Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Gene Expression Analysis Methods on Microarray Data -A Review

Prof G V Padma Raju[1]

[1] SRKR ENGINEERING COLLEGE AFFILIATED TO ANDHRA UNIVERSITY

Received: 16 December 2013 Accepted: 1 January 2014 Published: 15 January 2014

Abstract

In recent years a new type of experiments are changing the way that biologists and other specialists analyze many problems. These are called high throughput experiments and the main difference with those that were performed some years ago is mainly in the quantity of the data obtained from them. Thanks to the technology known generically as microarrays, it is possible to study nowadays in a single experiment the behavior of all the genes of an organism under different conditions. The data generated by these experiments may consist from thousands to millions of variables and they pose many challenges to the scientists who have to analyze them. Many of these are of statistical nature and will be the center of this review. There are many types of microarrays which have been developed to answer different biological questions and some of them will be explained later. For the sake of simplicity we start with the most well known ones: expression microarrays.

Index terms— micro array, classification

1 Introduction

icroarrays and other genomic data are different in nature from the classical data around which most statistical techniques have been developed. In consequence, in many cases it has been necessary to adapt existing techniques or to develop new ones in order to fit the situations encountered. We will examine some key components of microarray analysis, experimental design, quality control, preprocessing and statistical analysis. In the last section we will consider some topics where open questions still remain and which can be considered attractive for statisticians who wish to focus some of their research in this field. One of the handicaps for statisticians who may consider entering this field is how to start applying their knowledge to these problems. We will present some real examples, which we will use along the paper to illustrate some concepts [1][2][3][4][5][6][7][8][9][10][11][12][13][14][15]. The goal of this section is to present an integrated view of the whole process of analyzing microarray data (see figure 1). Many review papers discuss the statistical techniques available for the analysis at this level.

2 II.

3 Methods for Classification

Different strategies have been proposed over the last several years for feature/gene selection: filter, wrapper, embedded [16], and more recently ensemble techniques [17].

Filter techniques assess the discriminative power of features based only on intrinsic properties of the data. As a general rule, these methods estimate a relevance score and a threshold scheme is used to select the best-scoring features/ genes. Filter techniques are not necessarily used to build predictors. As stated in [18], DEGs may also be good candidates for genes which can be targeted by drugs. This group of techniques is independent of any classification scheme but under particular conditions they could give the optimal set of features for a given

classifier. Saeys et al. [1] also stress on the practical advantages of these methods stating that "even when the subset of features is not optimal, they may be preferable due to their computational and statistical scalability." Wrapper techniques select the most discriminant subset of features by minimizing the prediction error of a particular classifier. These methods are dependent on the classifier being used and they are M changing the way that biologists and other specialists analyze many problems. These are called high throughput experiments and the main difference with those that were performed some years ago is mainly in the quantity of the data obtained from them. Thanks to the technology known generically as microarrays, it is possible to study nowadays in a single experiment the behavior of all the genes of an organism under different conditions. The data generated by these experiments may consist from thousands to millions of variables and they pose many challenges to the scientists who have to analyze them. Many of these are of statistical nature and will be the center of this review. There are many types of microarrays which have been developed to answer different biological questions and some of them will be explained later. For the sake of simplicity we start with the most well known ones: expression microarrays.

mainly criticized because of their huge computational demands. More than that, there is no guarantee that the solution provided will be optimal if another classifier is used for prediction.

Embedded techniques represent a different class of methods in the sense that they still allow interactions with the learning algorithm but the computational time is smaller than wrapper methods.

Ensemble techniques represent a relatively new class of methods for FS. They have been proposed to cope with the instability issues observed in many techniques for FS when small perturbations in the training set occur. These methods are based on different sub sampling strategies. A particular FS method is run on a number of subsamples and the obtained features/genes are merged into a more stable subset [19]. a) Filter Methods -A Ranking Approach Most filter methods consider the problem of FS as a ranking problem. The solution is provided by selecting the top scoring features/genes while the rest are discarded. Generally these methods follow a typical scenario described below.

1. Use a scoring function S(x) to quantify the difference in expression between different groups of samples and rank features/genes in decreasing order of the estimated scores. It is supposed that a high score is indicative for a DEG. 2. Estimate the statistical significance (e.g., p-value, confidence intervals) of the estimated scores. 3. Select the top ranked features/genes which are statistically significant as the most informative features/ genes (alternatively one could be interested in selecting the top ranked features/genes only as opposed to the top ranked significant ones). 4. Validate the selected subset of genes.

In the above-mentioned generic algorithm one can identify two aspects specific to this type of methods which play an important role in identifying informative features/genes: first, the choice of a scoring function to compute the relevance indices (or scores) and second, the assignment of statistical significance to computed scores. They will receive further consideration in order to be able to reveal the main differences between different methods and therefore helping to categorize them. As an additional remark, the reader should note that ranked lists of features/genes can also be obtained via wrapper/embedded methods not only for filters, e.g., SVM.

Recursive Feature Elimination (SVMRFE) [20] or Greedy Least Square Regression [21].Here we also outline the fact that any combination of a scoring function and a statistical significance test designed to quantify the relevance of a feature/gene for a target annotation can be transformed into a ranking method for FS. Since all steps in the generic algorithm described above are independent one from another, the users do have a lot of freedom in the way they wish to perform the selection.

4 b) Scoring Functions -Assigning Relevance Indices to Features

Scoring functions represent the core of ranking methods and they are used to assign a relevance index to each feature/gene. The relevance index actually quantifies the difference in expression (or the informativeness) of a particular feature/gene across the population of samples, relative to a particular target annotation. Various scoring functions are reviewed and categorized here. They cover a wide range of the literature proposed for DEGs or biomarkers discovery. The scoring functions are enumerated and categorized according to their syntactic similarities. A similar approach presenting a very comprehensive survey on distance measures between probability density functions has been employed in [22].

Several groups of scoring functions for gene ranking have been identified. In the first group, we gathered scoring functions which estimate an average rank of genes across all samples. Scoring functions from the second group quantify the divergence (or the distance) between the distributions of samples corresponding to different classes associated to a target annotation per feature/gene. The third group contains information theory-based scoring functions while the fourth group measures the degree of association between genes and a target annotation. The last group gathers a list of miscellaneous scoring functions which cannot be included in the previous four. The big majority of scoring functions presented here are usually defined to rank single genes but some of them can be easily adapted for pairs or groups of genes.

5 i. Ranking Samples across Features

This group is represented by two scoring functions: rank sum and rank-product. Supposing x1 and x2 are the expression levels of a certain gene in class c1 and class c2, respectively, the rank-sum method first combines all

the samples in x1 and x2 and sorts them in ascending order. Then the ranks are assigned to samples based on that ordering. If k samples have the same value of rank i, then each of them has an average rank. If n1 and n2 denote the numbers of samples in the smaller and larger group, respectively, then the rank-sum score is computed by summing up the ranks corresponding to samples in c1. For a GEM data set, the rank-product method consists in ordering the genes across all samples in the value ascending order and then for each gene the rank product score is obtained by taking the geometrical average of the ranks of that gene in all samples. Another direction toward the identification of informative features/genes is to quantify the difference between the distributions of groups of samples associated to a target annotation. These scoring functions can be generically described as a function f(x1; x2) with x1; x2. For this purpose, some simple measures rely only on low-order statistics, in particular the first and second moment (mean and variance) of the distribution of expression levels in different groups. This is the simplest way to compare the distributions of two populations and implicitly imposes some more or less realistic assumptions on the distributions of samples in each population (e.g., normal distributed samples). Despite this obvious drawback they are still the most popular scoring functions used to create filters for FS in GEM analysis due to their simplicity. These scoring functions can be grouped in two families: fold-change family and t-test family. A different strategy in comparing the distributions of different populations is to rely on different estimates of the probability density function (pdf) or the cumulative density function (cdf) of populations but these methods are more expensive computationally. The different families of scoring functions mentioned here will be further presented in this section.

6 a. Fold-change family

Relative indices are assigned to features/genes based only on mean estimates of the expression levels across different groups of samples per gene. According to [23] two forms are encountered for the fold-change scoring functions: fold-change ratio and fold change difference. However, the fold-change difference is less known and usually researchers who mention foldchange in this context actually refer to fold change ratio. In practice, many packages for GEM analysis typically provide the log2 of the ratio between the means of group 1 and group 2. The numbers will be either positive or negative preserving the directionality of the expression change. t-test family. Several forms derived from the ordinary two-sample t-test are used to measure the difference in expression of genes. In the same family, we include the Z-score or the signal to noise ratio (SNR) defined as the ratio between the fold-change difference and the standardized square error of a particular gene. These scoring functions make use of both the first and second moments to assign relevance indices to genes.

7 b. Bayesian scoring functions

In several studies, the authors have defined scoring functions for informative features discovery in a Bayesian framework. The main motivation behind this is the difficulty in obtaining accurate estimates of the standard deviation of individual genes based on few measurements only. In order to cope with the weak empirical estimation of variance across a single feature/gene, several authors proposed more robust estimations of the variance by adding genes with similar expression values.

8 c. PDF-based scoring functions

Scoring functions in this category rely on different estimates of the pdfs of populations, from simple histograms to more complex estimators such as the Parzen window estimator [24]. Only few scoring functions based on this idea are used to discover informative features/genes. Here we identified Kolmogorov-Smirnov (K-S) tests [25], Kullback-Leibler divergence [26], or Bhattacharyya distance [27], but the mathematical literature abounds in measures quantifying the distance between pdfs revealing new possibilities to look for informative features/genes. We invite the reader to consult for a very comprehensive survey on this topic. Note that the use of these scoring functions for DEGs discovery is limited by the low number of samples in GEMexperiments which results in unreliable estimates of the pdf.

iii. Information Theory-Based Scoring Functions These scoring functions rely on different estimates of the information contained both in the target feature c and in the gene expression x.

9 iv. Measuring the Dependency between Features and

Target Feature as a Function Scoring functions in this group have the advantage that they allow features/genes ranking when the target annotation is a continuous variable (which is not the case of the previous mentioned scoring functions). They measure the dependency between the gene's expression profile x and the target feature c as a function f(x,c). Pearson's correlation coefficient (PCCs),Its absolute value equals 1 if x and c are linearly correlated and equals 0 if they are uncorrelated. Note that PCCs is only applied if c is a continuous variable. When c is binary, PCCs comes down to the Z -score. A similar measure used for this purpose is Kendall's rank correlation coefficient (KRCCs). A variant of this measure adapted to a two-class problem is proposed in [28].

10 v. Other Scoring Functions

A list of scoring functions mentioned in the literature for informative gene discovery which cannot be grouped in the above-mentioned families is presented here. The list presented in Table 1 includes: Area Under ROC Curve (AUC), Area Between the Curve and the Rising diagonal (ABCR), Between-Within class Sum of Squares (BWSS), and Threshold Number of Miss classifications (TNoM). The reader is encouraged to consult the associated references in Table 1 for further details about these scoring functions. Estimating the statistical significance for the relevance indices assigned to each feature/gene has been long addressed in the quest for DEGs. It is argued that statistical significance tests quantify the probability that a particular score or relevance index has been obtained by chance. It is common practice that features/genes ranked high in the list according to the relevance index, will be discarded if the computed scores are not statistically significant. There are different ways one can assign statistical significance despite many criticisms the most commonly used statistical significance test is the p-value. Many researchers advocate for alternative measures such as confidence intervals, especially due to the fact that p-values only bring evidence against a hypothesis (e.g., the null hypothesis of no "correlation" between features/genes and target annotation) and "confirm" a new hypothesis by rejecting the one which has been tested without bringing any evidence in supporting the new one [32]. Without entering into this debate, it is important to notice that statistical significance tests can be run either by exploring gene-wise information across all samples, either by exploring the large number of features in GEM experiments. Regardless the manner the statistical significance tests are performed, a permutation test is generally employed. It consists of running multiple tests which are identical to the original except that the target feature (or the class label) is permuted differently for each test. An important concept for estimating the statistical significance for DEGs discovery is the multiple hypotheses testing which will be described at the end of this section.

11 i. Exploring Feature-Wise Information to Asses Statistical Significance

This strategy assumes a large enough number of samples in order to infer upon the statistical significance of computed relevance indices of genes. The statistical significance is estimated for each feature/gene individually based on its intrinsic information. p-values. In statistics, the p-value is the probability of obtaining a test statistic (in our case a relevance index) at least as extreme as the one that was actually observed. The lower the p-value the more significant the result is (in the sense of statistical significance). Typical cutoff thresholds are set to 0.05 or 0.01 corresponding to a 5 or 1 percent chance that the tested hypothesis is accepted by chance. Pvalues can be estimated empirically by using a permutation test. However, standard asymptotic methods also exist, reducing substantially the computational time required by permutation tests. These methods rely on the assumption that the test statistic follows a particular distribution and the sample size is sufficiently large. When the sample size is not large enough, asymptotic results may not be valid, with the asymptotic p-values differing substantially from the exact p-values.

ii. Exploiting the Power of Large Number of Features An alternative strategy to overcome the drawback of the small number of samples in GEM experiments is to take advantage of the large number of features/genes [33]. In order to illustrate this idea we will consider the following: a GEM data set containing gene information about samples originating from two populations c1 and c2, and a filter algorithm to search for DEGs between c1 and c2.

12 iii. Multiple Hypothesis Testing Approach

The study of Dudoit et al. [34] was the first work describing the multiple hypothesis testing for GEM experiments in a statistical framework. In the context of DEGs discovery, multiple hypothesis testing is seen as simultaneously testing for each gene the null hypothesis of no association between the expression level and the responses or target features [34]. According to them, any test can result in two type of errors: false positive or Type I errors and false negative or Type II errors. Multiple hypothesis testing procedures aim to provide statistically significant results by controlling the incidence rate of these errors. In other words, provide a way of setting appropriate thresholds in declaring a result statistically significant. The most popular methods for multiple hypothesis testing focus on controlling Type I error rate. This is done by imposing a certain threshold for the Type I error rate and then applying a method to produce a list of rejected hypothesis until the error rate is less than or equal with the specified threshold.

p-value with Bonferroni correction is an improved version of the classical p-value and consists in increasing the statistical threshold for declaring a gene significant by dividing the desired significance with the number of statistical tests performed [35]. False discovery rate (FDR) is a recent alternative for significance testing and has been proposed as an extension of the concept of p-values [36]. The FDR is defined as FDR =[F/G] , where F is the number of false positive genes and G is the number of genes found as being significant. In order to overcome the situations where FDR is not defined (when G = 0), Storey [37] proposed a modified version of the FDR called positive false discovery rate (pFDR) defined as Pfdr= [E/F|G > 0].

A less accurate alternative to the FDR for significance testing is the family-wise error rate (FWER) which is defined as the probability of at least one truly insignificant feature to be called significant. q-value is an extension of FDR which has been proposed to answer the need of assigning a statistical significance score to each gene in

the same way that the p-value does [38]. The q-value is defined as being the minimum pFDR at which a test may be called significant. The reader should be aware that the q-value can be defined either in terms of the original statistics or in terms of the pvalues.

13 d) Ranking Methods for FS -Examples

In this section, we discuss and review ranking methods for FS by extending the taxonomy presented in Fig. 1.

14 i. Univariate Methods

According to [16], univariate methods for FS can be either parametric or nonparametric. Here, we provide a brief description of both groups.

15 a. Parametric methods

These methods rely on some more or less explicit assumption that the data are drawn from a given probability distribution. The scoring functions used to measure the difference in expression between groups of samples for each gene provide meaningful results only if this assumption holds. In particular, many researchers state that the t-test can be used to identify DEGs only if the data in each class are drawn from some normal distribution with mean and standard deviation.

16 b. Nonparametric methods

These methods assume by definition that the data are drawn from some unknown distribution. The scoring functions used to quantify the difference in expression between classes rely either on some estimates of the pdfs or on averaged ranks of genes or samples. Obviously, these methods have a higher generalization power but for most of them (especially those relying on estimates of the pdfs), the computational cost is higher. In [16], univariate nonparametric filter techniques are split in two groups: pure model-free methods and methods based on random permutation associated to parametric tests. Pure model free methods use nonparametric scoring functions to assign a relevance index to each gene and then the statistical relevance of that index is estimated in terms of either p-value, FDR or q-value. Methods based on random permutations associated with a parametric test take advantage on the large number of genes/features in order to find genes/features which present significant changes in expression. In a first instance, they make use of a parametric scoring function to assign a relevance index to each gene and then employ a nonparametric statistical significance test to check for DEGs. The nonparametric significance test consists in comparing the distribution of relevance indices of genes estimated in the previous step and the null distribution of the test statistic (or relevance index). The null distribution of the test statistic is usually estimated using a permutation test.

ii. Bivariate Ranking Methods Ranking pairs of genes according to their discrimination power between two or more conditions can be performed either using a "greedy strategy" or "all pair strategy." Greedy strategies. Methods in this group first rank all genes by individual ranking (using one of the criteria employed by univariate ranking methods); subsequently the highest scoring gene gi is paired with the gene gj that gives the highest gene pair score. After the first pair has been selected, the next highest ranked gene remaining gs is paired with the gene gr that maximizes the pair score, and so on. In [39], a greedy gene pair ranking method has been proposed where initially the t-test was employed to first rank genes individually while the pair score measures how well the pair in combination distinguishes between two populations. Concretely, the gene pair score is the t-test of the projected coordinates of each experiment on the diagonal linear discriminant (DLD) axis, using only these two genes. For further details we invite the reader to consult [39].

All pairs strategies. Unlike greedy pairs methods, all pairs strategies examine all possible gene pairs by computing the pair score for all pairs. The pairs are then ranked by pair score, and the gene ranking list is compiled by selecting non overlapping pairs, and selecting highest scoring pairs first. This method is computationally very expensive.

17 e) Filter Methods -Space Search Approach

The second direction to create filters for FS is to adopt an optimization strategy which will come up with the most informative and least redundant subset of features among the whole set. This strategy implies three main steps described as follows: 1. Define a cost function to optimize. 2. Use an optimization algorithm to find the subgroup of features which optimizes the cost function. 3. Validate the selected subset of genes.

18 Global Journal of Computer Science and Technology

Volume XIV Issue III Version I

19 Our Contribution

This work categorizes the algorithms into different categories to emphasize the data structure that drives the matching. We will give in this section some characteristics of standard clustering methods in relation to microarray

data analysis. Hierarchical clustering has been mainly used to find a partition of the samples more than of the genes because there are much less samples than genes so that, with genes, the resulting dendrogram is often difficult to interpret.

20 Algorithms Designed After 2000

In this section we survey the most classical micro array algorithms that have been designed after year 2000. In particular the algorithms based on comparisons and the algorithms based on micro array. Most of the comparison-based algorithms presented in the last ten years are obtained by improving or combining the ideas of previously published algorithms.

In the following we briefly review the state-of-the-art until 2014 and the main ideas and the algorithms to which the new solutions refer. a) During 2010 Leila Muresan et.al [40] developed an approach for the analysis of high-resolution microarray images. First, it consists of a single molecule detection step, based on undecimated wavelet transforms, and second, a spot identification step via spatial statistics approach (corresponding to the segmentation step in the classical microarray analysis). Proposed approach relies on two independent steps. First, present a wavelet-based method to detect single molecules in each subimage. Wavelet transform offers an attractive solution for the detection of small bright features, e.g., in astronomical images or in the case of microscopy, for the detection of subcellular structures. The detection is based on the property of the wavelet transform to concentrate the information in a few wavelet coefficients, and subsequently thresholding the pixels corresponding to the signal from background. Second, separate the detected molecules inside the spot of interest (the hybridization signal) from the unspecifically bound ones. This concentration estimation approaches based on spatial statistics. The first algorithm matches the empirical moments with the moments of a mixture of two Poisson distributions representing counts of molecules outside and inside the spot. The second algorithm separates spot-bound single molecules from dirt, based on nearest neighbor distances of all the detected peak locations, via an expectation-maximization (EM) approach. Since the surface was made antiadsorptive for target molecules, we can assume that the concentration of peaks outside the spot is lower than the concentration of the hybridized molecules inside the spot. The detection method was tested on simulated images with a concentration range of 0.001 to 0.5 molecules per square micrometer and signal-to-noise ratio (SNR) between 0.9 and 31.6. For SNR above 15, the false negatives relative error was below 15%. Separation of foreground/background is proved reliable, in case foreground density exceeds background by a factor of 2. The method has also been applied to real data from high-resolution microarray measurements.

Yoshinori Tamada et.al [41] presents a novel algorithm to estimate genome-wide gene networks consisting of more than 20 000 genes from gene expression data using nonparametric Bayesian networks. Due to the difficulty of learning Bayesian network structures, existing algorithms cannot be applied to more than a few thousand genes. Present algorithm overcomes this limitation by repeatedly estimating sub networks in parallel for genes selected by neighbor node sampling. Through numerical simulation, finally confirmed that proposed algorithm outperformed a heuristic algorithm in a shorter time. Proposed algorithm to microarray data from human umbilical vein endothelial cells (HUVECs) treated with siRNAs, to construct a human genome-wide gene network, which compared to a small gene network estimated for the genes extracted using a traditional bioinformatics method. The results showed that genome-wide gene network contains many features of the small network, as well as others that could not be captured during the small network estimation. The results also revealed master-regulator genes that are not in the small network but that control many of the genes in the small network. These analyses were impossible to realize without our proposed algorithm. Analysis of the result, we also constructed a gene network with 527 genes extracted. These 527 genes are selected based on the ordinal bioinformatics analysis with SAM (Significance Analysis of Microarrays) by applying it to another drug-response microarray data which were observed for HUVECs stimulated by anti-hyperlipidemia drug Fenofibrate. For this smaller gene network, performed the bootstrap method. The number of the bootstrap iterations is 1000. The final 527 gene network is generated by removing edges whose bootstrap probabilities are less than 0.5. Tianwei Yu et.al [42] proposes an imputation scheme based on nonlinear dependencies between genes. By simulations based on real microarray data, show that incorporating non-linear relationships could improve the accuracy of missing value imputation, both in terms of normalized root mean squared error and in terms of the preservation of the list of significant genes in statistical testing. In addition, studied the impact of artificial dependencies introduced by data normalization on the simulation results. Our results suggest that methods relying on global correlation structures may yield overly optimistic simulation results when the data has been subjected to row (gene) -wise mean removal. Six datasets were used in the simulation study. They included the B-cell lymphoma profiling data , the dataset of yeast transcriptome/translatome comparison, the NCI60 cell line gene expression data, and the GSE19119 dataset on Atlantic salmon. Two yeast cell cycle time series, the alpha factor dataset and the elutriation dataset, were used to probe the effect of data normalization on simulation results in imputation studies. Four popular imputation methods were used for comparison. They included the K-nearest neighbor (KNN) method, the Bayesian PCA (BPCA) method, the local least square (LLS) method, and the SVD method. Different percentages of missing (1%, 5%, 10%, 15% and 20%) were simulated.

Jianxing Feng et.al [43] propose a novel Graph Fragmentation Algorithm (GFA) for protein complex identification. Adapted from a classical maxflow algorithm for finding the (weighted) densest subgraphs, GFA first finds large (weighted) dense sub graphs in a protein-protein interaction network and then breaks each

such subgraph into fragments iteratively by weighting its nodes appropriately in terms of their corresponding log fold changes in the microarray data, until the fragment subgraphs are sufficiently small. Tests on three widely used protein-protein interaction datasets and comparisons with several latest methods for protein complex identification demonstrate the strong performance of proposed method in predicting novel protein complexes in terms of its specificity and efficiency. Given the high specificity (or precision) that method has achieved, finally conjecture that our prediction results imply more than 200 novel protein complexes. In this paper authors retrieved 51 sets of microarray gene expression data concerning yeast from the GEO database where the log fold changes of expression levels are provided. Each dataset contains multiple samples (or conditions). Totally, 824 samples are contained in the 51 datasets. Since the genes expressed in each sample are different and they could also be different from the genes contained in a PPI network, use a sample of the microarray data on a PPI network if it covers at least 90% of the genes in the network under consideration. For genes that have no expression data in a certain sample, treat their (log transformed) expression values as 0. Finally, chose (randomly) 500, 600, and 700 samples to be applied on the MIPS, DIP, and BioGRID PPI networks, respectively.

Jong Kyoung Kim et.al [44] develop a hybrid generative/discriminative model which enables us to make use of unlabeled sequences in the framework of discriminative motif discovery, leading to semisupervised discriminative motif discovery. Numerical experiments on yeast ChIP-chip data for discovering DNA motifs demonstrate that the best performance is obtained between the purely-generative and the purelydiscriminative and the semi-supervised learning improves the performance when labeled sequences are limited. This examined the yeast ChIP-chip data published to investigate the effect of ? on identifying TFBSs, and the benefit of semi-supervised learning for motif discovery. The data included the intergenic binding locations of yeast TFs which were profiled under various environmental conditions. For each TF under a particular condition, defined its original positive set to be probe sequences that are bound with P-value ? 0.001, where the binding P-value is evaluated according to relative intensities of spots on a microarray. To establish the importance of blending generative and discriminative approaches for discovering DNA motifs, examined the ability of DMOPSH to find true motifs by varying the size of the positive set with different values of ?. The top K sequences with smallest P values from the original positive set were chosen to define a positive set and the remaining sequences were defined to be unlabeled. Similarly, chose the 3K probe sequences with largest P-values for the negative set. We ran each experiment three times with different initializations and reported the means with ± 1 standard error.

Xin ZHAO et.al [45] Identifying significant differentially expressed genes of a disease can help understand the disease at the genomic level. A hierarchical statistical model named multi-class kernelimbedded Gaussian process (mKIGP) is developed under a Bayesian framework for a multi-class classification problem using microarray gene expression data. Specifically, based on a multinomial probit regression setting, an empirically adaptive algorithm with a cascading structure is designed to find appropriate featuring kernels, to discover potentially significant genes, and to make optimal tumor/cancer class predictions. A Gibbs sampler is adopted as the core of the algorithm to perform Bayesian inferences. A prescreening procedure is implemented to alleviate the computational complexity. The simulated examples show that mKIGP performed very close to the Bayesian bound and outperformed the referred state-of-the-art methods in a linear case, a non-linear case and a case with a mislabeled training sample. Its usability has great promises to problems that linear model based methods become unsatisfactory. The mKIGP was also applied to four published real microarray datasets and it was very effective for identifying significant differentially expressed genes and predicting classes in all of these datasets. This work builds a unified kernel-induced supervised learning model under a hierarchical Bayesian framework to analyze microarray gene expression patterns. With a multinomial probit regression setting, the introduction of latent variables, and a prescreening procedure, the mKIGP model was developed for a multi-class classification problem. An algorithm with a cascading structure was proposed to solve this problem and a Gibbs sampler was built as the mechanical core to do the Bayesian inference. Given a kernel type (such as a Gaussian kernel) with the training data as input, the fitted parameter(s) of the kernel and a Alfredo Benso et.al [46] presents a new cDNA microarray data classification algorithm based on graph theory and able to overcome most of the limitations of known classification methodologies. The classifier works by analyzing gene expression data organized in an innovative data structure based on graphs, where vertices correspond to genes and edges to gene expression relationships. One of the main contributions of the classifier stems in the ability of combining in a single algorithm high accuracy in the classification process together with the ability of detecting samples not belonging to any of the trained classes, thus drastically reducing the number of false positive classification outcomes. To validate the efficiency of the proposed approach, the paper presents an experimental comparison between the GEG-based classifier and several generic state-of-the-art multi-class and one-class classification methods on a set of cDNA microarray experiments for fifteen well known and documented diseases. Experimental results show that the GEG-based classifier is able to reach the same performances reached by multi-class classifiers when dealing with samples belonging to the considered class library, while it outperforms one-class classifiers in the ability of detecting samples not belonging to any of the trained classes. To demonstrate the novelty of the proposed approach, the authors present an experimental performance comparison between the proposed classifier and several state-of-the-art classification algorithms.

Yu-Cheng Liu et.al [47] proposed a temporal dependency association rule mining method named 3D-TDAR-Mine for three-dimensional analyzing microarray datasets. The mined rules can represent the regulatedrelations between genes. Through experimental evaluation, our proposed method can discover the meaningful temporal

dependent association rules that are really useful for biologists. In this paper, define the Frequently Coherent Pattern as gene expressions reaction. Furthermore, Coherent Pattern is focus on one gene in one continuous time segment to compute the gene expression value similarity between any two samples. Hence, user can depend on their required feature of Coherent Pattern to choice the similarity measure method. If user wants to discover the Coherent Pattern between two samples that have identical shape in gene expression value series. They can use the PCC (Pearson correlation coefficient). But, in the real life reaction, it not always has identical shape. The expression value series between samples also have Shifting, Scale and Trend relation. Therefore, it proposed the TS3 similarity measurement to estimate the Coherent Pattern that considers the Shifting, Scale and Trend factors.

Hong-Dong Li et.al [48] presented a new approach, called Margin Influence Analysis (MIA), designed to work with support vector machines (SVM) for selecting informative genes. The rationale for performing margin influence analysis lies in the fact that the margin of support vector machines is an important factor which underlies the generalization performance of SVM models. Briefly, MIA could reveal genes which have statistically significant influence on the margin by using Mann-Whitney U test. The reason for using the Mann-Whitney U test rather than two-sample t test is that Mann-Whitney U test is a nonparametric test method without any distribution-related assumptions and is also a robust method. Using two publicly available cancerous microarray datasets, it is demonstrated that MIA could typically select a small number of margininfluencing genes and further achieves comparable classification accuracy compared to those reported in the literature. The method reported here, named margin influence analysis (MIA), is quite different from previous work. it is developed based model population analysis (MPA), which is a general framework for designing bioinformatics algorithms. The MIA method is currently proposed by strictly implementing the idea of MPA and specially designed for variable selection of support vector machines. It works by first computing a large number of SVM classifiers using randomly sampled variables. Each model is associated with a margin. Then the nonparametric Mann-Whitney U test is employed to calculate a p-value for each variable, aiming at uncovering the variable that can increase the margin of a SVM model significantly. The rationale behind MIA is that the performance of SVM depends heavily on the margin of the classifier. As is known, the larger the margin is, the better the prediction performance will be. For this reason, variables that can increase the margin of SVM classifiers should be regarded as informative variables or possible biomarker candidates. On the whole, the main contributions of MIA are two folds. Firstly, it is originally from model population analysis which helps statistically establish variable rank by analyzing the empirical distributions of margins of related SVM classifiers. Secondly, it explicitly utilizes the influence of each variable on the margin for variable selection. The results for two publicly available microarray datasets show that MIA typically selects a small number of margin-influencing informative genes, leading to comparable classification accuracy compared to that reported in the literature. The distinguished features and outstanding performance may make MIA a good alternative for gene selection of high dimensional microarray data.

Yang Chen, and Jinglu Hu [49] presents a constructive heuristic algorithm, featuring an accurate reconstruction guided by a set of well-defined criteria and rules. Instead of directly reconstructing the original sequence, the new algorithm first builds several accurate short fragments, which are then carefully assembled into a whole sequence. The eSBH algorithm can achieve relatively high accuracy in reconstruction from a large spectrum, than other constructive heuristics and some meta heuristics, especially for real DNA sequences in the benchmark instance sets. The experiments on benchmark instance sets demonstrate that the proposed method can reconstruct long DNA sequences with higher accuracy than current approaches in the literature.

Jong Kyoung Kim and Seungjin Choi [50] develop a hybrid generative/discriminative model which enables us to make use of unlabeled sequences in the framework of discriminative motif discovery, leading to semi-supervised discriminative motif discovery. Here the authors, assume that each subsequence is generated by a finite mixture model with two components corresponding to motif and background models. While this generative approach is useful for finding overrepresented motifs in a given target set of sequences, our simple generative model has a limitation to capture the nature of labeled sequences. Numerical experiments on yeast ChIP-chip data for discovering DNA motifs demonstrate that the best performance is obtained between the purely-generative and the purely discriminative and the semi-supervised learning improves the performance when labeled sequences are limited.

Gene selection methods aim at determining biologically relevant subsets of genes in DNA microarray experiments. However, their assessment and validation represent a major difficulty since the subset of biologically relevant genes is usually unknown. To solve this problem a novel procedure for generating biologically plausible synthetic gene expression data is proposed by Marco Muselli et.al [51]. It is based on a proper mathematical model representing gene expression signatures and expression profiles through Boolean threshold functions. Here authors showed from a statistical standpoint that we may obtain artificial data reasonably close to real gene expression data. As a consequence, we may generate biologically plausible virtual gene expression data that may be easily used to evaluate gene selection methods, since, in this case, know in advance the set of ”relevant” genes. On the basis of the mathematical model, we proposed an algorithmic procedure to generate artificial gene expression data, and we showed how to apply the algorithm to the analysis of the performance of statistical and machine learning based gene selection methods. The results show that the proposed procedure can be successfully adopted to analyze the quality of statistical and machine learning-based gene selection algorithms.

Leila Muresan et.al [52] developed an approach for the analysis of high-resolution microarray images. First, it consists of a single molecule detection step, based on undecimated wavelet transforms, and second, a spot identification step via spatial statistics approach (corresponding to the segmentation step in the classical microarray analysis). The detection method was tested on simulated images with a concentration range of 0.001 to 0.5 molecules per square icrometer and signalto-noise ratio (SNR) between 0.9 and 31.6. For SNR above 15, the false negatives relative error was below 15%. Separation of foreground/background is proved reliable, in case foreground density exceeds background by a factor of 2. The method has also been applied to real data from high-resolution microarray measurements.

Banu Dost et.al [53] introduce here a new method, TCLUST, for clustering large, genome-scale data sets. The algorithm is based on measures of coconnectedness to identify dense subgraphs present in the data. The authors have applied this method to a large reference gene expression data set, and showed that the resulting clusters show strong enrichment in known biological pathways. Although TCLUST has been shown to perform as good as or better than existing methodologies, as with any methodology, certain caveats must be noted. A possible shortcoming might be that once two vertices end up in different clusters, they are never reconnected. On the one hand, this makes the algorithm converge faster, on the other hand, it might lead to some loss of sensitivity for higher errorrates. In principle, this could be adjusted, by applying the tcg thresholds more judiciously, gaining some FN edges at the cost of some FP edges, and increasing the number of iterations.

Giorgio Valentini [54] proposed a new hierarchical strategy, inspired by the true path rule, for gene function prediction extended to the overall functional taxonomy of genes. TPR-w ensembles significantly outperform both the basic TPR and Topdown ensembles in the genome and ontology wide prediction of gene functions in S. cerevisiae. The analysis of the experimental results and a theoretical investigation of the flow of information that traverses the hierarchical ensemble show the reasons why TPR-w are well-suited to the prediction of gene functions, and suggest new research lines for the development of new hierarchy-aware gene function prediction methods. The overall results show that using a single source of evidence we can obtain a high precision and recall for specific trees of the FunCat forest.

The prevalence of chronic diseases is increasing at an alarming rate. Among them the incidence of Type-2 Diabetes is rapidly increasing globally. Although genetics could play an important role in the higher prevalence of this disease, it is not clear how genetic factors interact with environmental and dietary factors to increase their incidence. In the current study, Gene Expression Analysis was performed by the authors [55,56] Mohak Shah and Jacques Corbeil [57] propose a general theoretical framework for analyzing differentially expressed genes and behavior patterns from two homogenous short time-course data. The framework generalizes the recently proposed Hilbert-Schmidt Independence Criterion (HSIC)-based framework adapting it to the time-series scenario by utilizing tensor analysis for data transformation. The proposed framework is effective in yielding criteria that can identify both the differentially expressed genes and time-course patterns of interest between two time-series experiments without requiring to explicitly cluster the data. The parameters used in the framework give the user explicit control on the type of analysis to be performed. For instance, identifying genes pertaining to the time-course patterns of interest can be done simply by choosing and adjusting an apt weight vector and does not require clustering all the genes in predefined profile sets unlike traditional clustering-based methods. Moreover, the criterion is a generalization of the integer fold-change-based methods. It is more sensitive in discerning relatively small differential expressions. Hence, it enables the user to identify the cases when genes undergo less than twofold change but are or can potentially be biologically important in our understanding of a certain treatment or condition. The results, obtained by applying the proposed framework with a linear kernel formulation, on various data sets are found to be both biologically meaningful and consistent with published studies.

Xin Zhao and Leo Wang-Kit Cheung [58] developed a hierarchical statistical model named multiclass kernel-imbedded Gaussian process (mKIGP) under a Bayesian framework for a multiclass classification problem using microarray gene expression data. Specifically, based on a multinomial probit regression setting, an empirically adaptive algorithm with a cascading structure is designed to find appropriate featuring kernels, to discover potentially significant genes, and to make optimal tumor/cancer class predictions. A Gibbs sampler is adopted as the core of the algorithm to perform Bayesian inferences. A prescreening procedure is implemented to alleviate the computational complexity. The simulated examples show that mKIGP performed very close to the Bayesian bound and outperformed the referred state-of-the-art methods in a linear case, a nonlinear case, and a case with a mislabeled training sample. Its usability has great promises to problems that linear-model-based methods become unsatisfactory. The mKIGP was also applied to four published real microarray data sets and it was very effective for identifying significant differentially expressed genes and predicting classes in all of these data sets. Comparing to a regular SVM, the most popular kernel-induced learning method, the mKIGP has three key advantages. First, the probabilistic class prediction by the mKIGP could be insightful for borderline cases in real applications. Second, the mKIGP method has implemented specific procedure for tuning the kernel parameter(s) (such as the width parameter of a GK) and the model parameters (such as the variance of the noise term). Tuning parameters have always been one of the key issues for nonlinear parametric learning methods. As the gene selection procedure is imbedded into the learner, the mKIGP is also more consistent in identifying significant genes when comparing to regular UR or RFE method with a cross-validation procedure. In the simulated studies, The authors showed that the mKIGP/GK significantly outperformed its SVM or PLR counterparts with either RFE or UR as gene

selection strategy in the nonlinear example and in the example with a mislabeled training sample. We also demonstrated that mKIGP functioned much better in a multiclass classification problem when comparing to another established Gaussian-Processesbased gene selection method, GP_ARD, for the real data sets. Third, the mKIGP method can provide more useful information, such as the posterior PDF of the parameters, for further statistical analysis and inference.

Argiris Sakellariou, Despina Sanoudou, and George Spyrou [59] investigate the minimum required subsets of genes, which best classify neuromuscular disease data. For this purpose, we implemented a methodology pipeline that facilitated the use of multiple feature selection methods and subsequent performance of data classification. Five feature selection methods on datasets from ten different neuromuscular diseases were utilized. Our findings reveal subsets of very small number of genes, which can successfully classify normal/disease samples. Interestingly, we observe that similar classification results may be obtained from different subsets of genes. The proposed methodology can expedite the identification of small gene subsets with high-classification accuracy that could ultimately be used in the genetics clinics for diagnostic, prognostic, and pharmacogenomic purposes. This study reveals that using appropriate bio-informatical tools, researchers can identify subsets with very small number of genes, which achieve high-classification results, as demonstrated for the neuromuscular disease datasets analyzed herein. Toward this goal, we applied five different feature selection methods on neuromuscular disease data (rare conditions for which only limited numbers of samples and microarray datasets are available), and investigated the minimum number of gene probes for highly accurate patient/sample classification.(D D D D D D D D) Year 2014 c

Microarray analysis is a method for analyzing expression levels of multiple genes at once. This method is especially suitable for identifying and classifying genes whose expression level differs in two samples. The present work focuses [60,61] on identifying and classifying genes that cause type-II diabetes with two different samples, one with parental history and other without parental history. Mahalanobis Distance, Minimum Co-variance Determinant are the statistical methods used for identifying multivariate and univariate outliers for the identified inflammatory genes, the functional classification is performed by using Gene Ontology and pathway analysis. It is observed that 38 differentially expressed genes were identified out of 39400 genes tested between diabetes with and without parental history.

21 c) During 2012

Pradipta **??**aji [62] proposed supervised attribute clustering algorithm is based on measuring the similarity between attributes using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised and unsupervised gene clustering and gene selection algorithms based on the class separability index and the predictive accuracy of naive bayes classifier, Knearest neighbor rule, and support vector machine on three cancer and two arthritis microarray data sets. The biological significance of the generated clusters is interpreted using the gene ontology. An important finding is that the proposed supervised attribute clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability. The main contribution of this paper is threefold, namely, 1. Defining a new quantitative measure, based on mutual information, to calculate the similarity between two genes, which incorporates the information of sample categories or class labels.

22 Development of a new supervised attribute

clustering algorithm to find coregulated clusters of genes whose collective expression is strongly associated with the sample categories. 3. Comparing the performance of the proposed method and some existing methods using the class separability index and predictive accuracy of support vector machine, K-nearest neighbor rule, and naive bayes classifier. For five microarray data, significantly better results are found for the proposed method compared to existing methods, irrespective of the classifiers used. All the results reported in this paper demonstrate the feasibility and effectiveness of the proposed method. It is capable of identifying coregulated clusters of genes whose average expression is strongly associated with the sample categories. The identified gene clusters may contribute to revealing underlying class structures, providing a useful tool for the exploratory analysis of biological data.

Ola ElBakry, M. Omair Ahmad, and M.N.S. Swamy [63] presents a general statistical method for detecting changes in microarray expression over time within a single biological group and is based on repeated measures (RM) ANOVA. In this method, unlike the classical F-statistic, statistical significance is determined taking into account the time dependency of the microarray data. A correction factor for this RM Fstatistic is introduced leading to a higher sensitivity as well as high specificity. We investigate the two approaches that exist in the literature for calculating the p-values using resampling techniques of gene-wise pvalues and pooled p-values. It is shown that the pooled p-values method compared to the method of the genewise p-values is more powerful, and computationally less expensive, and hence is applied along with the introduced correction factor to various synthetic data sets and a real data set. These results show that the proposed technique outperforms the current methods. The real data set results are consistent with the existing knowledge concerning the presence of the genes. The algorithms presented are implemented in R and are freely available upon request. In this work, RM Fstatistic, which considers the dependency of measurements across the time course, has been employed for gene

identification. The p-values have been computed using both the gene-wise and pooled pvalues methods. Since the gene-wise p-values procedure is based on the number of permutations for each gene, this number has to be large to achieve the granularity of the pooled p-values. The synthetic data results have shown that the pooled p-values procedure is able to detect more true positives than the gene-wise p-values method does, and hence, is preferred for microarray data analysis.

Alok Sharma, Seiya Imoto, and Satoru Miyano [64] propose a feature selection algorithm in gene expression data analysis of sample classifications. The proposed algorithm first divides genes into subsets, the sizes of which are relatively small (roughly of size h), then selects informative smaller subsets of genes (of size r $<$ h) from a subset and merges the chosen genes with another gene subset (of size r) to update the gene subset. It repeats this process until all subsets are merged into one informative subset. It illustrates the effectiveness of the proposed algorithm by analyzing three distinct gene expression data sets. The proposed algorithm explores this phenomenon and provides a way to investigate important genes. It is observed that the algorithm finds a small gene subset that provides high classification accuracy on several DNA microarray gene expression data sets. These subsets contain top-r genes. The small number of (r) genes would help to This method shows promising classification accuracy for all the test data sets. We also show the relevance of the selected genes in terms of their biological functions.

Andrew Janowczyk et.al [65] presents a system for accurately quantifying the presence and extent of stain on account of a vascular biomarker on tissue microarrays. It demonstrate their flexible, robust, accurate, and high-throughput minimally supervised segmentation algorithm, termed hierarchical normalized cuts (HNCuts) for the specific problem of quantifying extent of vascular staining on ovarian cancer tissue microarrays. The high-throughput aspect of HNCut is driven by the use of a hierarchically represented data structure that allows us to merge two powerful image segmentation algorithms-a frequency weighted mean shift and the normalized cuts algorithm. HNCuts rapidly traverses a hierarchical pyramid, generated from the input image at various color resolutions, enabling the rapid analysis of large images (e.g., a 1500 $\times$ 1500 sized image under 6 s on a standard 2.8-GHz desktop PC). HNCut is easily generalizable to other problem domains and only requires specification of a few representative pixels (swatch) from the object of interest in order to segment the target class. Across ten runs, the HNCut algorithm was found to have average true positive, false positive, and false negative rates (on a per pixel basis) of 82%, 34%, and 18%, in terms of overlap, when evaluated with respect to a pathologist annotated ground truth of the target region of interest. By comparison, a popular supervised classifier (probabilistic boosting trees) was only able to marginally improve on the true positive and false negative rates (84% and 14%) at the expense of a higher false positive rate (73%), with an additional computation time of 62% compared to HNCut.

Blaise Hanczar and Avner Bar-Hen [66] propose a new measure of classifier performance that takes account of the uncertainty of the error. We represent the available knowledge about the costs by a distribution function defined on the ratio of the costs. The performance of a classifier is therefore computed over the set of all possible costs weighted by their probability distribution. This method is tested on both artificial and real microarray data sets. The costs are represented by a distribution function defined on the ratio of the costs. Seven new classification cost functions have been used in experiments based on both artificial and real data sets. These experiments showed that the selection of the best classifier is very depending on the used cost functions. In many cases, the best classifier can be identified by our new measure whereas the classic error measures fail.

Pradipta Maji and Chandra Das [67] proposed a gene clustering algorithm is to group genes from microarray data. It directly incorporates the information of sample categories in the grouping process for finding groups of co-regulated genes with strong association to the sample categories, yielding a supervised gene clustering algorithm. The average expression of the genes from each cluster acts as its representative. Some significant representatives are taken to form the reduced feature set to build the classifiers for cancer classification. The mutual information is used to compute both gene-gene redundancy and gene-class relevance. The performance of the proposed method, along with a comparison with existing methods, is studied on six cancer microarray data sets using the predictive accuracy of naive Bayes classifier, K-nearest neighbor rule, and support vector machine. An important finding is that the proposed algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

23 d) During 2013 & 2014

Zidong Wang et.al [68] investigates the uncertainty quantification and state estimation issues. The polytopic uncertainty model (PUM) is exploited for describing the GRNs where the parameter uncertainties are constrained in a convex polytope domain. To cope with the high-dimension problem for GRN models, the principal component plane (PCP) algorithm is proposed to construct a pruned polytope in order to use as less vertices as possible to maintain the essential information from original polytope. The so-called system equivalence transformation is developed to transform the original system into a simpler canonical form and therefore facilitate the subsequent state estimation problem. For the state estimation problem, a robust stability condition is incorporated with guaranteed performance via the semi-definite programme method, and then a new sufficient condition is derived for the desired estimators with several free slack matrices. Such a condition is vertex-dependent and therefore possesses less conservatism. It is shown, via simulation from real-world microarray time-series data, that the

designed estimators have strong capability of dealing with modeling and estimation problems for short but high-dimensional gene expression time series.

Anirban Mukhopadhyay [69] proposed a novel interactive genetic algorithm-based multi objective approach that simultaneously finds the clustering solution as well as evolves the set of validity measures that are to be optimized simultaneously. The proposed method interactively takes the input from the human decision maker (DM) during execution and adaptively learns from that input to obtain the final set of validity measures along with the final clustering result. The algorithm is applied for clustering real-life benchmark gene expression datasets and its performance is compared with that of several other existing clustering algorithms to demonstrate its effectiveness. The results indicate that the proposed method outperforms the other existing algorithms for all the datasets considered here. The performance of IMOC has been demonstrated for two real-life gene expression datasets and compared with that of several other existing clustering algorithms. Results indicate that IMOC produces more biologically significant clusters compared to the other algorithms and the better result provided by IMOC is statistically significant.

Ujjwal Maulik et.al [70] proposed a novel approach to combine feature (gene) selection and transductive support vector machine (TSVM). We demonstrated that 1) potential gene markers could be identified and 2) TSVMs improved prediction accuracy as compared to the standard inductive SVMs (ISVMs). A forward greedy search algorithm based on consistency and a statistic called signal-to-noise ratio were employed to obtain the potential gene markers. The selected genes of the microarray data were then exploited to design the TSVM. Experimental results confirm the effectiveness of the proposed technique compared to the ISVM and low-density separation method in the area of semi supervised cancer classification as well as gene-marker identification.

Gui-Fang Shao [71] presented a fully automatic gridding technique to break through the limitation of traditional mathematical morphology gridding methods. First, a preprocessing algorithm was applied for noise reduction. Subsequently, the optimal threshold was gained by using the improved Otsu method to actually locate each spot. In order to diminish the error, the original gridding result was optimized according to the heuristic techniques by estimating the distribution of the spots. Intensive experiments on six different data sets indicate that our method is superior to the traditional morphology one and is robust in the presence of noise.

Xiaoxiao Xu [72] analyze the statistical performance of these arrays in imaging targets at typical low signal-to-noise ratio (SNR) levels. We compute the Ziv-Zakai bound (ZZB) on the errors in estimating the unknown parameters, including the target concentrations. We find the SNR level below which the ZZB provides a more accurate prediction of the error than the posterior Cramér-Rao bound (PCRB), through numerical examples. We further apply the ZZB to select the optimal design parameters of the microsphere array device and investigate the effects of the experimental variables such as microscope point-spread function. An imaging experiment on microspheres with protein targets verifies the optimal design parameters using the ZZB.

Pablo A. Jaskowiak [73] investigate the choice of proximity measures for the clustering of microarray data by evaluating the performance of 16 proximity measures in 52 data sets from time course and cancer experiments. This method considered six correlation coefficients, four "classical" distances, and six proximity measures specifically proposed for the clustering of gene time-course data. Given their differences, we evaluated proximity measures separately for cancer and time-course experiments. Apart from the comparison of proximity measures, we introduced a set of 17 timecourse benchmark data along with a new methodology (IBSA) to evaluate distances for the clustering of genes. Both data sets and methodology can be used in future research to evaluate the effectiveness of new proximity measures in this particular scenario. IBSA can be employed to evaluate proximity measures regarding any gene clustering application, i.e., it is not restricted to gene time-course data, the scenario addressed here. Results support that measures rarely employed in the gene expression literature can provide better results than commonly employed ones, such as Pearson, Spearman, and euclidean distance. Given that different measures stood out for time course and cancer data evaluations, their choice should be specific to each scenario. To evaluate measures on time-course data, we preprocessed and compiled 17 data sets from the microarray literature in a benchmark along with a new methodology, called Intrinsic Biological Separation Ability (IBSA). Both can be employed in future research to assess the effectiveness of new measures for gene time-course data.

Cosmin Lazar [74] propose GENESHIFT, a new nonparametric batch effect removal method based on two key elements from statistics: empirical density estimation and the inner product as a distance measure between two probability density functions; second we introduce a new validation index of batch effect removal methods based on the observation that samples from two independent studies drawn from a same population should exhibit similar probability density functions. This evaluated and compared the GENESHIFT method with four other state-of-the-art methods for batch effect removal: Batch-mean centering, empirical Bayes or COMBAT, distance-weighted discrimination, and crossplatform normalization. Several validation indices providing complementary information about the efficiency of batch effect removal methods have been employed in our validation framework. The results show that none of the methods clearly outperforms the others. More than that, most of the methods used for comparison perform very well with respect to some validation indices while performing very poor with respect to others. GENESHIFT exhibits robust performances and its average rank is the highest among the average ranks of all methods used for comparison.

Telmo Amaral [75] presents a computational pipeline for automatically classifying and scoring breast cancer TMA spots that have been subjected to nuclear immunostaining. Spots are classified based on a bag of visual

words approach. Immunohistochemical scoring is performed by computing spot features reflecting the proportion of epithelial nuclei that are stained and the strength of that staining. These are then mapped onto an ordinal scale used by pathologists. Multilayer perceptron classifiers are compared with latent topic models and support vector machines for spot classification and with Gaussian process ordinal regression and linear models for scoring. Intra-observer variation is also reported. The use of posterior entropy to identify uncertain cases is demonstrated. Evaluation is performed using TMA images stained for progesterone receptor.

Wenjie You et.al [76] focuses on extracting the potential structure hidden in high-dimensional multi category microarray data, and interpreting and understanding the results provided by the potential structure information. First, we propose using PLSbased recursive feature elimination (PLSRFE) in multi category problems. Then, we perform feature importance analysis based on PLSRFE for highdimensional microarray data to determine the information feature (biomarkers) subset, which relates to the studied tumor subtypes problem. Finally, PLS-based supervised feature extraction is conducted on the selected specific genes subset to extract comprehensive features that best reflect the nature of classification to have a discriminating ability. The proposed algorithm is compared with several state-ofthe-art methods using multiple high-dimensional multi category microarray datasets. Our comparison is performed in terms of recognition accuracy, relevance, and redundancy. Experimental results show that the algorithm proposed by us can improve the recognition rate and computational efficiency. Furthermore, mining potential structure information improves the interpretability and understandability of recognition results. The proposed algorithm can be effectively applied to microarray data analysis for the discovery of gene co-expression and co-regulation.

24 IV.

25 Conclusions

Micro array is a ubiquitous problem that arises in a wide range of applications in computing, to full fill this we need efficient techniques. In this study we concentrate on micro array data and this article gave an overview of micro array models as well as programming tools. Micro array classification will always be a challenge for programmers. Higher-level programming models and appropriate programming tools only facilitate the process but do not make it a simple task. In this we say that, this study will help the researchers to develop the better techniques in the field of microarray. [1] [2]

1

Figure 1: Figure 1 :

[1] © 2014 Global Journals Inc. (US)

[2] © 2014 Global Journals Inc. (US)Gene Expression Analysis Methods on Microarray Data -A Review

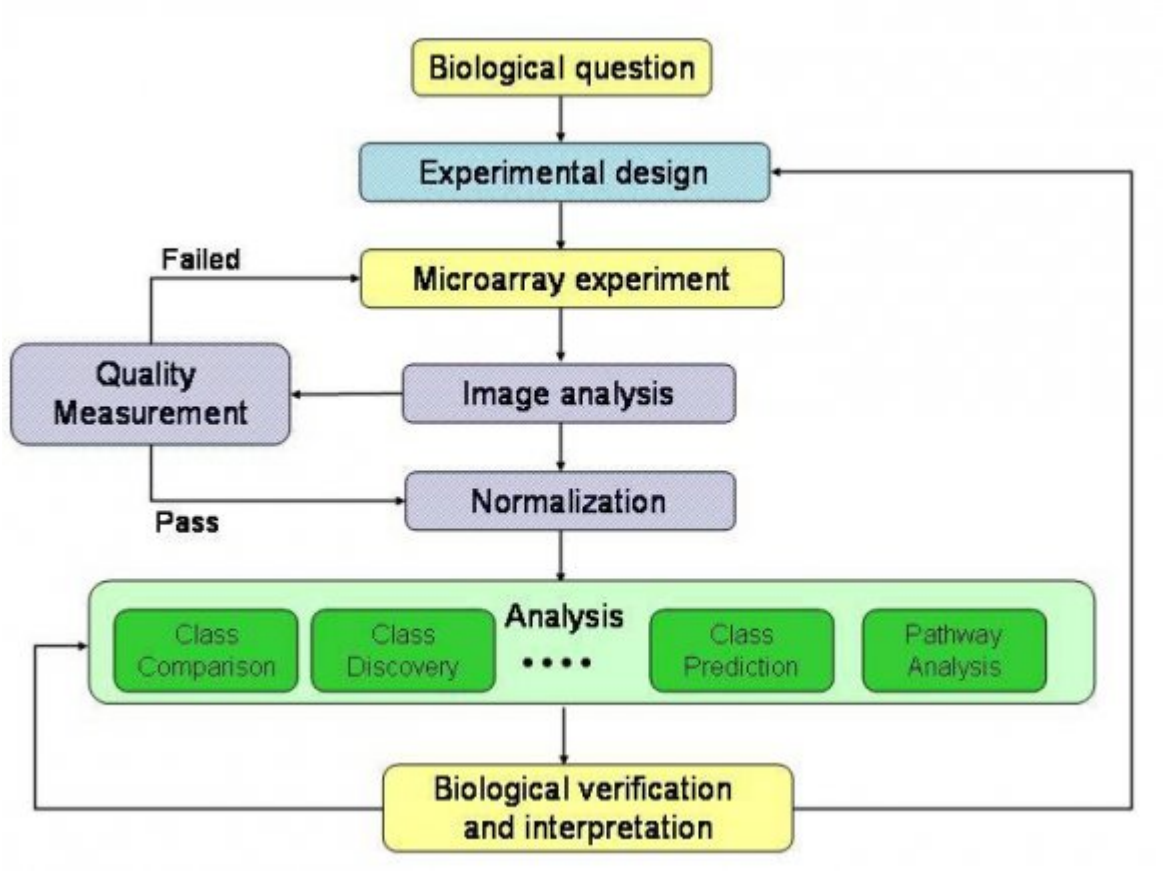

Figure 2:

1

AUC	??	=	??	??????	??	n 0 Number of individual values of gene x [29]	
	??????		0				
	= ?		??=1				
ABCR??	= ????????	= ?	??		??????? ??	?	
			0			??	
			??=1			??	
						||	

BWSS	?? = ???? = ? ? (?? ?? =??)(??? ?? ????)2 ?? ?? ? ? (?? ?? =??)(?? ?? ??? ??
TNoM	S=TNoM= min d,t Err(d,t|x,c) [31]

Figure 3: Table 1 :

[Cha] , S.-H Cha . *Comprehensive Survey on*

[Yang ()] , P Yang . *A Review of Ensemble Methods in Bioinformatics* 2010. 5 (4) p. . (Current Bioinformatics)

[Benso et al. ()] 'A cDNA Microarray Gene Expression Data Classifier for Clinical Diagnostics based on Graph Theory'. Alfredo Benso , Stefano Di Ieee Senior Member , Ieee Carlo , Gianfranco Member , Politano . *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 2010.

[Witten and Tibshirani ()] *A Comparison of Fold-Change and the t-Statistic for Microarray Data Analysis*, D Witten , R Tibshirani . 2007. Stanford Univ. (technical report)

[Storey ()] 'A Direct Approach to False Discovery Rates'. J D Storey . *J. Royal Statistics Soc.: Series B* 2002. 64 (3) p. .

[Shah and Corbeil ()] 'A general framework for analyzing data from two short timeseries microarray experiments'. Mohak Shah , Jacques Corbeil . *Computational Biology and Bioinformatics* 2011. 8 (1) p. . (IEEE/ACM Transactions on)

[Museli et al. ()] 'A mathematical model for the validation of gene selection methods'. Marco Museli , Member , Alberto Ieee , Marco Bertoni , Alessandro Frasca , Francesca Beghini , Giorgio Ruffino , Valentini . *IEEE ACM TRANS. ON COMP. BIOL. AND BIOINFORMATICS* 2010.

[Feng et al. ()] 'A Max-Flow Based Approach to the Identification of Protein Complexes Using Protein Interaction and Microarray Data'. Jianxing Feng , Rui Jiang , Tao Jiang . *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 2010.

[Pan et al. ()] 'A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data'. W Pan , J Lin , C T Le . *Functional and Integrative Genomics* 2003. 3 (3) p. .

[Hanczar and Bar-Hen ()] 'A new measure of classifier performance for gene expression data'. Blaise Hanczar , Avner Bar-Hen . *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2012. 9 (5) p. .

[Liu et al. ()] *A Novel Method for Mining Temporally Dependent Association Rules in Three-Dimensional Microarray Datasets*, Yu-Cheng Liu , Chao-Hui Lee , Wei-Chung Chen , J W Shin , Hui-Huang Hsu , Vincent S Tseng . 2010. IEEE.

[Saeys et al. ()] 'A Review of Feature Selection Techniques in Bioinformatics'. Y Saeys , I Inza , P Larran? . *Bioinformatics* 2007. 23 (19) p. .

[Sharma et al. ()] 'A top-r feature selection algorithm for microarray gene expression data'. Alok Sharma , Seiya Imoto , Satoru Miyano . *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2012. 9 (3) p. .

[Chen and Hu ()] 'Accurate Reconstruction for DNA Sequencing by Hybridization Based on A Constructive Heuristic'. Yang Chen , Jinglu Hu . *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 2010.

[Thomas ()] 'An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles'. J G Thomas . *Genome Research* 2001. 11 (7) p. .

[Liu et al. ()] 'An Entropy-Based Gene Selection Method for Cancer Classification Using Microarray Data'. X Liu , A Krishnan , A Mondry . *BMC Bioinformatics* 2005. 6.

[Mukhopadhyay et al. ()] 'An interactive approach to multiobjective clustering of gene expression patterns'. Anirban Mukhopadhyay , Ujjwal Maulik , Sanghamitra Bandyopadhyay . *Biomedical Engineering* 2013. 60 (1) p. . (IEEE Transactions on)

[Guyon ()] 'An Introduction to Variable and Feature Selection'. I Guyon . *J. Machine Learning Research* 2003. 3 p. .

[Blekas et al. ()] *an unsupervised artifact correction approach for the analysis of dna microarray images*, K Blekas , P Nikolas , Galatsanos , Georgiou . 2003. IEEE.

[Dietmar and Moeller ()] 'Business Objects as Part of a Preprocessing based Micro Array Data Analysis'. P F Dietmar , Moeller . *IEEE conference on EIT* 2005.

[Amaral et al. ()] *Classification and Immunohistochemical Scoring of Breast Tissue Microarray Spots*, Amaral , Stephen J Telmo , Katherine Mckenna , Alastair Robertson , Thompson . 2013. p. .

[Cohen ()] J Cohen . *The Earth is Round ($p < .05$)*, 1994. 38 p. .

[Dudoit et al. ()] 'Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data'. S Dudoit , J Fridlyand , T P Speed . *J. Am. Statistical Assoc* 2002. 97 (457) p. .

[Benjamini and Hochberg ()] 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. Y Benjamini , Y Hochberg . *J. Royal Statistical Soc. Series B (Methodological)* 1995. 57 (1) p. .

[Xiong et al. ()] 'Data-Dependent Kernel Machines for Microarray Data Classification'. Huilin Xiong , Ya Zhang , Xue-Wen Chen . *IEEE/ACM transactions on computational biology and bioinformatics* 2007. 4 (4) .

[Yan ()] 'Detecting Differentially Expressed Genes by Relative Entropy'. X Yan . *J. Theoretical Biology* 2005. 234 (3) p. .

[Sekhar et al. ()] 'Differential Gene Expression Analysis for Diabetes with and without parental history'. Chandra Sekhar , V Allam Appa Rao , P. Srinivasa Rao . *InComputer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, 2010. IEEE. 9 p. .

[Teng et al. ()] 'Dimension Reduction of Microarray Data Based on Local Tangent Space Alignment'. Li Teng , Hongyu Li , Xuping Fu , Wenbin Chen , I-Fan Shen . *IEEE* 2005.

[Distance/Similarity Measures Between Probability Density Functions ()] *Distance/Similarity Measures Between Probability Density Functions*, 2007. 1 p. .

[Tamada et al. ()] 'Estimating Genome-Wide Gene Networks Using Nonparametric Bayesian Network Models on Massively Parallel Computers'. Yoshinori Tamada , Seiya Imoto , Hiromitsu Araki , Masao Nagasaki , Cristin Print , D Stephen Charnock-Jones , Satoru Miyano . *IEEE TRANSACTIONS ON COMPUTA-TIONAL BIOLOGY AND BIOINFORMA-TICS* 2010.

[Michal et al. ()] 'Finding a Common Motif of RNA Sequences Using Genetic Programming: The GeRNAMo System'. Shahar Michal , Tor Ivry , Omer Schalit-Cohen , Moshe Sipper , Danny Barash . *IEEE/ACM transactions on computational biology and bioinformatics* 2007. 4 (4) .

[Vasamsetty et al. ()] 'Gene Expression Analysis for Type-2 Diabetes Mellitus–A Study on Diabetes With And Without Parental History'. Chandra Vasamsetty , Sekhar , Allam Appa Srinivasa Rao Peri , K Rao , Chinta Srinivas , Someswararao . *Journal of Theoretical & Applied Information Technology* 2011. 27 (1) .

[Vasamsetty et al. ()] 'Gene Expression Analysis for Type-2 Diabetes Mellitus-A Case Study on Healthy vs Diabetes with Parental History'. Chandra Vasamsetty , Sekhar , Allam Appa Srinivasa Rao Peri , K Rao , Chinta Srinivas , Someswararao . *IACSIT International Journal of Engineering and Technology* 2011. 2012. 3 (3) p. . (IEEE Transactions on)

[Wilinski et al. ()] 'Gene Selection for Cancer Classification through Ensemble of Methods'. A Wilinski , S Osowski , K Siwek . *Proc. Ninth Int'l Conf. Adaptive and Natural Computing Algorithms (ICANNGA '09)*, (Ninth Int'l Conf. Adaptive and Natural Computing Algorithms (ICANNGA '09)) 2009. p. .

[Guyon ()] 'Gene Selection for Cancer Classification Using Support Vector Machines'. I Guyon . *Machine Learning*, 2002. 46 p. .

[Zhang and Deng ()] 'Gene Selection for Classification of Microarray Data Based on the Bayes Error'. J.-G Zhang , H.-W Deng . *BMC Bioinformatics* 2007. 8 (1) p. 370.

[Maulik et al. ()] 'gene-expression-based cancer subtypes prediction through feature selection and transductive SVM'. Ujjwal Maulik , Anirban Mukhopadhyay , Debasis Chakraborty . *Biomedical Engineering* 2013. 60 (4) p. . (IEEE Transactions on)

[Lazar et al. ()] 'GENESHIFT: A Nonparametric Approach for Integrating Microarray Gene Expression Data Based on the Inner Product as a Distance Measure between the Distributions of Genes'. Cosmin Lazar , Jonatan Taminau , Stijn Meganck , David Steenhoff , Alain Coletta , Y Weiss David , Colin Solis , Robin Molter , Hugues Duque , Ann Bersini , Nowé . *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2013. TCBB. 10 (2) p. .

[Haury et al. ()] A.-C Haury , P Gestraud , J.-P Vert . *The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures*, 2011. 6 p. e28210.

[Janowczyk et al. ()] 'High-throughput biomarker segmentation on ovarian cancer tissue microarrays via hierarchical normalized cuts'. Andrew Janowczyk , Sharat Chandran , Rajendra Singh , Dimitra Sasaroli , George Coukos , Michael D Feldman , Anant Madabhushi . *Biomedical Engineering* 2012. 59 (5) p. . (IEEE Transactions on)

[Sekhar et al. ()] 'Identification of differentially expressed genes for diabetes with parental history vs healthy using Microarray data analysis'. V Sekhar , Chandra , P S Allam Appa Rao , K Rao , Srinivas . *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, 2010. IEEE. 4 p. .

[Elbakry et al. ()] 'Identification of Differentially Expressed Genes for Time-Course Microarray Data Based on Modified RM ANOVA'. Ola Elbakry , M Omair Ahmad , M N S Swamy . *Computational Biology and Bioinformatics* 2012. 9 (2) p. . (IEEE/ACM Transactions on)

[Chuang et al. ()] 'Identifying Significant Genes from Microarray Data'. Han-Yu Chuang , Hongfang Liu , Stuart Brown , Cameron Mcmunn-Coffran , Cheng-Yan Kao , D. Frank Hsu . *Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering*, (the Fourth IEEE Symposium on Bioinformatics and Bioengineering) 2004.

[Wei and Pan ()] 'Incorporating Gene Functions into Regression Analysis of DNA-Protein Binding Data and Gene Expression Data to Construct Transcriptional Networks'. Peng Wei , Wei Pan . *IEEE/ACM transactions on computational biology and bioinformatics* 2008. 5 (3) .

[Yu et al. ()] 'Incorporating nonlinear relationships in microarray missing value imputation'. Tianwei Yu , Hesen Peng , Wei Sun . *IEEE TRANSACTIONS ON COMPUT-ATIONAL BIOLOGY AND BIOINFORMATICS* 2010.

[Noman and Iba ()] 'Inferring Gene Regulatory Networks Using Differential Evolution with Local Search Heuristics'. Nasimul Noman , Hitoshi Iba . *IEEE/ACM transactions on computational biology and bioinformatics* 2007. 4 (4) .

[Peng and Li ()] 'IntClust: A Software Package for Clustering Replicated Microarray Data'. Wei Peng , Tao Li . *IEEE conference on BIBE*, 2006.

[Lu et al. (2007)] 'Interactive Semisupervised Learning for Microarray Analysis'. Yijuan Lu , Qi Tian , Feng Liu , Maribel Sanchez , Yufeng Wang . *ieee/acm transactions on computational biology and bioinformatics* april-june 2007. 4 (2) .

[Sakellariou et al. ()] 'Investigating the minimum required number of genes for the classification of neuromuscular disease microarray data'. Argiris Sakellariou , Despina Sanoudou , George Spyrou . *Information Technology in Biomedicine* 2011. 15 (3) p. . (IEEE Transactions on)

[Muresan et al. ()] 'Microarray Analysis at Single-Molecule Resolution'. Leila Muresan , Jaros?aw Jacak , Erich Peter Klement , Jan Hesse , Gerhard J Schutz . *IEEE TRANSACTIONS ON NANOBIOSCIENCE* MARCH 2010. 9 (1) .

[Muresan et al. ()] 'Microarray Analysis at Single-Molecule Resolution'. Leila Muresan , Jaros?aw Jacak , Erich Peter Klement , Jan Hesse , Gerhard J Schutz . *IEEE TRANSACTIONS ON NANOBIOSCIENCE* MARCH 2010. 9 (1) .

[Blekas et al. ()] 'Mixture Model Analysis of DNA Microarray Images'. K Blekas , Member , N P Ieee , Senior Galatsanos , Member , A Ieee , Senior Likas , Member , I E Ieee , Lagaris . *IEEE TRANSACTIONS ON MEDICAL IMAGING* JULY 2005. 24 (7) .

[Zhao and Wang-Kit ()] 'Multi-Class Kernel-Imbedded Gaussian Processes for Microarray Data Analysis'. Xin Zhao , Leo Wang-Kit , Cheung . *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFOR-MATICS* 2010.

[Zhao and Cheung ()] 'Multiclass Kernel-Imbedded Gaussian Processes for Microarray Data Analysis'. Xin Zhao , Leo Wang-Kit Cheung . *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2011. 8 (4) p. .

[Dudoit et al. ()] 'Multiple Hypothesis Testing in Microarray Experiments'. S Dudoit , J P Shaffer , J C Boldrick . *Statistical Science* 2003. 18 (1) p. .

[Maji ()] 'Mutual information-based supervised attribute clustering for microarray sample classification'. Pradipta Maji . *IEEE Transactions on* 2012. 24 (1) p. . (Knowledge and Data Engineering)

[Bø and Jonassen ()] 'New Feature Subset Selection Procedures for Classification of Expression Profiles'. T Bø , I Jonassen . *Genome Biology* 2002. 4 (4) p. .

[Xuan et al. ()] 'Normalization of Microarray Data by Iterative Nonlinear Regression'. Jianhua Xuan , Eric Hoffman , Robert Clarke , Yue Wang . *IEEE conference on IBBE*, 2005.

[Parodi et al. ()] 'Not Proper Roc Curves as New Tool for the Analysis of Differentially Expressed Genes in Microarray Experiments'. S Parodi , V Pistoia , M Museli . *BMC Bioinformatics* 2008. 9 (1) p. 410.

[Parzen ()] 'On Estimation of a Probability Density Function and Mode'. E Parzen . *The Annals of Math. Statistics* 1962. 33 (3) p. .

[Wang et al. ()] 'On modeling and state estimation for genetic regulatory networks with polytopic uncertainties'. Zidong Wang , Huihai Wu , Jinling Liang , Jie Cao , Xiaohui Liu . *IEEE Transactions on* 2013. 12 (1) p. . (NanoBioscience)

[Zhang ()] 'On the Consistency of Feature Selection Using Greedy Least Squares Regression'. T Zhang . *J. Machine Learning Research* 2009. 10 p. .

[Xu et al. ()] 'Performance analysis and design of position-encoded microsphere arrays using the Ziv-Zakai bound'. Xiaoxiao Xu , Pinaki Sarder , Nalinikanth Kotagiri , Samuel Achilefu , Arye Nehorai . *IEEE Transactions on* 2013. 12 (1) p. . (NanoBioscience)

[Wang et al. (2007)] 'Poisson-Based Self-Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data'. Haiying Wang , Huiru Zheng , Francisco Azuaje . *IEEE/ACM Transactions on Computational Biology and Bioinformatics* April 2007. 4 p. .

[Kyoung et al. ()] 'Probabilistic Models for Semi-Supervised Discriminative Motif Discovery in DNA Sequences'. Jong Kyoung , Kim , Seungjin Choi , Member , Ieee . *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* FEBRUARY 2, 2010.

[Kyoung et al. ()] 'Probabilistic Models for Semi-Supervised Discriminative Motif Discovery in DNA Sequences'. Jong Kyoung , Kim , Seungjin Choi , Member , Ieee . *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* FEBRUARY 2, 2010.

[Jaskowiak et al. ()] 'Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis'. Pablo A Jaskowiak , Jgb Ricardo , Ivan G Campello , Costa Filho . *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2013. 10 (4) p. .

[Li et al. ()] 'Recipe for Uncovering Predictive Genes using Support Vector Machines based on Model Population Analysis'. Hong-Dong Li , Yi-Zeng Liang , Qing-Song Xu , Dong-Sheng Cao , Bin-Bin Tan , Bai-Chuan Deng , Chen-Chen Lin . *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFOR-MATICS* 2010.

[Liu et al. ()] 'Recursive Feature Addition for Gene Selection'. Qingzhong Liu , Student Member , Andrew H Ieee , Sung . *IEEE conference on Nueral network* 2006.

[Dost et al. ()] 'TCLUST: A fast method for clusterin genome-scale expression data'. Banu Dost , Chunlei Wu , Andrew Su , Vineet Bafna . *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMA-TICS* 2010.

[Storey ()] 'The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value'. D Storey . *Annals of Statistics* 2003. 31 p. .

[Ben-Dor ()] 'Tissue Classification with Gene Expression Profiles'. A Ben-Dor . *J. Computational Biology* 2000. 7 p. .

[You et al.] *TotalPLS: Local Dimension Reduction for Multicategory Microarray Data*, Wenjie You , Zijiang Yang , Mingshun Yuan , Guoli Ji . p. .

[Uehara and Kakadiaris ()] 'towards automatic analysis of DNA microarrays'. Christian Uehara , Ioannis Kakadiaris . *Proceedings of WACV*, (WACV) 2002.

[Valentini ()] 'True Path Rule hierarchical ensembles for genome-wide gene function prediction'. Giorgio Valentini . *IEEE ACM TRANS. ON COMP. BIOL. AND BIOINFORMATICS* 2010.

[Shao et al. ()] 'Using the Maximum Between-Class Variance for Automatic Gridding of cDNA Microarray Images'. Gui - Shao , Fan Fang , Qian Yang , Qi-Feng Zhang , Lin-Kai Zhou , Luo . *Computational Biology and Bioinformatics* 2013. 10 (1) p. . (IEEE/ACM Transactions on)