Applying Data Mining and Gene Selection for Cancer Research

Wannapa Kay Mahamaneerat[1)], Chi-Ren Shyu[2)], Jaturon Harnsomburana[3)], Pearlly S. Yan[4)], and Tim H.M. Huang[5)]

[1), 2), 3)] **Computer Engineering and Computer Science Department**
{wannapa, shyu, jaturon}@diglib1.cecs.missouri.edu
[4), 5)] Department of Pathology and Anatomical Sciences
{YanP, HuangH}@health.missouri.edu
University of Missouri, Columbia, MO, 65211, USA

Abstract

In this paper, we propose a novel decision support and data mining mechanism to assist biological and medical researchers to identify genes related to various cancers. This approach will streamline future biological experiments in bio-pharmaceutical research by elimination of irrelevant genes thereby helping to determine a minimum set of useful genes to study. The analysis of genes and cancers is based on the state-of-the-art microarray technology that is capable of holding information from more than 10K genes in one DNA chip. However, the rich information obtained from a microarray experiment results in an underdetermined problem when we have only a small number of cancer cases. To overcome this problem, we have developed a suite of customized feature selection algorithm, a binarize-scoring algorithm, and a dynamic threshold determination algorithm to effectively reduce the dimensionality to a small number of genes and still maintain the discrimination power among different cancer types.

The unique contribution of this work is the development of the dynamic threshold determination (DTD) algorithm. This DTD algorithm is designed to compute the thresholds of methylation levels in microarray. It is based on a statistical tool – a minimum risk classifier using Bayes decision theorem. The thresholds play an important role in this domain because only those genes that pass the thresholds are considered relevant for future study. Almost all of the current microarray research depends on predetermined thresholds. This human-driven approach is subjective and purely based on a particular domain expert's experiences. Therefore, it is an urgent call for us to develop this systematic approach – the DTD algorithm – to dynamically identify methylation thresholds. Those genes that survive the DTD thresholds are classified as "hypermethylated" genes, and the rest of genes are "hypomethylated". We further study the "hypermethylated" genes by applying the binarize-scoring approach that has been developed in our previous work for ovarian cancer.

This collaborative research mines a set of data that has 81 cancer cases collected from the Ellis Fischel Cancer Center at the University of Missouri, Washington University at Saint Louis, Brigham and Women's Hospital at Harvard Medical School, Indiana University School of Medicine, CRC Beaston Laboratories at the University of Glasgow in the UK, University of Bonn, Bonn, Germany, and Campus Rhineland-Palatinate, Mainz, Germany. Our ultimate goal is to help biological researchers discover genes that significantly correlate with certain cancers and are potentially useful for predicting treatment outcome for various cancer patients. An extensive study, based on experiments performed for determining the classification rates of selected genes is presented. The experimental results show that genes selected by our new approach are more informative than those selected by the traditional feature selection and static threshold approaches.

1. Introduction

With the advent of microarray technologies, it is now possible to identify simultaneously multiple genes "down- or up-regulated" during tumorigenesis and classify tumors based on their global patterns of gene expression. This type of research provides an unprecedented opportunity to improve our understanding of the molecular mechanisms leading to the development of cancer [1] [2]. Motivated by the same concept, we have recently developed a novel microarray technique, called differential methylation hybridization (DMH) [3], which allows for the first time a global analysis of another type of molecular alteration, i.e., DNA methylation, in cancer. DNA methylation is known to be a frequent

epigenetic event in cancer cells and has profound effects on the silencing of tumor-suppressor genes and genes responsible for genomic stability. Methylation-associated silencing of tumor suppressor genes could result in cells with a growth advantage, and clonal expansion of these proliferating cells may bear specific epigenetic signatures reflecting different types or stages of various tumor types. Unlike cDNA-based arrays, DMH analysis requires hybridization templates that are GC- and CpG-rich and should contain specific methylation-sensitive restriction sites. The CpG island microarray consists of individual clones from a human genomic library (CGI, distributed by the UK Human Genome Mapping Project) that are enriched in CpG island fragments. Tumor DNA and its paired normal DNA are restricted into short fragments by employing a frequent cutter that preserves the GC-rich regions. PCR-linkers are legated to these fragments and are later restricted with methylation-sensitive restriction enzymes. CpG islands in tumor DNA are often methylated and thus unaffected by the restriction enzyme. CpG islands in normal DNA are often unmethylated and are thus restricted away by the enzyme. The remaining intact fragments in both tumor and normal DNA will be amplified by linker-PCR and ready for fluorescent-labeling with Cy5 and Cy3, respectively. The hybridization signals between the tumor and the normal DNA need to be normalized due to the differential labeling efficiency between the two dyes. Having achieved that, CpG island loci having higher fluorescent signals are marked as hypermethylated in tumor and further characterized by genomic sequencing.

Microarray technology has been widely used in biological and medical researches during the last several years. In cancer research, information residing in a microarray is a collection of supervised data that have class label (cancer type) for each sample (patient). In the context of our research presented in this paper, we categorize the existing microarray research into two major focuses: 1) understanding the relationship between methylation levels of genes and cancer types [4] [5] [6] [9] [10], and 2) selecting genes from the overwhelming number of genes [4] [5] [6] [7] [8] [11] [12] [13] [14]. Almost all of these systems overlook the importance of threshold determination, especially in research with experiments based on all available genes in a microarray using continuous methylation levels. It is more meaningful to consider only the “hypermethylated” genes and discard the “hypomethylated” genes since the former is believed to have a high probability of responding to some treatments in cancer research and the latter is not. To identify which gene is “hypermethylated” or “hypomethylated”, we need to determine the thresholds for methylation levels. Therefore, instead of finding a subjective number from a domain expert, it is necessary to have a systematic approach to calculate the threshold. To do that, we developed a dynamic threshold determination (DTD) algorithm that can logically identify the methylation thresholds based on statistical measurements. We believe that the DTD algorithm will support and/or suggest the predetermined thresholds given by the domain experts.

We also extend our effectively implemented work [4], a binarize-scoring algorithm (BSA), and combine it with our newly invented DTD approach. The steps of classifying the cancer types by using the BSA and the DTD include 1) supplying the thresholds of methylation level of each gene obtained from the DTD algorithm to the BSA so that we can separate the “hypermethylated” and the “hypomethylated”, and 2) eliminating the irrelevant genes that cannot effectively classify the cancer types based on the score calculated by the BSA.

By using the state-of-the-art microarray technology, we have faced a known and difficult research problem — an underdetermined problem — caused by a small number of the cancer cases with a large number of genes to be studied. There are several approaches, presented in the current microarray research, to efficiently select the subset of genes, such as 1) statistical approach, i.e. analysis of variance (ANOVA) [6] and principal component analysis (PCA) [7] to mine the relationship between some certain genes with a specific cancer type, 2) support vector machine (SVM) approach [5] [13] based on a binary classifier, 3) feature selection approach [5] [8] [12], etc. We believe the reduction of dimensionality is the first task for gene selection and the biological explanation of one gene at a time is more significant than any linear combination of multiple genes. Therefore, to incorporate the existing approaches into our new method, we chose to extend a traditional feature selection technique called sequential forward selection (SFS) [17] [18] [19] by using a criterion function based on Bayesian classifier [20]. The idea of SFS is to iteratively include the next best gene with respect to the set of currently selected genes as long as the true positive gain [21] of the new set of genes is greater than or equal to a preset value.

This paper is organized as follows: in Section 2, we review the criterion function used in our algorithms, minimum risk classifier using Bayes decision theorem. In Section 3, we discuss a well-known feature selection algorithm – sequential forward selection (SFS). Section 4, we discuss our previous approach in gene selection and cancer classification – binarize-scoring algorithm (BSA). Then, we discuss the main contribution presented in this paper – the dynamic threshold determination (DTD) algorithm in Section 5. We have conducted four experiments that study the

performance of the DTD with the BSA for selecting relevant genes in comparison to the SFS and the plain BSA. Detailed discussion and experimental results are reported in Section 6. This paper is concluded in Section 7.

2. Minimum Risk Classifier Using Bayes Decision Theorem

In this section, we briefly present the criterion function used in the feature selection and the dynamic threshold determination approaches, minimum risk classifier using Bayes decision theorem [15] [16] [20].

Let r_i be the risk function that measures the probability that a sample x misclassified into class W_i. We can obtain an optimal classification result by minimizing r_i $(1 \quad i \quad n)$. Each r_i is calculated by

$$r_i = \sum_{j=1}^{n} l_{ij} P(W_j \mid x), \tag{1}$$

where l_{ij} is the loss for classifying x into class W_j when the actual class of x is W_i. It is reasonable to apply 0-1 loss function to determine the value of l_{ij}. With this setting, the loss of true positive (TP) (l_{ij} where $i = j$) is 0, and the loss of false positive (FP) (l_{ij} where $i \quad j$) is 1. To compute r_i, we need to know the value of $P(W_j \mid x)$. By using Bayes theorem, we get

$$P(W_j \mid x) = \frac{P(x \mid W_j) P(W_j)}{\sum_{k=1}^{n} P(x \mid W_k) P(W_k)}, \tag{2}$$

where n is the number of classes, $P(W_k)$ is the prior probability of class k that can be calculated by

$$P(W_k) = \frac{\| W_k \|}{\sum_{m=1}^{n} \| W_m \|}, \tag{3}$$

where $\| W_k \|$ is the number of samples in class W_k. In Eq. (2), $P(x \mid W_j)$ is the probability density function of class W_j. It is calculated by

$$P(x \mid W_j) = \frac{1}{2p^{d/2} \, | s_j |^{1/2}} e^{-\frac{1}{2}(x - m_j)' s_j^{-1} (x - m_j)}, \tag{4}$$

where d is the dimension of x. m_j, s_j, and $| s_j |$ are the mean, the covariance, and the determinant of the covariance matrix of class j, respectively.

Therefore, after we know the value of $P(W_k)$ in Eq. (3) and the value of $P(x \mid W_j)$ in Eq. (4), we will be able to calculate r_i in Eq. (1).

3. Feature Selection Algorithm

To compare our approach to the traditional feature selection approach, we would like to briefly introduce the concept of the sequential forward selection (SFS) algorithm in this section. SFS is a well-known feature selection algorithm [17] [18] [19] that is capable of selecting a subset of the features while effectively maintaining the classification power. Generally, using the entire set of original genes in classifying the cancer type is not as accurate as using only the effective ones. The details of SFS and its algorithm are to follow.

SFS is a greedy algorithm that starts with an empty set of candidate features. The feature selection procedure first

picks up the most significant feature from the entire feature set by testing each feature individually in search of the one that provides the best performance based on a criterion function. The procedure then picks up a best feature from the remaining feature set. This newly selected feature gives the best performance when it combines with the previously selected one. The same process will be iteratively executed until the new feature set fails to provide a sufficient gain value.

```
Sequential Forward Selection Algorithm (Y){
  1  //Acquire the set of features, Y
  2  //Initialize the set of selected features, X_k, to be an empty set
  3  X_k = f
  4  while gain( X_{k+1} ) >= e {   //e is a certain user-defined threshold of gain value
  5        X_{k+1} = X_k + argmax_{x ∈ Y − X_k} (X_k + x)
  6        X_k = X_{k+1}
  7  }
  8  return X_{k+1}
  9  }
```

We extend the original SFS algorithm by utilizing Bayesian classifier [20] in calculating `gain(` X_{k+1} `)` at line 4 and in finding $\underset{x \in Y - X_k}{\operatorname{argmax}}(X_k + x)$ at line 5. SFS runs on each subset of features created to find the best feature (x) with respect to the currently selected set of features (X_k) from the set of the remaining features ($Y - X_k$). Then, x will be combined with X_k to form the new set of selected features (X_{k+1}). The while loop from lines 4-7 will be executed as long as the performance of the new selected feature set, `gain(` X_{k+1} `)`, is greater than or equal to a preset value (e). After the algorithm terminates, it will return a set of selected features— X_{k+1} . In Section 6, we will demonstrate the selected genes and their classification rates by applying the extended SFS.

4. Binarize-Scoring Algorithm (BSA)

The BSA approach [4] uses an idea based on the assumption that an "hypermethylated" gene is biologically meaningful in cancer studies. If a gene is "hypermethylated", its methylation level should be higher than a threshold. A perfect gene that is "hypermethylated" to only a certain cancer *C* has all 1's associated with *C* and 0's with other cancers. This kind of perfect gene is defined as an ideal vector denoted by V^{ideal}. As an example shown in Table 1, the fourth row of this table illustrated V^{ideal} when *C* = "Breast cancer". We will use V^{ideal} as a benchmark to select genes in a microarray.

The BSA approach first computes a binary vector called the after-thresholding vector V^{gene} by applying the following threshold process:

$$V^{gene}(i) = \begin{cases} 1 & \text{if } V^{raw}(i) \quad \text{threshold} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Then, the BSA calculates the score of each V^{gene} by comparing with the ideal vector V^{ideal}. The score is obtained by the following equation:

$$binarize-score = \frac{\dfrac{|Min(V^{ideal}, V^{gene})|}{Max(|V^{ideal}|, |V^{gene}|)} + \dfrac{|Min(\bar{V}^{ideal}, \bar{V}^{gene})|}{Max(|\bar{V}^{ideal}|, |\bar{V}^{gene}|)}}{2}, \tag{6}$$

where $\bar{V}^{ideal}$ is the compliment of V^{ideal} and $\bar{V}^{gene}$ is the compliment of V^{gene}. $|X|$ is the 1st norm of *X*. $Max(x, y)$ and $Min(x,y)$ return the maximum and minimum values of two scalars (x and y), respectively. If X and Y are vectors, $Max(X, Y) = V^{max}$ returns a vector that has $V^{max}[i] = Max(X[i],Y[i])$, while $Min(X,Y) = V^{min}$ returns a vector that has $V^{min}[i] = Min(X[i],Y[i])$. The dividend from Eq. (6) is composed of two terms, $\frac{|Min(V^{ideal}, V^{gene})|}{Max(|V^{ideal}|, |V^{gene}|)}$ and $\frac{|Min(\bar{V}^{ideal}, \bar{V}^{gene})|}{Max(|\bar{V}^{ideal}|, |\bar{V}^{gene}|)}$. The first term is to measure how close the after-thresholding vector is to the ideal vector. The second term is to measure how close the compliment of the after-thresholding vector is to the compliment of the ideal vector. This is because we are interested in a gene that is "hypermethylated" only when it associates with a certain cancer type, but is "hypomethylated" with the

others. The denominator of Eq. (6) is set to 2 for the purpose of averaging out the summation of these two scores. It is not difficult to see that the higher the score obtained from Eq. (6), the more selected the gene is. A perfect gene should have *binarize-score* = 1.0.

We illustrate the BSA process using the example listed in Table 1. In this table, there are 10 samples from three cancer categories. The first row contains the raw methylation level of a specific gene V^{raw}; the second row is the after-thresholding vector V^{gene} of the first row with threshold value 0.8. If we are interested in understanding a specific cancer type, we assign 1's to the elements of the ideal vector $V^{ideal}_{Breastcancer}$ when the samples have the same cancer label, breast cancer in our example.

Table 1 A Binarize-Scoring Example with Threshold = 0.8

Cancer Type	Breast cancer	Breast cancer	Breast cancer	Breast cancer	Lung cancer	Lung cancer	Lung cancer	Ovarian cancer	Ovarian cancer	Ovarian cancer
V^{raw}	1.2	0.5	1.4	0.8	1.1	0.3	0.6	0.04	0.9	0.3
V^{gene}	1	0	1	1	1	0	0	0	1	0
$V^{ideal}_{Breastcancer}$	1	1	1	1	0	0	0	0	0	0

From Table 1, we calculate *binarize-score* by applying Eq. (6) as shown below.

$$|Min(V^{ideal}, V^{gene})| = |Min([1, 1, 1, 1, 0, 0, 0, 0, 0, 0], [1, 0, 1, 1, 1, 0, 0, 0, 1, 0])| = |[1, 0, 1, 1, 0, 0, 0, 0, 0, 0]| = 3$$

$$|Min(\overline{V}^{ideal}, \overline{V}^{gene})| = |Min([0, 0, 0, 0, 1, 1, 1, 1, 1, 1], [0, 1, 0, 0, 0, 1, 1, 1, 0, 1])| = |[0, 0, 0, 0, 0, 1, 1, 1, 0, 1]| = 4$$

$$Max(|V^{ideal}|, |V^{gene}|) = Max(|[1, 1, 1, 1, 0, 0, 0, 0, 0, 0]|, |[1, 0, 1, 1, 1, 0, 0, 0, 1, 0]|) = Max(4, 5) = 5$$

$$Max(|\overline{V}^{ideal}|, |\overline{V}^{gene}|) = Max(|[0, 0, 0, 0, 1, 1, 1, 1, 1, 1]|, |[0, 1, 0, 0, 0, 1, 1, 1, 0, 1]|) = Max(6, 5) = 6$$

$$binarize-score = \frac{\frac{3}{5}+\frac{4}{6}}{2} = 0.633$$

```
Binarize-Scoring Algorithm (D, ct){
  //D is the data matrix with dimension N_P (N_G+1) where the last
  //column is the cancer type. ct is the cancer type to be studied.
  N_P = number of patient
  N_G = number of gene
  T = user specify threshold
  V^raw = D // read data
  assign the ideal vector V_ct^ideal
  for i=1 to N_G{
      for j=1 to N_P{
```

```
10          if V_i^raw[j]  T   then  V_i^gene[j]=1
11          else V_i^gene[j]=0
12       }
13     binarize-score[i] = CalcBinarizeScore(V_i^gene, V_ct^ideal)
14  }
15  return binarize-score
16  }
```

This approach is an efficient tool, yet a simple and straight forward approach. The computational complexity is O($N_P N_G$) dominated by the for-loop from line 8-14. Please note that the `CalcBinarizeScore(`V_i^{gene}`,`V_{ct}^{ideal}`)` function at line 13 is calculated by applying Eq. (6).

However, the limitations of the BSA are: 1) A methylation threshold must be subjectively assigned by the domain expert so that we can calculate the after-thresholding vector V^{gene}, and 2) we can only compare two sets of cancer types: a set of patients with a specific cancer type to be studied and another set of patients that have the rest of the cancer types. Therefore, instead of obtaining the predetermined thresholds, we implement dynamic threshold determination (DTD) algorithm in order to dynamically provide the thresholds to the BSA. Furthermore, the DTD algorithm can also analyze multiple cancer types simultaneously. The details are explained in the next section.

5. Dynamic Threshold Determination (DTD) Algorithm and BSA with DTD Algorithm

Threshold determination is an important issue in analyzing microarray data because a threshold is used to classify whether a gene is "hypermethylated" or "hypomethylated". From the existing approaches cited in Section 1, almost all of these systems overlook the importance of threshold determination that should be determined by utilizing statistical information calculated from the data. The thresholds that are currently used are normally acquired from the domain experts. A threshold obtained from an expert is a specific number that will be applied for all genes throughout an entire microarray to determine which gene is "hypermethylated". If a gene is "hypermethylated", it is believed to have a high probability of responding to some treatments in cancer research. We believe it is not a good idea to use only one threshold for all genes since the methylation level to identify a gene to be "hypermethylated" or "hypomethylated" can vary based on each gene's characteristic.

Our fundamental idea is to develop an algorithm that, for each gene, can determine a valid interval that will be able to classify samples of each cancer type. For this reason, we design a unique approach – dynamic threshold determination (DTD) algorithm – that dynamically calculates a pair of gene-cancer thresholds that will then determine the corresponding valid interval. The DTD algorithm utilizes the statistical information from a given microarray. We assume that each cancer type has a probability density function obtained in Eq. (4). To dynamically determine the thresholds and classify cancer types at the same time, we implement a criterion function in the DTD algorithm based on the Bayesian classifier [13] previously explained in Section 2 with the assumption that the distribution of the data is Gaussian. We use 0-1 loss function that minimizes the classification error rate: 0 for true positive and 1 for false positive. While minimizing the error rate, we maximize the true positive rate which then yields an optimal solution for classification.

Let C be the set of cancer types in a microarray used in this paper where C={*"Oligoastrocytoma(OA)"*, *"Glioblatoma multiforme(GM)"*, *"Pilocytic astrocytoma(PA)"*, *"Gangliogloma(GG)"*, *"Breast(Br)"*, *"Colon(Colo)"*, *"Endometrium(Endo)"*, *"Ovary(Ova)"*, *"Aplastic Anemia(AN)"*}, and $|C| = N_C$. The computational effort of applying the DTD algorithm for each gene is $N_C - 1$ comparisons that result in a set of $N_C - 1$ thresholds. We further calculate the gene-cancer threshold, $T_{c_j}^{g_i}$ for (g_i, c_j) pair, by the following procedures. Let $c_{remaining}$ be the samples that have cancer types other than c_j and $t_{c_j, c_{remaining}}^{g_i}$ be the threshold that is obtained by solving the intersection of two probability density functions of c_j and $c_{remaining}$ of a gene g_i. We obtain $T_{c_j}^{g_i}$ from the following equation:

$$T_{c_j}^{g_i} = \begin{cases} t_{c_j, c_{remaining}}^{g_i} & \text{if } t_{c_j, c_{remaining}}^{g_i} \text{ is solvable and } \mathrm{avg}(m_{c_j}^{g_i}) > \mathrm{avg}(m_{c_{remaining}}^{g_i}) \\ \min(m_{c_j}^{g_i}) & \text{if } t_{c_j, c_{remaining}}^{g_i} \text{ is unsolvable and } \min(m_{c_j}^{g_i}) > \min(m_{c_{remaining}}^{g_i}) \text{ and } \max(m_{c_j}^{g_i}) > \min(m_{c_{remaining}}^{g_i}) \\ avg(m_{c_j}^{g_i}) & \text{otherwise} \end{cases}, \quad (7)$$

where avg(.), min(.), and max(.) are functions to obtain the mean, minimum, and maximum of methylation levels $m_{c_j}^{g_i}$ for (g_i, c_j) pair. The pseudo code to describe the DTD algorithm is listed below:

```
Dynamic Threshold Determination Algorithm (D){
   1  //D is the data matrix with dimension N_P  (N_G+1) where the last column is the
   2  //cancer type.
   3  N_P = total number of patients
   4  N_G = number of genes
   5  N_C = number of cancer types
   6  N_max = a large positive number that is greater than the highest methylation level
   7  NP[j]=0                //Initialize NP--number of patients for cancer type c_j
   8  Calculate NP[j]        //Count number of patients for cancer type c_j
   9  for i=1 to N_G{         //Calculate the statistics of the data
  10      for j=1 to N_C{
  11          find min[i][j]
  12          find max[i][j]
  13          calculate avg[i][j]
  14          calculate cov[i][j]
  15      }
  16      for j=1 to N_C{     //Calculate the threshold for each cancer type
  17          for k=1 to N_C{
  18             threshold[i][j][k] = N_max //Initialize threshold[i][j][k]
  19             if j=k skip
  20             calculate threshold[i][j][k] = Bayes(avg[i][j], avg[i][k],
  21                                                  cov[i][j], cov[i][k])
  22          }
  23      }
  24 }
  25 for i=1 to N_G{         //Determine gene-cancer thresholds, gThres,
  26      for j=1 to N_C{       //as shown in Eq.(7)
  27          gThres[i][j] = avg[i][j];
  28          for k=1 to N_C{
  29             if j=k skip
  30             if((threshold[i][j][k]     N_max){
  31                 if (avg[i][j] > avg[k][j]){
  32                     gThres[i][j] = threshold[i][j][k]
  33                 } else skip
  34             } else if ((min[i][j] > min[k][j]) && (max[i][j] > max[k][j])){
  35                 gThres[i][j] = min[i][j]
  36             } else skip
  37          }
  38      }
  39 }
  40 return gThres
  41 }
```

The DTD algorithm has the computational complexity of $O(N_G N_C N_P)$ where N_G, N_C, and N_P are the number of genes, the number of cancer types, and the number of patients, respectively. However, it is always the case that $N_G >> N_P > N_C$. The detailed explanation of the DTD algorithm is to follow. There are two modules in this algorithm. In lines 9-15, the

algorithm calculates the statistical values of the microarray data set *D*; minimum, maximum, average, and covariance. The second module is to dynamically determine thresholds in lines16-39. It calculates the possible threshold for each pair of the probability density functions (Eq. (4)) of cancer types by using minimum risk classifier (Eq. (1)) based on Bayes theorem (Eq. (2)) in lines 16-23. We call the values calculated by a for-loop in lines 16-23 as "candidate thresholds" since we need to verify if these values are valid; not equal to the initialize value N_{max}, in lines 30. The final threshold for (g_i,c_j) pair is calculated by applying Eq. (7) in lines 25-39.

There are many benefits to applying the DTD algorithm in the context of microarray analysis in cancer research. Fig. 1 shows the advantages of the DTD over the existing approaches —the sequential forward selection (SFS) algorithm and the binarize -scoring algorithm (BSA). The remainder of this section will discuss the performance comparisons among the DTD, the SFS, and the BSA in details.

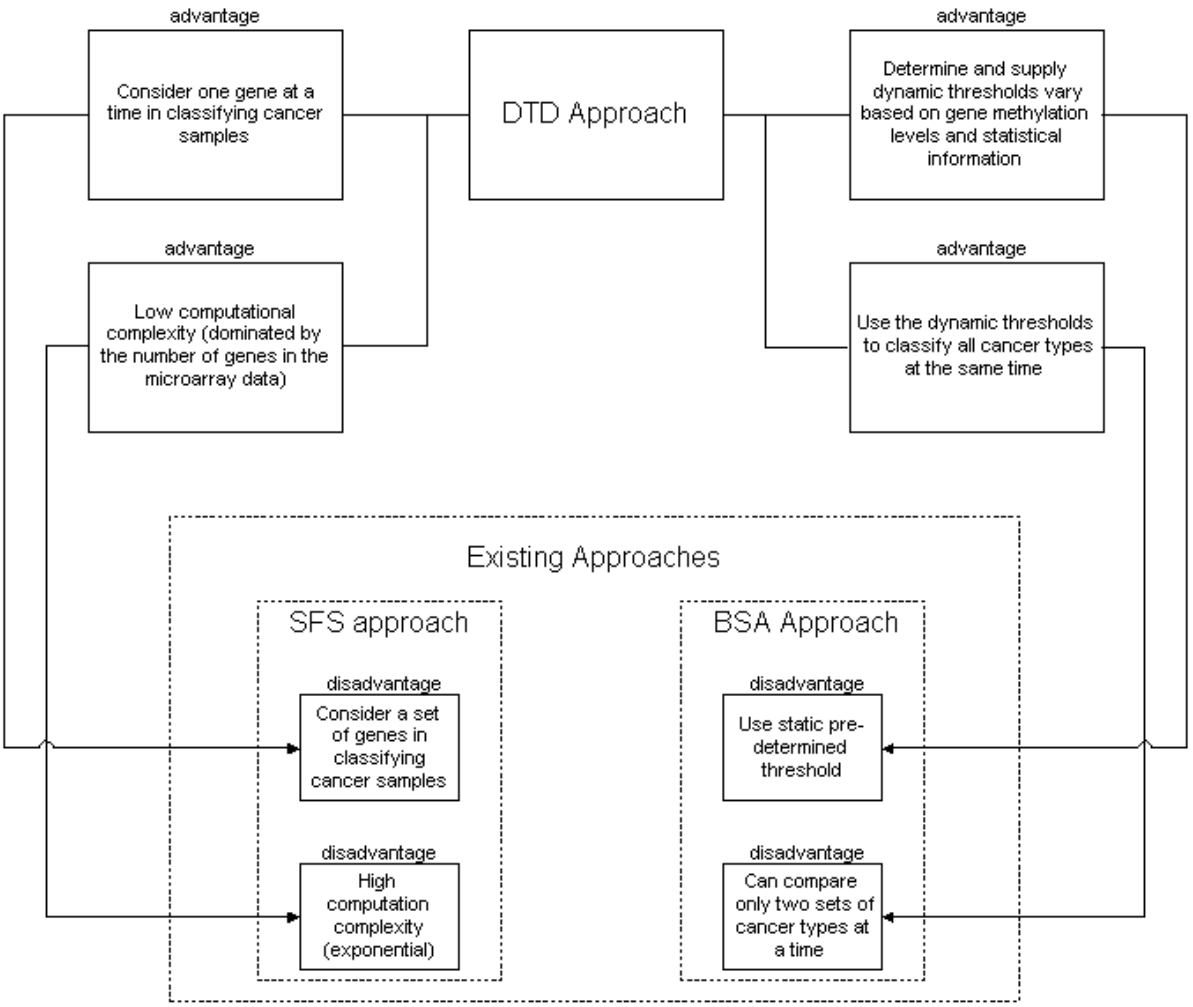

Fig. 1 The DTD overcomes disadvantages of the existing approaches, the SFS and the BSA

Biologically, the results from DTD algorithm can be more meaningfully explained than the results from the SFS because the latter considers a combination of genes in order to classify the cancer samples, but the former considers only an individual gene. Computationally, the DTD algorithm is much more efficient than the SFS. The computational complexity of DTD is $O(N_G N_C N_P)$ where N_G, N_C, and N_P are the number of genes, the number of cancer types, and the number of patients, respectively. It is always the case that N_G dominates the complexity since $N_G >> N_P > N_C$ as we already discussed previously. On the other hand, the computational complexity of SFS is dominated by the process of finding the $\underset{x \in Y - X_k}{\operatorname{argmax}}(X_k + x)$ at line 5 of the SFS algorithm which results in $\frac{N_G(N_G-1)}{2}$ computational time. In addition, it also needs to evaluate the `gain(` X_{k+1} `)` at line 4 by utilizing the Bayesian classifier [16] which has the computational complexity $N_G N_P$. In conclusion, the required computational time of the SFS algorithm is $(\frac{N_G(N_G-1)}{2})(N_G N_P)$.

As mentioned previously, the BSA relies on a static threshold that cannot be adapted to each individual gene's characteristics. Therefore, by applying the DTD algorithm, the BSA can utilize a set of reasonable and computable self-determination thresholds based on the likelihood of the data set in its gene selection process.

6. Experimental Results

We conducted the experiment on prescreened microarray data that contains 97 genes from 81 patients which are grouped into 9 cancer types — *C*, where *C*={*"Oligoastrocytoma(OA)"*, *"Glioblatoma multiforme(GM)"*, *"Pilocytic astrocytoma(PA)"*, *"Gangliogloma(GG)"*, *"Breast(Br)"*, *"Colon(Colo)"*, *"Endometrium(Endo)"*, *"Ovary(Ova)"*, *"Aplastic Anemia(AN)"*}. In order to compare the results among the different approaches presented in this paper, we benchmark the experimental results based on the binarize-score in Eq (6). The experimental results to be followed are designed to compare the selected genes and their classification power among the three approaches: the SFS algorithm discussed in Section 3, the BSA discussed in Section 4, and the BSA with the DTD algorithm discussed in Section 5.

6.1 Genes Selected by Sequential Forward Selection (SFS) Algorithm

We applied the SFS method, mentioned in Section 3, to our data set. We obtained the following genes: 13 (AutoGen18{SC#5}E9), 15 (AutoGen25{SC#13}C7), 35 (AutoGen1{PY#1}D5), 52 (AutoGen20{SC#7}D10), and 72 (AutoGen69{CpG-18}G8).

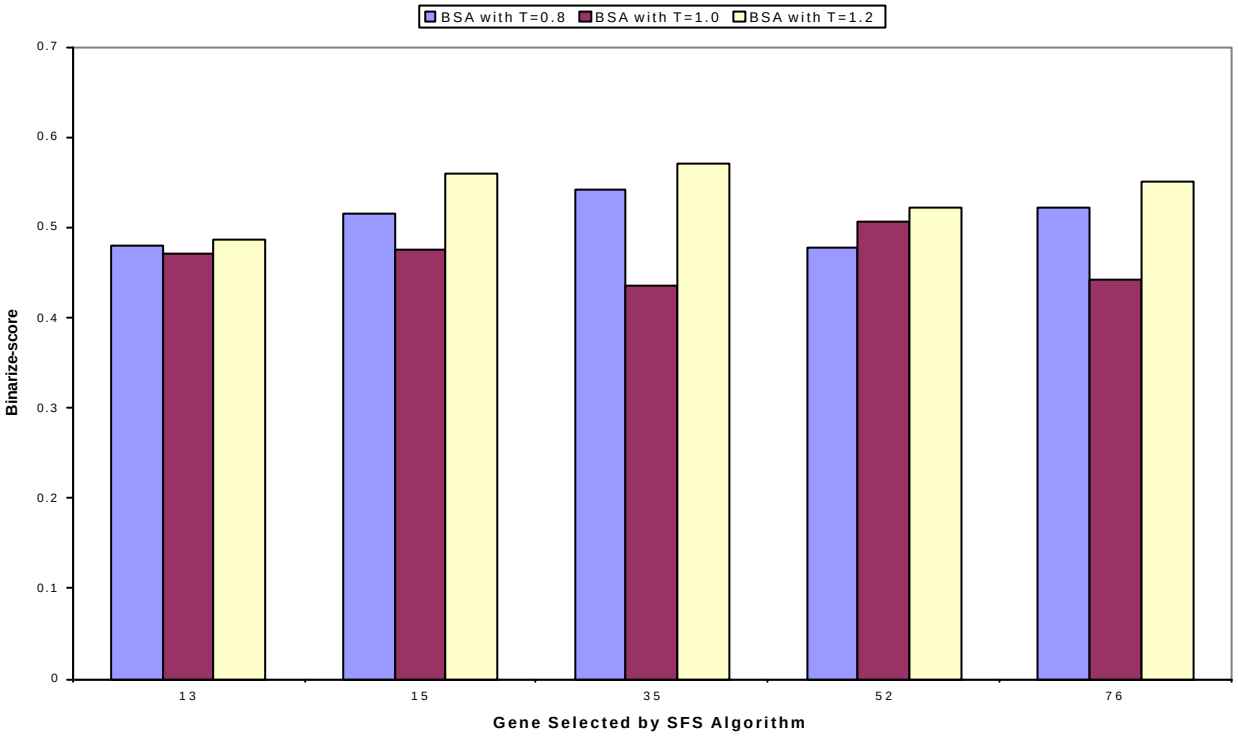

Fig. 2 The comparison of the binarize-scores of the selected genes by SFS algorithm

The experiment was then conducted to compare all cancer types simultaneously by using the selected set of genes based on their binarize-scores computed by the binarize-scoring algorithm (BSA) with three static thresholds of 0.8, 1.0, and 1.2. The range of possible scores is between 0 (minimum) and 1 (maximum). From Fig. 2 and comparing the results from the next experiment, no matter how the static thresholds were selected (0.8-1.2), we found that the genes selected by the SFS algorithm do not have significant classification power individually.

6.2 The Comparison between the Binarize-scoring Algorithm (BSA) and the Binarize-scoring with the Dynamic Threshold Determination (DTD) Algorithm

To compare the experimental results among the BSA with three static thresholds (T) of 0.8, 1.0, and 1.2 and the BSA with the DTD algorithm, we present the selected genes sorted by their binarize-scores of both approaches in Table 2. To make our discussion compact, we present only the result of the best 10 genes from each setting.

Table 2 The 10 best genes selected from BSA with static thresholds and BSA with DTD

Genes from BSA with T=0.8	Binarize-score	Genes from BSA with T=1.0	Binarize-score	Genes from BSA with T=1.2	Binarize-score	Genes from BSA with DTD	Binarize-score
73	0.58710	73	0.52628	73	0.59755	40	0.68100
40	0.56403	40	0.57398	46	0.57341	53	0.66426
28	0.54292	34	0.48216	35	0.57169	30	0.63910
35	0.54148	95	0.48718	21	0.57072	86	0.63569
69	0.53788	35	0.43584	40	0.57066	32	0.62040
95	0.53620	28	0.53006	95	0.56285	5	0.61304
11	0.53309	70	0.48070	15	0.55998	29	0.61242
0	0.53265	69	0.48662	8	0.55180	4	0.61124
34	0.53173	0	0.49242	76	0.55127	54	0.61020
46	0.53013	94	0.47586	70	0.55123	57	0.61016

From the experimental results shown in Table 2, we found that gene number 73 (AutoGen36(SC#26}F11) is the top ranked gene across three settings with static thresholds: T=0.8, T=1.0, and T=1.2 with the maxima binarize-scores 0.59755 that is comparable to the binarize-score from the 13-th ranked gene selected by our approach –the BSA with DTD algorithm. The best gene selected by the BSA with the DTD algorithm is gene number 40 (AutoGen7{MP#2}B9) with a higher binarize-score of 0.681. Overall, the top ten genes selected by our approach outperform those selected by the static settings. It is noteworthy to mention that the results from our approach also give much higher classification power and identify more relevant genes than the SFS algorithm.

Fig. **3** illustrates the binarize-scores of all genes in classifying each cancer using the dynamic thresholds generated by the DTD algorithm. The range of the scores is between 0.46754 (the worst gene – AutoGen1{PY#1}D5 – gene number 35) and 0.681 (the best gene – AutoGen7{MP#2}B9 – gene number 40)

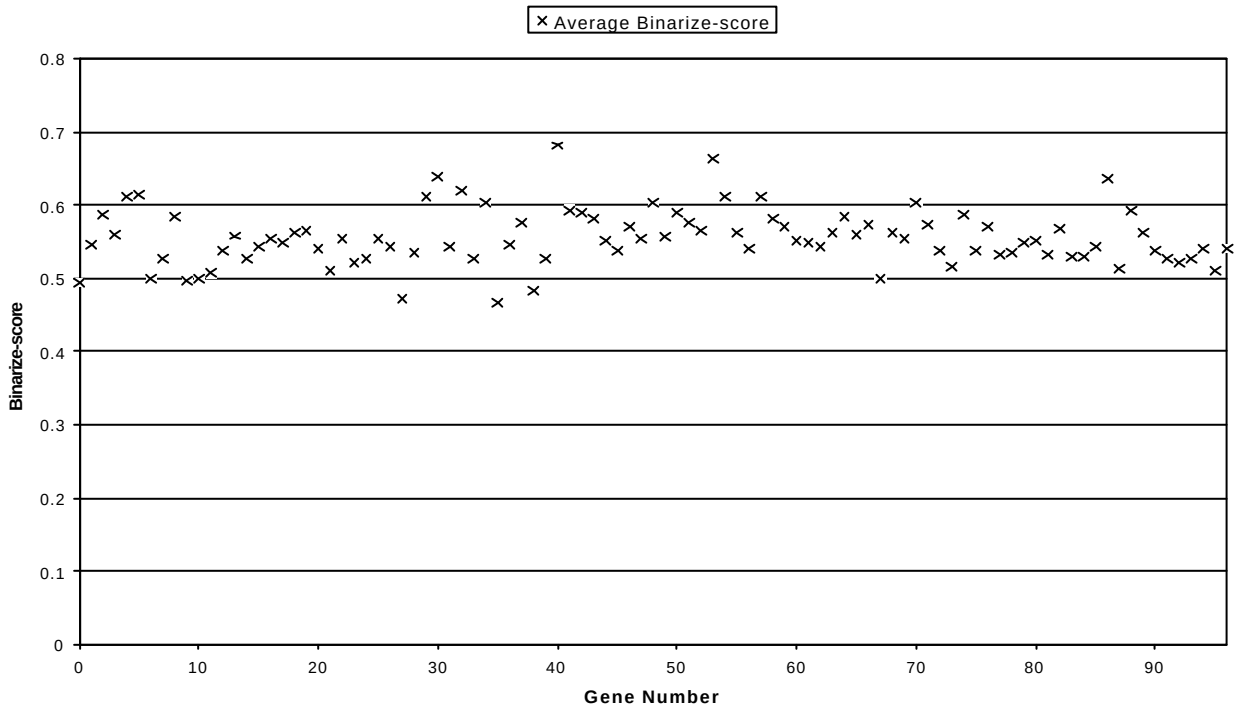

Fig. 3 Binarize-score of each gene (gene number 0-96) using the BSA with DTD algorithm

7. Summary and Future Work

In this paper, we have presented a new technology to mine microarray data in cancer research to assist biological researchers understanding the relationships between genes and cancer types. This unique approach — dynamic threshold determination (DTD) algorithm — can help researchers determine the cut-off value that categorizes an element in a microarray into either "hypermethylated" or "hypomethylated." It is also capable of eliminating irrelevant genes that have

low true positive rates with high false positive rates. It is critically important since the inclusion of irrelevant genes to be studied will result in a lengthened verification process in the biological community. We believe the genes selected by our approach closely reflect the methylation patterns for certain cancer types.

Research in mining microarray data is ongoing. In addition to the work reported in this paper, there are many research issues and interesting directions for future work. The first future work is to mine the microarray data without applying feature selection algorithm. To achieve this, we will face a well-known research issue – underdetermined system that has too many features (genes), but too few samples (cancer patients). To solve this problem, statistical tools, such as the Linear Mixed Model[22], will be studied and evaluated. The second future work is to study the pathways of genes[14] by applying association rules [23]. The outcome of the association rule research will provide a concrete idea about the relationships among genes. Last but not least, to biologically/clinically evaluate genes selected by our approach, we will follow up the results of the improvement from cancer treatment and findings in pharmaceutical research.

References

[1] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Jr. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, and L.M. Staudt: Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling, Nature, Vol. 403, pp503-11, 2000.

[2] J.B. Welsh, P.P. Zarrinkar, L.M. Sapinoso, S.G. Kern, C.A. Behling, B.J. Monk, D.J. Lockhart, R.A. Burger, and G.M. Hampton: Analysis of Gene Expression Profiles in Normal and Neoplastic Ovarian Tissue Samples Identifies Candidate Molecular Markers of Epithelial Ovarian Cancer, Proceedings of the National Academy of Science USA, Vol. 98, pp1176-1181, 2001.

[3] P. S. Yan, C-M. Chen, H. Shi, F. Rahmatpanah, S. H. Wei, C. W. Caldwell, and T. Huang: Dissecting Complex Epigenetic Alterations in Breast Cancer using CpG Island Microarrays, Cancer Research, Vol. 61, pp8375-8380, 2001.

[4] S. H. Wei, C-M Chen, G. Strathdee, J. Harnsomburana, C. R. Shyu, F. Rahmatpanah, H. Shi, S-W Ng, P. S. Yan, K. P. Nephew, R. Brown, and T. Huang: Methylation Microarray Analysis of Late-Stage Ovarian Cancinomas Distinguishes Progression-Free Survival in Patients and Identifies Candidate Epigenetic Markers, Clinical Cancer Research, Vol. 8, No. 7, July 2002.

[5] F. Model, P. Adorján, A. Olek, and C. Piepenbrock: Feature Selection for DNA Methylation Based Cancer Classification, Bioinformatics, Vol. 17, ppS157-S164, 2001.

[6] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing: Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblasitc Leukemia by Gene Expression Profiling, Cancer Cell, Vol. 1, pp133-143, March 2001.

[7] S. Raychaudhuri, J. M. Stuart, and R. B. Altman: Principal Component Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series, Pacific Symposium on Biocomputing 2000, Honolulu, Hawaii, pp452-463, 2000.

[8] S. Raychaudhuri, P. D. Sutphin, J. T. Chang, & R. B. Altman: Basic Microarray Analysis: Grouping and Feature Reduction. TRENDS in Biotechnology, Vol. 19, No. 5, pp189-193, 2001.

[9] J. H. W. Chen, R. Wu, P. C. Yang, J. Y. Huang, Y. P. Sher, M. H. Han, W. C. Kao, P. J. Lee, T. F. Chiu, F. Chang, Y. W. Chu, C. W. Wu, and K. Peck: Profiling Expression Patterns and Isolating Differentially Expressed Genes by cDNA Microarray System with Colorimetry Detection, Genomics, Vol. 51, pp313-324, 1998.

[10] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Boststein: Cluster Analysis and Display of Genome-wide Expression Patterns, Proceedings the National Academy Sciences USA, Vol. 95, pp 14863-14868, 1998.

[11] **J. Andrews, G. G. Bouffard, C. Cheadle, J. Lü, K. G. Becker, and B. Oliver:** Gene Discovery Using Computational and Microarray Analysis of Transcription in the Drosophila Melanogaster Testis, Genome Research, Vol. 10, Issue 12, pp2030-2043, December 2000.

[12] **M. Xiong, X. Fang, and J. Zhao:** Biomarker Identification by Feature Wrappers, Genome Research, Vol. 11, Issue 11, pp1878-1887, November 2001.

[13] M P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, Jr., and D. Haussier: Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines, Proceedings of the National Academy of Science USA, Vol. 97, pp262-267, 2000.

[14] **J. Theilhaber, T. Connolly, S. Roman-Roman, S. Bushnell, A. Jackson, K. Call, T. Garcia, and R. Baron:** Finding Genes in the C2C12 Osteogenic Pathway by k-Nearest-Neighbor Classification of Expression Data, Genome Research, Vol. 12, Issue 1, pp165-176, January 2002.

[15] S. Kumar, J. Ghosh, and M. Crawford: A Bayesian Pairwise Classifier for Character Recognition, Cognitive and Neural Models for Word Recognition and Document Processing, World Scientific Press, 2000.

[16] S. Theodoridis, K. Koutroumas: Pattern Recognition, First Edition Academic Press, 1999.

[17] P. Pudil, and J. Novovicova: Novel Methods for Subset Selection with Respect to Problem Knowledge, IEEE Transactions on Intelligent Systems, Vol. 13, No. 2, pp66-69, 1998.

[18] P. Pudil, J. Novovicova, and J. Kittler: Floating Search Methods in Feature Selection, Pattern Recognition Letters, Vol. 15, pp119-1125, 1994.

[19] A. W. Whitney: A Direct Method of Nonparametric Measurement Selection, IEEE Transactions on Computers, Vol. 20, pp1100-1103, 1971.

[20] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin: Bayesian Data Analysis, Chapman & Hall/CRC, 1995.

[21] J. S. Milton, and J. C. Arnold: Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences, McGraw-Hill, 1994.

[22] R. Natarajan, and R. E. Kass: Reference Bayesian Methods for Generalized Linear Mixed Models, Journal of the American Statistical Association, Vol. 95, pp227-237, 2000.

[23] R. Agrawal, T. Imielinski, and A. Swami: Mining Association Rules Between Sets of Items in Large Databases, Proc. ACM SIGMOD Conf. Management of Data, pp 207-216, May 1993.