ABSTRACT

JUN LU. Analysis on Microarray Data and DNA Regulatory Elements Prediction (Under the direction of Dr. Spencer Muse)

Transcription profiling with microarray technology has significantly accelerated our understanding of complex biological processes by allowing the genome-wide measure of message RNA levels. Microarrays are commonly used for identifying genes with expression differing between two or more samples (e.g. treatments vs. controls), searching for gene expression patterns among a set of samples or genes, and studying gene regulation networks. Here, we first address the variation intrinsic to microarray experiments. The analysis of variance technique was applied to partition and quantify several sources of variation likely to be present in a typical cDNA microarray experiment. Based on a pilot experiment with intensive replication at several levels, we showed that significant amounts of variation can be attributed to slide, plate and pin differences. The origin of these sources of variation was discussed and suggestions were made on how to minimize or avoid them when a future microarray experiment is designed.

Next, we demonstrated that molecular cancer classification could be approached by discriminant analysis. We analyzed a public Affymetrix chip dataset and selected the predictor genes based on the t-values and stepwise discriminant analysis, and evaluated the resulting model's performance in predicting 34 test samples by discriminant analysis. Only two samples were not correctly predicted with 25 predictor genes we chose. We also evaluated the parsimony of our model by evaluating, through a stepwise method, the

minimum number of genes required to maintain a high level of accuracy in predicting cancer types.

The accumulation of microarray data can help elucidate the gene regulation mechanisms in cells. Here, we attempted to find an improved matrix description for transcription factor binding site. We applied a genetic algorithm (GA) to derive matrices that were trained from a set of true binding sequences and random sequences. Preliminary results indicate that the matrix derived shows a higher specificity in binding site prediction than the regular position weighted matrix (PWM) within a range of cutoff scores. The binding site of the cell-cycle related transcription factors, E2Fs, was taken as an example to illustrate our method. When both the GA-derived and regular matrices were applied to scan the human gene upstream sequences, the matrix we derived gave significant less predictions than the regular matrix, given the same false negative rate observed in the training dataset.

ANALYSIS ON MICROARRAY DATA AND DNA REGULARTORY ELEMENTS PREDICTION

by

JUN LU

A dissertation submitted to the Graduate Faculty of

North Carolina State University

in partial fulfillment of the

requirements for the Degree of

Doctor of Philosophy

BIOINFORMATICS

Raleigh

2002

APPROVED BY:

Spencer V. Muse Chair of Advisory Committee	Sandra E. Dunn
Bruce S. Weir	Gregory C. Gibson
Zhao-Bang Zeng	

To my wife, Mingyan and my daughter, Christina

BIOGRAPHY

LU, JUN was born on October 4, 1970 in Jiangpu, Jiangsu province, P. R. China. He entered the Department of Biology at Nanjing University in 1988 and received his B.S. in Biology in 1992. Following college, he worked as a research assistant in China National Rice Research Institute (CNRRI) at Hangzhou, China. In August 1997 he came to the United States for his graduate study at North Carolina State University (NCSU). He was in Forest Biotechnology Group at NCSU, and later transferred to the newly established Bioinformatics program to pursue his Ph.D. degree. During his graduate study, he also worked as a research intern at DNA Science Laboratory, Morrisville, NC, and National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, NC.

ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude to Dr. Bruce Weir, and Dr. Spencer Muse, for their support, encouragement and guidance I need to stay on track during the graduate study. They have helped shape my research projects and provide numerous suggestions to this dissertation.

I would like to say a special thanks to Susan Spruill, Dr. Sandra Dunn, and Dr. Leping Li for their tremendous help on the projects, and for being mentors and friends. Without them, this dissertation would not have happened.

Thanks also to the other committee members, Dr. Gregory Gibson, Dr. Zhao-Bang Zeng, and my graduate representative, Dr. Brian Wiegmann for the reviewing of this work and continuous support.

I am truly grateful to Dr. Ross Whetten and Dr. Ron Sederoff for the generous support and mentoring when I was in their lab. Thanks are extended to many people in the Forest Biotechnology Group for their friendship.

I would also like to express my appreciation to DNA Science Laboratory and Biostatistics Branch at NIEHS for the generous financial support and many helpful suggestions from Dr. Clare Eisenberg and others in the microarray discussion group at NIEHS.

I would also like to say thanks to many good people at BRC, in particular, Debra Hibbard, Amy Elkins, Lisa Barefoot, Dr. Sarah Hardy and fellow graduate students for all the assistance with everything. I appreciate all the help they gave to me. Thanks.

Finally, I would like to say thanks to my wife, Mingyan, for the encouragement, endless support and love. I am grateful to her for always standing by with me during the course of this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Literature Review of Microarray Technology and Data Analysis

Introduction

Understanding the mechanisms of gene regulation has been a main focus of molecular biologists for many years. The control mechanisms for gene expression in eukaryotes are complex and the regulation can occur at the chromatin DNA level, during mRNA synthesis (transcription), RNA processing, protein synthesis (translation) and after protein translation (Lewin 1997). Among these possibilities, control at the transcription level is one of the key steps and has been studied intensively (Ptashne and Gann 1997, 1998). The amount of expressed RNA of a gene is traditionally measured by Northern blot or dot blot (Sambrook and Russell 2001). Briefly, the total RNA or mRNA samples are transferred and UV cross-linked to nylon membranes (with or without size separation). Then, a DNA fragment that represents the gene is radioactively or non-radioactively labeled and hybridized with the RNA samples on the membrane. After the hybridization and washing steps, the hybridization signal is often detected by autoradiography. The difference on signal intensity between a control and a treatment sample indicates the differential expression of a gene. One thing should be pointed out is that the Northern blot or dot blot can only measure the expression level of one gene in each hybridization process.

During the last decade, the application of high-throughput technology has attracted tremendous interests in biological science. Current genome projects have been

driven by the advance of new technology, such as high-throughput sequencing (Venter et al 2001; Lander et al 2001), large-scale genotyping (Syvanen 2001; Cutler et al 2001) and transcription profiling (Brown and Botstein 1999; Lipshutz et al 1999). In particular, a number of techniques have been developed to measure the cellular mRNA levels on a global scale, including serial analysis of gene expression (SAGE) (Velculescu et al 1995), cDNA microarray (Schena et at 1995) and oligonucleotide chips (Lockhart et al 1996; Mcgall et al 1996). Such high-throughput methods significantly accelerate our understanding of the complex biological process by allowing the identification of gene expression patterns from the broad assessment of gene transcription. Compared to traditional methods such as Northern or dot blots, the microarray and oligonucleotide technologies have been developed to measure the RNA transcript levels of hundreds or thousands of genes at a time. In principle, the way of measuring the RNA levels with microarrays or oligo-nucleotide chips is the same as that with Northern or dot blot. The major difference is that the positions of the probes (DNA fragments representing genes) and the targets (i.e. total RNA or mRNA samples) are reversed in microarray experiments. In another words, the total RNA (instead of the DNA fragment) is labeled and a large number of genes are spotted on a slide (or membrane) in an array experiment rather than vice versa in Northern or dot blot. This type of setting allows measuring RNA levels for many genes in each labeling reaction because those genes are spotted on one slide. The further development of the two-dye system allowed the comparisons of RNA levels between two samples (for instance, a treatment and a reference) for many genes simultaneously (Schena et al 1995; Shalon et al 1996).

In this chapter, I will first give a brief review on microarray techniques. The focus will then be on how the microarray data is analyzed, including data normalization, identification of differentially expressed genes, and recognition of gene expression patterns

Part I: Review of microarray techniques

In general, there are two types of microarrays that are commonly used: cDNA microarrays (Schena et al 1995) and Affymetrix oligo-nucleotide chips (Lockhart et al 1996). Although the strategies of constructing the arrays (chips) are different, the basic idea is the same: the relative concentration of a target mRNAs in a given sample is measured through hybridizing a labeled RNA sample to tens of thousands probes on an array (chip). Since the cDNA arrays can be made in a lab, there is more flexibility in choosing genes for spotting on an array. In contrast, the oligo-nucleotide chips are only commercially available (from Affymetrix Inc.), the sets of genes on a chip have been pre-chosen and the DNA chips have been made for users, which is less flexible than cDNA arrays. On the other hand, preparing the cDNA arrays in labs can introduce unexpected variations and the comparison of experiments conducted in different labs can be difficult due to the different sets of genes and the protocols used.

cDNA microarrays

The cDNA microarray technology was first developed at Brown's lab (Schena et al 1995; Brown and Botstein 1999). The whole procedure can be broadly divided into two stages: (1) the preparation of arrays and (2) the process of labeling sample RNA, hybridization and data collection. The fist stage is to prepare a set of glass microscope

slides in which tens or thousands of genes (represented by DNA fragments) have been spotted with high density. After the slides are made, they can be stored for a period of time. The second stage starts from the extraction of total RNA (or mRNA) to the data collection, which is often conducted and finished within a week.

(1) Array preparation

The Source of DNA The first step is to prepare DNA for spotting. Although the synthesized oligo-nucleotides can be used, in eukaryotes, the source of DNA is often from a constructed cDNA library that contains tens and thousands of clones. Each clone includes one species of expressed DNA fragment representing one gene. The cDNA clones can be sequenced, and the single run, partial sequence of a cDNA is called an expressed sequence tag (EST). Since the amount of DNA in each clone is very small, the DNA fragment needs to be amplified, usually through polymerase chain reaction (PCR). The size of PCR amplified products generally ranges from 0.3 to 3 kilobases (Kb). Usually the PCR reactions are conducted in a high-throughput way. For example, a 96-well plate can hold 96 PCR reactions and each reaction amplifies one cDNA clone (representing one gene). After the amplification, the quality and quantity of PCR products are checked by gel electrophoresis (Sambrook and Russell 2001). Any failure or contamination reactions will be excluded in next steps. Before the printing step, the PCR products need to be cleaned to have the unincorporated nucleotides removed. The ethanol precipitation and filter filtration are commonly used methods for cleaning PCR products. The final cleaned DNA is either dissolved in distilled water or dried, and stored in refregiretor ready for printing (Schena 1999; Eisen and Brown 1999).

One important issue that needs to be addressed is gene redundancy. Ideally, the DNA fragments spotted on the glass slides should represent as many genes as possible. However, the highly expressed genes in a cell would appear more frequently than the low-abundance genes if the library clones were randomly chosen. In order to avoid such problems, the partial sequence information of the clones may be necessary to identify each clone. For instance, if a given cDNA clone has been partially sequenced, the EST sequence information in UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/) can be helpful to determine whether the cDNA clone is unique on an array (Pruitt and Maglott 2001). For human cDNAs, there is another database called Human Gene Index (http://www.tigr.org/tbd/hgi/hgi.html), in which all the human ESTs have been clustered with one cluster representing one gene (Quackenbush et al. 2000).

Array Printing The glass slides on which the DNA probes are printed are coated with poly-L-lysine or aminosilane, both of which could be commercially available. The even surface of the coated slides is essential for the reproducibility of the microarray experiments. It was reported that the aminosilane-coated slides had more consistent surface than poly-L-lysine-coated slides. The background fluorescence produced from aminosilane-coated slides was only half of that from poly-L-lysine-coated slides with even higher signal intensities (Hegde et al 2000). Another factor affecting the quality of the printed slides is the solution for diluting the spotted DNA. Two types of solution are commonly used: 3XSSC and 50% dimethyl sulfoxide (DMSO). Hegde et al (2000) reported that PCR products diluted in 50% DMSO showed the highest intensities when aminosilane-coated slides were used. The DMSO printing solution provides another two advantages: the DMSO can denature the DNA fragments, and the single stranded DNA

has strong binding on the slide surface. Also DMSO has lower evaporation rate than 3XSSC, which allows the spotting DNA to be stored for a longer period of time (Hegde et al 2000).

The DNA is spotted on coated glass slides by robots. Brown's lab designed the original robot at Stanford University (Schena et al. 1995; Eisen and Brown 1999). Currently, a number of companies are selling such high-speed robotic systems, such as Intelligent Automation Systems (IAS) (http://www.ias.com), Genomic Solution (http://www.genomicsolutions.com) and others. The robots array DNA samples (dissolved in spotting solution) from 96- or 384-well microplates to slides using printing pins. One critical point for evaluating the quality of the robotic system is the consistency among these pins. The geometry of the pins greatly affects the size and the shape of spots on a slide, along with some other factors including humidity and temperature (Hegde et al. 2000; Eisen and Brown 1999). The printed DNA is bound to slide by ultraviolet (UV) cross-linking, and the slides can be stored in a dessicator at room temperature.

(2) RNA labeling, hybridization and data collection

RNA labeling The total RNA is extracted from the treatment and control samples through standard methods (Sambrook and Rusell 2001) or using RNA extraction kits that are commercially available for instance, RNeasy kit from Qiagen (http://www.qiagen.com), and TRIzol from Life Technology (**http://www.lifetechnologies.com**). The quality of RNA is critical for the labeling efficiency and the whole microarray experiments. Either message RNA (mRNA) or total RNA can be used for labeling. The first step of the labeling reaction is to synthesize the

first strand cDNA by reverse transcriptase with the oligo-dT as primers. The labeled nucleotides, such as Cy3 or Cy5 –dUTP, are added into the reaction mixtures and incorporated into the first-strand cDNA. After the labeling reactions are finished, the unincorporated Cy3 or Cy5 –dUTP needs to be removed by filter filtration or ethanol precipitation that is similar to the PCR product cleaning. Since the Cy3 and Cy5 –dUTP are light sensitive, the reactions and all the following steps should avoid light as much as possible. The final product is the first-stand cDNA labeled with Cy3 or Cy5. An equal amount of the Cy3 and Cy5 labeled products are mixed and denatured just before hybridization (Schena 1999; Eisen and Brown 1999; Hegde et al. 2000).

Pre-hybridization and hybridization For aminosilane slides, the free amino groups on the slide surface could bind non-specifically to labeled cDNA products that lead to high background in later steps. The prehybridization of slides is necessary for blocking the free amino groups and removing the unbound DNA. It has been shown that 1% BSA solution can effectively decrease the non-specific binding of labeled targets to the slide (Hegde et al. 2000). The hybridization step allows the single stranded target cDNA (labeled) to pair with the single stranded DNA that is spotted on slides. Besides the mixture of dye-labeled cDNA products, the hybridization solution also includes non-specific DNA, such as COT1-DNA and poly(A)-DNA, which are used for blocking nonspecific hybridization. The hybridization solution is added on the DNA surface of slides and covered with slide coverslip. The hybridization is usually conducted at 42 degrees and stays for 16-20 hours. Then the unhybridized target DNA will be removed by a few washing steps starting from low stringency to high stringency, where the high

stringency condition refers to washing steps with solutions containing low concentration of salt (Eisen and Brown 1999; Hegde et al. 2000).

Data collection First, the confocal laser scanner is used to scan the hybridized arrays. The scanner uses a laser to excite Cy3 and Cy5 dyes and the emission signals are then recorded. Two separate TIFF images will be obtained with one from the Cy5 and the other from the Cy3 channel. Next, the two image files are further processed and the images are transferred to quantitative data through spot identification (gridding), signal, and background estimation from the pixel intensities (Schena 1999; Eisen and Brown 1999). Although a number of software packages for image processing are available, for instance ScanAlyze (http://rana.lbl.gov/EisenSoftware.htm; Eisen and Brown 1999), in many cases human intervention is necessary due to the irregularities of the spots. Both the distance of spots and the spot size need to be adjusted to cover the true spot area. Finally, the foreground and the background intensities of each spot are estimated by either the median or the mean of the pixel values within the spot ellipse (Hegdes et al 2000). Sometimes, the spot quality is evaluated based on the pixel values and listed as well (Yang et al 2000). It has been noticed that the image processing can vary substantially from one lab to the other, which can introduce significant amount of variation into the data (Yang et al 2002; Eisen and Brown 1999; Hegde et al 2000). The final data for one hybridization experiment is often in the form of a table in which the rows list the names of genes and the columns represent the values of background intensities, the signal intensities, and possibly the quality measurements for each spot.

Affymetrix chips

The DNA chips produced from Affymetrix Inc. have been widely used as well (Lockhart et al 1996; Mcgall et al 1996). The major difference between DNA chips and cDNA arrays is that, instead of a long fragment of cDNA, 20 pairs of 25-mer oligonucleotides represent one gene on each DNA chip. Among the 20 pairs of oligonucleotides represented for each gene, 10 pairs are negative sets with only one base pair (in the middle) different from the corresponding positive sets that have the exact matches to the gene sequence. The DNA chip arrays are created by Affymetrix's light-directed chemical synthesis process, which combines solid-phase chemical synthesis with photolithographic fabrication techniques employed in the semiconductor industry (http://www.affymetrix.com; Mcgall et al 1996). The detail of this process is rather technical and will not be described here. Basically, the known sequences of oligonucleotides can be synthesized directly on the chip at a predefined location. The process is conducted automatically and generally has higher consistency among chips than does spotting cDNA on arrays. The process of hybridization with DNA chips is very similar to that using cDNA arrays, as we discussed above. One difference is that, rather than using the first labeled first-strand cDNA as final targets, the DNA chips require the synthesis of double stranded cRNA that is used as the hybridization targets (Lockhart et al 1996; Wodicka et al 1997). The expression level of a gene is usually represented by the mean of differences between 20 pairs of perfect match (PM) and mismatch (MM) probes (Wodick et al 1997), and recent studies suggests that model-based methods can provide further sensitivity by using the probe-level data (Li and Wong 2001; Schadt et al 2000).

Conclusion

Although microarray technology has been developed for nearly a decade, the laboratory techniques are still evolving. Many factors can affect the accurate measurement and comparison of expression intensities in different samples, in particular, the comparisons between arrays or DNA chips. Reproducible and reliable data is essential for further data analysis and for drawing scientific conclusions.

Part II. Review of gene expression data analysis

Currently, microarray (chip) technology has two primary types of application: (1) identifying the differently expressed genes among two or more experimental conditions and (2) finding gene expression patterns. In the first situation, the main interest is on finding genes that are up or down regulated between two or more samples (for example, between cancer and normal samples). Usually the question is answered through designed experiments, in which the sources of variation are generally under control. Identifying differentially expressed genes is simply a task of conducting hypothesis tests of whether the levels of mRNA transcripts are equal between two or more samples. Examples of this type of application include the identification of target genes of the c-Myc transcription factor (Coller et al 2000), stain and region-specific genes in mouse (Sandberg et al 2000), and genes responding to ionizing radiation (Tusher et al 2001).

Another powerful application with array technology is to extract gene expression patterns from a large amount of expression data (Eisen et al 1998; Alon et al 1999; Golub et al 1999; Alizaheh et al 2000). This area is more relevant to data mining, in which the data is collected from various sources, therefore we may have little control over the

variation in experiments (since they may have been conducted in different labs). The main interest is more on finding the relationships among a group of genes or samples rather than finding a specific gene. The typical examples of finding gene expression patterns include cancer sample classification (Golub et al 1999; Alon et al 1999) and identification of co-regulated gene clusters (Brazma et al 1998; Spellman et al 1998; Eisen et al 1998).

There are other types of applications on using microarray technology, such as SNP identification (Cutler et al 2001) and DNA-binding site identification (Iyer et al 2001). These areas are beyond the discussion here where the focus is on studying gene transcription.

Data normalization

Current microarray technology is far from giving precise measures of mRNA transcript concentration, especially for those genes with low-level expression. As we discussed in Part I, a number of steps are involved in a microarray experiment, and each step could potentially introduce systematic variation into the final data. For example, variation can be introduced during the dye labeling reaction, slide preparation (for instance, pin, slide) and slide hybridization (Kerr et al 2000; Schuchhardt et al 2000). The sources of variation can also arise from physiological and biological sampling, which have been nicely illustrated by Novak et al (2002) and Pritchard et al (2001). Normalization is a process attempting to remove systematic variation without affecting the detection of the biological difference of interests. Ideally, all published data should be appropriately normalized to remove any systematic variation from the raw data.

Unfortunately, there is no consensus on how the data should be normalized, and the choice of the normalization methods is also depends on the experimental design of an experiment (Kerr et al 2000; Wolfinger et al 2001).

Normalization can be either conducted first as a preliminary step or integrated into the data analysis process (Yang et al 2002; Wolfinger et al 2001; Kerr et al 2000). In general, there are two types of normalization methods: global and intensity-dependent normalization. Global normalization means that the same scaling factor will be applied to every gene on a slide, regardless of the intensity values of genes. Examples of global normalization include ANOVA-based method (Kerr et al 2000), the normalization by linear regression (Golub et al 1999), and other mean or median based normalization methods (Spellman et al 1998; Quackenbush 2001). In contrast, intensity-dependent (or local) normalization takes the difference of individual genes into account, and assumes that the intensities are non-linear. Local normalization is usually carried out through local regression, such as LOWESS (Locally Weighted Scatterplot Smoothing) regression (Yang et al 2002).

For cDNA microarray data obtained from a two-dye labeling system, the mixed samples (one control and one treatment sample carrying different dye labels) can hybridize with one slide simultaneously. The advantage of using such system is that the spot difference (such as the shape or the amount of DNA probes) has less effect on the comparisons of interests. However, using two dyes can introduce additional variation due to the difference on dye incorporation in the labeling reaction (Kerr et al 2000). Another major source of variation comes from the unequal labeling efficiency between two samples due to differences of RNA quality, which can result in different amount of

labeled hybridization products (Schuchhardt et al 2000; Tseng et al 2001). Such systematic variation has to be accounted for in the following data analysis. Currently, three normalization methods are widely used for data generated from a single slide hybridization. The first method is based on the total intensity values of two samples. The method depends on the assumptions that the total intensities from each labeled samples should be equal, and genes that are up-regulated are balanced out by the down-regulated genes in the same array. A rescaling factor can be calculated by assuming either equal median or equal mean between two samples (Cy3 and Cy5 channels). This factor is then assigned to each gene on the array (a strategy of global normalization). The second normalization method commonly used is based on regression. Again, the assumption is also made that the slope of the regression line between two samples should be one if there is no systematic variation involved. The basic steps include finding the least-square regression line on the scatter-plot of Cy5 versus Cy3 channel, and the slope is used to rescale the intensity values. In some cases, non-linear intensity is assumed and a local regression is conducted, such as LOWESS (Yang et al 2002). Another way of normalizing the data is described by Chen (1997). They derived a probability density function for the ratio of two channel intensities of the housekeeping genes on the array, and use that to iteratively adjust the mean expression ratio to one. The density of the ratio can also used to construct confidence intervals and to identify the differentially expressed genes (Chen et al 1997).

Although the methods discussed above have been commonly used in literature, recently developed ANOVA models can put normalization into a formal statistical framework if replicate experiments are conducted, and they can provide the error

estimates for statistical inferences (Kerr et al 2000). After normalization steps, the data are either ratio values (log ratios) or intensity values, and further analysis such as classification is based on these normalized data matrices.

Identifying differentially expressed genes

A common application of microarrays is to identify genes that are differentially expressed between two samples. The ratio-based approaches have been widely used for two-dye cDNA microarrays (Chen et al 1997; Chu et al 1998). The expression ratio is the normalized value of a treatment sample divided by the normalized value of a control sample, and is calculated for every individual gene on a slide. The base 2 logarithm of the ratio is often used. Typically the genes that show log-ratio greater than 1 (two-fold increase or decrease of the gene expression) are considered as differentially expressed. The advantage of calculating the log-ratio of expression value is easy to understand for biologists since 2-fold change has a log-ratio of 1. However, use of the ratio statistic has been drawn criticism since the ratio totally ignores the absolute values of the expression levels and the variation in the data (Wittes and Friedman 1999; Wolfinger et al 2001). The absolute value matters because it has been well known that low-intensity values often associate with high variation. Thus, the low intensity value can give high ratio values, however, those values are most likely unreliable (Tanaka et al 2000).

Another issue is replication. Early studies usually conducted the experiments without replication. Recently, more and more studies have shown that replication of the experiments is critical for both cDNA and DNA chip experiments. Lee et al (2000) applied a normal linear mixture model to fit the single-channel data from one slide. What

they found is that, even for the simple design, a large amount of variation within a gene was detected. They recommended a minimum of three replicates for this simple design. Another example is given by Pritchard et al (2001), in which they addressed the biological variation problem. By using analysis of variance (ANOVA) techniques, they found that 0.8, 1.9, and 3.3 percent of genes were normally variable in the mouse liver, testis and kidney, respectively. Among the variable genes, the stress-related, immune-modulated, and hormone-controlled genes were highly represented. Their experiment demonstrates that the biological variation could be introduced easily by handling the mouse differently and some genes often show variability without any relation to the biological difference of interests. Without replication, there is no way to distinguish such variation from the true treatment differences. A recent study by Tanaka et al (2000) gave another example of the danger of strictly using the fold-change to evaluate the change of the gene expression.

There are at least two purposes for conducting the replicated experiments. One is to accurately estimate the systematic variation, which is necessary for normalization. Another reason of having replicates is to estimate the random error, which can then be used to conduct statistical inference on changes of gene expression (Kerr and Churchill 2001). For replicated microarray experiments, the task of identifying differentially expressed genes becomes a statistical testing problem. A variety of methods supported by statistical theory can be applied to analyze the array data. A well-known example is the ANOVA model first introduced by Kerr and Churchill (2000). They showed that the ANOVA model could be used to combine the data normalization step into the procedure of identifying differentially expressed genes. This is an advantage since the normalization

parameters are estimated based on all the data, rather than a piece of information. Instead of using log ratios, the model was derived from the logs of original intensity values, which avoided the drawback of using ratios and preserved the data properties necessary for an additive effect model. Since the ANOVA model is tightly linked to the issue of experimental design, the authors also illustrated how different designs affected the parameter estimation. Finally they applied a resampling-based method (bootstrap) to derive the confidence intervals for estimates of differential expression of each gene, as opposed to assuming normal distributions for the residual errors (Kerr et al 2000; Kerr and Churchill 2001). Wolfinger et al (2001) further extended the ANOVA model by assuming some effects to be random, such as the array effect. The mixed model can provide broader inference on the gene expression. Also, the normalization process was conducted separately from the gene identification step. Briefly, they constructed two interconnected ANOVA models, the 'normalization model' and the 'gene model'. The normalization models are fit to the data first (to normalize the data) and the residuals derived were the input data for the gene models. Advantage is that separate gene models allow inference to be made with individual error estimates for each gene (heterogeneity), which is consistent with the observation that genes with low intensity values display high intra-gene variability.

Ideker et al (2000) constructed error models to account for the multiplicative and additive errors of the intensity values from each channel. The parameters in the model were estimated by maximum likelihood. Given the error structure and the estimators obtained, the generalized likelihood ratio tests were applied to test for differentially expressed genes (Ideker et al 2000).

A unique feature of microarray data is that usually there are very few replicates for each gene in a experimental sample but a large number of genes on the array, which makes the traditional t-test or rank-based nonparametric tests not effective. Efforts have been made to draw statistical inferences based on the distributions of quantities including the whole data on the array. For example, Pan et al (2001) proposed to estimate the distributions of two t-statistic-type scores using normal mixture models. The differentially- expressed genes were identified through the comparison of two distributions (one from each channel) by likelihood ratio tests. Baldi et al (2001) applied a Bayesian approach to make statistical inference on gene expression changes. Other methods for identifying differentially expressed genes directly model the ratio values (log-ratios) (Chen et al 1997; Newton et al 2001).

Finding gene expression patterns

Given a normalized gene expression matrix, one broad area of the application of microarray data is to identify the gene expression patterns existing in a large amount of data (Eisen et al 1998; Alon et al 1999; Brazma and Vilo 2000). Here the patterns simply refer to a number of common features shared either by a group of genes or a group of samples. The methods used in finding patterns can be divided into two categories: unsupervised learning and supervised learning. We call an unsupervised learning method if no expert knowledge is involved in the data learning process. Examples of unsupervised methods include hierarchical clustering (Eisen et al 1998) and self-organizing map (SOM) (Tamayo et al 1999). On the other hand, a supervised method can be applied if prior knowledge on the data was used during data analysis. Specifically, a

training data set will be constructed, including one or more classes of known functionally related genes (positive set) and one or more groups of genes not belonging to those classes. Supervised methods then learn to discriminate between the known group members and non-group members of a given class based on the gene expression data (Brown et al 2000). We discuss some of the methods commonly used in each category.

Unsupervised learning

The typical example of unsupervised learning is sample (gene) classification. Most work has focused on grouping genes into clusters based on expression data. The motivation of conducting cluster analysis is that genes in a particular pathway or involved in a specific process may be co-regulated and show similar expression patterns.

Various methods have been applied in cluster construction based on gene expression data. The most popular one is hierarchical clustering (Eisen et al 1998), which is described as follows. We define each gene by an expression vector, and the values in the vector are the expression measurements under different experiment conditions. The measurements can be the exact values or the relative ratios to a reference sample. The distance between two gene-expression vectors reflects how similar the two genes are. There are various methods for measuring distance, and the widely used one is Euclidean distance. Let D represents the distance between gene X and Y, the Euclidean distance is calculated by the following equation:

$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where x_i and y_i are the expression values of gene X and Y at condition i respectively. Hierarchical clustering can be conducted based on the gene distance with various

algorithms, including single-linkage algorithm, complete-linkage clustering, average-linkage clustering and others (Quackenbush 2001). For example, the average-linkage clustering algorithm finds the closest two genes first from a pairwise distance matrix of all n gene vectors. A node is created joining these two genes, and the node is then represented by a new expression vector calculated by the average of the expression vectors of two genes. The pair-wise distance matrix is then calculated again among *n-1* gene (node) vectors included, and the vectors with lowest distance are joined and form a new node. This process is carried out iteratively until all gene-vectors are in one cluster.

Hierarchical clustering is simple to implement and the results can be easily visualized. The major problem with cluster analysis is that phylogenetic-type cluster structures do not reflect the actual relations among genes in biological systems. A difficulty would arise if one gene participates in more than one regulatory or metabolic pathways. Another problem with this method is that once an incorrect assignment is made at an early stage, there is no way to correct it later on. Other clustering algorithms include k-means clustering (Tavazoie et al 1999) and self-organizing maps (Tamayo et al 1999). Both methods rely on other sources of information to pre-determine the number of clusters in the available data, which may not be realistic in situations where we know nothing about the whole structures of the data.

Another way of searching for biological meaningful expression patterns is using a relevance network (Butte et al 2000). The pairwise similarity among genes or features like chemotherapeutic susceptibility of cancer cells discussed in the paper can be measured with some methods (distance, mutual information and et al). A modified version of the Pearson correlation coefficient is used by Butte et al, in which a sign is

added into the correlation in order to capture the negative relations between genes (features).

$$\hat{r}^2 = \frac{r}{abs(r)} r^2$$

r^2 is representing the Pearson correlation coefficient. High $\hat{r}^2$ represents the hypothesis of a biological relation. Instead of grouping all genes into one cluster by hierarchical clustering, a relevance network only retains the significant relationships among features by threshholding. The cutoff $\hat{r}^2$ value is empirically determined by repeated random permutation study. For each feature pair, the measurement is permuted 100 times, and $\hat{r}^2$ is calculated for each permuted set. In Butte's paper, a threshold value 0.8 is chosen because no permutated data set can produce the absolute $\hat{r}^2$ value greater than 0.8 (so the p-value is less than 0.01 in this case). Groups of features with an $\hat{r}^2$ value greater than the threshold will aggregate and form the relevance network. The threshold value can be adjusted to include biological meaningful relations in the relevance network. In a relevance network, a gene can be directly or indirectly connected to several genes or other phenotypic measurements, which is an obvious advantage over tree-type clustering methods. The cross-connected networks represent not only pair-wise but also aggregated associations, which are the most trustable relationships between features. The example given in the paper successfully clustered the base line expression measurements in cancer cell lines and measurements of anticancer agent susceptibility in the same set of cell lines.

The limitation of the relevance network method is that the determination of threshold value is still somewhat subjective. We may miss many real biological

associations if we choose a high cutoff value, and we may produce many false-positive hypotheses if we relax that criterion.

Gene or sample classification has been the most popular method to search for possible expression patterns in the data. A great deal of effort was made on deriving the best classification method. However, little attention has been paid to the reliability of the clustering results, given the inherent noise associated with microarray data. Kerr and Churchill (2001) addressed this question by using bootstrapping techniques to evaluate stability of results from a cluster analysis. Basically, they assume that the "true" clustering *C* can only be obtained if the true expression differences *r* are known. In practice, the true expression differences can only be estimated from observed data, $\hat{r}$. Thus, the true clustering *C* can only be estimated by $\hat{C}$. This means the results from any clustering methods have variation if the error of estimation is considered. In the paper, they applied the bootstrapping method to create a collection of bootstrapping clusters $\{C^*\}$, in addition to the original cluster $\hat{C}$ based on the estimates of relative gene expression $\hat{r}$. A gene was claimed "95% stable" in a cluster (profile) if it occurs in at least 95% of the bootstrap clusters. The final results they received were the number of genes in each profile having a pre-defined bootstrap stability (for instance 95% stability). Obviously, the number of genes in a cluster would be smaller with 95% stability than with 80% stability (Kerr and Churchill 2001).

There are model-based clustering methods, by which the number of clusters can be estimated from the data rather than determined beforehand as with other methods (for instance, SOM) (Yeung et al 2001; Fraley and Raftery 1998). In another example, Pan et al (2002) applied a normal mixture model to cluster genes based on a summary statistic,

the t-statistic. They found that the genes showing differential expression were grouped separately from the genes with no expression changes based on the assumed model. Another advantage of using model-based clustering is that it can calculate the posterior probability of an observation belonging to a cluster (Pan et al 2002).

Supervised learning

As has been mentioned, the supervised learning methods take prior knowledge on samples (genes) into account. Theyt attempt to derive models from a set of training samples, and then use the model to predict a new sample (or gene). Obviously, the main part is to derive a model that can discriminate between the known group and non-group members of a given class based on the gene expression data. Brown et al (2000) introduced a supervised method, support vector machine (SVM), which has been widely used in pattern recognition in computer science (Vapnik 1998). SVM uses the biological information in the training data to determine the important characteristics in each class, and then use this knowledge to determine whether a new gene should be classified into a specific class (Brown et al 2000).

For a two-class classification problem, assume we have a set of gene expression vectors x_i (i = 1, 2, …, n) measured at p experimental conditions. Also assume there exists two classes in the input genes: for example, a group of genes encoding cytoplasmic ribosome proteins (labeled with +1), and the other group of genes encoding histone proteins (labeled with –1). The goal is to find a hyperplane to separate these two classes. In many real situations there is no hyperplane that can completely separate one class from the other class members. One solution to this problem is to map the original input space

into a higher dimension feature space. Then, any training data set can be separable if we can choose an appropriate feature space with sufficient dimensionality. SVM takes this approach by mapping the input vectors into a high dimensional feature space and constructs an Optimal Separating Hyperplane (OSH). OSH maximizes the margin, the distance between the hyperplane and the nearest data points of each class in the feature space. The specification of a SVM requires two parameters: the kernel function which defines an inner product of two expression vectors in the feature space, and a regularization parameter which controls the trade off between margin and misclassification error. Letting x_i and x_j represent two expression vectors, two typical kernel functions are:

$$K(x_i, x_j) = (x_i \bullet x_j + 1)^d ,$$

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) .$$

The first one is the polynomial kernel function of degree *d*, and the second one is the radical basic function (RBF) kernel with parameter γ.

Brown et al (2000) construct the SVMs to recognize six functional classes: tricarboxylic acid cycle (TCA), respiration, cytoplasmic ribosomes, proteasome, histones, and helix-turn-helix proteins. The group of helix-turn-helix proteins is used as a negative control since this group of proteins is involved in various functions and should not show any similar expression pattern (i.e. SVM can not learn to recognize members of this class based on expression values). One to three degrees of polynomial SVMs and RBF SVM are constructed for each class separately, and the results are compared with those obtained from several alternative supervised methods (i.e. Parzen windows, Fisher's linear discriminant, and two decision tree based methods). The comparison shows that

RBF SVM and third-order SVM generally perform better than other methods. An interesting comparison is also conducted between higher-order SVM and hierarchical clustering methods. Among known 121 ribosome genes, SVM correctly identifies 118 genes with 7 false positives, and by hierarchical clustering method, the ribosome cluster found 112 genes with 14 false positives. The supervised learning is considered promising since it directly combines the known knowledge with the expression data. The limitation of SVM is that some biological classes may not be recognizable based on transcription data alone.

Besides the SVM, there are other classical supervised methods, including the *k* nearest-neighbor (KNN) classifier, linear or quadratic discriminant analysis, and classification and regression trees (CART) (Webb 1998). In addition, results have shown that combining predictors from perturbed versions of the leaning set could increase the prediction accuracy (Breiman 1998).

One of the most successful applications of pattern recognition approaches using expression data is on the classification and prediction of tumor types or subtypes. Alizadeh et al (2000) applied DNA microarray expression data to distinguish the subtypes of diffuse large B-cell lymphoma (DLBCL). Their results also show that the molecular classification of tumors on the basis of gene expression can identify previously undetected and clinically significant subtypes of cancer. The SVM has been applied to multiclass cancer classification problem (Ramaswamy et al 2001). A total of 218 tumor samples spanning 14 common tumor types are classified by SVM algorithm based on expression data, and the overall accuracy is 78%.

Prediction of gene regulatory elements

Gene (or sample) classification using either supervised or unsupervised methods is only the first step in expression data analysis. The final goal of conducting transcription profiling is to elucidate the functional roles of the respective genes and to further understand the underlying biological processes. An immediate next step of microarray data analysis is to seek possible regulatory elements in the genomic sequences. A number of studies on predicting gene regulatory elements have been carried out in yeast (Brazma et al 1998; Hughes et al 2000; Tavazoie et al 1999). An assumption is often made that genes with similar expression patterns may share a common regulatory mechanism, such as transcription factor-binding sites. Therefore, the main goal is to find motifs (DNA elements) that are over-represented in a group of genes.

One area of study is to identify the potential regulatory elements in a group of functionally related genes. Most studies on finding DNA motifs were conducted in yeast, since the yeast promoter region is relatively well defined and close to the coding regions. For instance, Vilo et al (2000) used a public data set to carry out cluster analysis of 6221 genes based on the measurements from 80 conditions. For each cluster, the 600 bp upstream sequences were extracted for each gene, and all substrings with variable length were searched and assigned a probability score according to binomial distribution if it appeared in more than ten sequences. A total of 1498 substrings were selected based on a significance threshold determined by a randomization test. The substrings were further grouped into 62 clusters of patterns, and were searched against the known transcription factor in the yeast database SCPD (The Promoter Database of *Saccharomyces cerevisiae*). What they found is that 48 out of 62 patterns had atches with known

transcription factor binding sites. The remaining patterns may be either the false positives or ones that have not been identified yet, which could be new targets for further experimental study (Brazma 1998). A similar example was given by van Helden et al (1998) in which they also searched for over-represented oligonucleotides in a group of sequences. Both methods conducted an exhaustive search for sub-strings that meet the defined level of statistical significance.

Another strategy for discovering the DNA motifs is through the Gibbs sampling algorithm, which was previously used to find motifs in protein sequences (Lawrence 1993, Neuwald et 1995). Gibbs sampling is one of the Monte Carlo Markov Chain techniques for sampling from posterior distributions. For the case of searching DNA motifs in a group of sequences, the goal of Gibbs sampling is to find the starting location of a motif within each sequence and estimate the residue frequencies at each position of the motif. There are two versions of Gibbs sampler: site sampler and motif sampler. The Site sampler considers the simple case, where each motif is assumed to occur in every sequence and only once. The motif sampler is more general by allowing 0 or more motifs in each sequence. Given a group of sequences and the length of motif, three steps are involved in Gibbs sampling (An example of site sampler).

Notation:

S: a group of sequences (known)

W: the width of a motif (known). The motif is assumed to be ungapped.

$c_{i,j}$: observed counts of residue j at position i of a motif. i ranges from 1 to W, and j ranges from 1 to M (M equals 4 for DNA, and 20 for protein sequences).

$q_{i,j}$: the frequency of the residue j at position i of a motif. $q_{0,j}$ represents the background frequencies of the four residues.

a_k: Vector of starting positions of a motif in sequences S. k ranges from 1 to N, where N is the number of sequences.

(1) Initialization

The site sampler is initialized by randomly assigning a starting location of the motif within each sequence, and the location of starting points is recorded as a vector a_k. The frequencies of residues at each position within a motif, along with the background frequencies (sequences outside the motif), can be calculated. Usually pseudocounts for each residue are needed to avoid problems with zero probability.

(2) Predictive update step

The first step is to select one sequence (among N sequences) and place the motif sequence (with length of W) within the selected sequence in the background. The motif frequency matrix is updated based on the alignment (with N-1 sequences), along with the background residue frequencies.

(3) Sampling step

The sampling step is to find the starting position of the motif in the sequence that has been selected in step 2. All possible starting positions will be considered. One way of drawing samples is as follows. For a fragment X (with length W) from a given start point, a weight A_x is calculated based on the likelihood ratio, in which the numerator is the likelihood assuming fragment X in the motif model and the denominator is the likelihood by assuming X in the background. In the process of

sampling, the starting position will be chosen with the probability proportional to the weight. The positions with higher weights will be more likely to be selected than the positions with lower weights. Once the iterative predictive update and sampling steps have been conducted for all of the sequences, a probable alignment will appear and an associated *F* score will be calculated as well, where *F* is given by the formula

$$F = \sum_{i=1}^{W} \sum_{j=1}^{J} C_{i,j} \log \frac{q_{i,j}}{q_{0,j}}$$

The predictive update and sampling steps will be conducted again on each sequence of *S*, given the data and the start points that have been chosen. The whole process will be run iteratively as a Markov chain. As the number of iteration is large enough, the joint distribution of $(a_1^{(i)}, a_2^{(i)}, ..., a_N^{(i)})$ will converge to joint posterior distribution $f(a_{1,} a_{2,} ..., a_N \mid S)$.

Gibbs sampling is a heuristic and not an exhaustive search, so the method will give an optimal result but may not be the best. The method is very sensitive to the subtle DNA patterns existing in a group of sequences, but requires the knowledge on the length of a motif. A program AlignACE was specifically written for scanning multiple motifs in a given set of DNA sequences by using a Gibbs sampling strategy with some modifications (Hughes et al 2000). Gibbs sampling has also been applied to search for co-operatively binding sites in the program Co-Bind (GuhaThakurta and Stormo 2001).

Conclusions

Appropriate analysis of microrray data is critical to draw scientific conclusions. The first data normalization step can have significant influence on downstream analysis. The identification of differentially expressed genes can provide clues on the function of genes in a particular biological process. The finding of gene expression patterns allows one to predict gene function, classify normal and disease samples for future diagnosis, and to help elucidate the mechanisms of gene transcription.

Outline of the present study

The remainder of this dissertation is organized into three chapters. The first chapter addresses the variation problem existing in a typical micorarray experiment. The objective is to identify sources of variation and provide knowledge for future experiment design. It is titled "Addressing the sources of variation in gene expression data". In the second chapter, we demonstrate that molecular cancer classification could be approached by discriminant analysis combined with T-statistics, and the prediction result is better than that published by Golub et al (1999). In the fourth chapter, we present the preliminary results on improving matrix representations of transcription factor binding sites, which is titled "Toward improved matrix description of E2F binding sites".

References

Alizadeh A. A., Eisen M. B., Davis R. E., Ma C., Lossos I. S., Rosenwald A., Boldrick J. C., Sabet H., Tran T., Yu X., Powell J. I., Yang L., Marti G. E., Moore T., Hudson J.,Jr, Lu L., Lewis D. B., Tibshirani R., Sherlock G., Chan W. C., Greiner T. C., Weisenburger

D. D., Armitage J. O., Warnke R., Staudt L. M., et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 403(6769):503-11.

Alon U., Barkai N., Notterman D. A., Gish K., Ybarra S., Mack D., Levine A. J.. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. 96(12):6745-50.

Baldi P., Long A. D.. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics. 17(6):509-19.

Brazma A., Jonassen I., Vilo J., Ukkonen E.. (1998) Predicting gene regulatory elements in silico on a genomic scale. Genome Res. 8(11):1202-15.

Brazma A., Vilo J.. (2000) Gene expression data analysis. FEBS Lett. 480(1):17-24.

Breiman, L.. (1998) Arcing classifiers. The Annals of Statistics. 26:801-824.

Brown M. P., Grundy W. N., Lin D., Cristianini N., Sugnet C. W., Furey T. S., Ares M.,Jr, Haussler D.. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A. 97(1):262-7.

Brown P. O., Botstein D.. (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet. 21(1 Suppl):33-7.

Butte A. J., Tamayo P., Slonim D., Golub T. R., Kohane I. S.. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A. 97(22):12182-6.

Chen Y., Dougherty E. R., and Bitter M. L.. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Optics 2, 364-374.

Chu S., DeRisi J., Eisen M., Mulholland J., Botstein D., Brown P. O., Herskowitz I.. (1998) The transcriptional program of sporulation in budding yeast. Science. 282(5389):699-705.

Coller H. A., Grandori C., Tamayo P., Colbert T., Lander E. S., Eisenman R. N., Golub T. R.. (2000) Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. Proc Natl Acad Sci U S A. 97(7):3260-5.

Cutler D. J., Zwick M. E., Carrasquillo M. M., Yohn C. T., Tobin K. P., Kashuk C., Mathews D. J., Shah N. A., Eichler E. E., Warrington J. A., Chakravarti A.. (2001) High-

throughput variation detection and genotyping using microarrays. Genome Res. 11(11):1913-25.

Eisen M. B., Spellman P. T., Brown P. O., Botstein D.. (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 95(25):14863-8.

Eisen M. B., Brown P. O.. (1999) DNA arrays for analysis of gene expression. Methods Enzymol. 303:179-205.

Fraley C., Raftery A. E.. (1998) How many clusters? Which clustering methods? Answers via model-based cluster analysis. Computer J, 41:578-588.

Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S.. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 286(5439):531-7.

GuhaThakurta D., Stormo G. D.. (2001) Identifying target sites for cooperatively binding factors. Bioinformatics. 17(7):608-21.

Hegde P., Qi R., Abernathy K., Gay C., Dharap S., Gaspard R., Hughes J. E., Snesrud E., Lee N., Quackenbush J.. (2000) A concise guide to cDNA microarray analysis. Biotechniques. 29(3):548-562

Hughes J. D., Estep P. W., Tavazoie S., Church G. M.. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol. 296(5):1205-14.

Ideker T., Thorsson V., Siegel A. F., Hood L. E.. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J Comput Biol. 7(6):805-17.

Iyer V. R., Horak C. E., Scafe C. S., Botstein D., Snyder M., Brown P. O.. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature. 409(6819):533-8.

Kerr M. K., Martin M., Churchill G. A.. (2000) Analysis of variance for gene expression microarray data. J Comput Biol. 7(6):819-37.

Kerr M. K., Churchill G. A.. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A. 98(16):8961-5.

Kerr M. K., Churchill G. A.. (2001) Statistical design and the analysis of gene expression microarray data. Genet Res. 77(2):123-8.

Lander E. S. et al.. (2001) Initial sequencing and analysis of the human genome. Nature. 409(6822):860-921.

Lee M. T., Kuo F. C., Whitmore G. A., Sklar J.. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci U S A. 97(18):9834-9.

Lewin B.. (1997) GENES VI. Oxford University Press, New York, NY.

Li C., Wong W. H.. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A. 98(1):31-6.

Lipshutz R. J., Fodor S. P., Gingeras T. R., Lockhart D. J.. (1999) High density synthetic oligonucleotide arrays. Nat Genet. 21(1 Suppl):20-4.

Lockhart D. J., Dong H., Byrne M. C., Follettie M. T., Gallo M. V., Chee M. S., Mittmann M., Wang C., Kobayashi M., Horton H., Brown E. L.. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol. 14(13):1675-80.

McGall G., Labadie J., Brock P., Wallraff G., Nguyen T., Hinsberg W.. (1996) Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci U S A. 93(24):13555-60.

Neuwald A. F., Liu J. S., Lawrence C. E.. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci. 4(8):1618-32.

Newton M. A., Kendziorski C. M., Richmond C. S., Blattner F. R., Tsui K. W.. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol. 8(1):37-52.

Novak J. P., Sladek R., Hudson T. J.. (2002) Characterization of variability in large-scale gene expression data: implications for study design. Genomics. 79(1):104-13.

Pan W., Lin J., Le C.. (2001) A mixture model approach to detecting differentially expressed genes with microarray data. Technical Report 2001-011, Division of Biostatistics, University of Minnesota, 2001. http://www.biostat.umn.edu/cgi-bin/rrs?print+2001.

Pan W., Lin J., Le C. T.. (2002) Model-based cluster analysis of microarray gene-expression data. Genome Biol. 3(2):RESEARCH0009.

Pritchard C. C., Hsu L., Delrow J., Nelson P. S.. (2001) Project normal: defining normal variance in mouse gene expression. Proc Natl Acad Sci U S A. 98(23):13266-71.

Pruitt K. D., Maglott D. R.. (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29(1):137-40.

Ptashne M., Gann A.. (1997) Transcriptional activation by recruitment. Nature. 386(6625):569-77.

Ptashne M., Gann A.. (1998) Imposing specificity by localization: mechanism and evolvability. Curr Biol. 8(24):R897.

Quackenbush J., Liang F., Holt I., Pertea G., Upton J.. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. 28(1):141-5.

Quackenbush J.. (2001) Computational analysis of microarray data. Nat Rev Genet. 2(6):418-27.

Ramaswamy S., Tamayo P., Rifkin R., Mukherjee S., Yeang C. H., Angelo M., Ladd C., Reich M., Latulippe E., Mesirov J. P., Poggio T., Gerald W., Loda M., Lander E. S., Golub T. R.. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A. 98(26):15149-54.

Sambrook J., Russell D.. (2001) Molecular Cloning, 3rd edition. Cold Spring Harbor laboratory, New York, NY.

Sandberg R., Yasuda R., Pankratz D. G., Carter T. A., Del Rio J. A., Wodicka L., Mayford M., Lockhart D. J., Barlow C.. (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. Proc Natl Acad Sci U S A. 97(20):11038-43.

Schadt E. E., Li C., Su C., Wong W. H.. (2000) Analyzing high-density oligonucleotide gene expression array data. J Cell Biochem. 80(2):192-202.

Schuchhardt J., Beule D., Malik A., Wolski E., Eickhoff H., Lehrach H., Herzel H.. (2000) Normalization strategies for cDNA microarrays. Nucleic Acids Res. 28(10):E47.

Shalon D., Smith S. J., Brown P. O.. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6(7):639-45.

Schena M., Shalon D., Davis R. W., Brown P. O.. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 270(5235):467-70.

Schena M.. (1999) DNA Microarrays: An practical approach. Oxford University Press, New York, NY.

Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D., Futcher B.. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 9(12):3273-97.

Syvanen A. C.. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet. 2(12):930-42.

Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E. S., Golub T. R.. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A. 96(6):2907-12.

Tanaka T. S., Jaradat S. A., Lim M. K., Kargul G. J., Wang X., Grahovac M. J., Pantano S., Sano Y., Piao Y., Nagaraja R., Doi H., Wood W. H.,3rd, Becker K. G., Ko M. S.. (2000) Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. Proc Natl Acad Sci U S A. 97(16):9127-32.

Tavazoie S., Hughes J. D., Campbell M. J., Cho R. J., Church G. M.. (1999) Systematic determination of genetic network architecture. Nat Genet. 22(3):281-5.

Tseng G. C., Oh M. K., Rohlin L., Liao J. C., Wong W. H.. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. 29(12):2549-57.

Tusher V. G., Tibshirani R., Chu G.. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 98(9):5116-21.

van Helden J., Andre B., Collado-Vides J.. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol. 281(5):827-42.

Vapnik, V. (1998) *Statistical Learning Theory* Wiley, New York, NY.

Velculescu V. E., Zhang L., Vogelstein B., Kinzler K. W.. (1995) Serial analysis of gene expression. Science. 270(5235):484-7.

Venter J. C. et al.. (2001) The sequence of the human genome. Science. 291(5507): 1304-51.

Vilo J., Brazma A., Jonassen I., Robinson A., Ukkonen E.. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. Proc Int Conf Intell Syst Mol Biol. 8:384-94.

Webb A. R.. (1998) *Statistical pattern recognition*. Oxford University Press, New York, NY.

Wittes J., Friedman H. P.. (1999) Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. J Natl Cancer Inst. 91(5):400-1.

Wodicka L., Dong H., Mittmann M., Ho M. H., Lockhart D. J.. (1997) Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol. 15(13):1359-67.

Wolfinger R. D., Gibson G., Wolfinger E. D., Bennett L., Hamadeh H., Bushel P., Afshari C., Paules R. S.. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol. 8(6):625-37.

Yang Y. H., Buckley M. J., Dudoit S., Speed T. P.. (2000) Comparison of methods for image analysis on cDNA microarray data. http://stat-www.berkeley.edu/users/terry/zarray/Html/papersindex.html .

Yang Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J., Speed T. P.. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30(4):e15.

Yeung K. Y., Fraley C., Murua A., Raftery A. E., Ruzzo W. L.. (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics. 17(10):977-87.

Chapter 2

Assessing Sources of Variation in Gene Expression Data

Abstract

Analyzing the sources and quantities of variation is the basis for data normalization, for statistical inferences about changes of gene expression, and for downstream pattern recognition. Although microarray technology has been routinely used in genomic research, Rigorous studies on variation that is intrinsic to microarray experiments are still lacking. In this chapter, we address the variation problem by validating and evaluating the quality of data from an intensively replicated experiment. Analysis of variance (ANOVA) techniques have been applied to partition and quantify several sources of variation likely to be present in a typical cDNA microarray experiment. The results show that significant amounts of variation can be attributed to slide, plate and pin differences. Replication of the experiment is absolutely necessary, especially at slide level. We discuss the origin of these sources of variation and provide suggestions for how to minimize or avoid them when designing a microarray experiment.

Introduction

The advent of microarray technology has significantly increased the understanding of complex biological processes (Lander 1999; Brown and Botstein 1999). It provides a rapid, parallel snapshot of gene transcription at the whole genome level.

Although the technology has been rapidly developed and applied in answering many biological questions, the appropriate analysis of the microarray data is still evolving (Chen et al 1997; Efron and Tibshirani 2000; Baldi and Long 2001). The difficulties arise from the large dimension of the data (because of multiple genes and experimental conditions measured) and the inherent variation existing in a microarray experiment. For example, the hundreds or thousands of genes on an array can lead to multiple testing problems if the differentially expressed genes need to be identified (Quackenbush 2001;Dudoit et al 2002). Also the variation in the data is directly connected to data normalization, which can have profound influence on the downstream analysis of array data, such as classification and DNA regulatory element identification (Wittes and Friedman 1999; Kerr and Churchill 2001).

In this chapter, the variation problems are the main focus. Generally, there are two types of variation existing in every microarray experiment: systematic variation and stochastic variation. Systemic variation refers to the variation that results from any lack of uniformity in the physical conduct of the experiment, such as those attributed to the differences between pins used in spotting process, different labeling reactions, or different slides used for hybridization. The systematic variation has to be removed from the raw data before conducting any further analysis, a process usually called data normalization. Stochastic variation represents the inherent random variation among replicated experiments (the random error), which needs to be estimated if we want to make statistical inferences on changes of gene expression (Kerr and Churchill 2001).

Because of a series of steps involved in a microarray experiment, systematic variation can be easily introduced into the final data. The possible sources of variation can be due to probe, target and array preparation, hybridization, background and overshining effects (Schuchhardt et al 2000). For example, dye effects have been shown to be significant in several studies (Kerr and Churchill 2000; Wolfinger et al 2001). Loos et al (2001) measured spot-to-spot variability within a slide and between slides with the cDNAs from the same labeling reactions. They found that their experiments were highly reproducible with spot-to-spot variability 3.8% within a slide and 5.0% between slides. Another example was given by Yang et al (2002), in which pin effects and slide effects have been addressed and integrated into the data normalization procedure. Other reported sources of variation include variability in RNA isolation, tissue-to-tissue variability, and within-tissue heterogeneity (Siedow 2001). For experiments where the sources of variation are known, several normalization methods can be taken to exclude such systematic variation from the raw data (see Chapter 1).

Although the reported sources of variation may be common in many experiments, the sources of variation really depend on how an experiment is conducted. As many steps are involved in microarray experiments, we are discovering an increasing number of factors that can contribute to variability in microarray data. For situations in which we do not have any information on where the variation comes from and how much the variation would be, a pre-experiment becomes critical to find the possible sources of variation. In this case, we have to begin with a list of possible sources of variation, conduct an experiment with a prior design, and apply statistical tools to estimate the contribution of each effect. This is the motivation of the current study, which attempts to identify the

major sources of variation in a pre-designed experiment. ANOVA techniques were used to partition the total variation in the data and quantify the variation contributed by each factor. In an ANOVA framework, one uses the data to estimate both the relative gene expression and the magnitude of variation. Finally, based on the results obtained from ANOVA, we provide suggestions to better control for noise through changes in techniques or considerations in future experimental design.

Materials and Methods

Microarray experiments

Microarrays were prepared by spotting human genes on glass slides with a robot. A total of 79 genes were spotted in each slide, and the genes selected were often related to drug metabolism. Each gene was represented by a 50-mer oligo-nucleotide purchased from BD Biosciences Clontech (http://www.clotech.com). Also, two plates of DNA of the same set of genes were used for spotting, with 79 genes on each plate. Each gene occupied one well on each plate except for three genes, glucose-6 phosphate dehydrogenase (G6PD), glucose-3 phosphate dehydrogenase (G3PD) and beta-actin, which were replicated 8 times on each plate. All three genes are constitutively expressed in all tissues and developmental stages and served as controls. Seventy-nine genes were spotted on slides by a two-pin robot. During each spotting process, two pins carrying DNA from two wells on a plate spotted twice within each slide. After one plate of DNA was spotted on 30 slides, a second plate was used for spotting on the same set of slides. Such spotting scheme produced 8 quadrants on each slide. The origin of each quadrant

was shown in the diagram (Table 1). It should be noticed that about half of the genes were spotted only by pin 1, and the other half spotted by pin 2.

Hybridization samples were prepared as follows: the total RNA was extracted separately from 10 cadaverous livers that were suitable for donor transplantation. Ten liver samples were from donors with a range of ages (15-53 years old) and ethnic background (5 Caucasians, 2 African Americans, 2 Hispanics and 1 Asian). The liver RNA samples were reverse-transcribed to cDNA and labeled with Cy-3 d-CTP, then hybridized with 3 slides for each liver sample. Only one dye was used in this experiment.

Data preparation and analysis

The final data were collected as raw readings as well as the background values. Some spots were treated as missing because of low quality signals. The raw intensities were corrected for background noise by subtraction. Subtracted values less than zero were set to zero. No efforts were made to remove outliers. All computations for the data analysis were carried using SAS software (SAS Institute, Cary, NC).

Results

The analysis of variance (ANOVA) models

A number of factors accounted for the sources of variation in this microarray experiment. Before performing further analysis, we first listed all possible factors and interaction terms to construct a full model.

$$Y_{ijklm} = L_i + S(L)_{ij} + G_k + P_l + R_m + G*P_{kl} + G*R_{km} + P*R_{lm} + L*G_{ik} + L*P_{il} + L*R_{im} + L*G*R_{ikm} + L*G*P_{ikl} + L*P*R_{ilm} + L*G*P*R_{iklm} + G*S(L)_{ijk} + P*S(L)_{ijl} + R*S(L)_{ijm} + G*P*S(L)_{ijkl} + G*R*S(L)_{ijkm} + P*R*S(L)_{ijlm} + G3PD + G6PD + Beta\_actin + E_{ijklm}$$

where L represents liver (i=1, 2…10), S represents slide (j=1,2,3), G represents gene (k = 1, 2, …, 76), P represents plate (l=1, 2), R represents replicate (m = 1, 2), E represents residual error and Y represents background adjusted intensity reading. All model terms, except for E were treated as fixed effects. In addition, three housekeeping genes, G3PD, G6PD, and beta-actin, were added as covariates into the model. All effects with their degrees of freedom in the full model are listed in Table 2. In this study we were interested in identifying factors that had large contributions to the total variation of the data as opposed to making statistical inference on gene expression. All the effects were treated as fixed because of its simple interpretation, although it may be more reasonable to treat some terms as random effects (for instance slide effect) (Wolfinger et al 2001). The SAS proc GLM procedure was applied to partition the total variation in the data.

Unfortunately, we met difficulty on running GLM procedure when all the data was used to fit the full model. Many high-order interaction terms in the model and a fairly large amount of data significantly increased the memory requirement for SAS. In order to solve this problem, a reduced model was derived by the following strategy. First, seventy-six genes were assigned to 5 sets with 10 genes in each set (Some genes were not selected). The full model was run on each 10-gene set and ANOVA tables were constructed from the full model fitting on the subsets of genes (A total of five ANOVA

tables were calculated). Then, we derived our reduced model by excluding those terms that consistently showed relative small mean square values across five runs. The "small" mean square values were based on the comparison with the model mean squared error (MSE). All the small variance terms were removed from the full model into the error term in the reduced model, which is as follows:

$$Y_{ijklm} = L_i + S(L)_{ij} + G_k + G*P_{kl} + L*G_{ik} + G*S(L)_{ijk} + G3PD + G6PD + Beta\_actin + E_{ijklm}$$

We fit the reduced ANOVA model on the full set of genes (76 genes), and the ANOVA table was shown in Table 3. As we can see, the reduced model derived here remained a good fitting one on the original data with an R square value 96.8%.

The sources of variation

The main goal of this study was to identify the major sources of variation at this early stage of microarray experiments, and provide knowledge for technique improvements and/or better design of future experiments. From the ANOVA table calculated from the reduced model, we saw that a significant amount of variation was attribute to liver, slide and gene main effects, gene by liver, gene by plate, and gene by liver by slide interactions.

The gene effect reflected the intrinsic expression differences among 76 genes across all livers, slides and plates (the main effect of genes). It accounted for the largest variation among all the effects, which is not surprising since different genes are often

expressed at different levels. Slide effects were nested in the liver effects in the current experimental design. It had the second largest contribution to variation in the data, which indicated that within each liver sample the total level of gene expression was different from one slide to the others (across all genes and plates), even though all the labeled targets were originated from a common liver RNA sample. In our ANOVA mode, we detected a significant liver main effect, which had a mean square error 29 times greater than residual error. Liver effects reflected the total gene expression differences among 10 liver samples. Among those interaction terms, the gene by slide within liver interaction showed variation15 times greater than residual error. This term is often referred to the spot effect, which reflects the spotting variation among slides. Another significant source of variation was the gene by plate interaction, which contributed almost equal amount of variation as slide effects. This indicated that at least for some genes, the spotting DNA concentration or quality was different between two plates. Finally, the gene by liver interaction term was also among the significant interaction terms. It represented the differences on expression levels among liver samples for some genes. The gene by liver interaction term was the one we were mostly interested in since most experiments search for genes showing differential expression among two or more samples.

The covariates in the model

Besides the listed possible sources of variation, there may be variation that was introduced from unknown sources, for instance, in the scanning process. In order to capture such variation, we added three covariates, G6PD, G3PD, and beta-actin genes, into our ANOVA model. All three genes are commonly used as positive controls because

their expression levels are considered constant in all tissue samples and at various developmental stages. Based on the assumption above, any conditions that change the expression level of these three genes (among ten liver samples) were considered variation introduced in the microarray experiment. The results showed that G6PD, beta-actin and G3PD had mean squared errors 534, 55, and 32 times greater than the residual error respectively, which demonstrated that a large amount of variation could be accounted for by these covariates. Including covariates in the model is a statistically efficient way of capturing variation since each covariate only costs one degree of freedom.

Pin effects

In this experiment, two pins were used for spotting genes on glass slides. However, seventy-six genes were spotted by either pin one or pin two only on each slide. Therefore, we cannot use such data to evaluate pin effects because of no replication within each pin for those genes. Fortunately, there are two genes that were spotted by two pins on each slide: G6PD and beta-actin, which were used as positive controls on each slide. Each gene was replicated in 8 wells on each plate, and each pin spotted 4 wells. We used the data from these two genes and constructed an ANOVA model to evaluate pin effects (see Table 4). The full partitioning of the total variation in the data indicated that the pin main effect could not be ignored. The magnitude was nearly comparable to that caused by gene by plate interactions. It should be noticed that the model fitting on two genes data also detected the sources of variation described in the reduced model, which confirmed our results derived above.

The ANOVA model needs the assumption that the residual errors have independent and identical normal distribution. These assumptions are examined by visual inspection of residue and quantile-quantile (QQ) plots after the model fitting on the data (see Figure 1).

Discussion

This study addressed the variation issue by an experiment with intensive replication. It should be noticed that this experiment was not a common microarray experiment in which searching for differentially expressed genes was the primary interest. Rather, this project addressed the variation problems existing in a typical microarray experiment, which could be considered a pre-experiment, and the whole data analysis was more relevant to data normalization. When a microarray experiment is first conducted in a lab, the initial question is how successful the experiment is. In another words, we need to find out where the variation comes from, how it affects the results of the experiment, and whether the same results could be obtained if the experiment were repeated. If the major sources of variation were identified, efforts could be made to eliminate the variation, if possible, by technique improvement. For variations intrinsic to the data, good experimental designs and formal statistical analysis would be critical to draw conclusions of a scientific question. There are cases where the sources of variation are obvious for researchers, and special attention could be paid to eliminate that part if necessary. However, in most situations, the sources of variation have to be identified by statistical models, such as ANOVA in this study.

We used ANOVA models to address the variation problems in array data since ANOVA has been applied very successfully in agricultural field study, and the questions asked are very similar between two areas. Examples of applying ANOVA in array data analysis have been published (Kerr and Churchill 2000; Wolfinger et al 2001). However, we consider that our work has two aspects different from what have been published. First, the goal is different. Identifying differentially expressed genes is not our primary interests at this stage. Instead, we are most interested in where the variation comes from. Therefore, intensive replication has been conducted at a number of levels such as slides, plates and pins in this experiment. These terms represent possible sources of variation, and their effects were tested through the ANOVA models. Secondly, we added covariates into ANOVA model (also called, Analysis of Covariance, ANCOVA), attempting to capture unknown sources of variation. Results based on this strategy showed that covariates explained a large amount of variation existing in the data. This is an efficient way to capture variation in the data since each covariate only costs one degree of freedom. However, there exist risk of using housekeeping genes as covariates. It was assumed that there was no differential expression of three covariate genes among liver samples, which may be arguable.

One difficulty rises when many terms are involved in the model and created many high order interaction effects. The memory requirement for the computation is huge, and the running speed is slow for commonly used statistical packages, especially for a relatively large data set. We avoided such problems by constructing a reduced model through random sampling on levels of gene effects. By this strategy, we can only include major effects in our model without significantly decreasing the model fitting.

The key questions of this project are what information could be obtained from this study, and what suggestion could be provided for future experiments in terms of design issue. First, the gene effects, which contributed the largest variation in the data, represented the intrinsic expression difference among genes. This effect is not our concern since usually we are only interested in some specific genes rather than the gene main effects. The gene effects would not exist if models were run on gene-by-gene basis. Slide effects, which had the second largest contribution to the total variation in the data, should be treated seriously since this term directly affect the power of the statistical inference on which genes are differentially expressed. In another words, if we want to test whether a gene is differentially expressed among livers in this experiment, slide effects would be the denominator of F-statistics since slide effects were nested within the liver effects. The differences among slides can simply be observed from the box-plots drawn for ten livers (Figure 2). From the ANOVA table we also saw significant liver main effects. The overall expression differences among ten liver samples are most likely due to unequal amount of labeled cDNA products produced from different labeling reactions conducted. Though the liver effects should be eliminated as much as possible during the microarray experiment, this type of variation usually can not be totally avoided. Therefore, normalization steps will be necessary, such as global and local normalization suggested by Yang et al (2002) or by ANOVA techniques (Wolfinger et al 2001). Another way of accounting for total variation among samples is to add housekeeping genes as covariates in the model as in this study. We also identified another source of variation, pin main effects. It introduced variation in the process of the DNA spotting.

Obviously, the two pins did not behave the same. Although technical improvement could be made to make pins more consistent, pin effects can be eliminated as long as one gene is spotted by one pin across all slides. The pin difference can also been seen in Figure 3.

Among those interaction terms, the gene by plate interactions contributed significant amounts of variation, which is somewhat unexpected. This implies that the two plates are not the same for some genes. One possible reason could be that the DNA concentration on one plate is higher (lower) than that on the other plate, which results in unequal amount of DNA spotted on slides. This suggests that for the future experiment we need to either make the two plates more consistent, or use one plate only to spot arrays. Another interaction term is the gene by slide interaction, which is commonly referred to as spot effects. Unless this effect is small (through improving technique), spot effects have to be involved in the ANOVA model to account for such variation. This step is often conducted during the data normalization.

Conclusion

Appropriate analysis of microarray data requires knowledge on variation existing in all microarray experiments, in particular, the cDNA microarray data. This requires the replication of experiments and statistical tools to extract such information. We applied the ANOVA method to identify the sources of variation existing in a microarray experiment. Intensive replication at different levels allowed the estimation of the magnitude of variation from each of the possible sources. This study provided an example on how statistical methods were used to address the variation problem in a typical microarray experiments. The knowledge gained from this study provides direction

for further lab technique improvements and the suggestion for better experimental design in the future.

Acknowledgement

Microarray data were obtained from a project funded by DNA Sciences Laboratories (formerly PPGx, a wholly owned subsidiary of DNA Sciences, Inc.) The analysis was part of collaboration with PPGx and NC State University's Bioinformatics Program. The authors would like to acknowledge the support of Marco Guida whose laboratory conducted the microarray experiments yielding the data herein.

References

Baldi P., Long A. D.. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics. 17(6):509-19.

Brown P. O., Botstein D.. (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet. 21(1 Suppl):33-7.

Chen Y., Dougherty E. R., and Bitter M. L.. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Optics 2, 364-374.

Dudoit S., Yang Y. H., Callow M. J., and Speed T. P.. (2002) Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Stat. Sin., in press.

Efron B., Tibshirani R., Goss V., Chu G.. (2000) Microarray and their use in a comparative experiment. Stanford: Stanford University Department of Statistics, 2000.

Kerr M. K., Martin M., Churchill G. A.. (2000) Analysis of variance for gene expression microarray data. J Comput Biol. 7(6):819-37.

Kerr M. K., Churchill G. A.. (2001) Statistical design and the analysis of gene expression microarray data. Genet Res. 77(2):123-8.

Kerr M. K., Churchill G. A.. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A. 98(16):8961-5.

Lander E. S.. (1999) Array of hope. Nat Genet. 21(1 Suppl):3-4.

Loos A., Glanemann C., Willis L. B., O.,Brien X. M., Lessard P. A., Gerstmeir R., Guillouet S., Sinskey A. J.. (2001) Development and validation of corynebacterium DNA microarrays. Appl Environ Microbiol. 67(5):2310-8.

Quackenbush J.. (2001) Computational analysis of microarray data. Nat Rev Genet. 2(6):418-27.

Schuchhardt J., Beule D., Malik A., Wolski E., Eickhoff H., Lehrach H., Herzel H.. (2000) Normalization strategies for cDNA microarrays. Nucleic Acids Res. 28(10):E47.

Siedow J. N.. (2001) Making sense of microarrays. Genome Biol. 2(2):REPORTS4003.

Wittes J., Friedman H. P.. (1999) Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. J Natl Cancer Inst. 91(5):400-1.

Wolfinger R. D., Gibson G., Wolfinger E. D., Bennett L., Hamadeh H., Bushel P., Afshari C., Paules R. S.. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol. 8(6):625-37.

Yang Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J., Speed T. P.. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30(4):e15.

TABLES AND FIGURES

Table 1 The schematic representation of the origin of eight quadrants on a slide

Pin 1, Plate 1	Pin 1, Plate 2	Pin 1, Plate 1	Pin 1, Plate 2
Pin 2, Plate 1	Pin 2, Plate 2	Pin 2, Plate 1	Pin 2, Plate 2

Table 2: The full ANOVA model and the associated degree of freedom (df)

Sources	**df**
Liver	9
Slides(Livers)	20
---	---
Genes	75
Plates	1
Replicates	1
Genes x Plates	75
Genes x Replicates	75
Plates x Replicates	1
Genes x Plates x Replicates	75
---	---
Liver x Genes	675
Liver x Plates	9
Liver x Replicates	9
Liver x Genes x Plates	675
Liver x Genes x Replicates	675
Liver x Plates x Replicates	9
Liver x Genes x Plates x Replicates	675
---	---
Genes x Slide(Liver)	1500
Plates x Slide(Liver)	20
Replicates x Slide(Liver)	20
Genes x Plates x Slide(Liver)	1500
Genes x Replicates x Slide(Liver)	1500
Plates x Replicates x Slide (Liver)	20
Error	1500
---	---
G3PDH	1
G6PDH	1
Beta-actin	1

Table 3. The ANOVA table based on the reduced model

Sources	**Df**	**MS relative to MSE**
Liver	9	29x
Slides(Livers)	20	42x
Genes	75	1838x
Genes x Plates	76	38x
Liver x Genes	675	20x
Genes x Slide(Liver)	1500	16x
Residue Error	6764	1x
G3PD	1	32x
G6PD	1	534x
Beta-actin	1	55x

Table 4: Sources of Variation from the ANOVA partition

Source	Degrees of Freedom	Relative Mean Squares
Liver	9	590
Slide (Liver)	20	468
Gene	1	6556
Pin	1	71
Plate	1	822
Gene x Pin	1	5
Gene x Plate	1	82
Pin x Plate	1	11
Well(Pin x Plate)	12	13
Gene x Well(Pin x Plate)	13	10
Liver x Gene	9	169
Liver x Pin	9	21
Liver x Plate	9	11
Gene x Liver x Plate	9	8
Gene x Slide(Liver)	20	101
Pin x Slide(Liver)	20	11
Plate x Slide(Liver)	20	9
Gene x plate x Slide(Liver)	20	9
Residual Error	1724	1

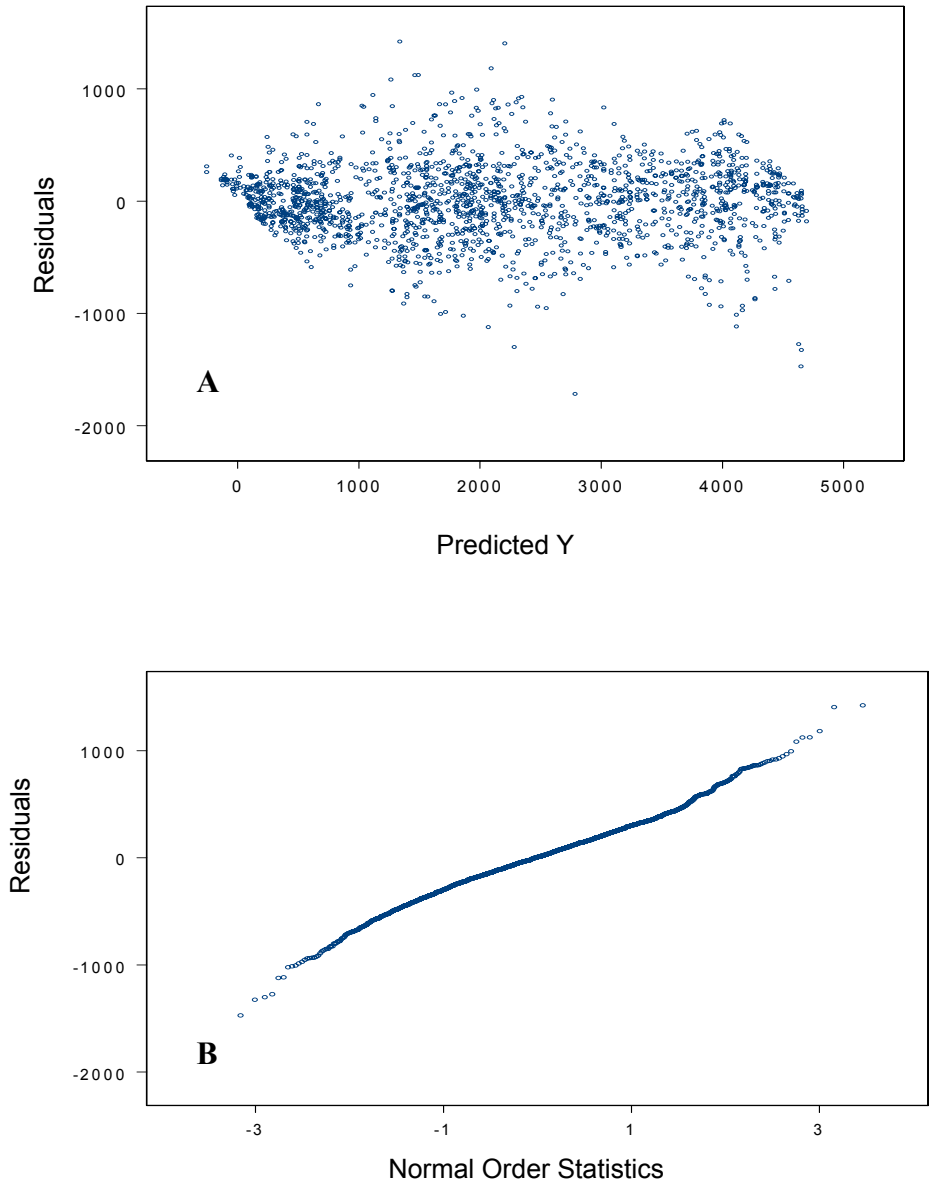

Figure 1 Observed residual plotted against predicted values (A) and normal order statistics (B) obtained from the ANOVA model fitting. A fan shaped dispersion in Plot A or an s-shaped curve in Plot B would indicate departure from the assumption that data are normally distributed with an identical error.

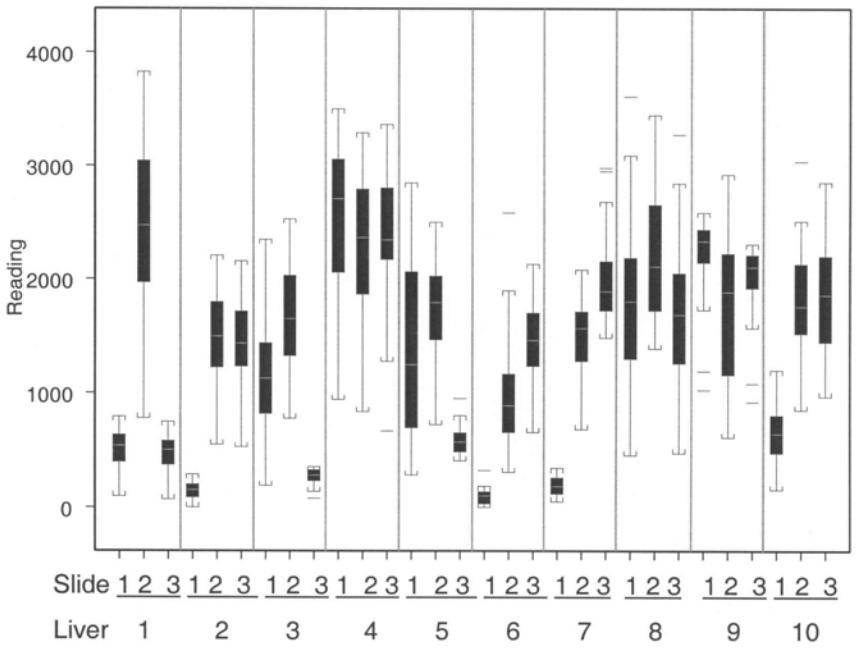

Figure 2. Box plots of the three slides hybridized in each liver.

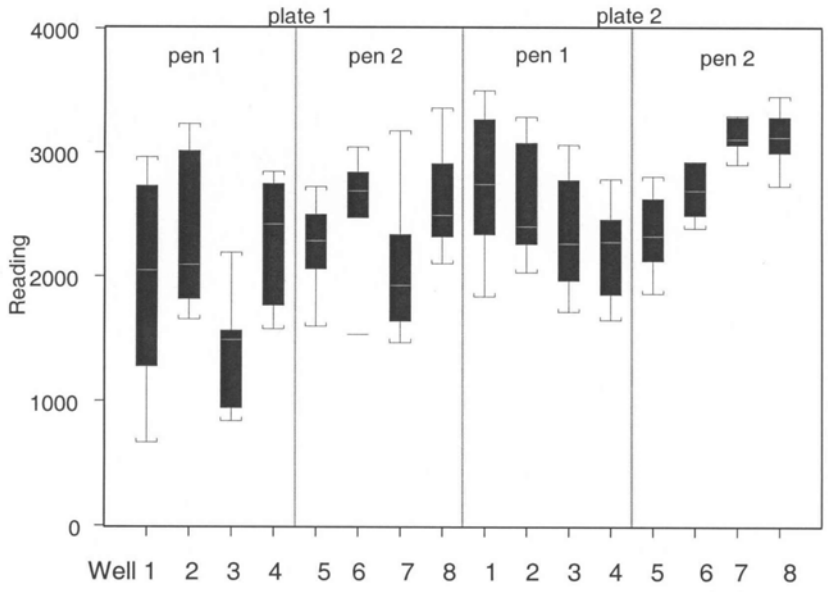

Figure 3. Box plot of the wells within each pen and plate arrangement. This illustration is for G6PD in Liver #3.

Chapter 3

Classical Statistical Approach to Molecular Classification of Cancer From Gene Expression Profiling

J. Lu[†], S. Hardy[†], W. Tao[†], S. Muse[†], B. Weir[†] and S. Spruill[‡]

[†]N.C. State University Bioinformatics Program and [‡]PPGx. Inc.

* Published as a conference paper in the book, *Methods of Microarray Data Analysis*, Kluwer Academic Publishers, Boston 2001.

Abstract

Recent literature regarding microarray technology has focused on the need to incorporate classical statistical practices in experimental design in order to utilize more robust, classical statistical methodologies in data analysis. We have demonstrated that classical statistical methods are applicable to analysis of data previously presented by Golub, *et al* (1999). Our preliminary analysis of all 6817 genes involves simple t-tests for statistically significant separation of means of gene expression level in two cancer types. We select those predictor genes based on the t-values and stepwise discriminant analysis, and evaluate the resulting model's performance in predicting 34 test samples by discriminant analysis. Only two samples were not correctly predicted (samples 61 and 66) with 25 predictor genes we chose. We also evaluate the parsimony of our model by evaluating, through a stepwise method, the minimum number of genes required to maintain a high level of accuracy in predicting cancer types.

Introduction

In an Nature Genetics article Bittner, *et al* (1999) acknowledged that the volume of data obtained from gene expression analysis using microarray presented a "mathematical challenge". This followed an earlier argument by, Duggan *et al* (1999) that in order for microarray research to achieve true understanding of genome function, it needed to recruit the assistance of statisticians and mathematicians to ponder the problems of data analysis. Since these publications were released more than 500 peer-reviewed articles referencing microarray technology have been published, and the number is rapidly climbing. Yet only 36 articles are currently available through PUBMED that seriously address the use of robust statistical methodology to analyze gene expression data.

While microarray technology is somewhat novel, the statistical methods appropriate for analyzing expression data need not be. Recent literature has focused on the need to incorporate classical statistical practices in experimental design in order to utilize more robust, classical statistical methodologies in the analysis of microarray data. We should clarify that for the purposes of this article, our use of the term "classical statistics" is meant to embody any statistical methodology that has been proven and accepted for use in other areas of science. Examples of the use of well-established statistical methodology can be found in articles such as Hilsenbeck, *et al* (1999) who illustrated the use of principal components. Findings of Kaminski, *et al* (2000) are among the many emerging examples of the use of cluster analysis to distinguish cell samples with array expression data. The simple use of t-tests with corrections for

multiple comparisons can be found in Dudoit, *et al* (2000), while Kerr and Churchill (2001) describe analysis of variance approaches.

Golub, *et al* (1999) are among the first to attempt to utilize expression results in a predictive modeling process. The objectives of Golub, *et al* are to use a quantitative process to identify discriminating genes and then develop a predictive model for classifying samples as coming from patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). Golub, *et al* present a molecular classification methodology for predicting tumor class of unknown samples. Their method utilizes a derived correlation coefficient in which the probability of a high density of genes distinguishes a class compared to a random pattern. From this "neighborhood analysis", the predictor is derived from the weight of each gene's "vote" in favor of one of the classes. The sum of these votes determines a prediction strength score, which in turn assigns the sample to a class based on a predetermined threshold.

While Golub and colleagues succeed in demonstrating that their predictive model was successful in distinguishing samples, their approach was a novel twist deviating from well-established statistical methodology. In addition, the final assignment of samples to class is based on an ambiguously determined threshold. The purpose of this article is to demonstrate that results similar to Golub, *et al* are obtainable by employing analyses available in commercial software. We will illustrate the use of standard t-tests and stepwise discriminant analysis as is available in SAS (SAS Institute, Cary, NC) to identify predictor candidate genes, develop a predictive discriminant model and test that model for accuracy.

Materials and Methods

Data manipulation

Golub's study consisted of two datasets: a training dataset containing 38 samples and a test dataset containing 34 samples. In an effort to allow for direct comparability to Golub's results, we maintained these datasets for training and testing in our analyses. Prior to analysis, data from both the training and tests dataset were filtered and standardized according to methods as described in detail on Golub's web site. Briefly, any genes with an "A" labeling call for all 72 samples were excluded from the further analysis. For the remaining genes, intensity readings were re-scaled based on background to allow for comparisons among the different samples.

We chose to exclude all the endogenous control genes from our data analysis, because we assumed that they would be less likely to differentiate samples. Additionally, because readings less than 20 were considered by Golub to be below the reliability level of the instrumentation, we also excluded genes with readings less than 20 in all 38 training samples and all 34 testing samples. After these exclusions, 4892 genes remained. In the remaining set, any individual intensity value less than 20 was set to 20 to avoid having readings below the lower bound. The logarithms of the Golub's re-scaled values were then standardized to mean zero and unit variance across all 72 samples.

Variable (gene) selection

Our ultimate goal was to identify the set of genes from the field of 4892 candidates that could most successfully be used to distinguish between the two leukemia types (ALL and AML). To achieve this goal we employed a three step screening process. Our first step was to reduce this large field of candidate genes to a smaller set for which subsequent statistical analysis would be more computationally feasible. The initial screen was accomplished by computing individual two sample t-statistics to compare the difference in expression levels between the ALL (n=27) and AML (n=11) groups in the 38-sample training set. The test statistics were computed under the assumption of independent samples with potentially different variation in the gene expression. We selected the 250 genes with the most positive t-values and the 250 genes with the most negative t-values to be retained as candidates for further analysis. In other words, the 250 genes exhibiting the greatest differences that tended to have high expression in the ALL group but low expression in the AML group were chosen, and visa versa. This strategy was adopted so that an equal number of genes favoring each of the two leukemia types would be represented.

The second step was to perform a stepwise discriminant analysis using the SAS procedure PROC STEPDISC (SAS, 1990). The stepwise discriminant analysis selects variables (genes) that contribute most to the discriminatory power of the model as measured by Wilks' lambda likelihood ratio criterion. The set of variables is assumed to be multivariate normal with a common covariance matrix. We used the default

significance level entry criterion of 0.15. Both the forward selection method and the stepwise method resulted in a final model with the same 25 variables.

The third step in our gene selection process was to create incrementally smaller subsets of these 25 genes by setting the STOP option in PROC STEPDISC to 20, 15, 10, and 5. This resulted in 5 subsets of genes, each with 25, 20, 15, 10 or 5 genes respectively, to be tested and compared in the final analysis.

Discriminant analysis

The final discriminant models (one for each of the 5 candidate datasets) were constructed using the SAS procedure PROC DISCRIM. A nonparametric method was employed, as opposed to the parametric model inherent in the stepwise discriminant analysis. This method was chosen in order to avoid making the assumption of multivariate normal distribution. The nonparametric method does not rely on this assumption.

Presented here is a basic overview of the nonparametric discriminant algorithm. For more details the reader is referred to the SAS Statistical User's Guide (SAS, 1990). First, a nonparametric multivariate probability density function is estimated for each group; in our case the two groups are the ALL and AML patients in the training dataset. The SAS software allows for different choices of kernel smoothers for generating the probability density functions, and we chose a normal smoothing kernel. The algorithm in this nonparametric method is simple; for each multidimensional observation, $\mathbf{x}$, in a test dataset, the estimated densities, based on the training dataset, are calculated separately for each group. Then, for each sample, a posterior probability of group membership is

calculated for each group based on prior probabilities, the previously discussed density estimates, and the unconditional or overall density estimate. However, in its simplest form, where the prior group probabilities are equal, the observation **x** is classified into the group with the highest density estimate at **x**. The POOL and CROSSVALIDATE options were also used in PROC DISCRIM. When POOL=yes, the generalized squared distances are calculated based on a pooled within-group covariance matrix. The CROSSVALIDATE option provides the cross-validation classification of the training dataset. Each sample in the training set is classified based on the information of estimated group-specific density from all observations in the training when the nonparametric method with normal kernel is specified.

Results and discussion

Selected t-test results appear in Table 1, which shows the 10 genes with largest t-values (highly expressed in AML group) and the 10 genes with the most negative t-values (highly expressed in ALL group). In genes that favored one group or the other, it was observed that the between group differences were largest when the genes favored AML. The 25 genes obtained from stepwise discriminant analysis are listed in Table 2, along with the partial R-square, F value, the p-value from the step where the gene first entered the model. The last column in the table is the average squared canonical correlation (ASCC), a measure of the degree of separation between the groups.

After 25 variables were selected, no other variables met the entry criterion. With these 25 variables, the ASCC is .9999 indicating that the two groups can be completely distinguished. Furthermore, the reader should note that after 6 variables have entered the

model the average squared canonical correlation is .9612 and each subsequent improvement is less than .01. Among 25 genes, eighteen had larger AML means and 7 genes had larger ALL means, which is consistent with the observation that more genes showed high expression in AML than in the ALL group. Figure 1 graphically illustrates the separation of genes. The diagonal reference line indicates where genes would fall if there was no distinction between leukemia types. Genes falling above the reference line have, on average higher expression in AML samples, while those falling below the line have, on average higher expression in ALL samples.

How does this affect the analysis? When classification is done between two groups, if an observation does not fit the profile of one group, it is, by default assigned to the other group. We believe that in this analysis, AML was the primary group in the sense that the genes associated with AML dominated the models. The scientific implication of this in terms of using these genes to classify other types of cancers is not clear, but it should be an important consideration because it is highly likely that while these genes may successfully classify AML leukemia, they could fail at distinguishing between ALL and other types of leukemia for example.

To evaluate how successful the 25 genes are in determining the group membership we first conducted a cross-validation analysis on 38 training samples and then used the model based on the 38 training samples to predict group membership for the 34 testing samples. The cross-validation classification based on these 25 genes was accurate for all 38 training samples. The cross-validation results from the PROC DISCRIM procedure were somewhat optimistic since the group-density estimation was based on all observations in the training set when the nonparametric method with normal

kernel was used. When the test data of 34 samples were classified based on density estimates from the training dataset the classification outcomes on the test dataset were correct for 32 of the 34 samples (see Table 3, column 3). We next ask the question: can we further decrease the number of genes without decreasing the prediction accuracy? Recall that based on using the stepwise discriminant analysis with the STOP option we also selected sets of 20, 15, 10, and 5 genes. After assessing the effectiveness of the 25-gene model, we also, in the same manner, assessed the effectiveness of using the sets of 20, 15, 10, and 5 genes. These evaluations indicated that the 20, 15 and 10 gene models were equally successful in group classification, and the 5 gene model was actually slightly more effective, yielding 33 out of 34 correct classifications (see Table 3, column 4). This indicates that there is some degree of over-fitting occurring in the models with more genes. Of the five genes chosen (highlighted in bold in Table 2), one favored larger ALL means and the other 4 favored larger AML means (See Figure 1, squares). The only case that has been misclassified was sample 66, which was not correctly classified by Golub as well. It has been suggested that the problem may come from the experimental data on sample 66, which had weak hybridization signal for all the genes.

The five genes we finally chose were: Cystatin C (accession M27891), Adipsin (accession M84526), Zyxin (accession X95735), PTH2 parathyroid hormone receptor (accession U25128), and proteasome subunit LMP7 (accession Z14982). The first three genes were also in Golub's gene list. All three of these genes are more highly expressed in AML group than in ALL group samples.

The reader should note that the genes chosen by stepwise discriminant analysis are not necessarily those genes with highest absolute t-values. This is not a surprising result; we would expect to see this only if there was no multicollinearity amongst the genes with highest absolute t-values. Because the variable selection, based on discriminant analyses, takes into account the correlation between genes, it results in a set of genes that collectively account for differences between the groups. If only t-tests were used to select genes, two genes that are highly correlated could easily both be selected. When two genes are highly correlated, adding the second to the model in the presence of the first will not provide any additional improvement because it is supplying redundant information. We did verify that the model we selected was vastly superior to one based only on genes with significant t-test results as is illustrated by the fact that so few genes identified in Table 1 survived the stepwise discriminant analysis. This indicates that the genes are indeed correlated. When there is a high degree of correlation, as exists here, there can easily be other sets of variables that can predict as well as the specific sets chosen by the methods we implemented.

Conclusions

Established statistical methods, found in commercial software, were applied to classify samples from two leukemia types with better prediction outcomes than the methods used by Golub, *et al*. It is notable that only 3 of the 5 genes used in the reduced model are the same as those identified by Golub, suggesting a high level of redundancy among levels of gene expression within the leukemia types.

Perhaps the message to take away from this exercise is that it is not always necessary or prudent to develop novel analytical methods in tandem with novel technology. When a new problem is addressed with a new analysis method, the validity of the technology and the analysis are somewhat confounded. In fact, there is a certain sense of comfort that comes from successfully applying proven analytical methods to new problems. The very foundation of statistics is based on answering the basic question: Is what we observe in the experiment any different from what we would expect to see by random chance? When we lose sight of this basic question, we overcomplicate the problem and then look for complex solutions. In doing so, we fail to recognize that the problem is analogous to other previously solved problems and thus we fail to consider the applicability of standardized analysis methods. For example, the problem of distinguishing tumor type based on gene expression levels is analogous to distinguishing crop types based on remote sensing data (Example 5 in SAS User's Guide). By recognizing this similarity, we can more readily see that discriminant analysis was worth consideration for analyzing the tumor data. We stand to gain a lot by looking to the past for methods to address the problems of the future.

Acknowledgements

This work was supported in part by grants from NIH GM45344 and NSF 99872631.

References

Bittner M., Meltzer P., Trent J.. (1999) Data analysis and integration of steps and arrows. Nature Genetics. 22: 213-215.

Dudoit S., Yang Y. H., Callow M. J., Speed, T. P.. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report #578, www.stat.Berkeley.EDU/users/terry/zarray/Html/matt.html

Duggan D. J., Bittner M., Chen Y., Meltzer P., Trent, J. M.. (1999) Expression profiling using cDNA microarrays. Nature Genetics. 21:10-14.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S.. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286: 531-537.

Golub's web site (www.genome.wi.mit.edu/MPR)

Hilsenbeck, S. G., Friedrichs, W. E., Schiff, R., O'Connell, P., Hansen, R. K., Osborne, K., Fuqua, S. A.W.. (1999) Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. J Nat. Cancer Inst, 91(5): 453-459.

Kaminski N., Allard J. D.., Pittet J. F., Zuo F., Griffiths M. J. D., Morris D., Huang

X., Sheppard D., Heller R. A.. (2000) Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis. PNAS, 97(4):1778-1783

Kerr, M.K. and Churchill, G.A.. (2001) Statistical design of the ananysis of gene expression microarray data. Genet. Res. 77(2):123-128.

PUBMED (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi)

SAS/STAT User's Guide (V6.04), 1990. SAS Institute, Inc., Cary, NC, USA

TABLES AND FIGURES

Table 1. **Results from T-Test**

10 Genes where t-value favors AML				
Gene Accession Number	**Gene**	**Standardized Mean ALL**	**Standardized Mean AML**	**t-value**
M27891_at	Cystatin C	-0.50819	1.24738	-9.49089
U50136_rna1_at	LTC4S	-0.50909	1.24958	-8.48094
X95735_at	Zyxin	-0.43795	1.07497	-8.11403
M55150_at	FAH	-0.42285	1.03791	-7.74088
Y12670_at	LEPR	-0.46825	1.14935	-7.69674
M23197_at	CD33	-0.50289	1.23435	-7.32637
U46499_at	GSTM	-0.46931	1.15193	-7.23349
M27783_s_at	ELA2	-0.48877	1.19972	-7.00355
M81933_at	CDC25A	-0.40481	0.99363	-6.89141
D88422_at	Cystatin A	-0.48973	1.20207	-6.69879
10 Genes where t-value favors ALL				
X66533_at	GUCY Beta-1	0.34207	-0.83963	6.12532
X82240_rna1_at	TCL1	0.34809	-0.85439	6.28326
M89957_at	IGB	0.35314	-0.86681	6.41754
L13278_at	CRYZ	0.38265	-0.93924	6.45551
Z15115_at	TOP2B	0.41605	-1.02121	6.52462
M77142_at	Nucleolysin TIA-1	0.41304	-1.01383	6.56631
M28170_at	CD19	0.41224	-1.01187	7.39117
M31523_at	TCF3	0.4596	-1.12810	7.47356
U22376_cds2_s_at	C-myb	0.49309	-1.2103	7.78074
X52142_at	CTPS	0.40777	-1.00089	8.05098

Table 2. **Genes obtains from Stepwise Discriminant Analysis**

Gene Accession Number	Gene	Partial R-Square	F Value	Pr > F	Average Squared Canonical Correlation
M84526_at	DF D component of complement (adipsin)	**0.6685**	**72.61**	**<.0001**	**0.6685**
Z14982_rna1_at	MHC-encoded proteasome subunit gene LAMP7-E1 gene	0.2677	13.16	0.0009	0.7893
X95735_at	Zyxin	**0.4835**	**33.7**	**<.0001**	**0.9050**
U25128_at	PTH2	0.2609	12.71	0.001	0.9337
M27891_at	Cystatin C	**0.651**	**67.16**	**<.0001**	**0.9476**
X03934_at	T-cell antigen receptor T3-delta	0.1336	5.55	0.024	0.9612
X64364_at	BSG	0.2706	13.36	0.0008	0.9699
X64072_s_at	Leukocyte adhesion protein beta subunit	0.317	16.71	0.0002	0.9768
M98399_s_at	CD36	0.4343	27.64	<.0001	0.9858
L08177_at	CMKBR7	0.5278	40.24	<.0001	0.9910
M62762_at	ATP6C	**0.2707**	**13.36**	**0.0008**	**0.9959**
D38524_at	NT5 5' nucleotidase (CD73)	0.276	13.72	0.0007	0.9969
L19872_at	AHR	0.3186	16.83	0.0002	0.9976
X16832_at	Cathepsin H	0.1674	7.24	0.0108	0.9981
D26579_at	Transmembrane protein	0.3624	20.46	<.0001	0.9987
M27783_s_at	ELA2	0.6022	54.51	<.0001	0.9990
Y00787_s_at	IL-8 Precursor	**0.5408**	**42.39**	**<.0001**	**0.9993**
X03363_s_at	ERBB2	0.1744	7.61	0.0091	0.9995
D79997_at	KIAA0175	0.1373	5.73	0.022	0.9997
S75256_s_at	HNL neutrophil lipocalin	0.2378	11.23	0.0019	0.9998
M12959_s_at	TCRA	0.3591	20.17	<.0001	0.9999
X62654_rna1_at	ME491	0.4282	26.96	<.0001	0.9999
M13792_at	ADA	**0.4184**	**25.9**	**<.0001**	**0.9999**
M95178_at	Alpha-Actinin 1	0.219	10.1	0.003	0.9999
Y00433_at	GPX1	0.2362	11.13	0.002	0.9999

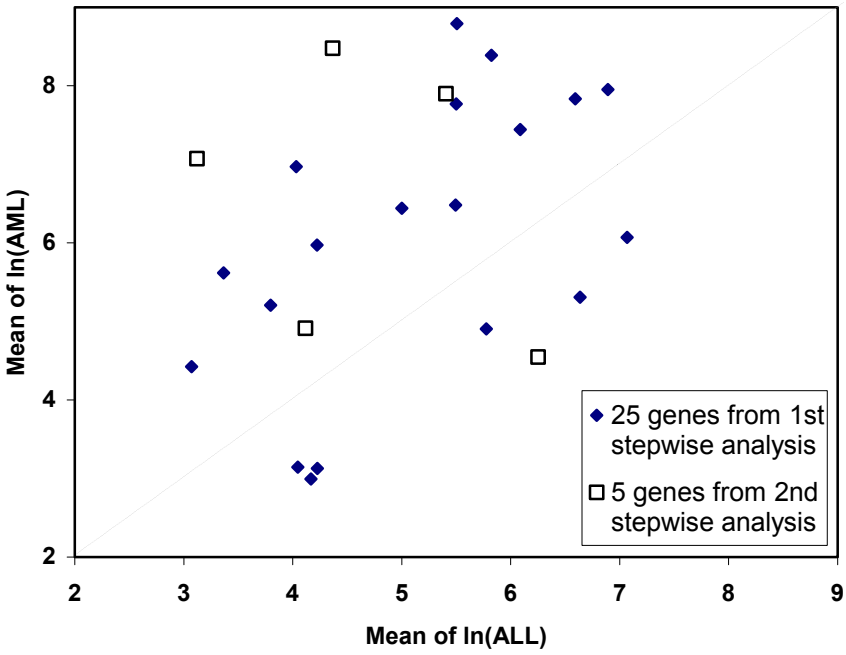

Figure 1. **Scatter plot of means of AML and ALL for 25 genes and 5 genes used in discriminant analyses**

Table 3. **Classification Results from Discriminant Analysis**

Test Sample	Actual Group Classification	25 Gene Model Classification	5 Gene Model Classification
39	ALL	ALL	ALL
40	ALL	ALL	ALL
41	ALL	ALL	ALL
42	ALL	ALL	ALL
43	ALL	ALL	ALL
44	ALL	ALL	ALL
45	ALL	ALL	ALL
46	ALL	ALL	ALL
47	ALL	ALL	ALL
48	ALL	ALL	ALL
49	ALL	ALL	ALL
50	AML	AML	AML
51	AML	AML	AML
52	AML	AML	AML
53	AML	AML	AML
54	AML	AML	AML
55	ALL	ALL	ALL
56	ALL	ALL	ALL
57	AML	AML	AML
58	AML	AML	AML
59	ALL	ALL	ALL
60	AML	AML	AML
61	AML	***ALL***	AML
62	AML	AML	AML
63	AML	AML	AML
64	AML	AML	AML
65	AML	AML	AML
66	AML	***ALL***	***ALL***
67	ALL	ALL	ALL
68	ALL	ALL	ALL
69	ALL	ALL	ALL
70	ALL	ALL	ALL
71	ALL	ALL	ALL
72	ALL	ALL	ALL

Chapter 4

Toward an improved matrix description of E2F binding sites

Abstract

In many cases, sequence analysis of the putative proteins provides no clues on gene function. Accurate identification of regulation regions becomes necessary to uncover the function of genes identified in genome sequencing projects. The first step toward identifying regulatory regions is to search for regulatory elements, the DNA sequences bound by transcription factors. Traditional methods to predict transcription factor binding sites using a position-weighted matrix (PWM) often yield too many false positives. To find an improved binding site predictor, we applied a genetic algorithm (GA) to derive matrices that were trained from a set of true binding sequences and random sequences. Initial studies indicate that the matrix derived show a higher specificity in binding site prediction than the regular PWM within a range of cutoff scores. The binding site of the cell-cycle related transcription factors, E2Fs, was taken as an example to illustrate our method. When both the GA-derived and regular matrices were applied to scan the human gene upstream sequences, the matrix we derived gave significant less predictions than the regular matrix, given the same false negative rate observed in the training dataset.

Introduction

Understanding the regulatory mechanisms of gene expression has been the focus in molecular biology for many years. Although the expression of proteins is regulated through a variety of mechanisms including DNA modification, transcription and protein synthesis, one of the most important steps is at the transcription (RNA) level through the differential binding of transcription factors to the DNA regulatory elements in promoter or enhancer regions. Many transcription factors have been identified with roles involved in tissue-specific or developmental regulation of gene expression. Recent studies have suggested that a single transcription factor is rarely sufficient to explain a pattern of gene expression. The complex, cooperative protein-protein interactions between transcription factors are required to regulate the gene expression (Ptashne and Gann 1997). Although numerous case studies have provided knowledge of the interaction between transcription factors, co-activators (co-repressors), and DNA regulatory elements, we are far from understanding such complex regulation process (Wasserman and Fickett 1998).

Identifying the transcription factors and the corresponding binding sites involved in controlling gene expression is laborious and challenging because of the complexity of gene regulation. During the last decade or so, computational tools have been applied to predict the sequence signals on the genomic DNA regions based on information from limited experimental data (Fickett and Hatzigeorgiou 1997; Crowley et al 1997; Wasserman and Fickett 1998; Kel et al 2001). Although computational prediction is still at the early stages in accurately modeling the whole process, significant progress has been made to assist the experimental identification of the regulatory factors and elements.

This is intriguing because computational prediction is rather cheap compared with conducting experiments in a laboratory. With a number of completely sequenced genomes available, there is an increasing need to automatically or semi-automatically predict the DNA regulatory elements. First, for a given DNA sequence arising from genome sequencing project, it would be beneficial to locate the regulatory regions in the genomic sequences and potential transcription binding sites that confer temporal and spatial expression patterns for the uncharacterized genes. Secondly, for genes with known expression patterns (e.g., from microarray experiments), identifying the regulatory regions or DNA elements may help us to understand the patterns of gene expression (Wasserman and Fickett 1998).

One area of research in identifying regulatory regions is to discover the transcription factor binding sites in a set of genes with similar expression patterns. For instance, there is great interest in finding whether all or some of the genes in a gene cluster resulting identified by a microarray experiment share known or novel transcription factor binding sites. A number of computational methods have been developed, e.g. Gibbs Sampler (Lawrence et al. 1993), MEME (Bailey and Elkan 1994), Consensus (Hertz and Stormo 1999). These methods are based on either Gibbs Sampling or Expectation and Maximization (EM) algorithm. The output of these programs is a set of ungapped multiple sequence alignments (also called motifs), which are presumably corresponding to transcription factor binding sites. Given these characterized or predicted binding sites, the next stage of promoter modeling is to combine all the binding site information, along with the spatial relationships among them, to predict regulatory regions conferring temporal or spatial control. For instance, Meta-MEME constructs

Hidden Markov Models (HMM) based on the motifs identified from MEME to model promoter regions (Grundy et al 1997). Crowley et al (1997) constructed a Bayesian model based on a Hidden Markov Chain for locating control regions in genomic DNA. A prerequisite condition of their model is the availability of a list of DNA binding site sequences that can act as signals in transcriptional control. FastM combines a search strategy for individual transcription binding sites with a distance correlation function (Klingenhoff et al 1999). This program allows fast detection of co-operative binding sites in a promoter or enhancer region. In addition, Wassermann and Fickett (1998) applied logistic regression methods to identify the regulatory regions of muscle-specific genes. Signals from several individual transcription factor binding sites were combined into one recognition function, which was subsequently used to identify muscle specific regulatory modules in uncharacterized genomic DNA sequences.

Statistical modeling of gene regulatory regions relies on accurate representation of transcription factor binding sites. Although the binding sites are very short compared to gene coding sequences, an accurate model representation of these binding sites is not an easy task. The difficulty is due to the limited number of experimentally confirmed binding sequences for each binding site, and the relatively large number of parameters that need to be estimated due to a number of positions in each binding site. The early method of searching for putative binding sites was through IUPAC coded string (Cornish-Bowden 1985), which is still implemented in current sequence software packages such as GCG (Womble 2000). For example, an IUPAC string SCAAK represents all possible combinations of GCAAG, CCAAG, GCAAT and CCAAT. An IUPAC-based search program runs very fast and only requires the input of IUPAC code.

However, the string representation of binding site has obvious shortcomings because no quality evaluation is given for each match (Quandt et al 1995). Currently, the most widely used method for representing transcription factor binding sites is position-weighted matrix (PWM). A PWM is derived from an ungapped multiple alignment on a group of known binding sites for a specific or a family of transcription factor(s). A weight is assigned to each position and a specific nucleotide, and the sequence match score is the sum of base weights of all positions. The match score can be interpreted as an estimation of the binding energy of a transcription factor to the corresponding binding site (Berg and vonHippel 1987; Stormo and Fields 1998). A number of programs are available for searching the potential transcription binding sites using PWMs, such as ConsInspector (Frech et al 1993), MATRIX SEARCH (Chen et al 1995), and MatInspector (Quandt et al 1995). The PWM has been successfully used in predicting transcription factor binding sites in bacteria (Stormo et al 1993) and in eukaryotic organisms (Frech et al 1997). For example, Tronche et al (1997) applied a PMW representing HNF1 binding sites to scan human genome sequences and found that 95% of the predicted sites could be bound by HNF1 in vitro. However, there are cases where no strong correlation was observed between PWM-based prediction and the experimental determined sites (Frech 1997). A common problem associated with PWM methods is the high rate of false positives (Lavorgna et al 1999). The PWM method assumes independence between positions within the binding site, which may not be valid in some cases. In addition, the known binding site sequences used to derive a PWM may not be independent of each other due to the phylogenetic relationships between genes. Therefore, the estimates of nucleotide frequencies in PWM can be biased.

In this study, we are interested in improving the matrix representation of transcription factor binding sites. We propose an alternative approach for deriving the PWM using genetic algorithm. The E2F binding site was take as an example to illustrate our method. The E2Fs are a family of transcription factors that play key roles in cell-cycle control, and they have been extensively studied in recent years (Dyson 1998). A computer-assisted identification of the E2F binding sites has also been reported (Kel et al 2001). Here, we took a different approach to finding the representing matrix of E2F, which does not depend on the multiple sequence alignment, but was instead derived from an optimization procedure through a genetic algorithm (GA). The matrix we derived can be used to scan the genomic sequences for potential E2F binding sites with lower false positive rate than the regular PWM in a range of cutoff values.

Datasets and Methods

Data sets

a. E2F binding sequences (target sequences)

A set of E2F binding sites was downloaded from the CYCLE-TRRD database (http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/) (Kel et al 2001). Three additional binding sites in two genes were added (Pearson et al 1991; Sears et al 1997). Sites with identical sequences were removed. All these binding sites have been experimentally confirmed. The binding site sequences were truncated to 12 base pairs with the core CG dinucleotides in the central position and five base pairs on each side (Kel et al 2001).

The total number of sequences obtained was 50. Next, we randomly divided the dataset into a training (30 sequences) and a testing (20 sequences) subset (dataset train_30 and test_20). To examine the effect of random split on the results, we repeated the random split two more times, resulting in 3 sets of training and testing subsets (dataset split_1, split_2, and split_3).

b. Random and exon sequences (non-target sequences)

We also created a set of random sequences containing 1000 short sequences, each of which is 12 base-pairs long (dataset random_1000). These random sequences were generated according to the base composition of nucleotides in the training sequences. frequencies equal to those in the set train_30. In addition, we downloaded the vertebrate second exon sequences from the TRANSFAC database (http://transfac.gbf.de/TRANSFAC/). So far, very few regulatory elements have been found in this region. Thus, the second exon sequences could serve as the negative controls (dataset exon_100k) (Pickert et al 1998).

Searching for E2F binding sites with PWM

We implemented the weight matrix method described by Kel et al (2001). Basically, a sequence score *q* was calculated for each position of a sliding window of size 12 within a given DNA sequence. The *q* score was computed for each subsequence *S* according to the following formula:

$$q(S) = q(b_1, b_2, ..., b_l) = \frac{\sum_{i=1}^{l} I(i) f_{ib_i} - \sum_{i=1}^{l} I(i) f_i^{\min}}{\sum_{i=1}^{l} I(i) f_i^{\max}}$$

where b_i is the nucleotide in the position i (i = 1, 2, …, l); l is the length of a matrix; f_{ib_i} is the frequency of the nucleotide b_i in the i^{th} position in the weight matrix; $f_i^{\min}$ and $f_i^{\max}$ are the minimum and maximum nucleotide frequency at position i respectively. $I(i)$ is the information vector which gives a relative value of position conservation. For each position i, the information score is calculated as follows:

$$I(i) = \sum_{B=\{A,C,G,T\}} f_{i,B} \ln(4 f_{i,B}) \qquad i = 1, 2, \ldots, l$$

where $f_{i,B}$ is the frequency of nucleotide B at position i. No restrictions were added on each position i, such as the forbidden nucleotide that was described by Kel et al (2001).

Matrix optimization by GA

A genetic algorithm is a search method that has been used in many optimization problems (Holland 1975; Forrest 1993). GA is based on the idea of population genetics. It starts with a population of "chromosomes". Each "chromosome" represents one possible solution to a problem desired. Through mutation, crossover, and selection from one generation to the next, a solution to the problem desired may be obtained. There are various ways to design and implement GA (Forrest 1993).

In this study, the form of GA is as follows (Figure 1). A weight matrix was represented by a string of binary digits (bits), which was subsequently decoded into integers. The integers were further scaled to values between 0 to1 so that the sum of the values at each position was 1. We used 8 bits to represent each element in the matrix. Thus, the total number of bits on a chromosome was 384 (4*12*8) for the matrix representing the E2F binding site. A total of 200 such chromosomes were generated to form the initial population. A list of parameters used is listed in Table 1. The mutation rate (0.01) was the probability that a mutation (from 1 to 0 or vise versa) occurred at any bit position on a chromosome. Next, for each chromosome, the corresponding matrix was used to scan the training set sequences applying the methodology described in the above section. Each sequence was scanned twice, one from each direction (5'-3' and 3'-5'). The larger score was taken as the score for the sequence. The sequence scores for all target and non-target sequences were then ranked in descending order separately. The difference between the average scores from the bottom 15% of the target sequences (the dataset, train_30) and 5% of the non-target sequences (the dataset, random_1000) was used as the fitness score for the corresponding matrix. The next stage was chromosome selection, in which the best chromosome from each population (a total of 20) at the current generation was selected and replaced the chromosomes with worst performance (i.e. with low fitness scores) in each population. This process evolved for 1000 generations and the best chromosome (matrix) from the last generation was saved and evaluated.

Extraction of upstream sequences of human genes

We extracted the upstream regions of genes on the NIEHS's Toxchip (http://dir.niehs.nih.gov/microarray/). First, the gene symbol and the chromosome number were extracted from the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/) according to the gene accession number. Next, the upstream sequences were extracted from human chromosome DNA sequences (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens, updated September 2001). We first searched for the annotated start position of mRNA, coding region (CDS) and first intron, and then a DNA fragment with pre-defined sequence length was extracted based on the mRNA start position. For genes on the complementary strand, the reverse complemented sequences were used. Among the 1900 cDNA clones printed on the Toxchip array, the upstream sequences of 1291 genes were extracted from GenBank database.

Computer programs

The implementation of regular PWM method described by Kel et al (2001) and the GA-optimization program were written in C. The sequence extraction programs were written in Perl. Other sequence processing programs were written in Perl. All programs were run on SGI workstations.

Results

An initial study on the cAMP-responsive element binding (CREB) protein binding site prediction

CREB proteins are transcription factors involved in controlling gene transcription in response to signals such as cAMP activators and growth factors (Mayr and Montminy

2001). For example, the CREB protein can be phosphorylated and activated by insulin-like growth factor-1 (IGF-1) induced signal transduction cascade (Oh et al 2002). A significant number of target genes of CREB have been identified and shown diverse roles ranging from metobolic regulation, transcription, to cell cycle control (Mayr and Montminy 2001). Earlier work has shown that the CREB is one of the targets for protein kinase AKT (Du et al 1998) and was phosphorylated in human breast epithelial cells treated with IGF-1 (Oh et al 2002). In this initial study, we are interested in identifying the CREB targets among genes differentially expressed in response to the IGF-1 treatment.

The design of the cDNA microarray experiment was described (Oh et al 2002). At each time point, a set of differentially expressed genes has been identified (at 95% confidence level). For example, 97 among 1920 genes were differentially expressed in cells treated with IGF-1 for 2 hours. Since other factors besides CREB can be activated by IGF-1/IGF-1R as well (for instance, FKHRL1, AFX and NFkB), the question is, how many genes are regulated by CREB?

We approached this problem by searching for CREB binding sites in sequences of the promoter or gene upstream regions. From the previous study on the target genes of CREB, most CREB binding sequences (about two-thirds) are located between -5 to -150 with the consensus "TGACGTCA" (Mayr and Montminy 2001). For a better coverage, the gene upstream region from 0 to -500 was extracted based on the GenBank annotation of human genome. Briefly, the symbol and chromosome number of each genes were extracted from the UniGene database according to the gene accession number first. Then based on the information of the chromosome number and the gene symbol, the position

of transcription start site was found and the 500 base-pair sequences upstream of that position were extracted from the human chromosome genome database (see details in Materials and Methods). For the group of genes differentially expressed at 2-hrs time-point, the upstream sequences of 67 genes were available in GenBank.

The original MatInspector software was used to scan for the CREB binding sequences in the gene upstream sequences. It should be noted that original MatInspector used a score function with subtle difference from the function described in Methods. The thresholds of both core and full site sequences may be defined to control the false positive rates (Quandt et al 1995). The weight matrix of CREB binding site in the TRANSFAC database (http://transfac.gbf.de/TRANSFAC/) was used to scan for the potential new CREB binding elements. Results showed that a total of 32 putative CREB binding sites in 67 genes differentially expressed at 2 hours time-point were identified (with cutoff values of 0.95 and 0.85 for core and full site, respectively). Although some of predicted sites are real (for instance, sites in V-fos and ornithine decarboxylase 1), many of the putative sequences are more likely to be false positives. This initial study on CREB binding site prediction provided an example that the information from single binding site is not adequate to predict the target genes of transcription factors. The modeling on regulatory regions (which contain multiple binding sites) becomes necessary, which is the also the motivation of the current study.

Construction of E2F PWMs

A list of the experimentally confirmed E2F binding sites from vertebrates is shown in Table 2. All the sequences have been manually aligned (Kel et al 2001). The

totally conserved CG di-nucleotides are in the central position, and the 5' and 3' end sequences are enriched with T and A respectively. Also, E2F sites were found preferentially close to the transcription starting site in the range of -300 to +100 (48 out of 50). From the alignment, a regular PWM based on one dataset train_30 was constructed and shown as an example (Figure 2). We also showed the information vector value indicating the relative conservation of each position.

To improve the performance of PWM method, matrices were trained from a set of true binding sequences (set train_30) and a set of random sequences (random_1000) through GA (see details in Methods). The parameters used in GA were listed in Table 1. In addition, we plotted the fitness scores as a function of number of generations from 4 independent runs with the same parameter settings (Figure 3). It can be seen that the fitness scores stabilized after around 800 generations. Considering the computation time required, we chose to stop the whole process at the1000^{th} generation. An example of GA-derived matrix, along with the information vector values, was shown in Figure 4. As expected, the GA is capable of identifying a matrix that is similar to that obtained from a multiple alignment and counting method (see Figure 1. B). The most conserved nucleotide was often assigned with the highest weight in each position. However, the difference is obvious at some positions. For instance, we found that the information value at the second position was relatively small in GA-derived matrices from several independent runs tested.

Matrix comparison

The performance of matrices derived was evaluated based on the dataset test_20 and exon_100kb. The second exon sequences were chosen as negative sets since very few active binding sites were found in second exons so far (Pickert et al 1998). Therefore, any binding sites identified in this region should be considered as false positives. We compared the performance of matrices derived from GA and from the regular count-based method using a selectivity ration statistic (Fujibuchi et al 2001). Selectivity for a particular matrix is measured as the fraction of correctly predicted binding sites out of all sites predicted when the matrix was applied. For example, in the first case of Table 3, the matrix derived from GA predicted 57 sites in the 100kb exon2 sequences and missed 2 of the 20 true sites; then the selectivity value is 0.24 ((20-2)/(57+20-2)). Similarly, the regular matrix derived from the same training set showed a selectivity value of 0.12 ((20-2)/(136+20-2)). Then, the selectivity ratio is about 2 (0.24/0.12) (see Table 3). Any selective values greater than 1.0 indicate that the GA derived matrix has better sensitivity than the regular matrix. All the comparisons were conducted between matrices from three independent data splitting (see Methods). From Table 3, we can see that the GA-derived matrix generally has better performance than regular matrix within a range of cutoff values at which the low false negative rates were observed. In some cases, using the GA-derived matrix can cut the false positive rate by half without significant loss of sensitivity. However, if high cutoff values (which were often associated with high false negative rates) were chosen, the performance of GA-derived matrix may not perform better than the regular method (data not shown).

Scan for potential E2F sites in the upstream regions of human genes

We are interested in comparing the regular matrix and GA-derived matrix when the gene promoter or upstream sequences were examined. The upstream sequences of a set of human genes were extracted according to the annotation in GenBank. We searched the E2F binding sites in the sequence region between –300 to +100 since more than 90% real E2F binding sites identified so far are located in this region. A new matrix was derived through GA training from the dataset including all 50 known binding sequences. When such matrix was applied, a total of 605 sites in 423 genes were predicted as potential binding sites in the sequences extracted (the total sequence length is 387.3Kb). We chose a cutoff value at which 10% false negative rate was allowed in the training sequences. A list of predicted binding sites, along with the corresponding gene name, accession number and the binding site positions relative to the transcription start sites are shown in Table 4. In contrast, applying a regular matrix with the same cutoff criteria, 1086 sites in 618 genes were predicted in those upstream sequences. Most likely, a significant number of binding sites predicted from both false positives. However, the GA-derived matrix gave significant less number of predictions, which implies that the matrix derived from GA may have better discrimination power than that derived from the regular count-based method.

Discussion

The accurate identification of regulatory regions is both experimentally and computationally challenging. The experimental study on gene regulation is rather time-consuming and costly. It would be impractical to study the regulation mechanisms for every gene, given the complexity of gene regulation and the numbers of genes in a

genome. The coding sequence prediction often provides no clues about the function of the gene. Therefore, the computational prediction of gene regulatory regions is becoming increasingly important in studying gene functions.

Very often, the first step of identifying regulatory regions is to search for potential transcription factor binding sites in genomic sequences. The methods based on position weight matrix have been shown to be effective in identifying sequences that may be targets of a specific transcription factor (Wasserman and Fickett 1998; Kel et al 1999). However, many binding sites predicted were not active in vivo (Tronche et al 1997). The individual binding of a transcription factor to a DNA element is rarely sufficient to confer temporal and spatial gene expression in eukaryotes. Cooperation between multiple transcription factors at multiple sites has been shown to be essential in a number of case studies of tissue specific gene regulation (Tronche et al 1997; Kel et al 1999; Roulet et al 2000). Thus, to model promoter region, one not only needs to know the binding sites for individual factors, but also the orientation, and space between the sites (Wasserman and Fickett 1998; Kel et al 1999). Recently, progress has been made in combining the information into a predictive model, and then using the model to predict a regulatory region in genomic sequences (Crowley et al 1997; Wagner 1999; Frith et al 2001). The basic elements in these models are the putative sites predicted from using PWMs, in which a low cutoff score was often chosen. The sites with low binding affinity (often with a low PWM score) can play a significantly role in gene regulation when the cooperation between multiple binding sites occurs. However, the PWM often gives many false positive when a small cutoff value is used (Kel et al 2001).

In this study, we attempt to improve the weight matrix method, thus, reduce the number of false positives in binding site prediction. We applied a genetic algorithm to train a weight matrix that could distinguish between the target and non-target sequences. Unlike the traditional PWM that is derived based on a set of target sequences only, the GA-derived matrix is trained using both target and non-target sequences. Our preliminary results show that the matrix derived from GA gives fewer false positives than the regular PWM without affecting the false negative rate. As expected, the GA-derived matrix is similar to the regular PWM. However, significant differences exist. For instance, the second position in the matrix has lower information score in the GA-derived matrix compared to the regular PWM, suggesting that this position is less important in GA-derived matrix in binding site prediction. It is possible that this position may be correlated with other positions and the GA may implicitly take such correlation into account. In addition, when both GA-derived and regular matrices were applied to scan the sequences in dataset train_30, we found that using GA-derived matrix gave fewer sequences with either extremely high or extremely low scores (data not shown). This suggests that more weights were assigned to those less represented sequences in the training set in the GA-derived matrix.

When the GA-derived matrix trained from all 50 known binding sequences was applied to scan for potential E2F binding sites, we identified, on average, 1.6 sites per kilobase in the upstream regions of a set of human genes. In contrast, when the dataset exon_100k sequences were scanned with the same matrix and the cutoff value, 0.37 binding sites were predicted in each kilobase sequences. This indicated that the frequency

of E2F binding sites is much higher in the gene upstream regions than in the second exon regions.

The GA-derived matrix reduces the number of false positives, however, the accurate identification of E2F regulatory elements needs other information and depends on the context of a promoter. Statistical modeling of regulatory regions is necessary (Crowley et al 1997; Grundy et al 1997). Cooperative binding between the transcription factors can be important, for instance, the interaction between E2Fs and Sp-1 (van Ginkel et al 1997). The transcription profiling experiments on E2F can be extremely helpful in determine which is one of the E2F targets and whether the E2F binding sites within the gene are functional (Wells et al 2002).

Although the GA-based method is promising, a number of issues remain to be addressed. It would be necessary to explore different fitness functions, which can have significant influence on the matrix and the computation time required. Although the matrices derived have improved performance, the further investigation on choosing the GA parameters (such as the population size and the number of generations) is needed to address the GA convergence problem.

Conclusions

The detection of functional transcription factor binding sites requires both a well-defined matrix representation for the binding sites and the context information in a regulatory region. This study attempts to address the first question. Preliminary studies indicate that a genetic algorithm could be applied to derive a matrix with better discrimination power than the regular count-based method. A significant number of false

positives could be removed when the GA-derived matrix was used. Combined with the context information in neighboring sequences, the GA-derived matrix might be useful in searching for the potential binding sequences with high sensitivity and specificity in genomic DNA.

References

Bailey T. L., Elkan C.. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 2:28-36.

Berg O. G., von Hippel P. H.. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol. 193(4):723-50.

Chen Q. K., Hertz G. Z., Stormo G. D.. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comput Appl Biosci. 11(5):563-6.

Cornish-Bowden A.. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 13(9):3021-30.

Crowley E. M., Roeder K., Bina M.. (1997) A statistical model for locating regulatory regions in genomic DNA. J Mol Biol. 268(1):8-14.

Du K., Montminy M.. (1998) CREB is a regulatory target for the protein kinase Akt/PKB. J Biol Chem. 273(49):32377-9.

Dyson N.. (1998) The regulation of E2F by pRB-family proteins. Genes Dev. 12(15):2245-62.

Fickett J. W., Hatzigeorgiou A. G.. (1997) Eukaryotic promoter recognition. Genome Res. 7(9):861-78.

Forrest S.. (1993) Genetic algorithms: principles of natural selection applied to computation. Science. 261(5123):872-8.

Frech K., Quandt K., Werner T.. (1997) Finding protein-binding sites in DNA sequences: the next generation. Trends Biochem Sci. 22(3):103-4.

Frith M. C., Hansen U., Weng Z.. (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics. 17(10):878-89.

Fujibuchi W., Anderson J. S., Landsman D.. (2001) PROSPECT improves cis-acting regulatory element prediction by integrating expression profile data with consensus pattern searches. Nucleic Acids Res. 29(19):3988-96.

Grundy W. N., Bailey T. L., Elkan C. P., Baker M. E.. (1997) Meta-MEME: motif-based hidden Markov models of protein families. Comput Appl Biosci. 13(4):397-406.

Hertz G. Z., Stormo G. D.. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.
Bioinformatics. 15(7-8):563-77.

Holland, J. H. (1975) Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, IL.

Kel A., Kel-Margoulis O., Babenko V., Wingender E.. (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. J Mol Biol. 288(3):353-76.

Kel A. E., Kel-Margoulis O. V., Farnham P. J., Bartley S. M., Wingender E., Zhang M. Q.. (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. J Mol Biol. 309(1):99-120.

Klingenhoff A., Frech K., Quandt K., Werner T.. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. Bioinformatics. 15(3):180-6.

Lavorgna G., Guffanti A., Borsani G., Ballabio A., Boncinelli E.. (1999) TargetFinder: searching annotated sequence databases for target genes of transcription factors. Bioinformatics. 15(2):172-3.

Lawrence C.E., Altschul S. F., Boguski M. S., Liu J. S., Neuwald A. F., Wootton J. C.. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262(5131):208-14

Mayr B., Montminy M.. (2001) Transcriptional regulation by the phosphorylation-dependent factor CREB. Nat Rev Mol Cell Biol. 2(8):599-609.

Oh J.S, Kucab J. E., Bushel P. R., Martin K., Bennett L., Collins J., DiAugustine R. P., Barrett J. C., Afshari C. A., Dunn S. E.. (2002) Insulin-like growth factor-1 inscribes a gene expression profile for angiogenic factors and cancer progression in breast epithelial cells. Neoplasia. 4(3):204-17.

Pearson B. E., Nasheuer H. P., Wang T. S.. (1991) Human DNA polymerase alpha gene: sequences controlling expression in cycling and serum-stimulated cells. Mol Cell Biol. 11(4):2081-95.

Pickert L., Reuter I., Klawonn F., Wingender E.. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering. Bioinformatics. 14(3):244-51.

Ptashne M., Gann A.. (1997) Transcriptional activation by recruitment. Nature. 386(6625):569-77.

Quandt K., Frech K., Karas H., Wingender E., Werner T.. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23(23):4878-84.

Roulet E., Bucher P., Schneider R., Wingender E., Dusserre Y., Werner T., Mermod N.. (2000) Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. J Mol Biol. 297(4):833-48.

Sears R., Ohtani K., Nevins J. R.. (1997) Identification of positively and negatively acting elements regulating expression of the E2F2 gene in response to cell growth signals. Mol Cell Biol. 17(9):5227-35.

Stormo G. D., Strobl S., Yoshioka M, Lee J. S.. (1993) Specificity of the Mnt protein. Independent effects of mutations at different
positions in the operator. J Mol Biol. 229(4):821-6.

Stormo G. D., Fields D. S.. (1998) Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci. 23(3):109-13.

Tronche F., Ringeisen F., Blumenfeld M., Yaniv M., Pontoglio M.. (1997) Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. J Mol Biol. 266(2):231-45.

van Ginkel P. R., Hsiao K. M., Schjerven H., Farnham P. J.. (1997) E2F-mediated growth regulation requires transcription factor cooperation. J Biol Chem. 272(29):18367-74.

Wagner A.. (1997) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics. 15(10):776-84.

Wasserman W. W., Fickett J. W.. (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol. 278(1):167-81.

Wells J., Graveel C. R., Bartley S. M., Madore S. J., Farnham P. J.. (2002) The identification of E2F1-specific target genes. Proc Natl Acad Sci U S A. 99(6):3890-5.

Womble D. D.. (2000) GCG: The Wisconsin Package of sequence analysis programs. Methods Mol Biol. 132:3-22.

TABLES AND FIGURES

Table 1. A set of GA parameters used

Population size	200
Number of niches	20
Number of generations	1000
Mutation rate	0.01
Bit length	8

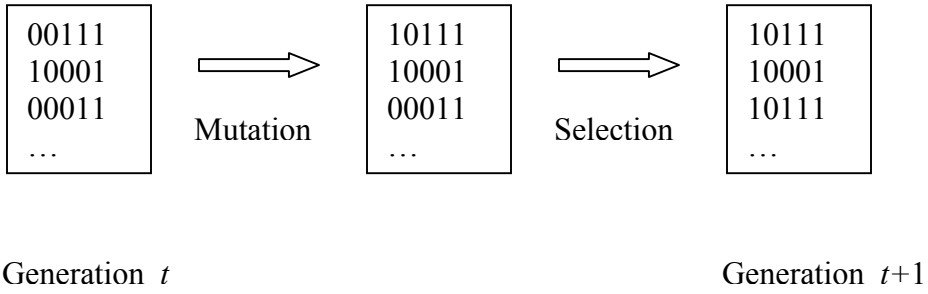

Figure 1. A diagram showing the evolution based on genetic algorithm.

Each string of 0 and 1 represents a chromosome, and multiple chromosomes form a population. Here only mutation was applied (no crossover).

A	2	1	0	1	0	0	0	0	1	15	17	15
C	3	1	1	17	2	30	0	20	10	7	1	3
G	1	1	0	12	28	0	30	10	16	3	5	9
T	24	27	29	0	0	0	0	0	3	5	7	3

A

A	0.07	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.03	0.50	0.57	0.50
C	0.10	0.03	0.03	0.57	0.07	1.00	0.00	0.67	0.33	0.23	0.03	0.10
G	0.03	0.03	0.00	0.40	0.93	0.00	1.00	0.33	0.53	0.10	0.17	0.30
T	0.80	0.90	0.97	0.00	0.00	0.00	0.00	0.00	0.10	0.17	0.23	0.10
I(i)	0.68	0.95	1.24	0.59	1.14	1.39	1.39	0.75	0.34	0.17	0.31	0.22

B

Figure 2. An example of PWM based on a dataset train_30. (**A**) The nucleotide counts each position (**B**) Frequency matrix and the calculated information vector values, I(i).

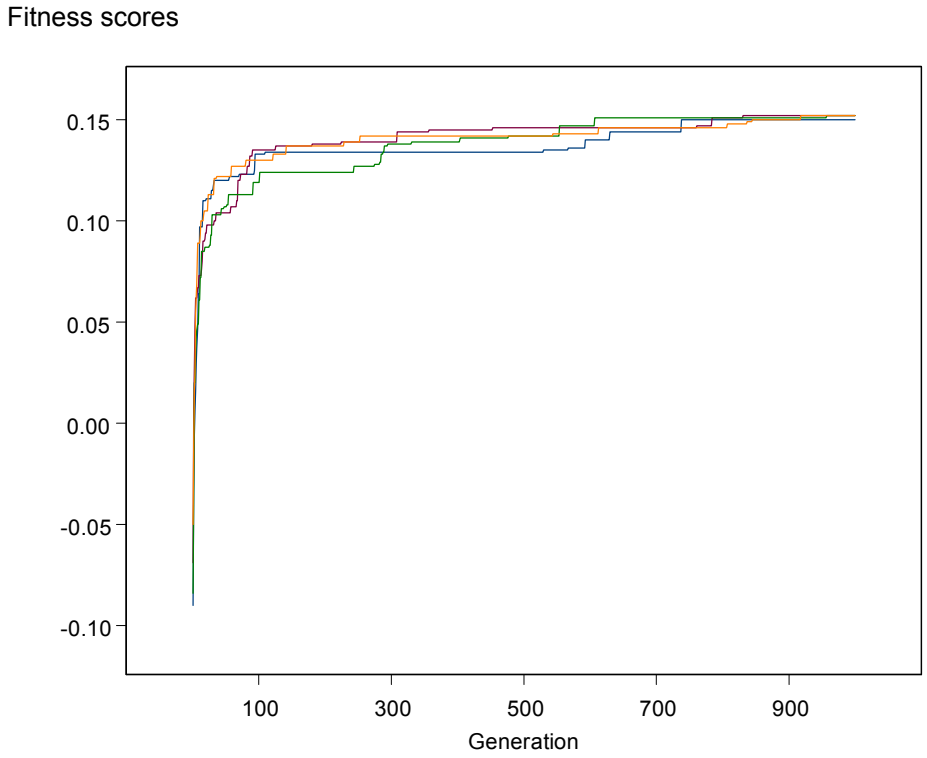

Figure 3. The increase of fitness scores based on 4 independent runs

Table 2. A list of experimentally confirmed E2F binding sites

Gene name	Site Acc. Name	Position[a]	Sequences of E2F sites	Organisms[b]
BMYB	S2353	-212	CTTGGCGGGAGA	Mm
CAD	S1338	+74	TTCGGCGCGCGG	Ma
CDC2	S1870	-131	TTTCGCGCTCTA	Hs
CDC2	S1874	-30	TTTAGCGCGGTG	Hs
CDC2	S1879	-130	TTTCGCGCTCTG	Rn
CDC6	S3147	-47	TTTGGCGGGAGG	Hs
CDC6	S3148	-14	TTTGGCGCGAGC	Hs
CDC25A	S3672	-63	TTTGGCGCCAAC	Hs
CMYB	S5928	-68	TTTGGCGGGAGG	Hs
CMYC	S1835	-67	CTTGGCGGGAAA	Hs
CMYC	S1836	-47	GATCGCGCTGAG	Hs
CYCA	S1853	-42	AGTCGCGGGATA	Hs
CYCD1	S1858	-53	TTTGGCGCCCGC	Hs
CYCE	S3256	-24	CCTCGCGCCCGC	Mm
CYCE	S1909	-20	TTCCGCGCGCAG	Hs
CYCE	S1910	+3	TGTCCCGCTCTG	Hs
DHFR	S2365	-10	TTTCGCGCCAAA	Mm
DHR	S2260	-14	TTTCGCGCCAAA	Hs
DHR	S88	-64	TTTCGCGCCAAA	Cs
E1A	S3676	-291	TTTCGCGCGGTT	Ad5
E1A	S3677	-228	TTTCGCGGGAAA	Ad5
E2AE1	S3673	-68	TTTCGCGCTTAA	Ad2
E2AE1	S3674	-48	TTTCGCGCCCTT	Ad2
E2F1	S1788	-31	TTTCGCGGCAAA	Hs
E2F1	S1789	-14	TTTGGCGCGTAA	Hs
E2F1	S1786	-46	TTTCGCGGCAAA	Mm
E2F1	S1787	-26	TTTGGCGCGTAA	Mm
E2F1	S3143	-21	TTTCGCGGCAAA	Cc
E2F1	S3144	+2	TTTGGCGCGCAA	Cc
E2F3	S5798	-175	TTTCGCGGGAGG	Mm
E2F3	S5798	-155	CTTGGCGCGTAA	Mm
EBNA1	S2187	+218	GATGGCGGGTAA	EBV
H2A1	S2288	-55	TTTCGCGCCCAG	EBV
H2AX	S2378	-256	TTTCGCGCGCTC	Hs
HTF9A	S5964	-121	TTTGGCGGGAAG	Mm
HTF9A	S5968	-35	TTTCCCGCCGCT	Mm
NMYC	S1926	-142	TTTGGCGCGAAA	Mm
NMYC	S1927	-127	TTTGGCGCCTCC	Mm
ORC1	S3257	-12	ATTGGCGCGAAG	Mm
P107	S1343	-21	TTTCGCGCGCTT	Hs
P107	S1345	-10	TTTGGCGCAGGT	Hs
PCNA	S2334	+646	TTTCGCGCCAAA	Hs
RBG	S1590	+84	TTTCCCGCGGTT	Hs
TK	S3421	-101	TCTCCCGCCAGG	Hs

Table 2 (*continued*)

Gene name	Site Acc. Name	Position[a]	Sequences of E2F sites	Organisms[b]
TK	S1914	-82	GTTCGCGGGCAA	Mm
UDG	S2338	+80	TTTCCCGGTTGA	Hs
UDG	S3495	-115	TTTGCCGCGAAA	Mm
POLA	Not listed	-135	TTTGGCGCCCTG	Hs
E2F2	Not listed	-46	TTTGGCGCTAAA	Hs
E2F2	Not listed	-61	TTTCGCGGCACG	Hs

[a] The binding site position is given relatively to transcription start site, except for S2353 and S88, which are the position relative to translation start sites.

[b] The abbreviations of organisms: Cc, *Coturnix coturnix* (quail); Cs, *Cricetulus sp*. (Chinese hamster); Ma, *Mesocricetus auratus* (golden hamster); Mm, *Mus musculus*; Hs, *Homo sapiens*; Rn, *Rattus norvegicus*. Ad2, Ad5 and EBV are abbreviated for adenovirus 2, 5 and Epstein-Barr virus respectively.

A	0.33	0.36	0.31	0.19	0.06	0.17	0.31	0.07	0.29	0.35	0.61	0.22
C	0.16	0.20	0.03	0.53	0.38	0.70	0.04	0.47	0.27	0.19	0.08	0.30
G	0.02	0.07	0.07	0.27	0.56	0.05	0.65	0.44	0.36	0.29	0.16	0.31
T	0.50	0.37	0.60	0.00	0.00	0.09	0.01	0.03	0.08	0.17	0.16	0.18
I(i)	0.30	0.14	0.45	0.38	0.53	0.48	0.58	0.38	0.10	0.04	0.30	0.03

Figure 4. A GA-derived matrix from the dataset train_30
(the same dataset as in Figure 1)

Table 3. Comparisons between the regular and the GA-derived matrices

Dataset[a]	Cutoff[b]	Regular matrix		GA-derived matrix		Selectivity ratio
		False positives[c]	False Negatives[d]	False positives	False negatives	
Split_1	1/30	136	2	57	2	2.05
	2/30	107	2	57	2	1.67
	3/30	100	2	44	4	1.75
	4/30	88	3	44	4	1.65
	5/30	37	5	44	4	0.92
	6/30	37	5	31	6	1.08
Split_2	1/30	627	0	77	3	5.85
	2/30	100	2	64	3	1.38
	3/30	97	2	64	3	1.34
	4/30	93	2	49	3	1.59
	5/30	51	4	48	3	1.10
	6/30	51	4	43	3	1.19
Split_3	1/30	254	2	89	1	2.66
	2/30	247	2	77	1	2.91
	3/30	104	3	34	3	2.37
	4/30	34	4	32	4	1.04
	5/30	30	4	29	6	0.94
	6/30	22	4	26	6	0.83

[a] Datasets from three independent splits on original 50 true binding sequences.

[b] The false negative rate observed in the dataset train_30 when a cutoff score was taken.

[c] The number of sequences with scores higher than the cutoff values in dataset exon_100k.

[d] The number of sequences missed in a dataset test_20 from each split.

Table 4. A list of genes containing potential E2F binding sites with –300 to +100 region

Gene name	Accession number	Position	Sequences
cleavage stimulation factor, 3' pre-RNA, subunit 2, 64kD	AA190813	-10	tttggcgcaccg
glutathione S-transferase theta 2	W68102	-26	tttcccggcgcg
transforming growth factor, beta 1	R24913	-27	cttcgcgccctg
CDC16 cell division cycle 16 homolog	AA010159	-76	tttcccgcgcca
cell division cycle 25A	H59260	-7	tttcgcggtaat
SWI/SNF related, matrix associated, actin dependent regulator of chromatin	W70150	-223	tttcccgcgcgc
hypothetical protein FLJ20156	FLJ20156	-122	tttcgcgcgggc
RAN binding protein 1	N91825	-111	tttggcgggaag
v-fos FBJ murine osteosarcoma viral oncogene homolog	R20750	-89	tctggcgccacc
B-cell CLL/lymphoma 3	H13606	-68	tttcgcgggcgc
cyclin-dependent kinase 6	H92463	-111	tttcgcggcggc
AD50 homolog (S. cerevisiae)	N24479	-14	tttcccggcgtg
activating transcription factor 7	W45393	-178	tctcccgctaga
mitogen-activated protein kinase 11	MAPK11	-107	tttcccgggctg
Retinoblastoma 1	AA045192	-49	tttcccgcggtt
MCM5 minichromosome maintenance deficient 5	W80586	-194	tttcgcgccaaa
cyclin-dependent kinase (CDC2-like) 10	N98775	-58	tttcgcgcctgc
uracil-DNA glycosylase	N63852	-112	tttcgcggcaaa
MADS box transcription enhancer factor 2 (MEF2)	W37622	+7	tttcccggttcg