [1*]
[1]Center for Data Analytics and Biomedical Informatics, Temple University
[2]Gastroenterology Section, School of Medicine, Temple University
[3]School of Pharmacy, Temple University
[4]Jayne Haines Center for Pharmacogenomics and Drug Safety, Temple University
Philadelphia, USA
{qianglou, henryp, mrjacobs, ekrynets, zoran.obradovic}@temple.edu

***Abstract*—Genome-wide analysis of single nucleotide polymorphisms (SNP) can potentially be helpful in exploring the role of genetic variability in drug therapy. However, two major problems with such an analysis are the need for a large number of interrogated genomes, and the resulting high-dimensional data where the number of SNPs used as features is much larger than the number of subjects. The aim of this study is to identify informative SNPs associated with clinical efficacy and side effects of domperidone treatment for gastroparesis from DNA microarray experiments by applying our feature selection method, which approximates the Markov Blanket in a kernel-induced space. DNA samples extracted from the saliva of 46 patients treated with domperidone were analyzed using Affymetrix 6.0 SNP microarrays. Experimental evaluations on this SNP microarray dataset provide evidence that our feature selection method can remove useless SNP features more accurately than existing Markov Blanket based alternatives.**

Keywords-genetic variability; feature selection; drug therapy; SNP

I. INTRODUCTION

Single nucleotide polymorphism (SNP) analysis with DNA microarrays holds great promises to elucidate the role of genetic variability on the clinical efficacy of drug therapy [1]. This technology collects millions of data points per genome, providing an unprecedented volume of genetic information. Therefore, arrays have dramatically changed the experimental strategy in molecular medicine. Unfortunately, assessment of data quality, statistical evaluation, and interpretation tools required for DNA microarray experiments far exceed the capacity of approaches accepted in traditional molecular studies [2]. A genome-wide search for SNP associations with a phenotype in question using DNA microarray experiments requires hundreds and thousands of interrogated genomes. Even then, the definitive identification of genes associated with the phenotype can be cumbersome.

We proposed a new feature selection method that is actually an improvement of the method we have recently developed [13] for approximation of the Markov Blanket. Instead of relying on the conditional independence test or network structure learning, our method uses the Hilbert-Schmidt Independence criterion as a measure of dependence among variables in a kernel-induced space. This allows effective approximation of the Markov Blanket that consists of multiple dependent features rather than being limited to a single feature. Our method was previously proven to perform better than three alternatives on benchmark data sets. However, we didn't test our previous method [13] on microarray data where there are few instances and a large number of features. Moreover, our previous method [13] checks only one Markov Blanket candidate for each feature, which is not accurate enough to remove the feature. The objective of this study is to demonstrate that our method is applicable to genome-wide SNP data analysis. Our feature selection method aims to find a minimum subset of the most informative SNPs associated with clinical efficacy and side effects. This is done by efficiently approximating the Markov Blanket, which is a set of variables that can shield a certain SNP from the target.

II. RELATED WORK

This paper focuses on the filtering-based feature selection where information theory is used to identify a minimum subset of the most informative features by searching the so-called Markov Blanket. The Markov Blanket of a variable is regarded as a set of variables that can shield it from other variables. The premise of feature selection using the Markov Blanket is the elimination of a feature for which we can find the Markov Blanket in the remaining features. Such a feature selection process has been shown to result in a theoretically optimal set of features [3]. Feature selection methods using the Markov Blanket can be categorized into test-based, network structure learning-based and approximate methods.

Existing test-based Markov Blanket feature selection methods [5] all use a 'Growing-Shrinking' (GS) [6] approach for discovering the Markov Blanket. In the growing phase of this approach, all features belonging to the Markov Blanket and possibly some false features are also included. Then, in the shrinking phase, all features in the current Markov Blanket are checked again to remove the false features introduced at the growing phase. In both phases, conditional independence

* Corresponding Author

978-1-4673-1191-5/12/$31.00 ©2012 IEEE

testing is used to judge if a feature belongs to the Markov Blanket or not. However, such conditional independence test-based methods require that the sample have a large number of instances to ensure the reliability of the independence test. Another limitation of test-based feature selection algorithms is that they are usually too aggressive in removing features [7].

In a structure learning-based method, heuristic Bayesian network structure learning is performed and then the Markov Blanket is discovered corresponding to the learned structure [7]. In such an approach, to restrict search space, two heuristics (called 'sparse candidate' and 'screen-based') are proposed for selecting the promising candidates. However, a Bayesian network structure is learned using heuristic methods, as the optimization here is different. These heuristics combine locally optimal structures, which results in a learned structure that is not a global optimal solution. An additional limitation of such an approach is that the learning network structure could be computationally prohibitively expensive in the presence of a large number of features.

In an approximate Markov Blanket method called FCBF, the redundant features are eliminated in a potentially relevant subset obtained by excluding the irrelevant features based on the correlation to the target variable [8]. In this approach, symmetrical uncertainty is used to measure the relationship between variables. For a pair of features, FCBF measures symmetrical uncertainty and also the symmetrical uncertainty between either of them and the target variable. If the measured value between these two variables is greater than the measured value between either one of them and the target variable, the variable with the larger symmetrical uncertainty to the target is regarded as the Markov Blanket of the other variable, which then is removed. FCBF assumes the Markov Blanket of a feature has only one feature, since it is based on pairwise comparison. Such an approach is often too restrictive in practical situations, as illustrated in the results section of this article. In addition, we found that FCBF is too aggressive in eliminating features, since it gives too high priority to dominant features.

The feature selection method we proposed and evaluated in a previous study [13] approximates the Markov Blanket without relying on the conditional independence test or network structure learning. This is achieved by efficiently measuring dependence among variables according to the Hilbert-Schmidt Independence Criterion. This is used to effectively find an approximation of the Markov Blanket that consists of multiple dependent features rather than being limited to a single feature as in FCBF [8]. However, our previous method [13] only checks one Markov Blanket candidate for each feature with limited accuracy. In this study, the proposed methods check multiple Markov Blanket candidates for each feature. Also, in the previous work, we didn't determined if our method is applicable to microarray data where there are few instances and large numbers of features. Another objective of this study is to compare our method against four alternatives when applied to genome-wide SNP data analysis. Two high impact classification problems related to the development of side effects of drug therapy, and to the clinical efficacy of drug therapy were assessed. We will show that our method outperforms the alternatives on this genome-wide SNP data analysis when selecting few informative SNPs from a large number of SNPs in the presence of few data instances.

III. Approximating Markov Blanket

A. Dependence Measurement

Feature selection requires use of an appropriate dependence measure to evaluate the relationship between features and the target variable, or between different features. We measure the dependence among variables in an appropriate kernel space. In this section, we will first describe a dependence measure called the Hilbert-Schmidt Independence Criterion (HSIC) and will discuss the reason why HSIC is used as the basis measure for the Markov Blanket discovery.

A Hilbert space F of functions in which point wise evaluation is a continuous linear function is called a Reproducing Kernel Hilbert Space (RKHS) [9]. In other words, in RKHS for an arbitrary feature set X, there is a mapping $\varphi : X \to F$ to a Hilbert space F such that $< \varphi(x), \varphi(x') >= k(x, x')$, where k is a unique positive definite kernel. Let X and Y be sets drawn from some joint probability distribution $\Pr_{xy}$. Let F be the RKHS on X with, $k : X \times X \to \Re$ and $\varphi : X \to F$ be the corresponding kernel and feature map. Similarly, let G be the RKHS on Y with kernel ℓ and feature map ψ. Then, the cross-covariance operator [10] $C_{xy} : G \to F$ is defined as:

$$C_{xy} = E_{xy}[(\varphi(x) - \mu_x) \otimes (\psi(y) - \mu_y)]$$
$$= E_{x,y}[\varphi(x) \otimes \psi(y)] - \mu_x \otimes \mu_y$$

where $\otimes$ is the tensor product. Then, HSIC is defined as the square of the Hilbert-Schmidt norm of the cross-covariance operator. A kernel version of HSIC [10] is computed as:

$$HSIC(F, G, \Pr_{xy}) = \| C_{xy} \|^2_{HS}$$
$$= Exx'yy'[k(x,x')\ell(y,y')] + Exx'[k(x,x')]Eyy'[\ell(y,y')]$$
$$-2Exy[Ex'[k(x,x')]Ey'[l(y,y')]]$$

Here, we regard $E_{x,x',y,y'}$ as the expectation of two pairs (x, y) and (x', y') which are independent to each other and both drawn from $\Pr_{xy}$.

Given a sample Z drawn from the distribution $\Pr_{xy}$, an empirical estimate of HSIC [9] is:

$$HSIC(F_i, G, Z) = (m-1)^{-2} trHKHL_c$$

where K and L_c are the kernel matrices of the feature F_i and the target variable C respectively, m is the number of instances and $H_{ij} = \sigma_{ij} - m^{-1}$ is used to center the features and targets in the feature space.

HSIC can detect any kind of dependence between two variable sets X and Y by using a universal kernel (such as a Gaussian Kernel). It has been proved in [9] that $\| C_{xy} \|^2_{HS} = 0$ if and only if x and y are independent to each other. This is our

direct motivation for choosing HSIC to measure the dependence. Actually, in this paper, we use $HSIC(F,G,Z)$ to replace the $HSIC(F,G,\Pr_{xy})$ to measure the independence between two variable sets. This is because $HSIC(F,G,Z)$ is easy to compute and is concentrated, as previously proven [9] by showing that with probability at least $1-\sigma$

$$|HSIC(F,G,Pr_{xy}) - HSIC(F,G,Z)| \le \sqrt{\frac{\log(6/\sigma)}{\alpha^2 m}} + \frac{C}{M}$$

where $\alpha^2 > 0.24$, $\sigma > 0$ and C is the constant.

There are three reasons HSIC is used as the measure of dependence among variables in the proposed algorithm. First, HSIC measures dependence in high dimensional kernel space and can detect any kind of dependence between two variable sets with a universal kernel, such as the RBF kernel, which is not possible with previously used measures. Second, HSIC can measure the dependence between both discrete and continuous variables. In contrast, most measures previously used to find the Markov Blanket are Entropy-based, and so they are not directly applicable to datasets with continuous variables. Third, HSIC is easy to compute from the kernel matrices without density estimation.

B. Identification of the Markov Blanket Candidates

The Markov Blanket MB_i of feature F_i ($MB_i \subset F, (F_i \notin MB_i)$) has the property that F_i is conditionally independent of the remaining features U and the target C.

Definition 1 (Markov Blanket). Let F be the whole set of features and C be the target variable, given a feature $F_i \in F$, let $MB_i \subset F (F_i \notin MB_i)$, MB_i is the Markov Blanket of F_i iff:

$$P(U, C \mid F_i, MB_i) = P(U, C \mid MB_i)$$
$$where: U = F - \{F_i\} - MB_i$$

If MB_i is the Markov Blanket of F_i, then the prediction model learned without considering F_i would be as accurate as the model learned on all features F. It is often difficult to find the exact Markov Blanket for a given feature. To address this problem for a given feature we proposed a novel method [13] of finding an approximate Markov Blanket. We then use this method to develop a feature selection algorithm based on the discovered approximated Blanket.

Given a set of features, we can determine if it is the Markov Blanket MB_i of feature F_i. However, evaluating all subsets of F for this property is prohibitively costly. To reduce the search cost we will evaluate some candidate subsets, as proposed in the following subsection.

Often, there might not be an exact Markov Blanket for a feature, but we can still try to identify an approximating Markov Blanket, which approximately subsumes the information about this feature. Thus we can remove this feature with little useful information lost. We now present a simple method to find an approximating Markov Blanket.

Intuitively, if a feature F_i has its Markov Blanket, described as MB_i, we assume that the features in MB_i have more dependence with F_i than those features that are not in MB_i. We choose a subset of k features that are strongly dependent to F_i as the candidate Markov Blanket of F_i. Then, for each feature, we only need to evaluate its candidate Markov Blanket rather than all possible subsets in the remaining features to see if such a candidate Markov Blanket is accurate enough to be regarded as the Markov Blanket.

The problem is to find the candidate Markov Blanket for each feature, that is, how to find the set of k features that are most dependent on the feature F_i. The naive method is to calculate exactly the candidate Markov Blanket for each feature. This means having to compute the dependence $HSIC(F_i, F_j) = (m-1)^{-2} trHK_iHK_j$ for each pair of features F_i and F_j (K_i and K_j are the kernel matrices of feature F_i and F_j respectively), which obviously is too computationally expensive and cannot be applied in high dimensional datasets.

In our method, in order to avoid such expensive computation, we don't compute the Candidate Markov Blanket exactly but rather approximate the Candidate Markov Blanket for each feature.

Definition 2 (Markov Blanket Candidate). Let B_i be the set of features that have higher dependence on the target variable C than does the feature F_i. A set of features MB_i is said to be the Markov Blanket candidate of the feature F_i, if each feature F_{MB_i} in MB_i satisfies:

$$F_{MB_i} = \arg\max_{F_c} HSIC(K_{Fc}, K_i),$$
$$\text{where}: F_c \in B_i - MB_i \cup \{F_{MB_i}\}$$

where K_{Fc} and K_i are Kernel matrices of feature F_c and F_i respectively.

We tend to find the Markov Blanket for a feature F_i in the features which have higher dependence with the target variable than F_i has, since we assume these high dependent features are more likely to subsume the information which low dependent features have.

In order to find the Markov Blanket Candidate for each feature, we define $\boldsymbol{S}$ as a list of all features ordered in descending value of dependence $HSIC(F_i, C) = (m-1)^{-2} trHK_iHL_c$ which is actually the dependence between each feature F_i and the target variable C (K_i and L_c are the kernel matrices of feature F_i and the target variable C). We can approximate the Markov Blanket Candidate of F_i as the set of k features that are most dependent on F_i and are listed before F_i in the list $\boldsymbol{S}$. For the features in the top of $\boldsymbol{S}$, which might not have enough k features before them, we can either choose the nearest features after them or assume that they don't have a Markov Blanket and keep these features in our optimal subset of features.

One possible problem of our previous method [13] in finding the Markov Blanket Candidate is the value of *k*, the number of features in MB_i. Basically, k should not be too large or too small. The larger the *k* is, the more likely that we

identify all features that should be in the Markov Blanket. But a larger *k* produces a greater possibility of identifying features not necessary in the Markov Blanket, which increases the chance of selecting unnecessary features in the Markov Blanket. Conversely, a small value of *k* reduces the possibility of choosing unnecessary features but increases the chance of getting an un-completed Markov Blanket. In this study, for a certain feature F_i, we regard all features listed before F_i in the list ***S*** as possible candidates of Markov Blanket. Therefore we proposed a new feature selection method described in section Ⅳ.

C. Screening Markov Blanket Candidates

After defining the candidate Markov Blanket, we show in this section how we approximate the Markov Blanket for each feature F_i

Definition 3 (Dependence-based Screening Test). Let MB_i be the approximate candidate Markov Blanket of feature F_i. We define the following dependence-based screen test to check whether MB_i can be regarded as an actual approximation of the Markov Blanket:

$$1.HSIC(MB_i, C) > HSIC(MB_i \cup F_i, C)$$
$$2.HSIC(MB_i, C) > HSIC(F_i, C), \text{and}$$
$$HSIC(MB_i, F_i) > HSIC(F_i, C)$$

where C is the target variable and HSIC(X, Y) is defined as the dependence measure between two variable sets X and Y.

We say MB_i passes the screening test if it satisfies at least one of two conditions. We remove the feature whose Markov Blanket candidate passes the Screening test. Unlike [9], we remove both irrelevant and redundant features at the same time rather than in two separate steps.

We can remove irrelevant features, regardless of independent irrelevant features or a set of irrelevant features that are dependent to each other. The independent irrelevant feature always satisfies $HSIC(MB_i, C) > HSIC(MB_i \cup F_i, C)$, since adding such an irrelevant feature F_i into MB_i will decrease the dependence between MB_i and C. If we have a set of irrelevant features which themselves are dependent on each other, adding such irrelevant features F_i into MB_i may not decrease the dependence on target variable C, and will not be removed based on condition 1. However, each will still be removed one by one according to condition 2. This is because in such cases, F_i is irrelevant to C, which means they are likely to be independent resulting in HSIC(F_i ,C) usually being smaller than HSIC(MB_i , F_i). Therefore such kinds of F_i will be removed one by one, and the last one in this set will be removed by condition 1 as we discussed before.

Condition 2 ensures that we remove the redundant features, since the corresponding MB_i of the redundant feature F_i can subsume the information this feature has about the target variable. $HSIC(MB_i, F_i) > HSIC(F_i, C)$ implies that F_i is more dependent on MB_i than to C; $HSIC(MB_i, C) > HSIC(F_i, C)$ means MB_i is more dependent on C than F_i and ensures MB_i has more deterministic information with the target variable C than F_i does.

IV. Feature Selection Algorithm

Having defined our approximate Markov Blanket, we now describe how to apply it to our feature selection method. The premise of feature selection using the Markov Blanket is to remove the feature for which we can find the Markov Blanket in the remaining features. In our method, we remove the feature for which we can identify its approximate Markov Blanket since we are actually looking for an approximate the Markov Blanket here. With the approximate Markov Blanket defined in section 3.1.2, we proposed a feature selection algorithm called the Markov Blanket based Feature Selection method (MBFS), as shown in Algorithm 1.

Algorithm 1. MBFS method

Input: ***F*** = $\boldsymbol{F_1}$, $\boldsymbol{F_2}$, ..., $\boldsymbol{F_N}$, ***C*** // training data set with N features and target ***C***
Output: ***MB*** // as set of selected features

For i = 1 to *N* do
  calculate HSIC($\boldsymbol{F_i}$, ***C***);
  insert $\boldsymbol{F_i}$ into list ***S*** based on *HSIC*($\boldsymbol{F_i}$, ***C***);
End For
For i = 1 to *N* do
  For each $\boldsymbol{F_j}$ listed in ***S*** before $\boldsymbol{F_i}$
    MB_{can} = $\boldsymbol{F_j}$
    If
      $\boldsymbol{F_i}$ and MB_{can} pass the screen test
    Then
      Remove $\boldsymbol{F_i}$ from ***S***
      Skip the rest $\boldsymbol{F_j}$, goto check next $\boldsymbol{F_i}$
    End If
  End For
End For
MB = ***S***

In the MBFS algorithm for each F_i in the whole feature set F, we first compute HSIC(F_i ,C), which is actually the dependence between F_i and the target variable C, and then sort the features into a list ***S*** in descending order based on the measured dependence. Then, for each feature F_i, we find the candidate Markov Blanket MB_{can} of F_i, which consists of the most dependent features before F_i in the sorted list ***S***. If F_i and MB_{can} pass the screen test, F_i will be removed from the sorted list ***S***. Repeating this process with all features one by one, we can remove all features that can find the approximate the Markov Blanket in the remaining set of features. No multi-iteration is needed in our algorithm.

The main computation lies in computing HSIC values, which has complexity of $O(M^2)$ in terms of the number of instances M, since it takes $O(M^2)$ to compute the HSIC value. However, this is more efficient than other kernel-based methods, which usually need $O(M^3)$ complexity. In terms of the number of features N, MBFS has linear complexity to compute the HSIC value between each feature and target. However, for each feature we have to find the candidate Markov Blanket in which there are k features. We need to search the features before F_i in the list ***S*** to find the candidate Markov Blanket with k features. Hence, in order to select the optimal subsets MB, the algorithm has a complexity of $O(p * N)$, where p is

TABLE I. SIDE EFFECTS PREDICTION OF DOMPERIDONE

	All feature	MBFS	HSMB	FCBF	GS	BAHSIC
Sensitivity	0	**0.8889**	0.6154	0	0	0.5714
Specificity	1.0000	**0.9730**	0.9697	1.0000	1.0000	0.9688
Accuracy	0.5000	**0.9309**	0.7925	0.5000	0.5000	0.7701

(a) classification accuracy

MBFS	HSMB	FCBF	GS	BAHSIC
3	1	3	4	Top 4

(b) Number of selected features

TABLE II. CLINIC EFFICACY PREDICTION OF DOMPERIDONE

	All feature	MBFS	HSMB	FCBF	GS	BAHSIC
Sensitivity	1.0000	**0.9667**	0.8750	0.7647	0.7429	0.8387
Specificity	0	**0.8750**	0.8571	0.6667	0.6364	0.7333
Accuracy	0.5000	**0.9208**	0.8661	0.7157	0.6896	0.7860

(a) classification accuracy

MBFS	HSMB	FCBF	GS	BAHSIC
6	1	3	5	Top 6

(b) Number of selected features

the number of features before a certain feature F_i in the list $\boldsymbol{S}$. In the worst case, p tends to become N, resulting in $O(N^2)$. However, in the best case, p tends to be a small constant if we remove enough features, resulting in a complexity of $O(N)$. Fortunately, in real datasets, we are more likely to tend to the best case. This is because in high dimensional datasets, we are likely to remove most features.

V. RESULTS

A. Data Description and Experiment Setup

Our SNP microarray data consists of 46 patients, each with 853,943 SNPs. We consider two classification problems, one related to Clinical Efficacy (1 and -1 correspond to responder and non-responder, respectively), and the other related to Side Effect (1 and -1 correspond to patients with and without side effects, respectively). The patient-reported response on the Clinical Patient Grading Assessment Scale (CPGAS) was used to categorize patients as definite responders (scores of 2 or 3) and non-responders (scores of 1 or less). To reduce data dimensionality, single nucleotide polymorphisms (SNPs) were coded as a single three valued variable (AA=0,BB=2, null=AB=1). Initially, data for fluorescence intensity for all 906,600 SNPs were included in the analysis. Data was then screened to ensure the presence of polymorphisms. Constant invariant SNPs were excluded, leaving 853,943 SNPs.

We compared our algorithm to the GS algorithm [6], which uses a test-based Markov Blanket discovering method, FCBF [8], which uses an approximate Markov Blanket method as well, HSMB[13], which checks only one Markov Blanket candidate for each feature, and BAHSIC [11], which directly uses HSIC as the feature selection criterion without finding the Markov Blanket.

For the learning algorithm, we performed a leave-one-out Support Vector Machine (SVM) using a Gaussian kernel. As in [11], we used the same SVM for all methods where δ of the Gaussian kernel was set as the median distance between points in the sample [12]. FCBF, GS, HSMB and MBFS do not need to predefine the number of selected features. For the BAHSIC method, we choose the same number of features as the largest number of features in the subset FCBF, GS, HSMB, and MBFS returns.

B. Results by MBFS and Alternatives

We first compared our method to alternatives according to classification accuracy when using selected SNPs and also according to the number of selected SNPs. The results on clinical efficacy and side effects of Domperidone are shown in Table 1 and Table 2, respectively. We compared the sensitivity, specificity, and accuracy, defined as the mean of sensitivity and specificity. Both Table 1 and Table 2 show that our MBFS method outperforms the alternatives since our method always achieved the best accuracy with the least number of SNPs. 'All feature' in Table 1 and Table 2 means using all SNPs to train the classifier.

To compare the quality of selected SNPs by different methods, we modified the stop criteria to let MBFS, HSMB, FCBF and GS select different numbers of SNPs, though they can choose optimal sets of SNPs automatically. We compared the quality of different numbers of selected features and present

the results in Figure 1 and Figure 2. Figure 1 and Figure 2 show that our feature selection method MBFS performs the best out of the four methods considered.

We also compared the stability among different feature selection methods. As mentioned before, we applied leave-one-out Support Vector Machine (SVM). Specifically, for our data, we applied 46 SVMs where each SVM was trained on 45 patients and tested on the remaining patient. Features selected are slightly different when training different SVM. We choose the most frequent features as the final optimal set. We compared the frequency of the optimal set selected by different methods to demonstrate the stability of the methods. The higher the frequency is, the higher stability is. The results are shown in Figure 3 and Figure 4. We clearly show that our method was the most stable of the four methods considered in this study.

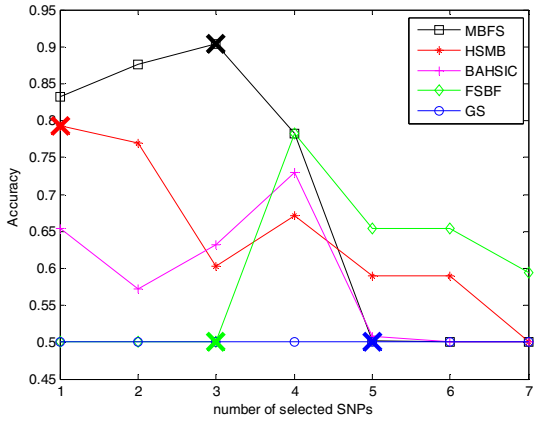

Figure 1. Accuracy for different numbers of features on the side effect prediction of Domperidone ('X' indicates the number of features selected automatically by each method except BAHSIC)

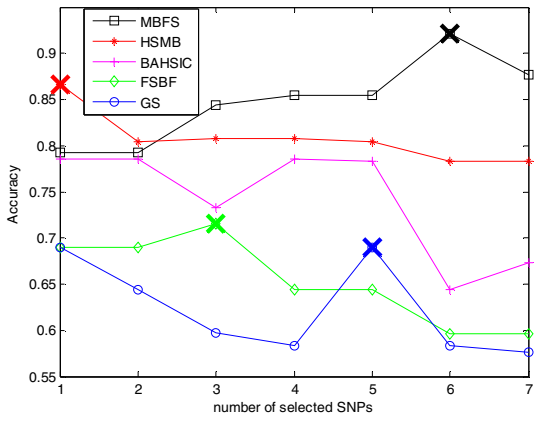

Figure 2. Accuracy for different numbers of features on the side effect prediction of Domperidone ('X' indicates the number of features selected automatically by each method except BAHSIC)

C. *Discussion*

The IDs of selected SNPs are shown in Table 3 and Table 4. Since BAHSIC cannot automatically select the optimal set of SNPs, we report 4 and 6 most relevant SNPs as the results of this method for side effect and efficacy problem, respectively. Table 3 shows the SNPs selected in the problem of Side Effect Prediction, and Table 4 lists the selected SNPs for Clinical Efficacy Prediction.

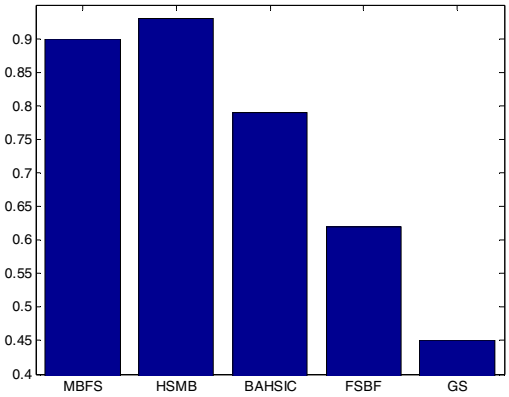

Figure 3. Frequency of optimal sets selected by different methods with the side effect of Domperidone.

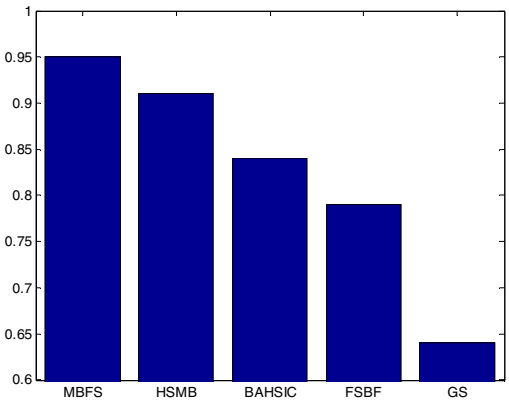

Figure 4. Frequency of optimal sets selected by different methods on the clinic efficacy of Domperidone.

The feature selection method MBFS was applied to 46 genomes categorized by the incidence of side effects, or clinical efficacy of domperidone. As a result, SNPs with the highest classification accuracy were selected in each of these analyzes. Our analytical tool identified several SNPs associated with efficacy and side effects of domperidone. Importantly,

TABLE III. 6 SNPS SELECTED BY MBFS FOR PREDICTION OF DOMPERIDONE SIDE EFFECTS

Selected SNP ID	Chromosome	Gene Symbol	Location
rs965739	7	C7orf25 /// GLI3	intergenic region
rs3891683	1	CACNA1E	intron
rs9964594	18	CDH2 /// CHST9	intergenic region
rs4511574	17	FBXL20	intron
rs10831474	11	MAML2	intron
rs9632703	7	SEMA3E	intron

TABLE IV. 3 SNPS SELECTED BY MBFS FOR PREDICTION OF DOMPERIDONE CLINICAL EFFICACY

Selected SNP ID	Chromosome	Gene Symbol	Location
rs9977558	21	JAM2 /// MRPL39	promoter region
rs585620	10	KCNMA1	intron
rs7637788	3	UBE2E1	putative promoter region

some of the identified SNPs are located in the promoter, intronic, and intergenic regions of the genes that could contribute to physiological effects of domperidone (Table 3 and 4). While in many cases the causal relation between physiological mechanism of domperidone and affected genes remains obscure, some genes provide attractive candidates with clear physiological significance. For example, the calcium-sensitive potassium channel KCMA1 plays a key role in regulation of the contraction of smooth muscle, and genetic polymorphism in the *KCNMA1* gene could modulate domperidone effects. Similarly, *JAM2* may have functions related to efficacy/side effects of domperidone. *JAM2* (junctional adhesion molecule 2) encodes a protein that is localized in the tight junctions between high endothelial cells. Because this protein forms a physical barrier to prevent solutes and water from passing through the paracellular space, genetic variations in its expression or functioning may contribute to domperidone gut absorption.

Clinical efficacy of domperidone pharmacotherapy has been predicted by SNP rs4511574, located within the first intron of the FBXL20 gene. Moreover, the same SNP rs4511574 was selected with the help of Least Angle Regression Analysis (LAR) [14]. Therefore, the feature selection method MBFS suggested in the present work allows an exploratory data analysis without a pre-specified hypothesis. By accomplishing MBFS analysis of a limited number of interrogated genomes, we were able to generate testable hypotheses about an association of phenotypes with genetic markers. Specific hypotheses formulated as a result of our analysis will be statistically evaluated in a larger prospective group of patients. Direct functional assays to address the clinical significance of our findings are needed using the *in vitro* and *in vivo* models.

Until the results of clinical validation of identified markers become available, we have to rely on circumstantial evidence that the results generated by the proposed approach are feasible. First, several selection methods provide overlapping (though non-identical) lists of genes. Second, if differential genetic markers located in the same gene are selected, this improves the chances that a true association is found.

VI. CONCLUSION

We applied our feature selection method (MBFS) on real SNP microarray data to select the informative SNPs for prediction of clinical efficacy and side effects of Domperidone. Experiments show that the classification model built on SNPs selected by our method is more accurate than the model built on all SNPs and the models built on SNPs selected by four alternative feature selection methods.

ACKNOWLEDGMENT

We are grateful to Drs. J.R. Testa and J.M. Pei (Genomics Facility, Fox Chase Cancer Center, Philadelphia, PA) for their help in performing SNP microarray analysis. This work was supported in part by the Jayne Haines Center for Pharmacogenomics and Drug Safety, Temple University School of Pharmacy, Seed Grant from Temple University (to HP, MJ, and EK), and in part by a grant from the Pennsylvania Department of Health (to ZO). The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. This work was also funded in part by DARPA grant DARPA-N66001-11-1-4183 negotiated by SSC Pacific (to ZO).

REFERENCES

[1] S.T.Sherry, M.H.Ward, M.Kholodov, J.Baker, L.Phan, E.M.Smigielski, K.Sirotkin. "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res*. 29 (2001) 308-311.

[2] H.Y.Benjamini Yoav. "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society 57 (1995) 289-300.*

[3] L. Yu and H. Liu. "Feature selection for high-dimensional data: a fast correlation-based filter solution", *In Proceedings of the twentieth International Conference on Machine Learning,* pages 856-863, 2003.

[4] L. Song, A. Smola, A. Gretton and K. M. Borgwardt. "A dependence maximization view of clustering", *In Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

[5] S. Yaramakala and D. Margaritis. "Speculative Markov Blanket Discovery for Optimal Feature Selection". *International Conference on Machine Learning,* 2005.

[6] D. Margaritis, and S. Thrun, "Bayesian network induction via local neighborhoods", *In NIPS* 1999, 12:505-511.

[7] J. Shen, L. Li and W. Wong. "Markov Blanket Feature Selection for Support Vector Machines", *In AAAI,* 2008.

[8] L. Yu and H. Liu. "Efficient Feature Selection via Analysis of Relevance and Redundancy", *Journal of Machine Learning Research*, 2004, 5, 1205-1224.

[9] A. Gretton, O. Bousquet, A. Smola,, and B. Scholkopf, "Measuring statistical dependence with Hilbert-Schmidt norms" *In ALT*, 2005, 63-78.

[10] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces", *Journal of Machine Learning Research*, 2004, 5, 73-99.

[11] A. V. Kozlov, and J. P. Singh , "An alternative to conditional probabilities for Bayesian belief networks" *In Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*., 1995, pp.376-385.

[12] B. Scholkopf, and A. Smola, "Learning with Kernels". Cambridge, MA, MIT Press, 2002

[13] Q. Lou and Z. Obradovic, "Feature selection by approximating the Markov Blanket in a kernel-induced space", *Europe Conference on Artificial Intelligence*, 2010, pp. 797-802.

[14] D. Wang, H. P. Parkman, M.R. Jacobs, A.K. Mishra, E. Krynetskiy, Z. Obradovic, "Least Angle Regression Analysis of a Genome-Wide Genotyping Study: Association with Clinical Efficacy and Side Effects of Domperidone Treatment for Gastroparesis". *Journal of Biomedical Informatics*, in press.